ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

A POSSIBLE FRAMEWORK FOR BIG DATA ANALYTICS FOR FRAUD DETECTION -IN VEHICLE INSURANCE

ABRHAM WORKU

OCTOBER, 2016

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

A POSSIBLE FRAMEWORK FOR BIG DATA ANALYTICS FOR FRAUD DETECTION -IN VEHICLE INSURANCE

A Thesis Submitted to the Adama Science and Technology University School of Graduate Studies

Department of Computing

In Partial Fulfillment of the Requirements for the Degree of Master of Science in Information System

By

ABRHAM WORKU

OCTOBER, 2016

DECLARATION

I declared that this thesis is my original work and has not been presented for a degree in any other Universities, and all sources of material used for the study have been accordingly acknowledged.

_____

Abrham Worku Negash

October, 2016

This Thesis work has been submitted for examination with my approval as a University advisor.

_____

Dileep Kumar G

October, 2016

ADAMA SCIENCE AND TECHNOLOGY UNIVERSITY

SCHOOL OF GRADUATE STUDIES

DEPARTMENT OF COMPUTING

APPROVAL SHEET

This Research Thesis entitled "A POSSIBLE FRAMEWORK FOR BIG DATA ANALYTICS FOR FRAUD DETECTION -IN VEHICLE INSURANCE" has been read and approved as meeting the preliminary research requirements of the Department of Computing in partial fulfillment for the award of the degree of Master Science in Information System, Adama Science and Technology University.

Adama, Ethiopia.

Abrham Worku Negash **Name of the Student**	**Signature** _____	**Date**_____
Dileep Kumar G **Advisor**	**Signature** _____	**Date**_____
Dr. Wondewossen Mulugeta **External Examiner**	**Signature** _____	**Date**_____
Dr. Dereje Yohannes **Internal Examiner**	**Signature** _____	**Date**_____
Mr. Mohammad Kemale **Head of the Department**	**Signature** _____	**Date**_____
_____ **Institute coordinator for Research**	**Signature** _____	**Date**_____
_____ **Department Chairman**	**Signature** _____	**Date**_____

DEDICATION

I would like to dedicate this paper to my sister -**SIS**, **MEKDES WORKU**, for her those six years and those 17 days that we all know **GASHE-TA-CH**……. And **SABAINA-Sept 17**! I thank you! And be always in my life.

ACKNOWLEDGMENT

Above all, I would like to express my gratitude and heartfelt thanks to my advisor **Dileep Kumar G**. for his constructive comments and overall guidance. Besides bringing the research area to my attention, his direction, guidance, and skillful pushes to get me explore had a huge impact both in this research and on my academic development. I also would like to thank all the **SIG** (Special Interested Group) members specially **Dr. Vuda Sreenivasa Rao** for giving me all the inputs every week in each Friday and also inviting me in each seminar they provided.

LIST OF ABBREVIATIONS

ASTU	Adama Science and Technology University
BD	Big Data
BDA	Big Data Analytics
DB	Data Base
CRAN	Comprehensive R Archive Network.
EIC	Ethiopian Insurance Company
HDFS	Hadoop Distributed File System
IoT	Internet of Things
IBM	International Business Machine
OLAP	On Line Analytical Processing
PM	Predictive Modeling
PDF	Portable Data Format
SCRM	Customer Relationship Management
SNA	Social Network Analysis
SWTO	Strength and Weaknesses Opportunity Threats

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES AND GRAPH

ABSTRACT

Big Data Analytics Technology has been a growing industry in solving our today's daily activity, meanwhile in enhancing the industry of insurance by detecting fraud in all claims of the client using information from customer driving history and social media status data. Fraud and any illegal activity are the main headache of the vehicle insurance industry that change its characteristics like virus depending on the technology that the fraudster ability of using it. So insurance companies must use an advanced technology that analysis the claim and predict the outcome for decisions. Traditional methods, Data Mining and expensive Algorithm of fraud detection use only stored data or structured data to analysis and identify claims but we have to use data from different sources, indifferent format i.e. Unstructured data for a better balanced decision.

The designed Possible Big Data analytics framework for fraud detection in vehicle insurance has three main parts: **Data Acquisition and Preparation:** The data about the insured person vehicle will gathered from different sources, then all the data will be in to Big Data Integration-Hadoop that use: Crawler: Is a program that visits websites and reads their page to create entries for a search engine with having: Flume, Sqoop. **Integration:** Is combine data from disparate source in to meaningful and valuable information i.e. Integration Big Data Analytical State: Integrate-Customer Behavior, Drinking Habit, Bank Statement,, claim history, Social Media status:- FB, Twitter, Instagram, google+ and others to protect insurance and claim Patterns. Then finally **Delivery/Visualization** of results, i.e. Data exploration, Visualization using query language either Hive, R for Fraud detection.

The proposed possible Big Data analytics framework for fraud detection in vehicle insurance will help Insurance companies to know all of the data to gain basic insights of the dataset and prepare analysis: called a **descriptive analysis** then **predictive analysis** based on results of dataset.

So using different V's of property of Big Data, different Analytics methods(descriptive and predictive) and with the help of Hadoop ecosystem, including R programming language we could to analysis datasets from different sources(Social media activity, telemetric software, personal profile and behavior and others) in insurance companies.

Key Words:

Big Data, Big Data Analytics, Big Data Analytics Framework, Fraud, Hadoop

CHAPTER ONE

1.1 Background and Motivation

"Big Data" was first introduced to the computing world by Roger Magoulas in 2005 in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data [1]. In IBM's view Big Data has four aspects: Volume, Velocity, Variety and Veracity.

Volume: refers to the quantity of data gathered by companies that used to obtain important knowledge;

Velocity: refers to the time in which Big Data can be processed. For important activities are that need immediate responses, that is why fast processing maximizes efficiency;

Variety: is the type of data either structured or unstructured (not in the traditional methods), that Big Data can comprise.

Veracity: it is the degree which a leader trusts the used information and take decision using the processed information from Big Data sources [1].

Big Data is a relative term describing a situation when the Volume, Velocity, Variety and Veracity of data exceed an organization's storage or computational capacity for accurate and timely decision making [2]. Big Data Analytic is a way of analyzing and understanding big data that involves data produced by different devices and applications using an advanced tools and technology for making an accurate balanced decision that came from the Big Data property: its huge quantity of data, its high rate, its different format and its truth fullness.

Big Data Analytics use different sources of Big Data generated in our day to day activity like Transport data, Stock exchange data, Power greed data, and Search engine data and then turn the data to data product to solve the problems and make our life easy.

Insurance is a contract between the insurer and insured where by the insurer undertake to pay the insured a fixed amount, in the exchange of a fixed premium, on the happening of a certain events or to compensate the actual loss.

Insurance is a way to protect you from financial loss, if the unexpected happens to the things you own, your health or your ability to work. We can buy insurance to repair or replace your home or car if they get damaged, if we become too ill, pay off the mortgage, medical and recover costs to work, provide money for our family- if we die [3].

Among all these insurances in today's world the most common one is Vehicle insurance, but this system is challenged by frauds. Fraudsters that orchestrate accidents undergo treatment for fictitious injuries and claim larger amount as insurance payouts [4].

Fraud is a white collar crime. And it is the outcome of three elements: motivation, opportunity, and worthwhile (rationalization): - known as the "fraud triangle". With a

highly skilled unemployed population, these causal factors make for a trained, motivated and potentially desperate group of people [5].

Insurance fraud can be defined as "knowingly making a fictitious claim, inflating a claim or adding extra items to a claim, or being in any way dishonest with the intention of gaining more than legitimate entitlement"[6]. And insurance companies are trying to handle fraud to protect their insurance system using traditional fraud detection techniques but, now a day's fraud activity is become as sophisticated as the improvement of technologies' so insurance companies need to use advanced methods to handle fraud activity using Big Data Analytics.

These emerged technologies can be applied in Insurance data, especially in vehicle Insurance to find the fraud claim in vehicles in different situations. With these huge number of vehicles, a lot of accident occur and then insurance claim is going to different insurance companies that may be truthful or fraudulent, even if vehicles accident has occurred: the way how it happens and why it happens is a big headache for insurance organization, so in order to identify fraud claim we have to use Big Data analytics and we need to a framework to developed the Big Data Analysis system, that will used to extract stored records to analyze and predict fraud. Then the insurance companies address the insurance to the right insurer, which create truthful, healthy and secured insurance system for the insurance industry.

1.2 Statement of the Problem

Vehicle Insurance fraud is any deception action committed against insurance companies for unfair financial gain. Fraud in insurance is illegal, so insurance companies try to identify and expose fraudsters based on the claim reports using different investigation techniques. But fraudsters get smarter and use technologies to their benefits to deceive the insurance company and insurance companies face greater difficulties in preventing and detecting deceptions. Because Vehicle insurance company use traditional methods of fraud detection that fail to identify today's fraud and it costs insurance providers millions of birr. So there must be an advanced new technology that the insurers must adopt to solve today's Vehicle Insurance Fraud using Big Data Analytics [4].

Traditional methods of fraud detection like Data Mining use the stored data or the structured one to analysis and identify the claim but it doesn't predict what will be in the future, other traditional method us an expensive algorithm but as we know "More data usually beats better Algorithm" [1], So we have to collect, aggregate and summarize unstructured data from different sources to make a good decision and visualized the result in a better way. In order to do this we have to use a big data and its analysis technology.

Traditional methods to counter new form of fraud are mostly rule based system where algorithms are used to search for anomalies, intrusions or unusual pattern for further investigation like verifying information presented by the claimant, referring suspicious claims to experts, investigating the situation, visiting the place where the accident occur to collect evidence and mapping of claims to check any link between the location of the

incident and the claimants but all of the methods are manual, time consuming, prone to slips and not comprehensive[4].

The problems of the current (existing system) Insurance System includes:

In the case of EIC It is manual or traditional and uses the old traditional claim investigation methods using the company policy and procedure. When we asked the insurance companies about their data sources, all of them are from people that work related to the insurance system that might be corrupted including the customer/claim person.

Because of all of the above reasons all insurance companies lose many of their resources like time and their budgets by trying to identify and paying to the fraud clams. So in order to solve and handle these problems we must have to create a holistic framework to detect fraud which integrates various aspects of insurance claims, premiums called Big Data Analytics for insurance fraud [4]. That uses the 4 V's and advanced way of hoax identify activates.

In today's insurance system there is no Big Data Analytics methods to detect fraud activity, this research identified the gaps that are found in the traditional insurance fraud detection methods then developed Big Data Analytics framework that help to detect all kinds of fraud types that's found in insurance system and it helps to save resources for the Insurance companies.

So by this research: "A Big Data Analytics Framework for Fraud Detection in Vehicle Insurance" attempted to answer the following research questions: -

1.3 Research Questions:

Question 1: How Insurance Company knows all of the data collected, to gain basic insights of the dataset and prepare analysis?

Question 2: How the Insurance Company will monitor total value of payment for the claim and how to visualize results?

Question 3: How Insurance Company wants to figure out the reasons for insurance payment increase and decrease from the data sets?

1.4 Objectives of the Study

1.4.1 General objective:

The objective of the research is proposing and designing a Framework that helps to control Vehicles Insurance Fraud using Data Analytics Technology for Vehicles Insurance System.

1.4.2 Specific objectives:

For the realization of the general objective stated above, the following specific objectives are formulated:

- Review and analysis the literature related to Big Data Analytics techniques and tools proposed by scholars to detect frauds in insurance.
- Classify the types of fraud that is found in the vehicle insurance system.
- Study and identify the limitation and gaps of traditional fraud detection methods and techniques.
- Develop a Framework for fraud detection using Data Analytic methods.
- Test and evaluate the proposed framework against the traditional one.

1.5 Scope and limitation of study

The scope of the proposed research work is to develop a framework that identifies and help to control all insurance system activity and detects frauds in vehicle insurance but the framework can also be used in different insurance situations that found around the industry, keeping in view the resource limitation tasks.

The proposed framework can only handle Vehicle Insurance; it doesn't solve other insurance claim other than vehicles.

Specifically, all data for knowledge discovery or experiment is obtained from the Ethiopian Insurance Corporation and Google data centers. As we know vehicle insurance are the most effective insurance working areas for any private and government insurance organization.

The limitation of the proposed research work are resources that include budgets and big data from social media networks status about the vehicle owner to do experiment using unstructured data format, so this research experimented using the structured data format.

1.6 Research Methodology

For this thesis work of design, a frame work for fraud detection using big data analytics for vehicle insurance, different methods are used to understand insurance fraud problem. Among them Literature Review of different journals, thesis and written articles about BDA and BDA approach are the main ones and then different Data collection methods: observation of real life scenario in Ethiopian Insurance Company (EIC), log book, Google data center, databases of the Ethiopian Insurance Company (EIC), questioner's managers to collect data about the existing system and the existing fraud types and how they are trying to detect it.

1.6.1 Research design

To overcome the major problems, we have mentioned above and for the purpose of achieving this research: we collect data from *http://www.data.gov.et/* in CSV data format and its size is 2182 records with 7 columns. Be in mind that there is a gap in big data sources.

We have tried to design the possible Framework of Big Data Analysis, we used Big Data platform of Apache Hadoop ecosystem to collect and analysis of structured and unstructured data format and then R programming languages used to make descriptive or predictive analysis to define and differentiate behavior of collected data.

1.6.2 Understanding of the Insurance problem domain

In this Big Data Analytics Knowledge discovery research look of the insurance environment is the first step, on the way of knowledge gained from this phase, the BDA problem is defined. Primary (observation &interview) and secondary (DB analysis of Google data center and EIC) data collection methods are used in order to clearly identify, understand, and analyze the problems in the insurance sector. Interview is conducted to define and understood the problems about fraud in the insurance that involved within it. Further, databases such as claim, vehicle, and policy along with insured's information are consulted to gather the pertinent data for the present research. The main BA goal for this research is identifying and detecting fraudulent insurance claims in order to attain the goals of the claims processing and underwriting departments.

For that matter, a model is developed using different data analysis technology, which helps to predict all insurance claims and its premium activity.

Analytics is traditionally described as being descriptive, predictive, or prescriptive. ***Descriptive*** analytics is retrospective; describing what happened in the past and is associated with the field of business intelligence. ***Predictive*** analytics seeks to forecast trends and to determine probabilities. It is associated with time-series analysis, econometrics, and the determination of statistical probabilities. ***Prescriptive*** analytics seeks to determine optimal systems states and is associated with the field of operations management and management science. But in this thesis work we used Descriptive analytics and Predictive analytics.

1.7 Significance of study

Meanwhile the main advantages of the study are to create a framework that used to understand to how detect fraud claim easily then insurance company will address the right premium at the right time to the right claim in order to create a healthy well organized, effective and easy Fraud detection insurance system.

This system gives various benefits for insurance company and their different stakeholders' which are: Insurance Company employees, Clients of the insurance company, low enforcement body that try to identify the claim together with the company.

- It will save Resources of the insurance company budgets for premium.
- Protect the insurance company from paying premium for wrong claim.
- Address the right premium to the right clients within a short time.
- Minimize the employees of the company from making frauds or errors.
- Easy for low enforcement body to make decision.

1.8 Organization of the Thesis

This research is prepared into five chapters.

The First chapter discusses the background of the problem area of vehicle insurance and its related Data technology, and then states the problem, objective of the study, research methodology, scope and its limitation, including the significance of the research outcome.

Chapter Two study about the Literature on Data Analysis and its technology, algorithms and different Methodologies, Big Data process, the different tasks of Big Data, and its application in the insurance sector.

The Third chapter provides the methodology that we used to analysis the data and its description, the Programming Language we used are specified along the frame work diagram.

Chapter Four provides elaboration and discussions about the different Data analytics techniques that are undertaken by the methodology used in this research work. We also provide detailed discussion about the experiment part of this work using R-programming. This includes answering all the research Questions and rise a common consistence of our work.

The Fifth chapter provides a conclusion and recommendation including future works. Moreover, concluding remarks and recommendations forwarded based on the research findings.

CHAPTER TWO

LITERATURE REVIEW

In this chapter, an attempt has been made to review literature on the concepts and techniques of Big Data, Big Data Analytics and its application in Insurance industry especially in vehicle insurance sector that states how it helps to detect and identify frauds with the aim to provide background about the Framework to be built.

2.1 Big Data

The evolution of Information Communication Technology and its phenomenal rate of data growth tend to bring wide availability of vast amount of data in our life; this huge amount of Data is known as Big Data: which came from different source with different format and it is affecting our activity in different ways.

The term 'Big Data' was included in an update of the Oxford English Dictionary (2013): Big data n. computing (also with capital initials) data of a very large size, typically to the extent that its manipulation and management present significant logistical challenges; (also) the branch of computing involving such data.

In Wikipedia: Big data is a blanket term for any collection of datasets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications. The challenges include capture, curtain, storage, search, sharing, transfer, analysis and visualization [7].

"Big Data" was first introduced to the computing world by Roger Magoulas in 2005 in order to define a great amount of data that traditional data management techniques cannot manage and process due to the complexity and size of this data [1].

Big Data is a relative term describing a situation when the Volume, Velocity, Variety and Veracity of data exceed an organization's storage or compute capacity for accurate and timely decision making [2].

The Big Data contain Petabytes to Zettabytes of large and complex data set that might be Unstructured, semi structured and Structured data from different sources and they are difficult to analysis using the traditional data management technology but when we came to this day, information do have its own great value specially for decision making process like in insurance company.

Big data has many different purposes fraud risk management, web display advertising, call center optimization, social media analysis, intelligent traffic management and among other things. Most of these analytical solutions were not possible previously because data technology was unable to store such huge size of data or processing technologies were not

capable of handling large volume of workload or it was too costly to implement the solution in a timely manner.[8]

Authore Elena, Florina, Anca, Manole, 2012 expaline some characterstics of Big Data, in there paper from the IBM's view Big Data point of view has four aspects:

Volume: refers to the quantity of data gathered by a company. This data must be used further to obtain important knowledge;

Velocity: refers to the time in which Big Data can be processed. Some activities are very important and need an immediate response, that is why fast processing maximizes efficiency;

Variety: Refers to the type of data that Big Data can comprise. This data can be structured as well as unstructured;

Veracity: refers to the degree in which a leader trusts the used information in order to take decision. So getting the right correlations in Big Data is very important for the business future. Similarly, this definition big data characteristics are explained in the SAS whitepaper called "Big Data meets big data anaytics".[1]And [2]

2.2 Big Data Analytics

Big data analytics refers to the strategy of analyzing large volumes of data, or big data. This big data is gathered from a wide variety of sources, including social networks, videos, digital images, sensors, and sales transaction records. The aim in analyzing all this data is to uncover patterns and connections that might otherwise be invisible, and that might provide valuable insights about the users who created it. Through this insight, businesses may be able to gain an edge over their rivals and make superior business decisions.[9], Similarly WhatIs.com define Big Data Analytics As the process of examining large data sets containing a variety of data types -- i.e., Big Data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The analytical findings can lead to more effective marketing, new revenue opportunities, better customer service, improved operational efficiency, competitive advantages over rival organizations and other business benefits.[10]

Big data analytics allows data scientists and various other users to evaluate large volumes of transaction data and other data sources that traditional business systems would be unable to tackle. Traditional systems may fall short because they're unable to analyze as many data sources. Sophisticated software programs are used for big data analytics, but the unstructured data used in big data analytics may not be well suited to conventional data warehouses. Big data's high processing requirements may also make traditional data warehousing a poor fit. As a result, newer, bigger data analytics environments and technologies have emerged, including Hadoop, MapReduce and No SQL databases. These technologies make up an open-source software framework that's used to process huge data sets over clustered systems.[9]

2.3 Big Data Analytics Types

Big data can be applied to real-time fraud detection, complex competitive analysis, call center optimization, consumer sentiment analysis, intelligent traffic management, and to manage smart power grids, to name only a few applications. But with the right analytics, Big Data can deliver richer insight since it draws from multiple sources and transactions to uncover hidden patterns and relationships means once you have enough data, you start to see patterns.

According to InformationWeek there are four types of Big Data Analytics:

1. Descriptive – What is happening now based on incoming data. To mine the analytics, you typically use a real-time dashboard and/or email reports. The purpose of descriptive analytics is to summarize what happened.

2. Diagnostic – A look at past performance to determine what happened and why. The result of the analysis is often an analytic dashboard.

3. Predictive – An analysis of likely scenarios of what might happen. The deliverables are usually a predictive forecast. Predictive analytics is the next step up in data reduction. It utilizes a variety of statistical, modeling, data mining, and machine learning techniques to study recent and historical data, thereby allowing analysts to make predictions about the future. The purpose of predictive analytics is NOT to tell you what will happen in the future, "It cannot do that. In fact, no analytics can do that. Predictive analytics can only forecast what might happen in the future, because all predictive analytics are probabilistic in nature.

4. Prescriptive – This type of analysis reveals what actions should be taken. This is the most valuable kind of analysis and usually results in rules and recommendations for next steps. Prescriptive analytics is a type of predictive analytics; it's basically when we need to prescribe an action, so the business decision-maker can take this information and act. Predictive analytics doesn't predict one possible future, but rather "multiple futures" based on the decision-makers actions. In addition, prescriptive analytics requires a predictive model with two additional components: actionable data and a feedback system that tracks the outcome produced by the action taken. Since a prescriptive model is able to predict the possible consequences based on different choice of action, it can also recommend the best course of action for any pre-specified outcome.

2.4 Big Data Analytics in Insurance for Fraud Detection

Different authors, scholars and article states about what is Big Data and Big Data Analytics with its advantage related to insurance sector to detect fraud so let's summaries what has been aside.

Big Data Analytic is a way of analyzing and understanding big data that involves data produced by different devices and applications using an advanced tools and technology for making an accurate balanced decision that came from the Big Data property: its huge quantity of data, its high rate, its different format and its truth fullness.

1. Using Data Mining to Predict Automobile Insurance Fraud JOÃO BERNARDO DO VALE, Portuguesa, September 2012

This paper states Data mining techniques and methodologies to extract meaningful information from large information systems or data sets using Logistic Regression and Classification and Regression Trees – CHAID=Chi-square Automatic Interaction Detector and software package **SPSS Modeler**(window based software package that enables the use of basic and advanced statistical procedures) to perform all techniques that might help us identify the most contributing factors and build models, through which one might estimate a fraud propensity for any given claim, given a set of its characteristics.

Results of the paper is just using information available in most insurance companies' databases, in variables that contained characteristics of the policy, the policy holder, the vehicle and the loss event, the Logit model and the CHAID model derived were able to correctly classify claims in the test sample, respectively, which are considered fair results.

Gap: = Sources of data is only the stored records in the data base and it doesn't handle the unstructured data from different sources and there is no advanced technology that analysis the clime like Hadoop [11]

2. "Survey of Insurance Fraud Detection Using Data Mining Techniques" - International Journal of Innovative Technology and Exploring Engineering (IJITEE) H. Lookman Sithic, T. Balasubramanian, February 2013

Financial fraud can be classified into four: bank fraud, insurance fraud, securities and commodities fraud. That contain wrongful or criminal trick planned to result in financial or personal gains

The researcher focuses on some common fraud in insurance and detection techniques. Their papers states Insurance generally classified into four types: Home Insurance, Life Insurance, Motor Insurance and Medical Insurance.

Among four the motor insurance has more fraud problems. Motor insurance is the most possible and weak fraud ridden sector in the sector in comparison to other line of

insurance. Motor own damage claims fraud committed at pre and post insurance stage. Auto mobile insurance data are usually binary indicator grouped into accident, claimant, and driver, and injury, treatment, lost wages, vehicle, and other categories.

The paper deeply state about fraud in generally the motor fraud have two types: *Hard frauds*-it includes total damage to the vehicle deliberately to get rid of the same or to earn money than its market value. Some of the examples are theft the vehicle, burnt by fire, fall into river, and loss under an excluded risk. A real accident may occur, but the dishonest owner may take the opportunity to incorporate a whole range of previous minor damage to the vehicle into the garage bill associated with the real accident. *Soft fraud*- It accounts for the majority of the motor insurance frauds. For instance, more than one claim for single loss, higher cost for repair, damage caused earlier, replacement of old spare parts.

So then two authors try to uses Data Mining as a methodology to find out the correct information. Data Mining is to find out information with special meaning from a great number of data by some technology as the procedure to discover knowledge from the database.

The paper presents some steps as follows to find and identify any claim. Data cleaning, Data integration, Data selection, Data Transformation, Data Mining, Pattern evaluation, and then Knowledge presentation [12].

Nevertheless, Data Mining is to adopt the mining procedure to discover unknown knowledge and rules from plentiful data. Data mining which part of an iterative process is called knowledge discovery in database can assist to extract knowledge automatically; it doesn't have the ability to handle data from both Semi Structured or Unstructured sources and different data format called Big Data- that limit the capacity to handle any different fraud other than from the database.

3. "Big Data and Specific Analysis Methods for Insurance Fraud Detection": by Ana-Ramona BOLOGA, Razvan BOLOGA, and Alexandra FLOREA.

According to Ana, Razvan, and Alexandra, performing Big Data Analysis, common repetitive errors that are "hidden" inside huge repositories of data can be identified and corrected. Such errors would go undetected in the absence of big data technologies because the human brain is not capable to correlate the huge quantities of data available in the medical insurance sector. Then the authors propose to use big data analytics technologies to prevent insurance fraud like: business rules, anomaly detection, text mining, database searches and social network analysis.

And the paper discusses about big data technologies funded by different companies such as IBM, ORACLE, SAS and Microsoft and open source Hadoop which is often integrated with commercial technologies. Our thesis agrees with the Hadoop technology because it's the open source platform with different cluster at low costs.

Ana, Razvan, and Alexandra explained two cores Hadoop main systems:

1. Hadoop Distributed File System (HDFS): self-healing high-bandwidth clustered storage.
2. MapReduce: distributed fault-tolerant resource management and scheduling coupled with a scalable data programming abstraction.

In this paper there are five criteria which make fraud prevention perfect suited application for big data analytics

- Data-restricted throttling
- Computation-restricted throttling
- Large data volumes
- Significant data variety
- Benefits from data parallelization.[6]

4. "Perspectives on Big Data and Big Data Analytics"- Database Systems Journal vol. III, no. 4/2012-by Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU , 2012

According to this journal by Elena, Florina, Anca, Manoleuse the slogan "More Data usually beats Better Algorithms", that is larger amount of data gives a better output than an expensively developed Algorithm even if working within it can become a challenge due to processing limitations. But the paper gives direction how to process the vast amount of data called Big Data and stress the importance of Big Data Analytics.

"More data usually beats better algorithms". Similarly, (Ana, Razvan, and Alexandra) also makes this point with more information is a driven force for any fraud detection mechanism using a good analytic technology.

This paper explains the source of big data that will be used for further analysis and it began by stating: Today we are living in an Informational Society and we are moving towards a Knowledge Based Society. In order to extract better knowledge, we need a bigger amount of data. The Society of Information is a society where information plays a major role in the economic, cultural and political stage that all data is came from different Social Networks and Medias

The above two papers describe when the term "Big Data" was present in research starting with 1970s but it was first introduced to the computing world by Roger Magoulas from O'Reilly media in 2005.

More over the paper by Elena, Florina, Anca, Manole, explain the challenges: understanding the data, the needs of new technology and IT specialists, Privacy and Security of the data and also Apache Hadoop Technology.

Likewise, author (Ana, Razvan, and Alexandra), Elena, Florina, Anca, Manole, states about Apache Hadoop Technology as big-data processing platform defined as "an open source software project that enables the distributed processing of large data sets across clusters of commodity servers and it do have two main subprojects called MapReduce and Hadoop Distributed File System(HDFS)

The paper by Elena, Florina, Anca, Manole, explain using four basic characteristics that ApacheHadoop become the default plat form to define Big Data this are Scalable, Cost effective, Flexible and Fault tolerant.[1].

5. "Fraud Detection and Mitigation via Advanced Analytics: Trends and Directions "by Scott Allen Mongeau.

This paper tries to summarize trends and implications associated with business analytics and Big Data for the benefit of practitioners working in the field of fraud detection and mitigation. The paper proposes a definition for big data and fraud analytics, in this form:

Big Data is a broad term implying analytics with very large data sets: data which contains many measurements over time and a breadth of variables. The confines of Big Data thus are associated with the engineering challenges of efficiently storing, retrieving, processing, and assessing very large sets of data. Big Data also introduces an advancement concerning the ability to rapidly identify reliable predictive models across large sets of variables.

And the author explains how to understood the fraud using a statement "All fraud is the outcome of three elements: motivation, opportunity, and worthwhile outcome" this statement is likewise Puneet Bharal and Amir Halfon stated with, Fraud is a white collar crime. And it is the outcome of three elements: motivation, opportunity, and worthwhile (rationalization): - known as the "fraud triangle". With a highly skilled unemployed population, these causal factors make for a trained, motivated and potentially desperate group of people [5].

Analytics is traditionally described as being descriptive, predictive, or prescriptive. Big Data Analytics Types Descriptive analytics is retrospective; describing what happened in the past and is associated with the field of business intelligence. Predictive analytics seeks to forecast trends and to determine probabilities. It is associated with time-series analysis, econometrics, and the determination of statistical probabilities. Prescriptive analytics seeks to determine optimal systems states and is associated with the field of operations management and management science. The Author defines what are Big Data and Big Data Analytics: [13].

6. "Using Analytics for Insurance Fraud Detection": by Ruchi Verma Sathyan Ramakrishna Mani.

The paper raises a concert point of view by comparing the traditional and the proposed system using Big Data Analytics.

Traditionally, insurance companies use statistical models to identify fraudulent claims nevertheless these models have their own disadvantages. First they use sampling methods to analyze data, which leads to one or more frauds going undetected. There is a penalty for not analyzing all the data. Second, this method relies on the previously existing fraud cases, so every time a new fraud occurs Insurance companies have to bear the consequences of the first time. Finally, the traditional method works in silos and is not quite capable of

handling the ever-growing sources of information from different channels and different function in an integrated way.

Handling fraud manually is highly costly for insurance companies, even if one or two low incidences of high-value fraud went undetected. Big Data trend (the growth in unstructured data) always leaves lot of room fraud going undetected if data is not analyzed thoroughly.

When they describe about the Big Data analytics methods they used three innovative fraud detection methods: Known as Social Network Analysis (SNA), Predictive Analytics for Big Data (PM) and Social Customer Relationship Management (CRM)

Author Lovro, Štefan, Marko also makes this point the in paper called "An expert system for detecting automobile insurance fraud using social network analysis".[14]

Then finally the paper provides a 10 step approach to implement analytics for fraud detection

1. Perform SWOT analysis (an examination of an organization's internal Strengths and Weaknesses, its Opportunities for growth and improvement, and the Threats the external environment presents to its survival"[15])
2. Build an educated fraud management team.
3. Decide whether to build or buy an analytical fraud detection solution.
4. Clean data.
5. Come up with relevant business rules.
6. Come up with pre-determined anomaly detection thresholds.
7. Use predictive modeling.
8. Use of SNA.
9. Build an integrated case management system leveraging social media.
10. Forward looking analytics solutions

Even if the authors wrote a nice article how to detect insurance fraud in insurance but the paper doesn't use any technology that handles big data like Apache Hadoop or there is no any explanation about the framework and its technology.[16].

7. "Capitalizing on Big Data Analytics for Insurance Industry: -StackIQWhite Paper".

The insurance industry is today awash in data about customers, trends, and competitors. The volume of data available to businesses from both internal systems and external channels has led to a new category of application known as "Big Data" For insurers; the benefits of using analytical applications that tap into the Big Data stream are significant.

"The paper also specifies Apache Hadoop, the leading software framework for Big Data applications, has until now required one team of administrators to install and configure cluster hardware, networking components, software, and middleware in the foundation of a Hadoop cluster. Another team has been responsible for deploying and managing Hadoop software atop the cluster infrastructure. These tasks have relied on a variety of legacy and newer tools to handle deployment, scalability, and management."

Now, however, a paradigm shift in the design, deployment, and management of Big Data applications is underway, providing a faster, easier solution for insurers and other organizations interested in reaping the benefits of Big Data. For the first time in the IT industry, best-of-breed Hadoop implementation tools have been integrated with Hadoop and cluster management solutions in StackIQ Enterprise Data.

This paper summarizes the benefits of Big Data applications for the insurance industry. It also presents the challenges to deploying and managing robust Hadoop clusters. Finally, the cost, efficiency, reliability, and agility features that make StackIQ Enterprise Data competitively unique plus reference architecture for Hadoop deployments using the product are included.

The amount and variety of data available to insurance companies today provide a wealth of new opportunities to increase revenue, control costs, and counter competitive threats. Applying analytical tools to these new huge volumes of data requires a distinctly different infrastructure from traditional database architectures and query products, "Big Data" applications can be run on hundreds or thousands of clustered computer servers instead of supercomputers.

In this paper there are some use cases for Big Data analytics in insurance include:Fraud detection, Risk avoidance, Product personalization, Cross selling and up selling, Catastrophe planning and Customer needs analysis. When the paper state about the Deploying and Managing Big Data Analytical Applications:

The complexity of deploying "Big Infrastructure" clusters has been somewhat lessened by a new generation of open-source software frameworks. The leading solution is Apache Hadoop, which has gained great popularity due to its maturity, ease of scaling, affordability as a non-proprietary data platform, ability to handle both structured and unstructured data, and many connector products. However, the papers state its difficulty when we use it one group installs and configures the cluster hardware, networking components, software, and middleware that form the foundation of a Hadoop cluster. Another group of IT professionals is responsible for deploying the Hadoop software as part of the cluster infrastructure. Until now, cluster management products have been mainly focused on the upper layers of the cluster (e.g., Hadoop products, including the Hadoop Distributed File System [HDFS], MapReduce, Pig, Hive, HBase, and Zookeeper) as a solution it prefers to use Horton works Data Platform makes it easier than ever to integrate Apache Hadoop into existing data architectures of StackIQ Enterprise Data Components. [17]

As a summery all the above papers states the advantages of the using the dataset we found in the insurance sectors to predict some output using their own technology and methodology but they didn't design the framework for understand and experiment to evaluate it, moreover some of the paper work on the data mining methods, that didn't handle data from different sources called unstructured data. So our research solves those issues for better understandings and designed a framework to do the experiment.

Summary of Literature Review:

No	Authors	Titles	Stated or Used Models	Place	Year	Results
1	JOÃO BERNARDO DO VALE	Using Data Mining to Predict Automobile Insurance Fraud	Data Mining techniques and SPSS **Modeler** Logistic Regression and Classification and Regression Trees – CHAID	Portuguese	2012	Results of the paper it uses data from databases, then Logistic model and the CHAID model derived used to classify claims in the test which are considered fair results. But Sources of data is only the stored records in the data base and it doesn't handle the unstructured data
2	H. Lookman Sithic, T. Balasubramanian	Survey of Insurance Fraud Detection Using Data Mining Techniques-(IJITEE)	Data Mining techniques	India	February 2013	States all the technology that can be used in the implementation of data mining to solve problems of fraud in insurance. Doesn't include information in unstructured format, and it works only in a single one not distributed base- unstructured data.
3	Ana-Ramona Bologa, Razvan Boloha, Alexandra Florea.	Big Data and Specific Analysis Methods for Insurance Fraud Detection	Business Rules, Data mining, PM, SNA, Social CRM tool for Hadoop	Romania	2010	It investigates the benefits of Big Data technology and main methods of analysis that can be applied to the particular case of fraud detection, but there must be an organized framework to do a coherent analysis

4	Lena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU	Perspectives on Big Data and Big Data Analytics	Hadoop	Romania	2012	It states Apache Hadoop as de facto Standard for a Big Data processing platform, because it is an open source, Scalable, cost effective, flexible and fault tolerance. But without explaining how it analysis the awash of Big data.
5	Scott Allen Mongeau	Fraud Detection and Mitigation via advanced Analytics: Trends and Directions.	Explain about descriptive, predictive analysis of the Big Data and Big Data Analytics		2014	summarize trends and implications associated with business analytics and Big Data for the benefit of practitioners working in the field of fraud detection and mitigation.
6	Ruchi Verma, Sathyan Ramakrishna Mani	Using Analytics for Insurance Fraud Detection	SNA, PM, SCRM and SWTO analysis tools models	India	-	SNA, PM, Social CRM and SWTO analysis for detecting frauds: - but there must be an organized framework to do a coherent analysis. The paper provides a 10 step approach to implement analytics for fraud.
7	StackIQWhite Paper	Capitalizing on Big Data Analytics for Insurance Industry	Specifies Apache Hadoop and environment as a software framework for Big Data application.	-	-	Summaries the benefits of Big Data application for the insurance industry and present the challenges of deployment of Hadoop.

Table2.1: Summary of Literature Review

So, based on the papers reviewed above, our research creates a framework that use Big Data Analytics to detect frauds in insurance system because all the above paper focused on the traditional methods to overcome or reduce the frauds in insurance systems even some so them give a direction of big data analytic there was on any experiment and the framework.

CHAPTER THREE

METHODOLOGY

3.1 Research Design

This part of the research defines our study types: i.e. descriptive experiments that define different research question with dependent and independent variables and apply data collection methods to do the experiment.

If our data set is in Data warehouse and OLAP (On Line Analytical Processing) we can work on Aggregation and Statistics or Indexing, Searching, and Querying using Keyword based search and Pattern matching, if it is for Knowledge discovery from data set we can use Data Mining and Statistical Modeling.

But when we came to huge data set with different file format from different source, Data Mining for knowledge Discovery cannot handle it. And Data Storage capacity, Large Scale Computing or Distributed computing is the main motivation for developing and using Hadoop technology that is not found in the data mining tools (clustering and classification) for fraud detection purpose.

The detection of vehicle insurance fraud can be done using Big Data Analytics rather than a using the company rules, procedures and data mining. The reason is data mining can only work on structured data formats but Big Data Analytics includes data from different source like Geospatial data, Audio and Video data, Unstructured text, Log files and social media Web clicks and Sensor Data.

Identifying clients claim from the structured data set is not enough to check every claim are right or wrong, moreover claims become more sophisticated because of the emerge of new technology that fraudster master on it.

Big Data doesn't always mean lots of records but Big Data is defined as data sets so large or complex that gaining anti-fraud insights becomes challenging. That certainly describes the amount and types of data that insurers gather in an environment that is growing increasingly data-rich and will continue to expand for all insurers.

3.2 Process Model

Fraud in vehicle insurance become common because peoples want to get unfair income and for a long time insurance company use the same methods to identify claims report from clients and this create the opportunity to commit fraud activity.

And it is much obvious that vehicle insurance frauds increase through time but as we know insurance company are in the same track that used to identify every claim: that is data base oriented technologies that handle all the data about their client information and claims. When they went to make analyses they refer the data base which includes only structured data format, this data base technology allows data mining tools to discover knowledge for making decision.

The DM techniques to identify fraud in insurance use both supervised learning methods (where a dependent variable is available for training the model) and unsupervised learning methods (where no prior information of dependent variable is available for use) can be potentially employed to solve this problem. The DM techniques mostly used for fraud detection are clustering and classification but only from the data base sources.[18].

Even if data mining is an effective way of making analysis in a structured data source, today's claims are more sophisticated so we need data from various sources. And "the more the data, the more we are assuring to make an effective decision". As we know we cannot handle huge amount of data set with different file format in A structured data base, the reason is it does not support Storage capacity, Large Scale Computing or Distributed computing but now we need to have this kind of technology if we want to have a balanced and efficient decision.

In the current Insurance sector, especially in Ethiopia clams are trying to ascertain from evidence that came from claims, police report, garage result, medical records result from hospital and some eye witness when the accident is occurred.

But when we came to the reality, even if the above sources are important and mandatory they are not enough to make a better decision because either of the above entity might be corrupted. So, in order to have the right decision insurance companies must build up data from different sources like: The client personal behavior, drinking habits, Health condition, Bank history, Age, Sex and also his or her social media activity statuses before the accident occurred. Then all the data that came from the above sources will help to identify right claims from the wrong one.

3.2.1. Practice to be preserved from the existing system

We retained some practices of the existing system. The following are the lists of what we have preserved from the old system.

1. Background information, i.e. information from the client about the incident.
2. Information from traffic police, fire brigade, Independent survey, Garage report and court decision.

3.2.2. The Proposed Framework Overview

The proposed framework is used to make a better and balanced decision using Descriptive and Predictive analytic based on data set collected from different sources. The Framework has three main parts:

Web Service for data exploration and visualization: this is based on the given query from user-that is called Data delivery.

Integrated Big Data Analytical State- Called Integration: that handles customer behavior, patterns and social networks to Big Data Integration –Hadoop and analysis insurances datasets and claim patter.

And finally Big Data Integration-Hadoop: it uses Crawlers and Adapter for the unstructured data from different sources and structured Data-Acquisition.

The proposed framework will help to solve the problems of the insurance company regarding its economic aspect; beside to this it solves the problem of claim specialist in acquiring or getting access to information and decision they need. The framework uses the following:

- *Hadoop ecosystem for managing the data from different entities and*
- *R programming language for analyzing the data using different graphs and others.*

3.2.3. Proposed System Requirements:

In this sub phase, the insurance systems claim analysts worked together with the target users of the new system, so that we will gather all the necessary information or get a clear understanding of what insurance system want from fraud detection framework that we proposed. For the valuable information gathered in this phase the sources were users of the current system, from the interview and questioner's answers.

A. Functional Requirements:

The functional requirements explain that the benefits the target user will get from the proposed system in terms of operation ability.

1) Detecting frauds easily.
2) Support the customer claim information by accessing data from Big Data sources.
3) Facilitate the customer or client need within a short period of time to give or not the insured amount in order to complicit he's/her loss after analyzing the claims.
4) Display results using various data exploration/ visualization methods.

Functional Services: It can be used for any customer claims services provided by the system we have developed. The new system is used to:

- ❖ To collect customer information (claim).
- ❖ Updating customer information to support and identify claims for investigation.
- ❖ Predicting the claims are write or wrong and then make dissemination using system rules from the stored information i.e. information from the Hadoop Ecosystem.

B. Nonfunctional requirements:

This types of requirements concentrate on the performance, ability and technicality of the system.

1. Claim will be safeguarded from unauthorized user; to manage fraudster body from the main entities.
2. Data retrieval will be performed in a short period of time from Big Data Sources
3. Producing reports will be done easily.
4. The system should give greater fraud free result.
5. The system should be user friendly for claim specialist in analyzing the data from different sources.

3.3 Data Source

To define this research, we need data to make the experiment but when we came to the having data in electronic way is unthinkable and this makes our research more difficult. But for the purpose of the research we collect the data from different website using the link http://www.data.gov.et/ dataset the data we collect from this source are in CSV formats.

When we are conducting the research, there are alternative ways to get the information we need in addition to collecting the data by the insurance company.

Primary data sources include information collected and processed directly by us, such as observations, surveys, interviews, and focus groups by working with the insurance company we apply this method because there no soft copy data or database in the organization we select for as a center by the name Ethiopian Insurance Corporation (EIC), because it has long years of experience with many clients in the industry.

Secondary data sources include information that we retrieved through pre-existing sources such as research articles, Internet.

3.4 Data Collection Techniques

Information we gathered come from a range of sources as we explain before to get big data from various sources. Likewise, there are a variety of techniques we used when we gathered the primary and secondary data. Listed below are some of the most common data collection techniques we used for collecting data: Interviews, Questionnaires, Observations and Documents and Records.

The data collection process to conduct this thesis is both the qualitative and quantitative. But, it much focused on the qualitative data. This is done through the use of instruments such as observations and interview. From these two data gathering tools, interview was used to collect data from the President Corporate Business Unit Manager of Ethiopian Insurance Corporation (EIC) and other officials and observation was also used to oversee the required things in the organization.

First we have a letter from ASTU that explain our aim to the EIC: that is, we are doing research in the insurance sector that used to find fraud. Then they accept our paper and given us an appointment.

In the appointment day we tell them the research objective, our need from them related to the data. We had administered a standard questionnaire on the 27th of June 2016 at the EIC for those responsible people that found in Addis Ababa head office, after they complete the questioner they given us on 6th of August 2016.

The Interview questions are adopted from ASTU-SIG in Big Data Science members' questions for their research that held at the Ethiopian Aviation Industry.

Interview and Questionnaire:

Interview is a conversation, or questioning, for the purpose of eliciting information for

publication; the published statement so elicited. We will use this method to gather necessary information for our thesis.

Interview and Questionnaire Procedures that we followed:

- First we decided which individuals would be most appropriate to the interview.
- We scheduled the interview and we confirm the meeting time and date a day before conducting the interview.
- We would critically look as much as possible about the topic of our interview before conducting the interview.
- We prepared a note book and pen for the interview to take notes.
- Then we have conducted the interview by fulfilling and criteria required for the interview.
- After conducting the interview, we examined the interview by preparing a written summary and outline of the key points that was discussed in the interview that are relevant to our topic.
- Finally, we have determined its usefulness by analyzing whether the information obtained from the interview is useful, contribute, and collect for the development of our topic.

We conduct personal interviews, using a semi-formal interview as attached in appendix A They don't have any database system that explain client claim, claim kind, source of data, and the methods they used to verify clams and the final decision but he provide us document and records. The company uses their policy and procedure to mine climes.

Observation, Documents and Records analysis:

Observation is the other instrument that we used to collect data which were very essential for our thesis work. In our observation process we have tried to investigate the information by making ourselves participate in the process. This observation helped us to relate the information obtained from the interviewee by looking the quality of the university. We perform Documents and Records analysis because all the data in the EIC are found in hard copy so we have to use it get data even they are not willing to give me the policy and procedure the follow.

Summary of the Data collection results:

EIC register around 300,000 vehicles until now, the companies registered around 2000 new client's vehicles per year. They do have INSIS and Agresso database for operation and account purposes respectively but they don't have any database, big data and big data technology for detection fraud.

Investigations techniques are as the policy and procedure of the company and the procedure to find the fraud is by cross check all the document and entities available on hand. The major failure they experienced are they sometimes cannot get the exact information from the supporting entities.

The kind of claim data available in the organizations are: Claim data for property insurance, liability insurance, life insurance:

According to the interview and questioners the main entities are: The insured itself, traffic police, courts, fire brigade and independent survey: among all of them the main fraudsters is the insured or the client even if, it is primary sources of data for the procedure they follow to detect fraud.

As we know fraud in insurance system is the global problem and in General there are two types of fraud:

1. *Opportunistic fraud,* when a person takes advantage of the deliberate padding or inflating of a legitimate insurance claim. This type of fraud is very common, but the incident is related to a reduced amount.

2. *Professional fraud*, usually done by organized groups of people who may have multiple, false identities. They know very well how to organize the system and often work together with people within the system. The incidence of these events is lower, but the amount related to an incident is much higher.[19]

3.5 Issues of Reliability and Validity

The greater the degree of error, the less accurate and ruthless of results for research. So our research was especially watchful of the sources of error when we planning, data collection, implementing and when we did the experiment.

For convenience sake the major sources of error can be categorized as follows:

A. The researcher.
B. The subjects participating in the project.
C. The situation or social context.
D. The methods of data collection and analysis.

And we give a great emphasis to erase errors in all of the above four categories to maintain Reliability and Validity of our research work.[21].

3.6The Development of the Prototype

Prototype is an early sample, model, or release of a product built to test a concept or process or to act as a thing to be replicated or learned from. It is a term used in a variety of contexts, including semantics, design, electronics, and software programming. A prototype is designed to test and try a new design to enhance precision by system analysts and users. Prototyping serves to provide specifications for a real, working system rather than a theoretical one. In some workflow models, creating a prototype (a process sometimes called materialization) is the step between the formalization and the evaluation of an idea.

Some businesses continue to operate in traditional ways. However, nowadays, all aspects of a business are collect data and are equipped for data collection. The wide availability of data has led to increasing interest in methods to extract useful information from Data Science. Organizations in almost every industry focus on exploiting data for competitive advantage.

Business Analytics Is a scientific process that transforms data into insight and It used for fact-based or data-driven decision making using tools such as reports and graphs (simple), optimization, data mining, and simulation (complex) to work on the four types of Big Data Analytics: Descriptive, Diagnostic, Predictive and Prescriptive.

3.6.1. The Architecture of the System

When a Big Data system is realized, the main important considerations are including architecture design of the system, and utilization of underlying technologies and products/services.

The goals of this thesis part are to design Reference Architecture for Big Data systems and Classify related technologies and products with respect to the reference architecture to detect vehicle insurance claim.

The Reference architecture would be useful to facilitate creation of concrete architectures, and increase understanding as an overall picture by containing typical functionality and data flows in a Big Data system. Moreover, the classification of technologies and products/services should facilitate decision making regarding realization of system functionalities. Also, it would be important to understand architecture and performance characteristics of related technologies.

Due to the complicated nature of Insurance, frauds have always found a favorable environment in the Vehicle Insurance system. And as we know knowing what the client wants is the key factor to success in any type of business, and the best way to find this information is to conduct a survey in the insurance sectors specially the vehicle system.

Insurance system use different information to address the correct premium to client that compensates their loss by any means if it is legal.

But the traditional systems, which identify the claims, correct or not are the corrupted one: Because of the lack of Big Data information, good Analytic tools and Technology including the skilled human power that work on it.

When we design Model of the Architecture of the System and do an Analytics we follow Model building is a well-defined process, for building our predictive model we follow the 6-six stages steps including: Problem Definition, Hypothesis Generation, Data Extraction/Collection, Data Exploration and Transformation, Predictive Modeling Finally Model Development/Implementation.

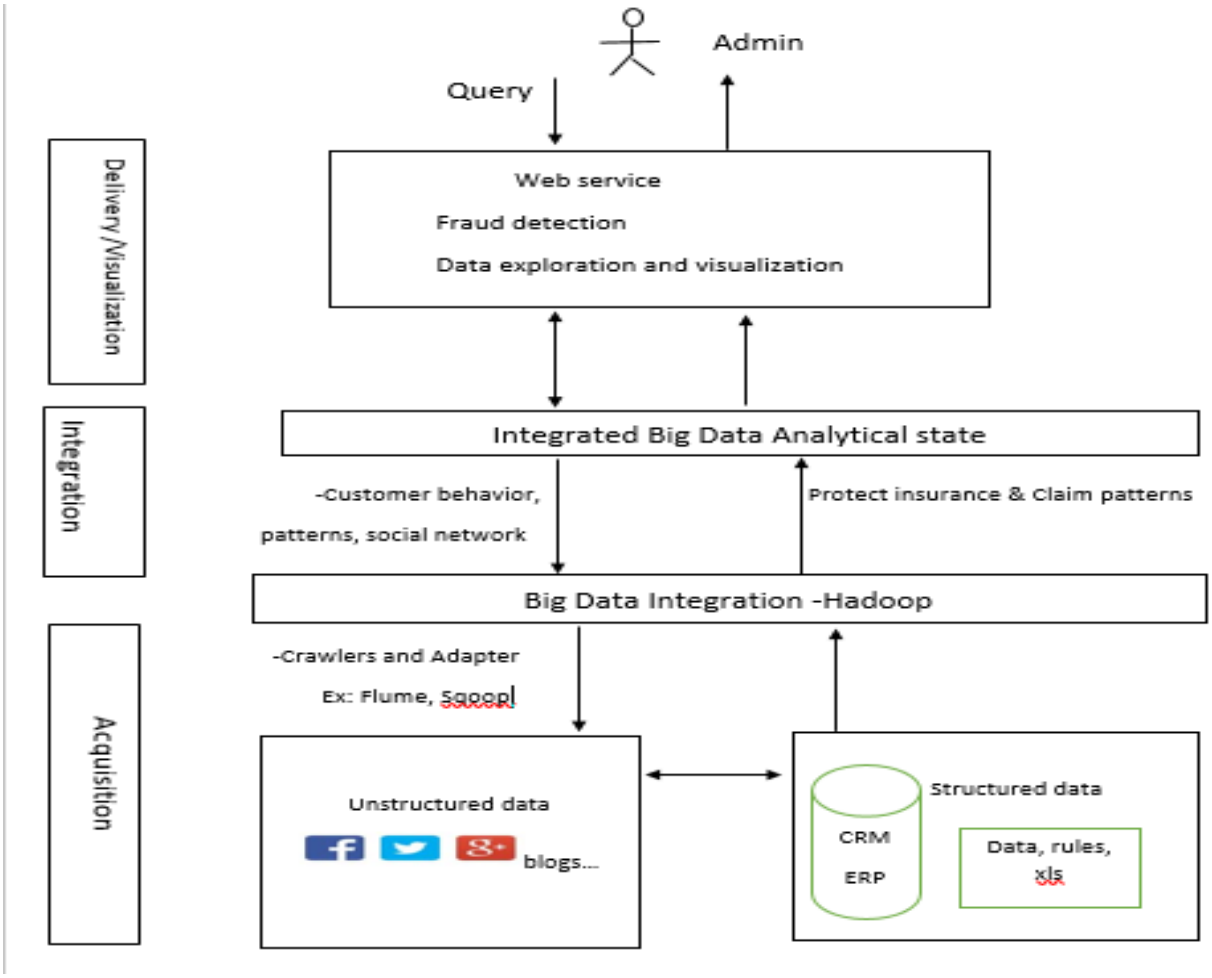


Figure3.1The Possible Big Data Analysis Framework

The designed Big Data analytics frameworks for fraud detection in vehicle insurance have three main parts: Web Service: data exploration and visualization, fraud detection based on the given query from user. That is Data delivery. Integrated Big Data Analytical State: that handles customer behavior, patterns and social networks to Big Data Integration -Hadoop and analysis insurances datasets and claim patter back from: Called Integration. And finally Big Data Integration-Hadoop: it uses Crawlers and Adapter for the unstructured data from different sources and structured Data-Acquisition.

- **Data Acquisition and preparation:**

The data about the insured person vehicle will gathered from different sources, then all the data will be in to Big Data Integration-Hadoop that use:

Crawler it is a program that visits websites and reads their page to create entries for a search engine. Ex: Flume, Sqoop.

- **Flume** is a distributed, reliable and available service for efficiently collecting, aggregating and moving large amount of log data.
- **Sqoop**: is a command Line Interface application for transferring data between RDB and Hadoop.

- **Integration:** Is combine data from disparate source in to meaningful and valuable information.

Integration Big Data Analytical State: Integrate-Customer Behavior, Drinking Habit, Bank Statement, claim history, Social Media status: - FB, Twitter, Instagram, google+, to protect insurance and claim Patterns

- **Delivery/Visualization:**

- Web service
- Using query in Hive or R
- Data exploration, Visualization for Fraud detection.

Various solutions have been presented for the Big Data Analytics which can be divided into:

1) Processing/Compute: Hadoop (MapReduce and HDFS)
2) Storage: HDFS
3) Analytics: Descriptive and Prediction Analytics using R

Although there exist commercial products for data analysis most of the studies on the traditional data analysis are focused on the design and development of efficient and/or effective "ways" to find the useful things from the data. But when we enter the age of big data, most of the current computer systems will not be able to handle the whole dataset all at once; thus, how to design a good data analytics framework or platform and how to design analysis methods are both important things for the data analysis process [22].

3.6.2. Programming Language Used and Justification

R was initially written by **Ross Ihaka** and **Robert Gentleman** at the Department of Statistics of the University of Auckland in Auckland, New Zealand and is currently developed by the R Development Core Team. R made its first appearance in 1993. R is a programming language and software environment for statistical analysis, graphics representation and reporting.

R is freely available under the GNU General Public License, and pre-compiled binary versions are provided for various operating systems like Linux, Windows and Mac. This programming language was named **R**, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language **S.**[23].

Justification and the Benefits of Using R

Of the many attractive benefits of R programming language, a few are easy to recognize. It's actively maintained, it has good connectivity to various types of data and other systems, and it's versatile enough to solve problems in many domains. Possibly best of all, it's available for free [24]. (http://www.tutorialspoint.com/r/index.htm)

I. R is free, open-source code

R is available under an open-source license, which means that anyone can download and modify the code. This freedom is often referred to as “free as in speech.” R is also available free of charge a second kind of freedom, sometimes referred to as “free as in beer.” In practical terms, this means that you can download and use R free of charge.

Another benefit, albeit slightly more indirect, is that anybody can access the source code, modify it, and improve it. As a result, many excellent programmers have contributed improvements and fixes to the R code. For this reason, R is very stable and reliable.

II. R runs anywhere

The R Development Core Team has put a lot of effort into making R available for different types of hardware and software. This means that R is available for Windows, UNIX systems (such as Linux), and the Mac.

III. R supports extensions

R performs a wide variety of functions, such as data manipulation, statistical modeling, and graphics. One really big advantage of R, however, is its extensibility. Developers can easily write their own software and distribute it in the form of add-on packages.

IV. R connects with other languages

As more and more people moved to R for their analyses, they started trying to combine R with their previous workflows, which led to a whole set of packages for linking R to file systems, databases, and other applications. Many of these packages have since been incorporated into the base installation of R.

Initially, most of R was based on FORTRAN and C. Code from these two languages easily could be called from within R. As the community grew, C++, Java, Python, and other popular programming languages got more and more connected with R.

Because many statisticians also worked with commercial programs, the R Development Core Team wrote tools to read data from those programs, including SAS Institute’s SAS (Statistical Analysis system) and IBM’s SPSS.

Many of the big commercial packages have add-ons to connect with R. Notably, SPSS has incorporated a link to R for its users, and SAS has numerous protocols that show you how to move data and graphics between the two packages.

V. R support different file formats

R languages support different data formats from different sources so in order to investigate all claims we can use from various data interfaces: CSV, Excel, Binary, XML, JSON and from Web data files.

CHAPTER FOUR

PRESENTATION, ANALYSIS, AND INTERPRETATION OF DATA

4.1 Introduction

To complete this study properly, it is necessary to analyze the data collected in order to test the hypothesis and answer the research questions. As already indicated in the preceding chapter, data is interpreted in a descriptive form.

This chapter comprises the analysis, presentation and interpretation of the findings resulting from this study. The analysis and interpretation of data is carried out in two phases. The first part which is based on the results of the questionnaire and observation of the data that deals with a quantitative analysis of data. The second, which is based on the results of the R programming codes, is a qualitative interpretation.

4.2 Description of the data collected

The main data to carry out the research is collected from www.data.gov , http://instruction.bus.wisc.edu/jfrees/jfreesbooks/Regression%20Modeling /BookWebDec2010/data.html and EIC through interview and questioners as database. And the insurance dataset holds several variables related to the insured one that the insurance company used for identification of their clients, among them we selected and crop from the huge dataset for confidentiality and the purpose of our works. From this data sets we used the Seven (7) variables and the description of these variables are given below: -

No	Attribute Name	Data type	Descriptions
1	Kilometer	int , continues	Kilometers travelled per year
2	Zone	varchar	Geographical zone in the country
3	Bonus	int	No claims bonus; equal to the number of years, plus one, since the last claim.
4	Make	int , Categorical	different common car models
5	Insured	int , Continues	Number of insured in policy-years
6	Claims	int , continues	Number of claims
7	Payment	int , continues	Total value of payments in Ethiopian birr

Table 4.1Description of the dataset variables

Attribute name	Representation	Zone Geographical
Zone	1	Addis Ababa, and its surroundings
	2	Other large cities with surroundings
	3	Smaller cities with surroundings in southern Ethiopia
	4	Rural areas in southern Ethiopia
	5	Smaller cities with surroundings in northern Ethiopia
	6	Rural areas in northern Ethiopia
	7	East and West of Ethiopia

Table 4.2Description of the Zone variables

Zone: Graphic zone of a vehicle, grouped into 7 categories.

"Zone" represents geographical area that found in the country and it relates with Insured, Claims and Payment in dataset as: if the vehicle travel frequently to the difficult accident area the insured amount, payment and number of claims will be high.

Attribute name:	Representation	Category in Km travelled per year
Kilometer	1	<1000
	2	1000-15000
	3	15000-20000
	4	20000-25000
	5	>25000

Table 2.3 Description of the Kilometer variables

Kilometer: Distance driven by a vehicle, grouped into five categories.

"Kilometer" represents category in Km that a vehicle travelled per year and it relates with Insured, Claims and Payment in dataset as: if the car travels long Kilometer the opportunity of an accident occur and calms amount will be high. So, insurance company should ask high amount of payment that the client should pay that complicate the insured amount when the calms came. Seems too expensive the payment will be high as the insured amount and the number of claims.

Attribute name:	Representation of car model	Car models
Make	1	BMW: 1,3,5,7 series
	2	Ford: F-150,fiesta, Mustang
	3	Toyota: corolla, Camry, Prius
	4	Chevrolet corvette, Silverado
	5	Mercedes-Benz: C, E, S-class
	6	Nissan: Altima, Maxima, Rogue
	7	Hyundai: Sonata, Santa fe, Genesis
	8	Lexus: IS,GS,LS
	9	All other models

Table 4.4Description of the Make variables

"Make" is representation of car model and it relates with Insured, Claims and Payment in dataset as: if the car model seems too expensive the payment will be high as the insured amount and the number of claims.

4.3 Data quality assurance

The collected data contains some missing, incomplete and irrelevant data. Some important information regarding the driver's age, marital status, and health condition, date of driving license, the place and time of the accident occurred, traffic police report and some others are missing. Generally, most of the attributes of the dataset have not that much use and it was too difficult to understand the data because of it's too complexity. Only some of the attributes are applicable for the problem at hand.

4.4 Preparation of the Data

While DM is a key stage in the knowledge discovery process, the data preprocessing process often require considerable effort. The purpose of the preprocessing stage is to cleanse the data as much as possible and to put it into a form that is suitable for use in later stages. Starting from the data extracted from the source a number of transformations are performed before a working dataset was built.

4.5 Experimentation of Research Questions and its Results

4.5.1 Experimentation I:

This part of experiment used to answer: *How Insurance Company knows all of the data to gain basic insights of the dataset and prepare analysis?* The insurance company is interested to know each fields of the data collected. In order to do have the results we had use a Descriptive Analysis: that used to get basic insights in to the data set for further analysis. That defined what happened; Because of the dataset is large, we have to know the pattern of the data- called the distribution of the data using Summary Statistics. To see all the records of the dataset just write view function in the R console;

>View (data)

Data: data

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment
21	1	1	3	3	32.02	3	4399
22	1	1	3	4	497.20	23	112992
23	1	1	3	5	73.48	6	14788
24	1	1	3	6	278.01	9	48713
25	1	1	3	7	66.36	9	52076
26	1	1	3	8	17.86	3	13161
27	1	1	3	9	5063.15	408	1707680
28	1	1	4	1	318.48	29	103866
29	1	1	4	2	57.21	7	77588
30	1	1	4	3	35.33	4	11839
31	1	1	4	4	374.28	20	98140
32	1	1	4	5	85.18	7	27919
33	1	1	4	6	199.70	7	103910
34	1	1	4	7	60.46	4	38065
35	1	1	4	8	12.74	0	0
36	1	1	4	9	4263.09	300	1267678
37	1	1	5	1	444.37	25	69203
38	1	1	5	2	86.65	6	14620
39	1	1	5	3	53.81	5	40258

Table4.5.1A sample of the dataset: Results of the view function

For example, the following row of data from the above data set describes: -

22	1	1	3	4	497.20	23	112992

The vehicle travel 1km categories that represent less than 1000 km per year, in zone 1 that is Addis Ababa and its surrounding and this driver got recognition three times as a Bonus from the insurance companies for paying regularly on time, the vehicle make company is 3 that is car model group called Chevrolet corvette, Silverado, and they made a total of 23 climes for which a total of 497.20 birr is paid.

It displays all the records but in order to get the summery of the data set we have to write Summary () function.

>summary (data)

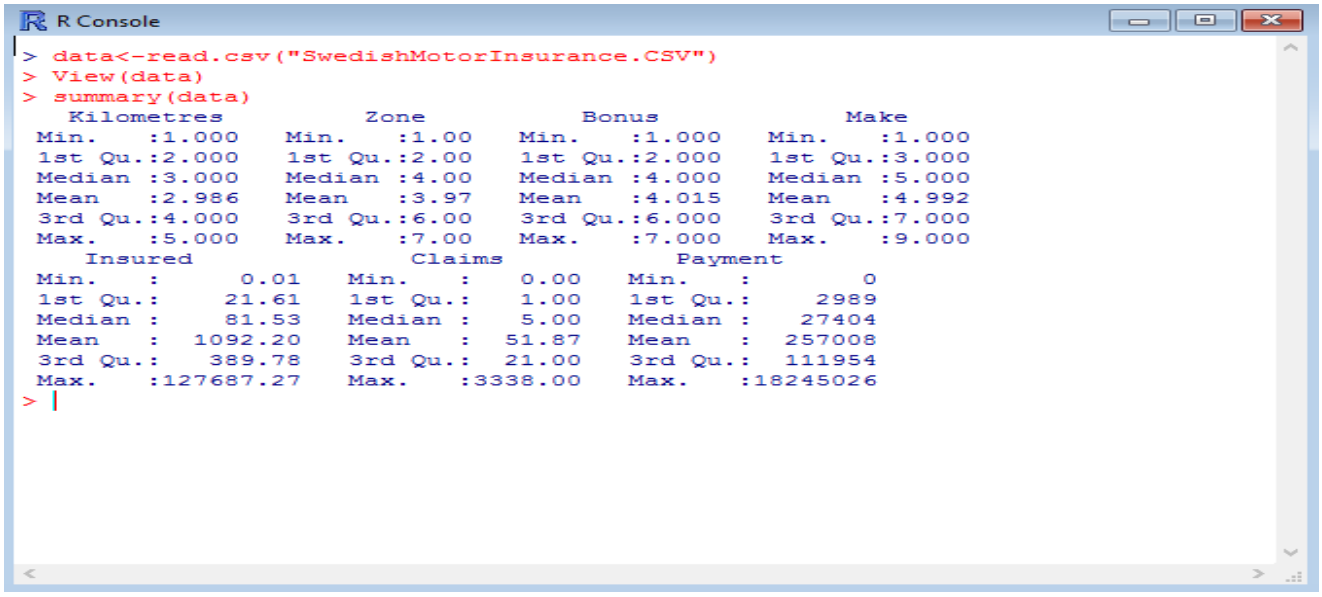


Table 4.5.1Bsummary of the dataset after results of the view function

Results of the Descriptive Analysis:

- As we see the summary(data) provide MIN and MAX of each Variables or columns
- Claims and Payment also have zero values and insured column does not have Zero value
- No claims or Payment has been made for that combination of car maker, Zone, and Kilometers
- There are some entries where the car has been Insured for a given period of time

The results provide the minimum and maximum values. It also provides the mean and median values of all variables. From this you can understand the spread of data. We can see that claims and payment also have null or zero values, however the insured column does not have a zero value. This specifies that there are few entries where the car has been insured for a given period of time. However, no claim or payment has been made for that combination of car Make, zone, and kilometers.

4.5.2 Experimentation II:

This part of experiment used to answer the research question*: How the Insurance Company will monitor total value of payment for the claim and how to visualize results?* To monitor the total value of payment we have to start from a Descriptive analysis, that relates Payment to the number of Claims and the number of Insured policy using correlation function between them (payment, claims and insured) that identify is it a positively or negatively correlated. Then the result is displayed using Scatter plot to understand the results.

To do so first we have to read the dataset in to the R database then analysis each variable of Claims and Payments.

>data<-read.csv ("MotorInsurance.csv")

>data$Claims

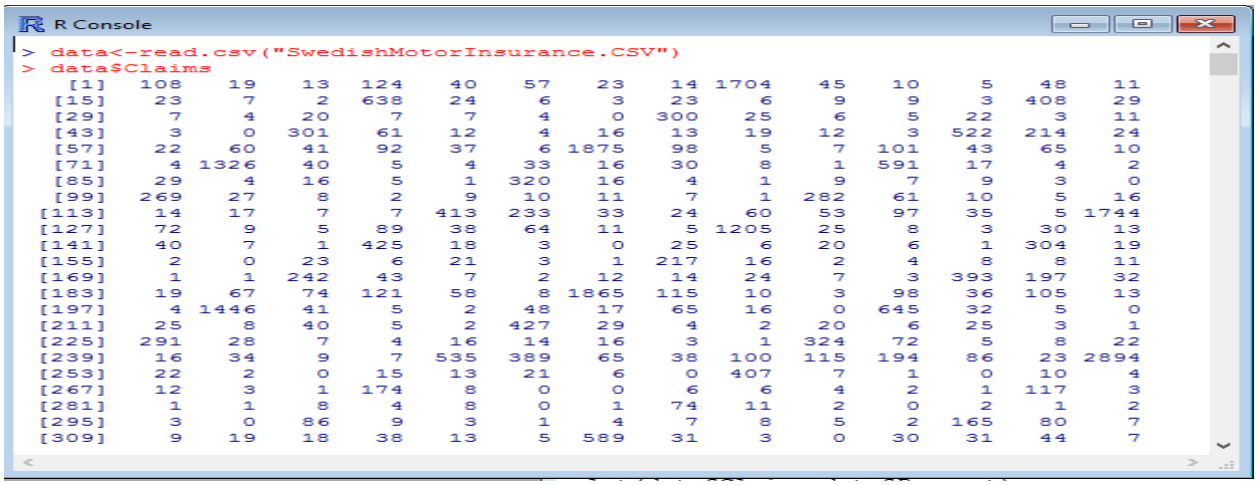


Table 4.5.2sample claim data in the dataset

>data<-read.csv ("MotorInsurance.csv")

>data$Payments

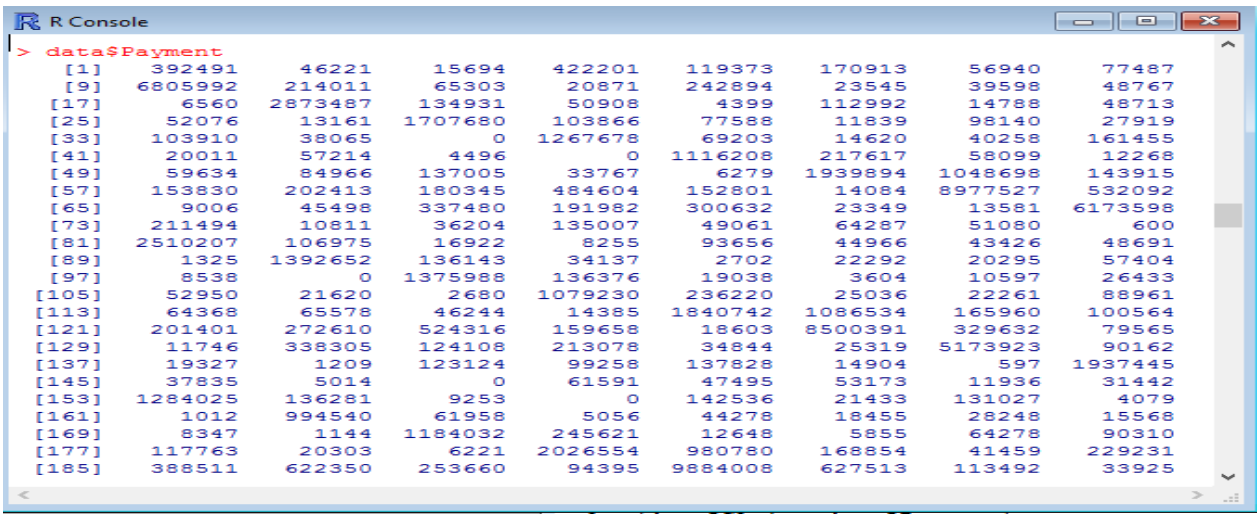


Table 4.5.2BSample Payment data in the dataset

>cor (data$ claims, data$ Payment)

[1] 0.9954003

>cor (data $Insured, data $ Payment)

[1] 0.933217

The results show that *claims* are 99 percent positively correlated with *payment* and *insured* is 93 percent positively correlated with *payment*. The scatter plot shows that the relationship between the variables are strong as there is a linear trend in the graph, that is, as the value of claims increases, the payment value also increases and the same trend will occur for the insured and the payment

Then we need scatter plot for Visualize the results for better understanding of the correlation between the dependent variable Payment and the most independent useful variable Claims and Insured.

>plot (data$Claims, data$Payment)

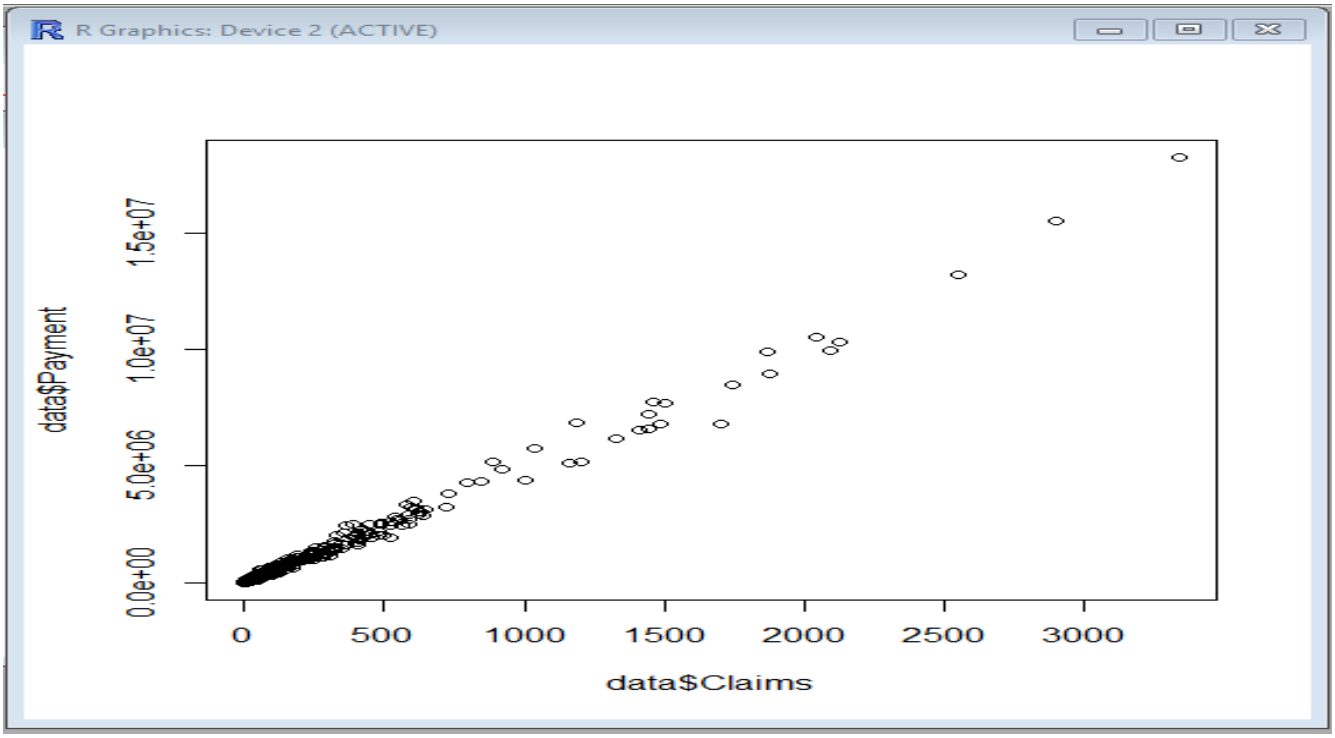


Figure 2 Relationship between claims and Payment

>plot (data$ Insured, data$Payment)

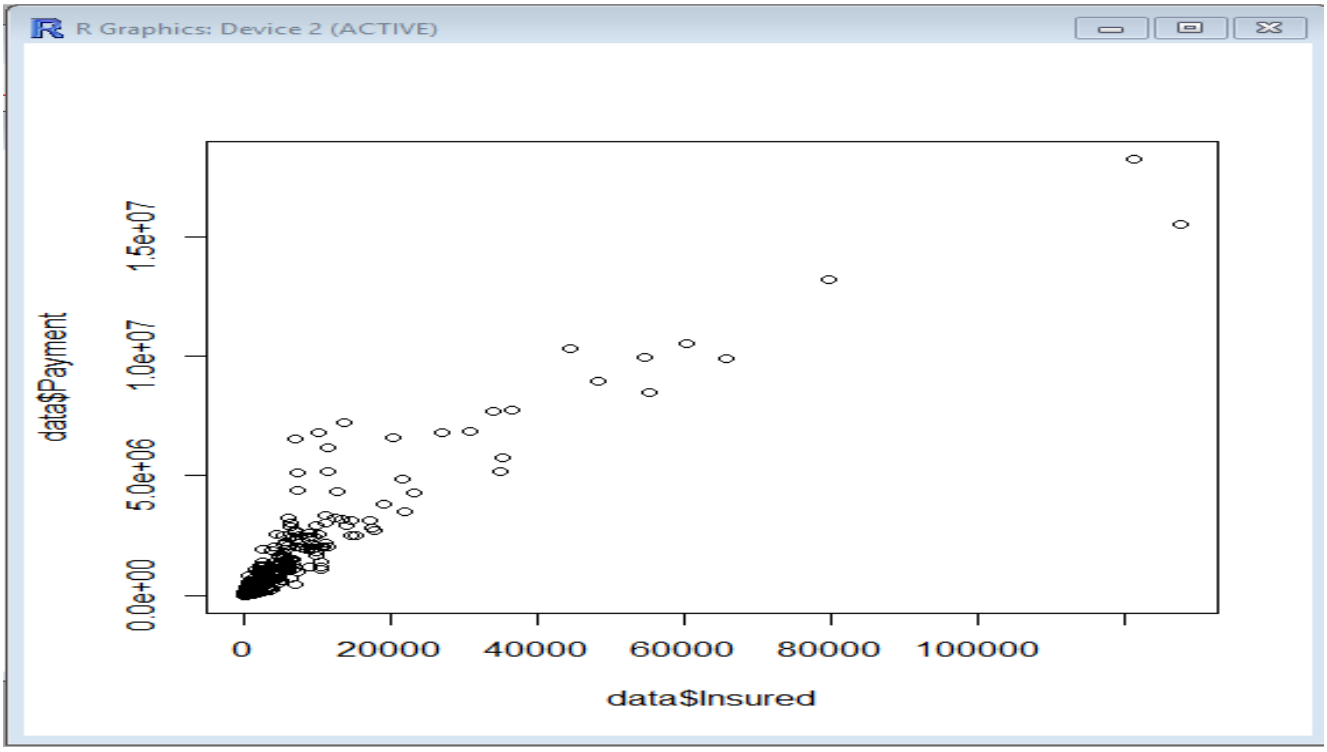


Figure 3 Relationship between Insured and Payment

4.5.3 Experimentation III:

This part of experiment answered the research question: How Insurance Company wants to figure out the reasons for insurance payment increase and decrease from the data sets? To answer this query, we had to decide which variable is more affecting the Payment Variable among distance, location, bonus, make, insured amount and claims however all of them or some of them might be also affecting it.

The relationship between Dependent variables (Payment) and Independent variables can be solved using linear regression function *lm ()*

>data<-read.csv ("MotorInsurance.csv")

> View (data)

>lreg<-lm (data$Payment~data$Insured+data$Claims+data$Make+data$Bonus+

data$Zone+data$Kilometres)

>summary (lreg)

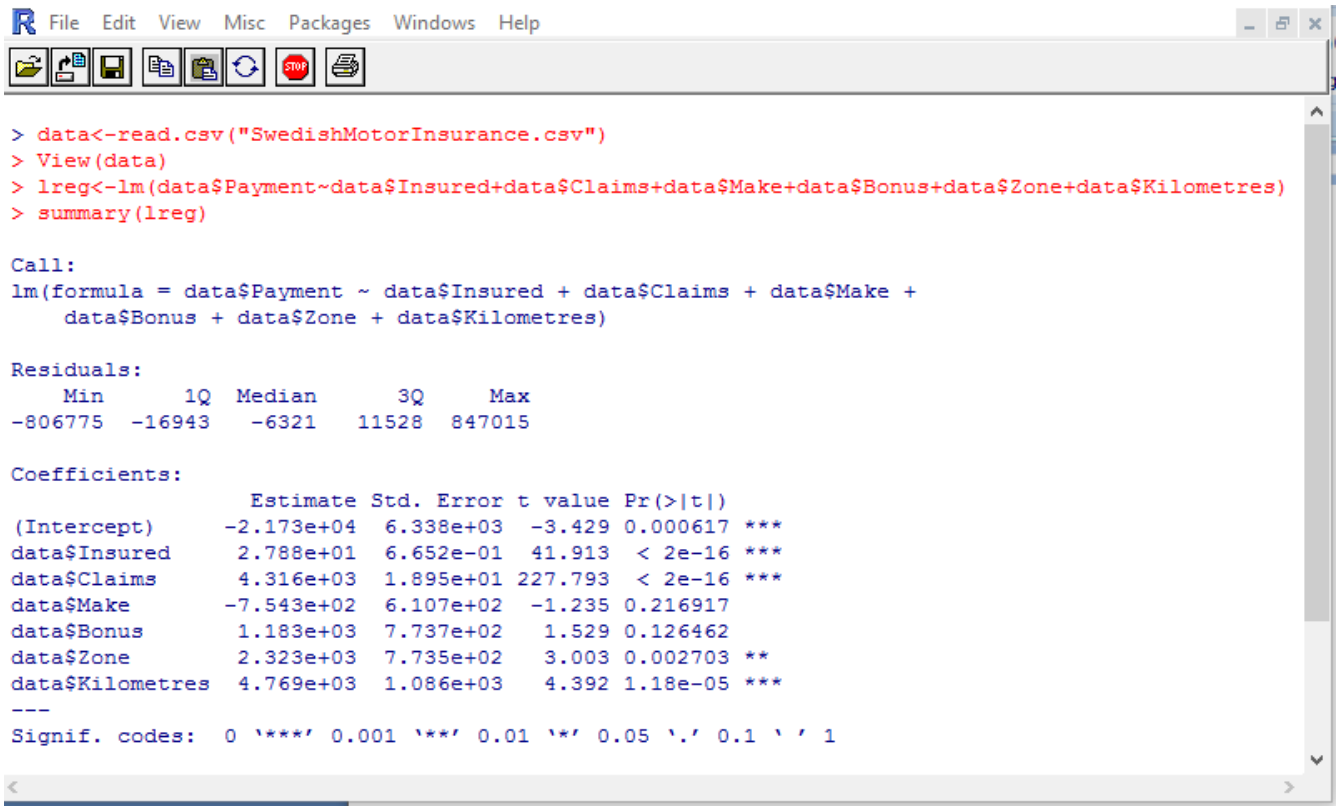
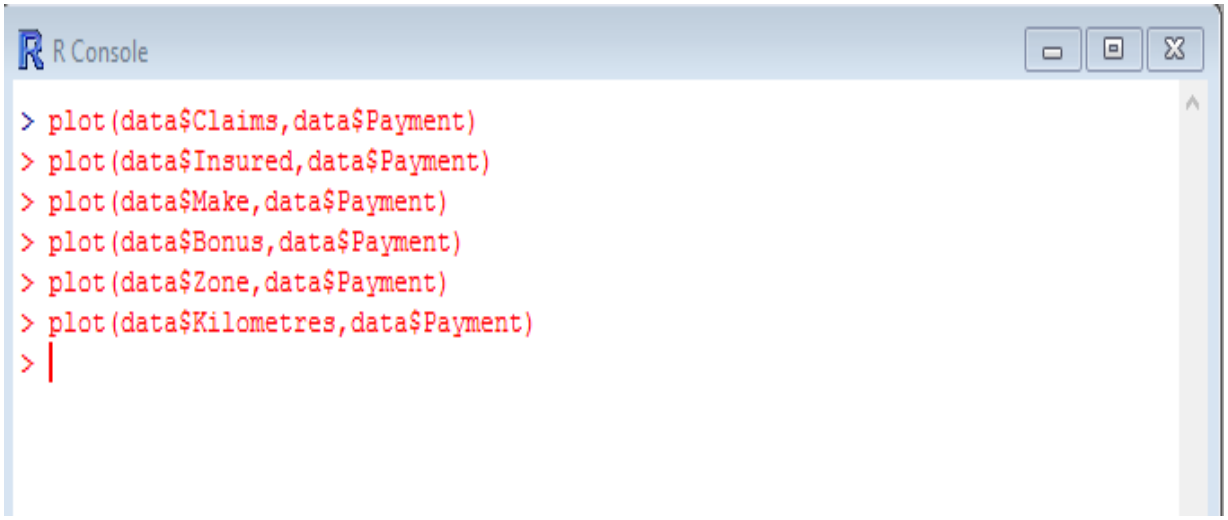


Table 4.5.3Summary of payment Vs Insured, Claims, Make, Bonus, Zone, Kilometer using the lm()

The result shows the intercept value and estimated values of all independent variables. From this we can derive the regression line and this would help us in predicting the future payment values. The high p-value of the make and bonus show that they do not make much impact on payment, as compared to all other variables.

To analyze the relation between claim that is independent variable and dependent one Payment we can plot the graph using *plot ()* function in the R console.



Plot (data$Claims,data$Payment)

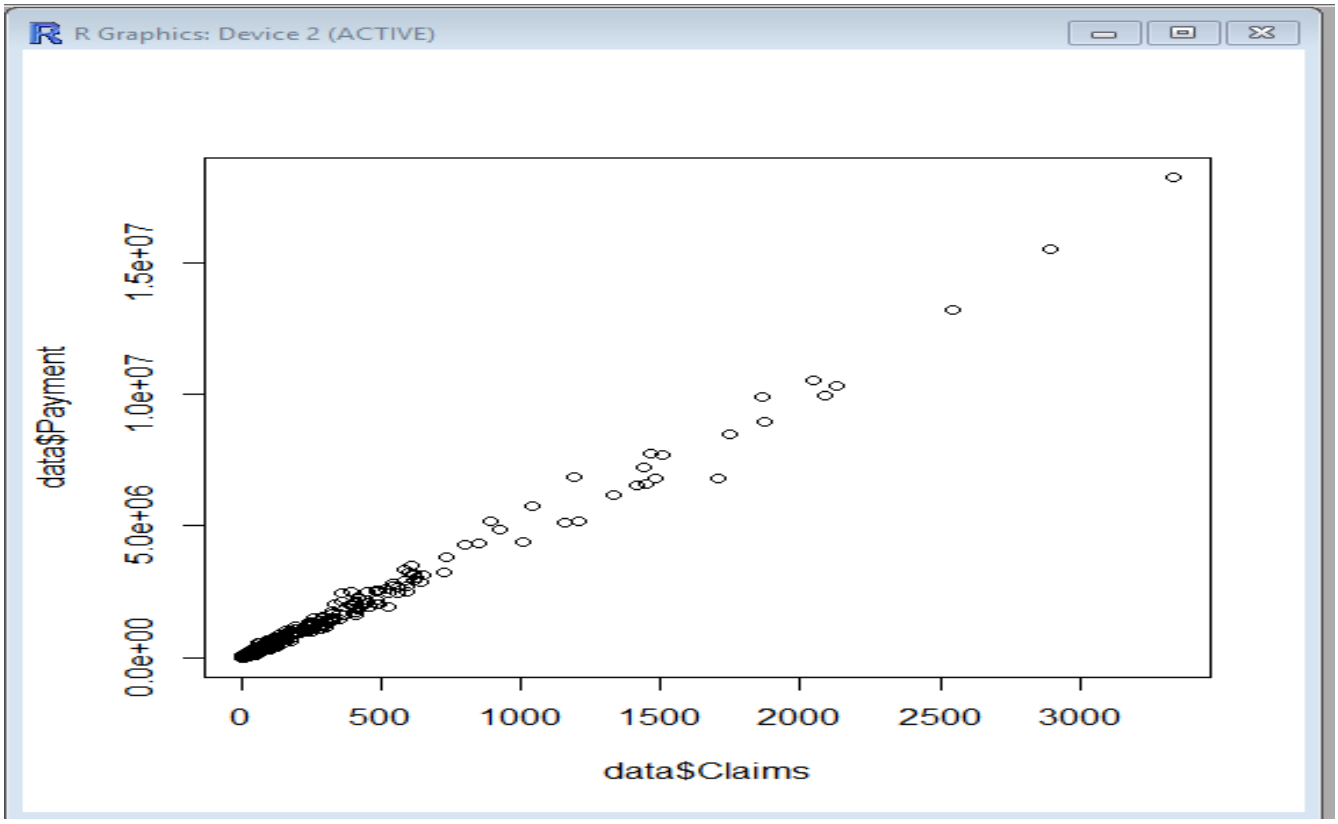


Figure 4 Relationship between Claims and Payment

As we see in the output the Payment will increase directly proportional to the number of Claims.

Plot (data$Insured,data$Payment)

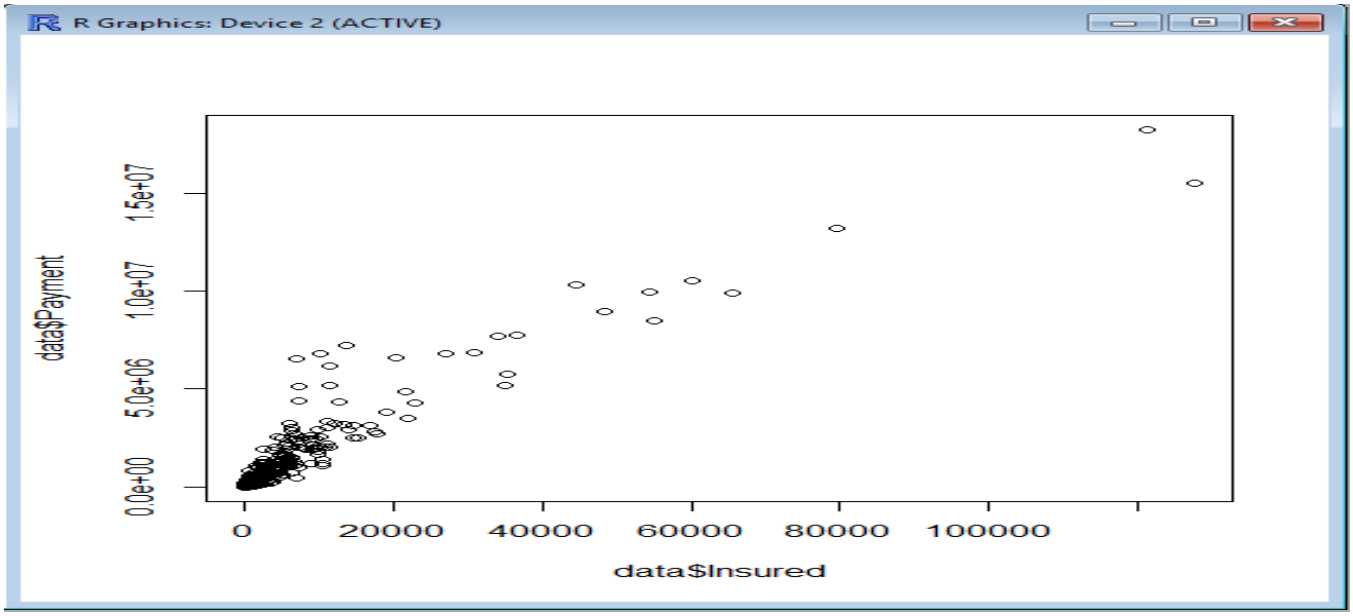


Figure 5 Relationship between Claims and Payment

As we see in the output the Payment will increase directly proportional to the number of Insured.

Plot (data$Make,data$Payment)

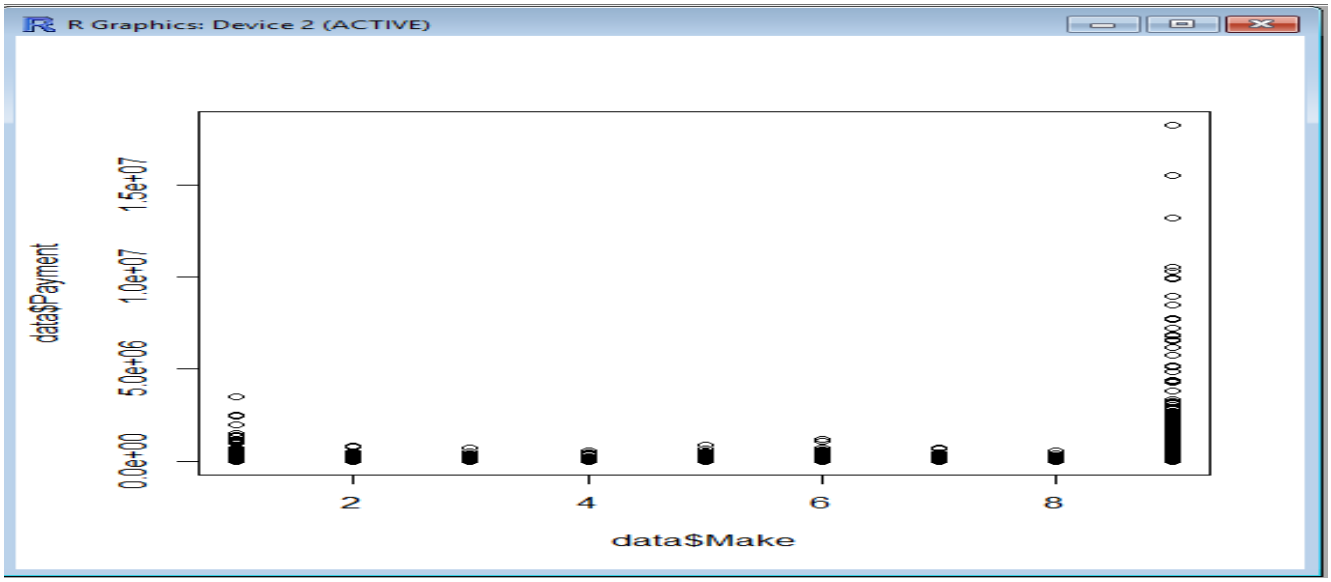


Figure 6 Relationship between Make and Payment

As we see in the output the Payment does not rise when the make is varied, so make doesn't affect the Payment and we can say as a generalization on models 1up to 8 have low impact on Payment but model nine have high Impact on Payment than others.

Plot (data$Zone,data$Payment)

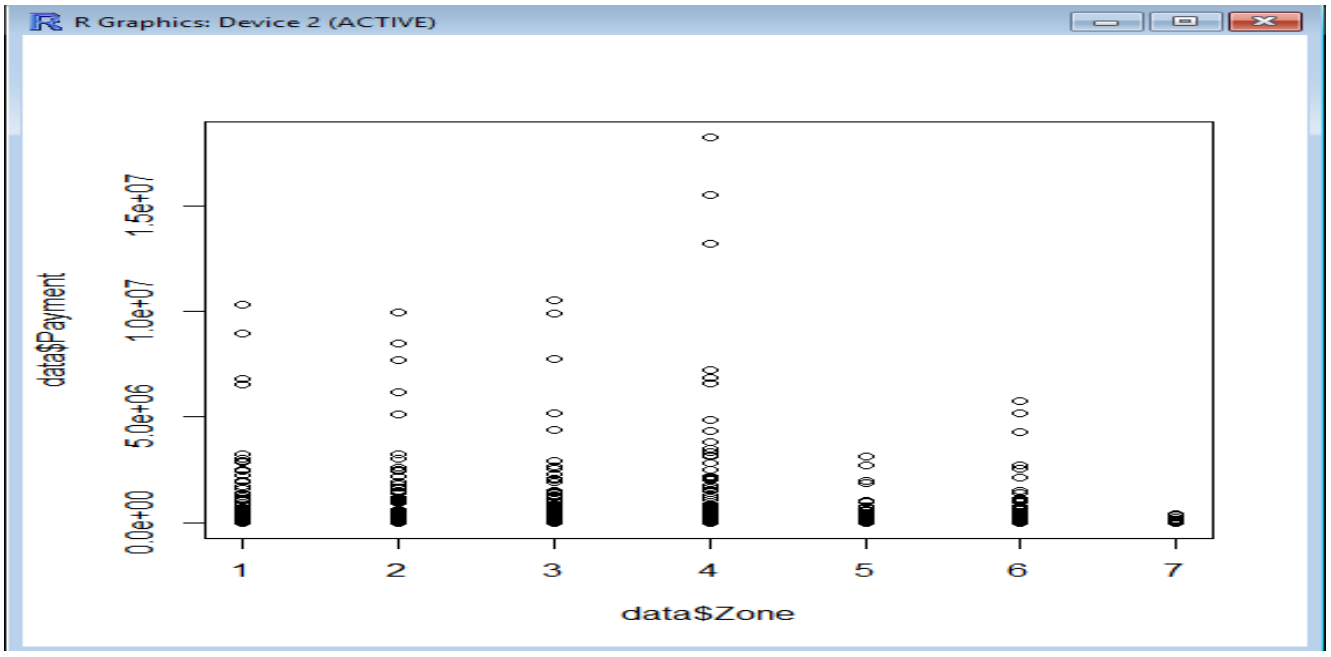


Figure 7 Relationship between Zone and Payment

As we see in the output the Payment is high in Gotland zone (or Zone 4) more than other states.

Plot (data$Bonus,data$Payment)

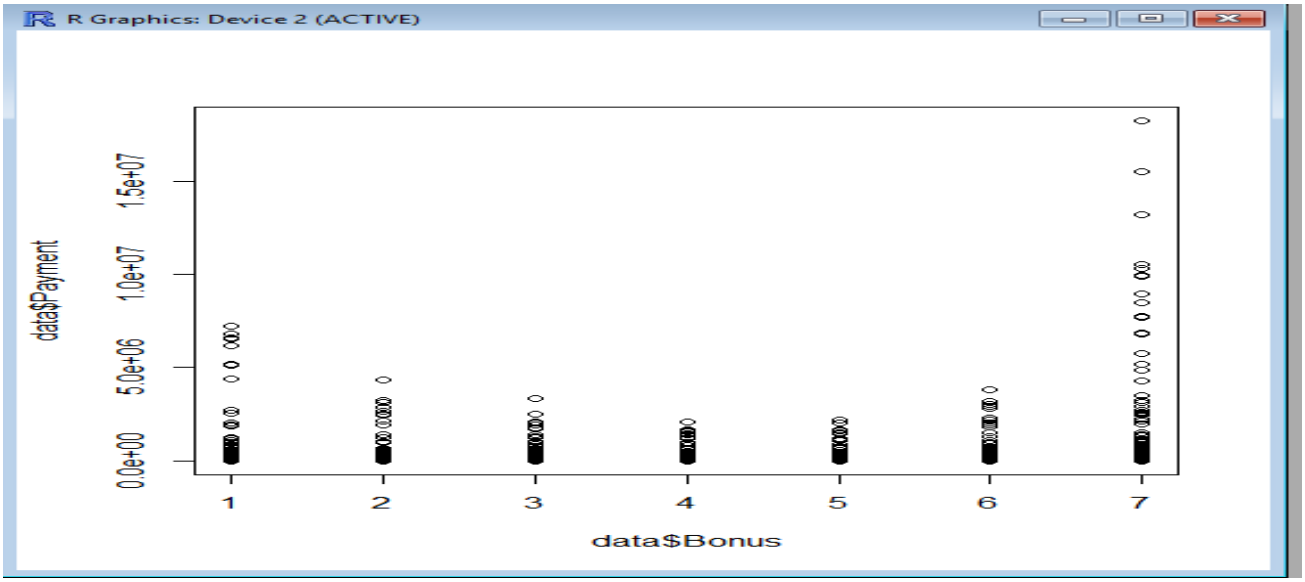


Figure 8 Relationship between Bonus and Payment

As we see in the output the Payment does decline at the bonus #4 and its high at bounce 7.

Plot (data$Kilometers,data$Payment)

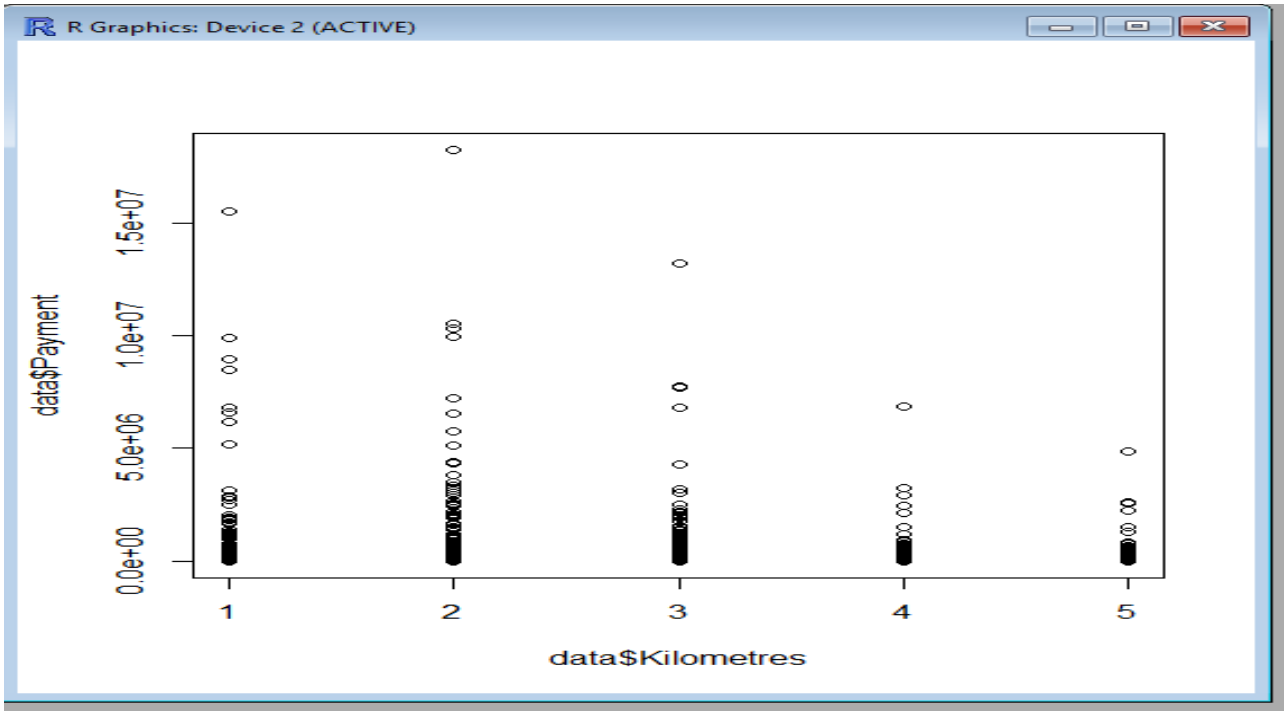
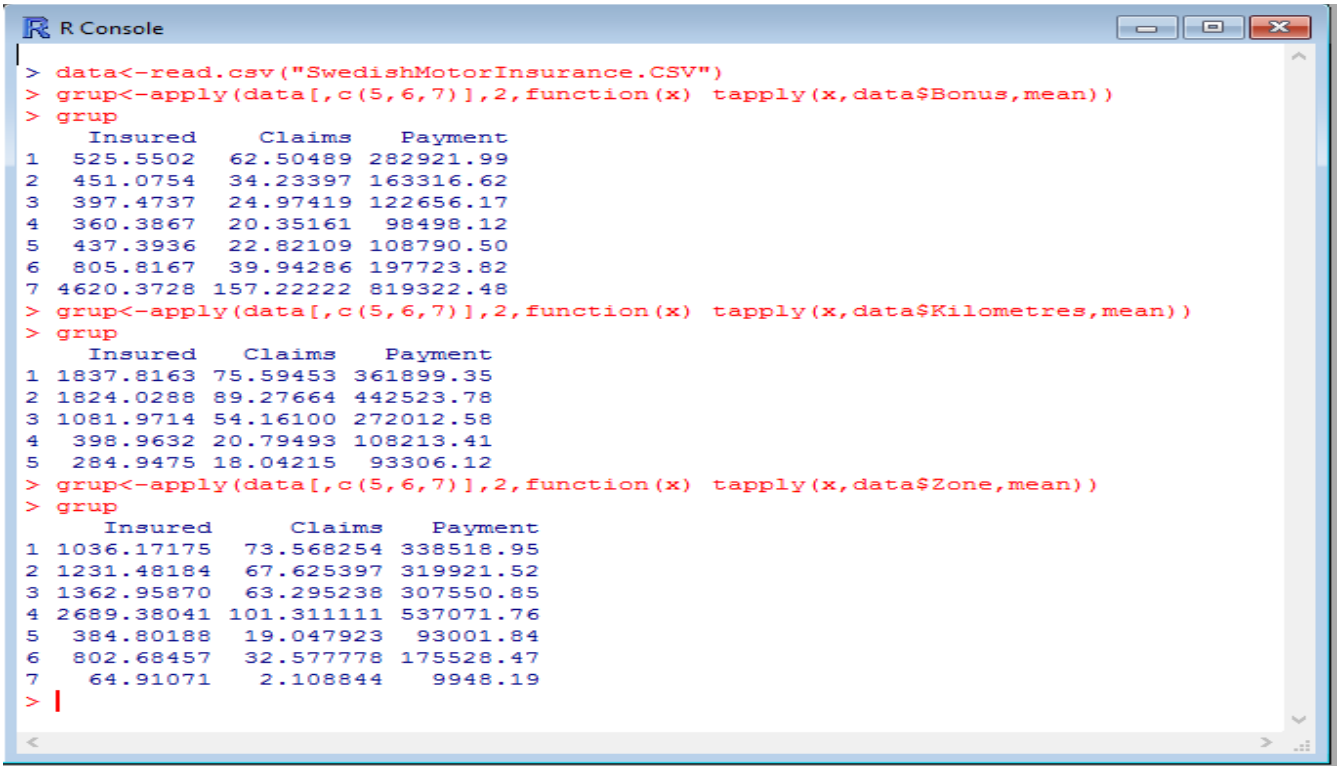


Figure 9 Relationship between Kilometers and Payment

As we see in the output the Payment does not have any change on kilometer variable.

4.5.4 Experimentation IV:

This part of experiment answered the research question: How the insurance company is planning to detect fraud; to do this we have to be interested to find at what location, kilometer, and bonus level their insured amount, claims, and payment get increased. Then the insurance company will give a lot of effort on the claim data to detect fraud, at that area. To perform this, we have to do an Aggregate Dataset in order to find the mean value of insured, payment, and claims based on zone, kilometer, and bonus variables, group all the result variables based on individual categorical variables.



As we can see the output result in the above we can make the following observations:

I. Zone 4 has the highest number of claims, and thus payment as well.
II. Zones 1-4 have more insured years, claims, and payments.
III. Kilometer group 2 has the maximum payments. Though the insured number of years is lesser than kilometer 1, the claims and payments are higher for group 2.
IV. There is not much variation in groups of bonus except for 7 with unusually high number of insured years, claims, and payments.

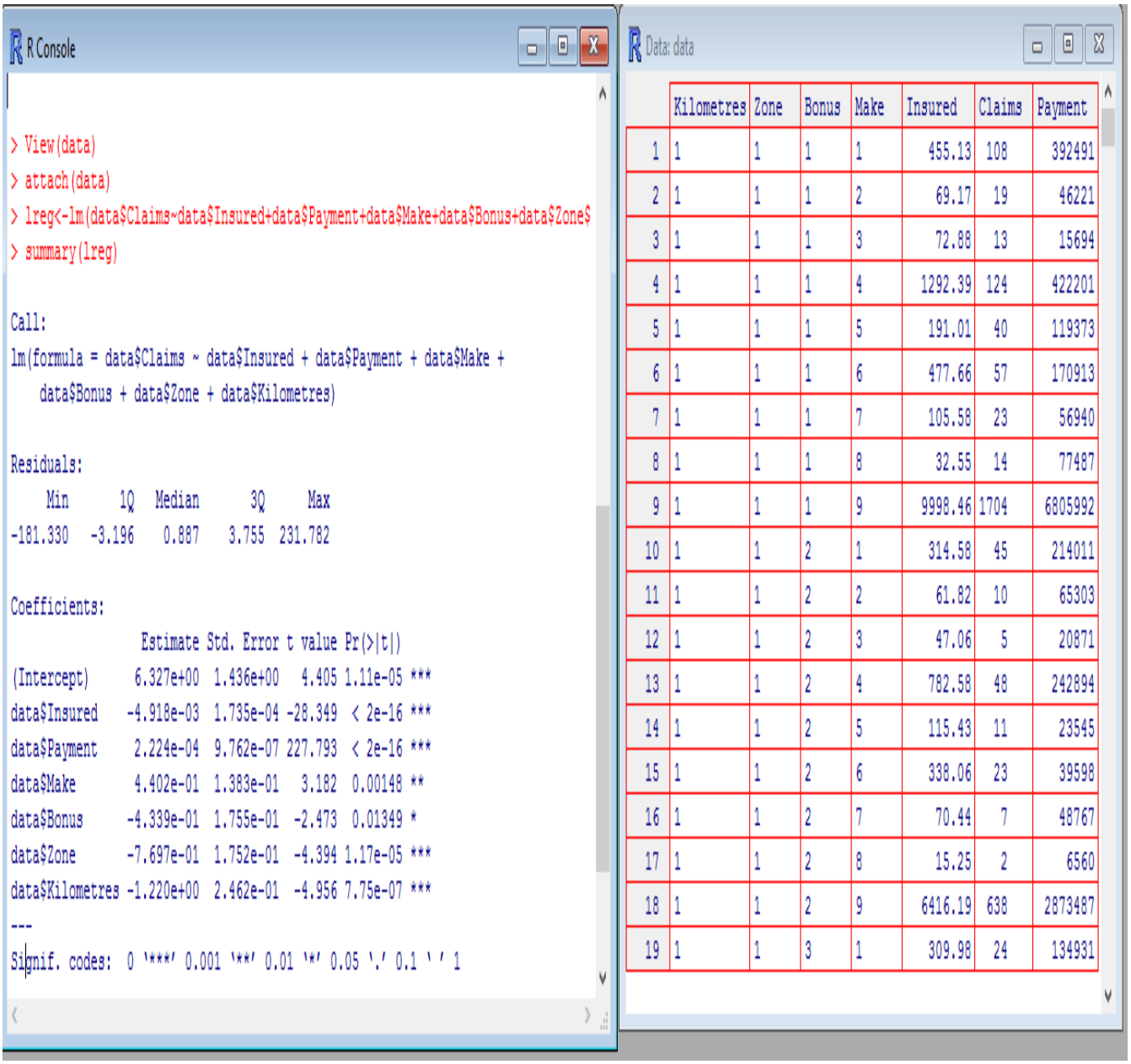
4.5.5 Experimentation V:

This part of experiment answered the research question: How the Insurance Company wants to understand what affects their claim rates so as to decide the right premiums for a certain set of situations?

So, they need to find whether the ***insured amount, zone, kilometer, bonus, or make*** affects the ***claim rates*** and to what extent. So, in order to find the dependency of claim variable by other variables build a linear regression model.

Dependent variable: ***claim***
Independent Variable: ***insured, zone, kilometer, bonus and make***

attach (data) ->this means that the database is searched by R



R Console

```
> View(data)
> attach(data)
> lreg<-lm(data$Claims~data$Insured+data$Payment+data$Make+data$Bonus+data$Zone$
> summary(lreg)

Call:
lm(formula = data$Claims ~ data$Insured + data$Payment + data$Make +
    data$Bonus + data$Zone + data$Kilometres)

Residuals:
     Min       1Q   Median       3Q      Max
-181.330   -3.196    0.887    3.755  231.782

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     6.327e+00  1.436e+00   4.405 1.11e-05 ***
data$Insured   -4.918e-03  1.735e-04 -28.349  < 2e-16 ***
data$Payment    2.224e-04  9.762e-07 227.793  < 2e-16 ***
data$Make       4.402e-01  1.383e-01   3.182  0.00148 **
data$Bonus     -4.339e-01  1.755e-01  -2.473  0.01349 *
data$Zone      -7.697e-01  1.752e-01  -4.394 1.17e-05 ***
data$Kilometres -1.220e+00  2.462e-01  -4.956 7.75e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Data: data

	Kilometres	Zone	Bonus	Make	Insured	Claims	Payment
1	1	1	1	1	455.13	108	392491
2	1	1	1	2	69.17	19	46221
3	1	1	1	3	72.88	13	15694
4	1	1	1	4	1292.39	124	422201
5	1	1	1	5	191.01	40	119373
6	1	1	1	6	477.66	57	170913
7	1	1	1	7	105.58	23	56940
8	1	1	1	8	32.55	14	77487
9	1	1	1	9	9998.46	1704	6805992
10	1	1	2	1	314.58	45	214011
11	1	1	2	2	61.82	10	65303
12	1	1	2	3	47.06	5	20871
13	1	1	2	4	782.58	48	242894
14	1	1	2	5	115.43	11	23545
15	1	1	2	6	338.06	23	39598
16	1	1	2	7	70.44	7	48767
17	1	1	2	8	15.25	2	6560
18	1	1	2	9	6416.19	638	2873487
19	1	1	3	1	309.98	24	134931

Result:

The results provide the intercept and estimated value and this in turn shows that all the payment values of independent variables, such as kilometers, zone, bonus, make, and insured are highly significant and are making an impact on the claims.

Therefore, to understand the results graphically we can plot the relationship between the Claims with all other variables using the *plot ()* function

To display the graph that visualizes the results.

```
>plot(data$Claims,data$Insured)
>plot(data$Claims,data$Make)
>plot(data$Claims,data$Bonus)
>plot(data$Claims,data$Insured)
>plot(data$Claims,data$Zone)
>plot(data$Claims,data$Kilometers)
```

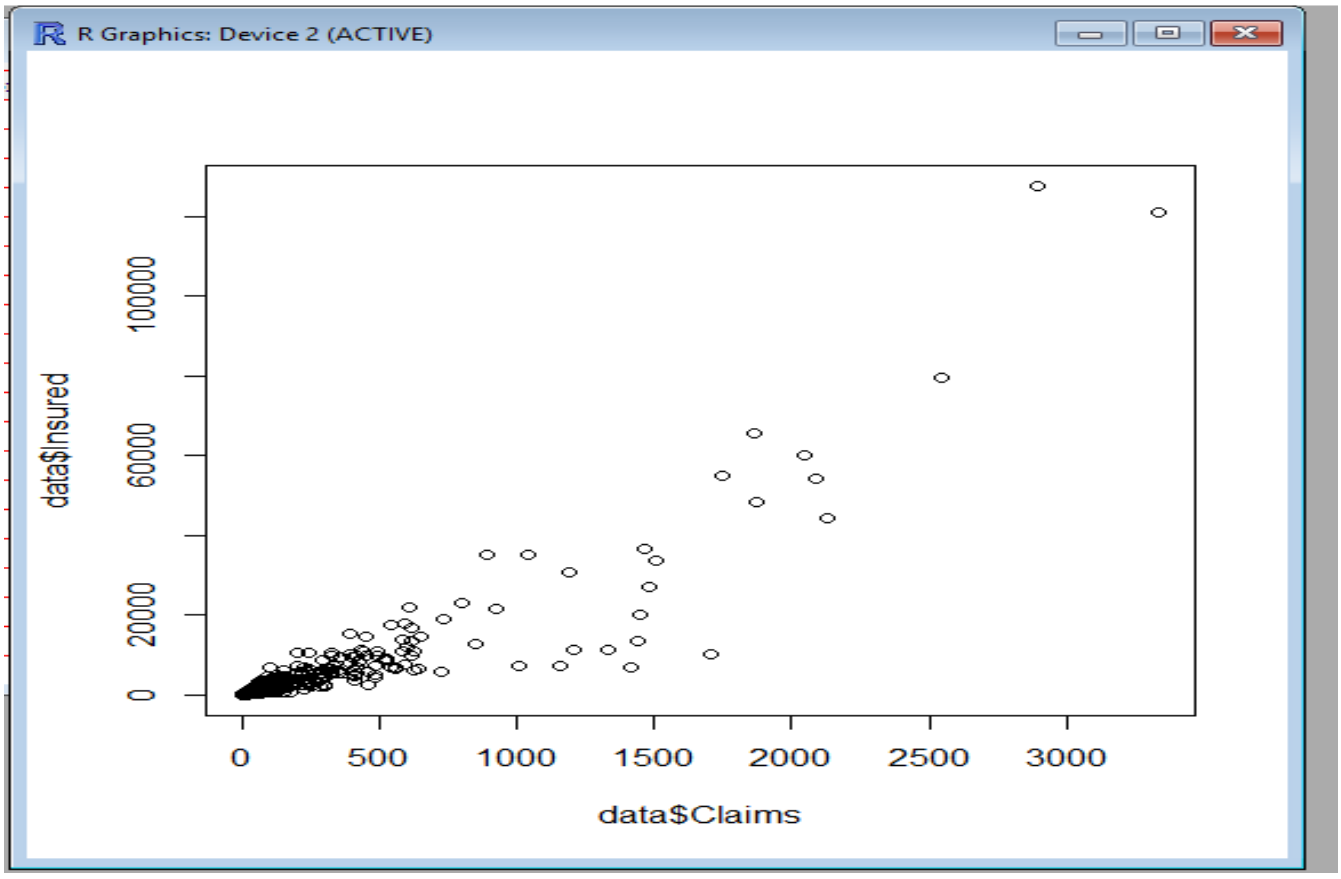


Figure 10Relationship between claims and Insured

The figure display that claims and Insured are directly proportional. So, insurance sectors must create great emphasis towards the claim type and what is the claim.

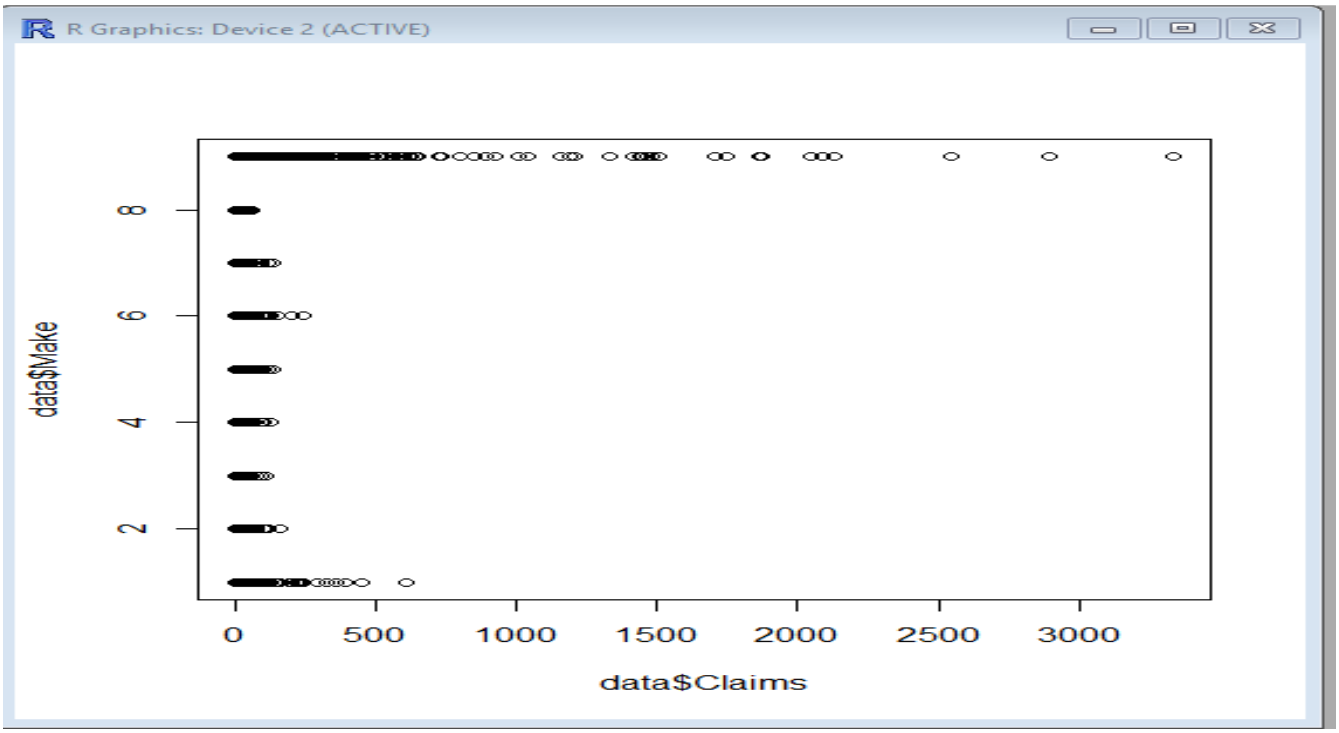


Figure 11Relationship between claims and Make

As we see the figure, it tells that claims and Make are not that much have relationship. And an insurance sector doesn't have to bother to find fraud based on the car make fields.

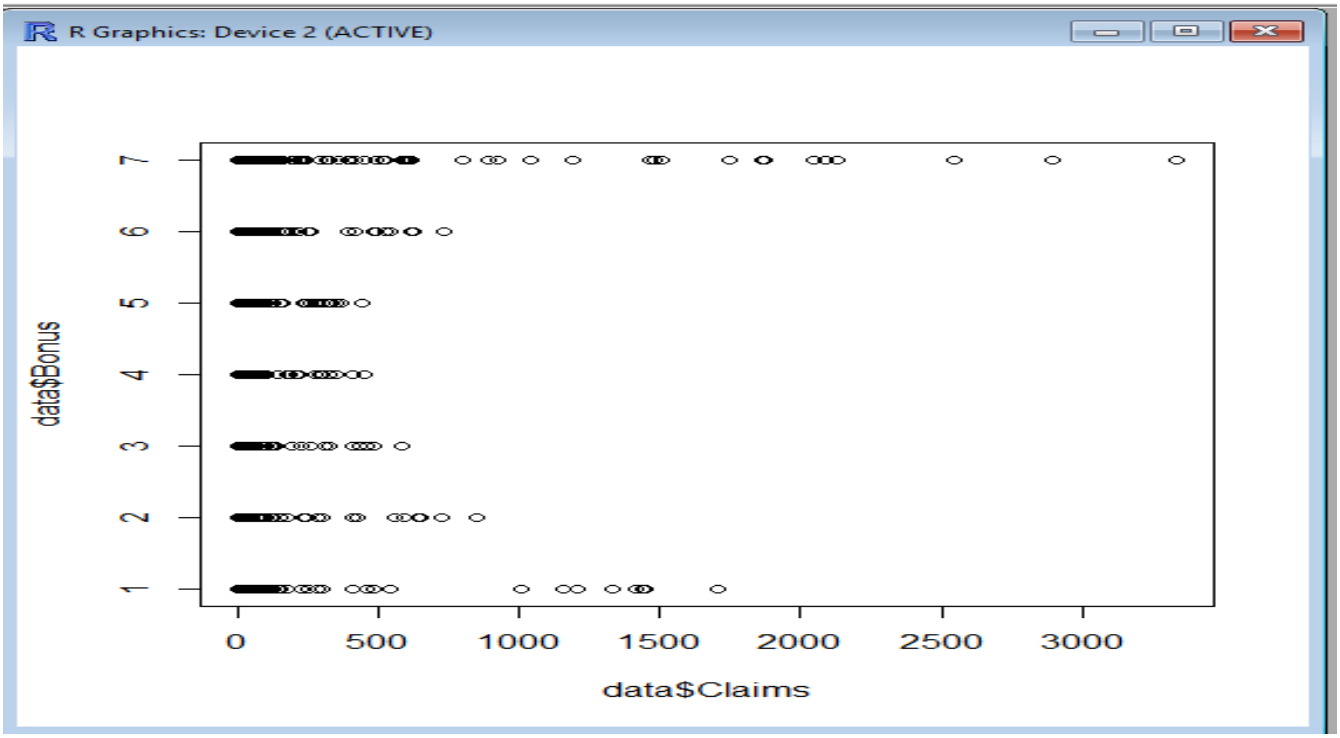


Figure 12Relationship between claims and Bonus

The figure tells that claims and Bonus are not that much have relationship. And an insurance sector doesn't have to bother to find fraud based on the Bonus fields.

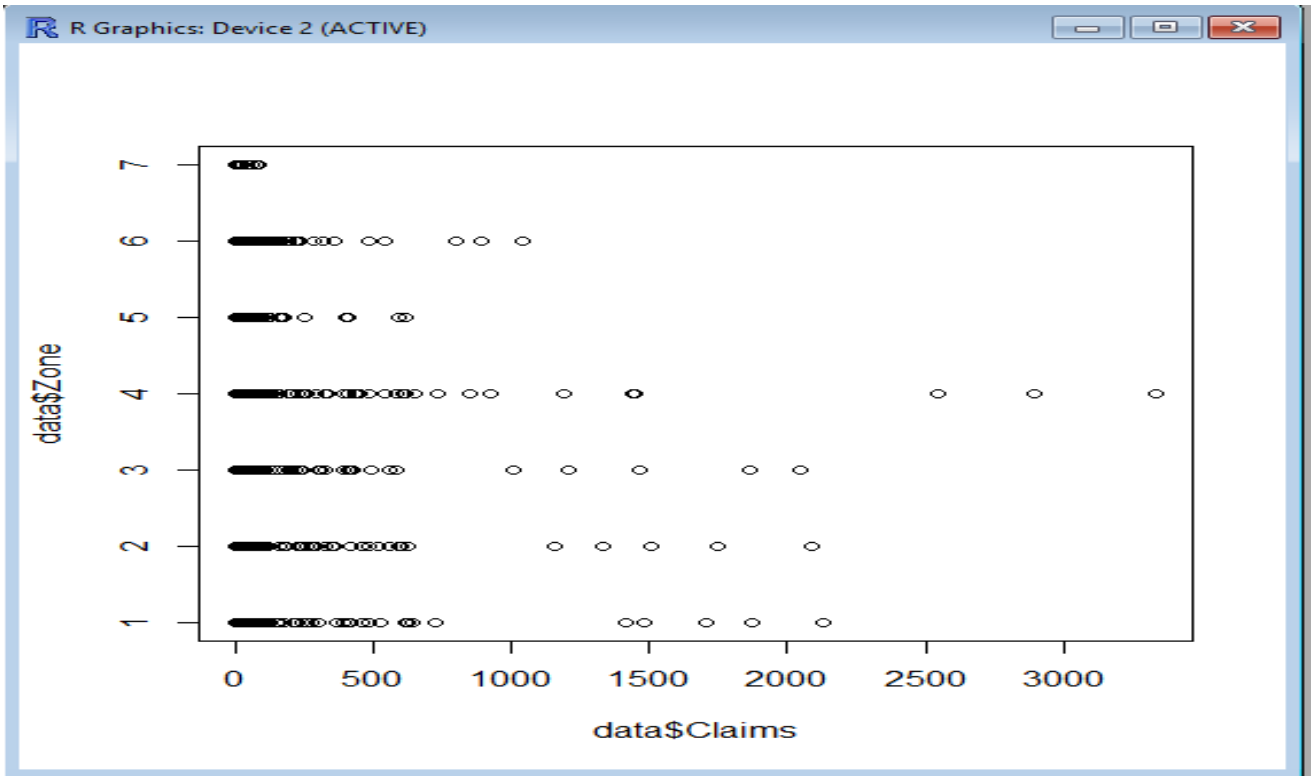


Figure 13Relationship between claims and Zone

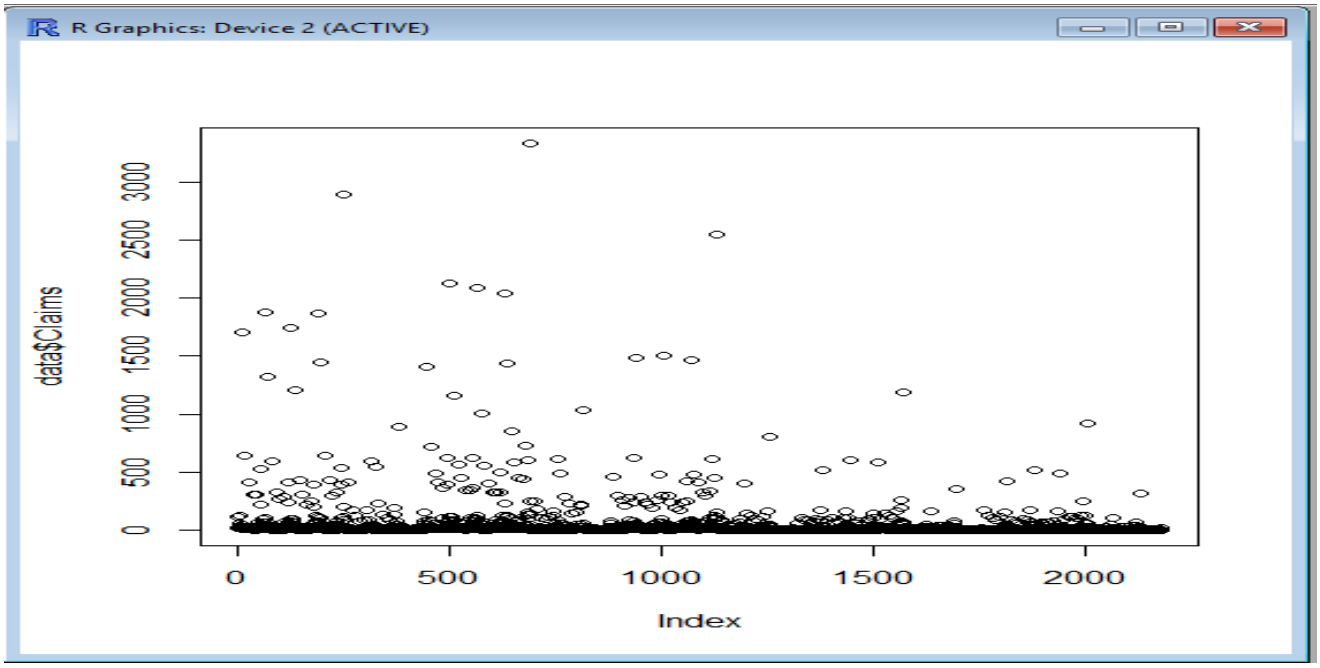


Figure 14Relationship between claims and Kilometers

In general, based on the plot function we decide that the main variable we have to deal to get fraud in the data set are Claim and Insured data field.

When the Insurance Company wants to understand, what affects their claim rates to decide the right premiums for a certain set of situations they only need to find whether the insured amount, zone, kilometer, bonus, or make affects the claim rates and to what extent. Then only have to perform to find the dependency of claim variable by other variables build a linear regression model. As we observe the Experiment V it provides the intercept and estimated value and this in turn shows that all the payment values of independent variables, such as kilometers, zone, bonus, make, and insured are highly significant and are making an impact on the claims. Therefore, to understand the results graphically to understood easily we can plot the relationship between the Claims with all other variables using the *plot ()* function.

4.6 Summary:

As we observed the results that we had perform in the above sections Data Analytics do have a huge number of advantages to understand the thing that's going on in the insurance sector and to have a big insight to make decision.

Having Data Analytics Framework for Fraud Detection in Vehicle Insurance over the traditional one is: -

- It is computerized handling of dataset to make descriptive analytics
- It maintains the quality of data using Preparation of the Data and Data quality assurance techniques
- It easily correlation the relationship between Variables to identify frauds and decide the most important variable to deal with for more analyzing and prediction of decisions.

CHAPTER FIVE

DISCUSSION AND CONCLUSION

This chapter summarizes the thesis, discusses its findings and contributions, points out limitations of the current work, and also outlines directions for future research. The recognition and interpretation of Data Analysis using R programming from the Hadoop ecosystem then understand the dataset to do descriptive and predictive analysis decision to support the insurance sector. However, still many extensions of this research deserve further consideration.

5.1. Summary of the thesis

The application of Big Data Technology has increasingly become very popular and proved to be relevant for many sectors such as insurance, airline, telecommunications, banking, and healthcare industries. Particularly in the insurance industry, DM technology has been applied for fraud detection but it is not enough to assure accuracy so we have to use Big Data Technology to implement and handle fraud activity and display descriptive and predictive analysis that helps to analysis all the data from different sources without bothering of the data format. As a matter of fact, insurance fraud is the most challenging problem in today's motor insurance business.

In this research, an attempt has been made to apply the analytics technology in support of detecting and predicting fraudulent insurance claims in the insurance industry.

This paper is Unique among other fraud detection in vehicle insurance which is developed: the designed Possible framework, the way data collection, data format and data sources and visualization including the

5.2. Contribution and limitations of the current work

- As a summary of Findings/Results:

5.2.1 If the insurance company is interested to know each fields of the dataset. We can work on Descriptive Analysis: used to get basic insights in to the data set for further analysis. That defined what happened; Because of the dataset is large, we have to know the pattern of the data- called the distribution of the data using summary statistics.

- From the Results of the Descriptive Analysis: we see the Summary (data) provide MIN and MAX of each Variables or columns.

5.2.2 Insurance company can monitor the total value of Payment from the Descriptive Analysis that relates Payment with the number of Claims and the number of Insured year using Correlation Function as we see in the chapter4 experiment II and visualized using Scatter Plot.

5.2.3The Insurance Company can figure out reasons for how insurance Payment increase and decrease, and they can decide which variable is more affecting the Payment among distance, location, bonus, make, and insured amount or claims or all or some of them are affecting it. That is relationship between Dependent variables (Payment) and Independent variables can be solved using linear regression function lm (): -the output the Payment will increase directly proportional to the number of Claims.

5.2.4 To answer how the insurance company is planning to detect fraud we had to find at what location, kilometer, and bonus level their insured amount, claims, and payment get increased. Then the insurance company will establish a new branch office to detect fraud, at that area. To perform this, we can perform to do an Aggregate Dataset in order to find the mean value of insured, payment, and claims based on zone, kilometer, and bonus variables, group all the result variables based on individual categorical variables.

As we can see in chapter 4, Experiment IV output we get the following observations:

V. Zone 4 has the highest number of claims, and thus payment as well.
VI. Zones 1-4 have more insured years, claims, and payments.
VII. Kilometer group 2 has the maximum payments. Though the insured number of years is lesser than kilometer 1, the claims and payments are higher for group 2.
VIII. There is not much variation in groups of bonus except for 7 with unusually high number of insured years, claims, and payments.

5.2.5 When the Insurance Company wants to understand what affects their claim rates to decide the right premiums for a certain set of situations they only need to find whether the insured amount, zone, kilometer, bonus, or make affects the claim rates and to what extent. Then only have to perform to find the dependency of claim variable by other variables build a linear regression model.

As we observe the Experiment V in the chapter 4 it provides the intercept and estimated value and this in turn shows that all the payment values of independent variables, such as kilometers, zone, bonus, make, and insured are highly significant and are making an impact on the claims. Therefore, to understand the results graphically to understood easily we can plot the relationship between the Claims with all other variables using the *plot ()* function

- As a limitation of the current work

Our research didn't provide data from social media sources and it doesn't solve other insurance activity rather than the Motor.

5.3. Recommendations for Future Research

This research is mainly conducted for an academic purpose. However, that the results of this study are found promising to be applied to address practical problems of insurance area. This research work can contribute a lot towards solving insurance company problem. The results of this study have also shown that the data Analytics technology particularly R programming and the Hadoop are well applicable in the efforts of insurance fraud detection, data analysis and decision making process. Hence, based on the findings of this study, the following recommendations are forwarded.

For this work, only a limited number of all possible attributes are available with their values in the database of the company. There are inconsistency and missing values in the database. There is no record related to age, sex, health status, driving experience, and marital status of driver and the place where and when the accident has occurred, and traffic police report, which are important fraud factors. Since big data is the most important component in this research, the company has to design a data warehouse where operational and non-operational data can be kept. So:

- Insurance company need to register all information from their clients came from email, face book, twitter and other social media networks.
- Insurance company need to register every information from thematic software that tell how many kilometers the vehicle travels the location of the accident, the day the claim done after the accident occurred, bank statement of the client and also drinking habit.
- Then the researcher will use all the above required data and information besides the data we used for this research to the Hadoop ecosystem to minimize the effort to work on the data and easily display the output using either Hive or R programming languages.
- And finally, insurance sectors can use it this research for all insurance activity beside the motor insurance.

BIBLIOGRAPHY

[1] Elena Geanina ULARU, Florina Camelia PUICAN, Anca APOSTU, Manole VELICANU , "Perspectives on Big Dataand Big Data Analytics," *Database Systems Journalvol. III,no. 4* , p. 3, 2012.

[2] t. p. t. k. SAS, "Big Data meets beg data anaytics," *white paper,* p. 2, 2012.

[3] s. i. m. guide, "Insurance," p. 2, december 2014.

[4] S. Pandhare, "Big Data Analytics :new whistleblower on insurance fraud," *Infosys,* p. 8, 2015.

[5] Puneet Bharal and Amir Halfon, "Making sense of Big Data in Insurance," *Jornal of Advanced Analytics IQ,* p. 3, 2013.

[6] Ana-Ramona BOLOGA, Razvan BOLOGA, Alexandra FLOREA, "Big Data and Specific Analysis Methods for," *Database Systems Journal vol. I, no. 1,* p. 10, 2010.

[7] Wikipedia, "wikipedia," 22 05 2014. [Online]. Available: http://en.wikipedia.org/wiki/big_data. [Accessed September 2015].

[8] Jidong Chen, Ye Tao, Haoran Wang, Tao Chen, "Big data based fraud risk management at Alibaba," *Ke Ai- The Journal of Finance and Data Science,* p. 10, 2015.

[9] Techopedia, "Techopedia Inc," 2016. [Online]. Available: www. Techopedia.html. [Accessed 18 may 2016].

[10] M. Rouse, "WhatIs.com," 2016. [Online]. Available: www.WhatIs.com. [Accessed 18 may 2016].

[11] J. B. D. VALE, "Using Data Mining to Predict Automobile Insurance," *MSc thesis,* p. 45, 2012.

[12] H.Lookman Sithic, T.Balasubramanian, "Survey of Insurance Fraud Detection Using Data," *International Journal of Innovative Technology and Exploring Engineering (IJITEE),* p. 4, 2013.

[13] S. A. Mongeau, "CONTINUOUS FRAUD MONITORING AND DETECTION VIA ADVANCED ANALYTICS," in *24th Annual ACFE Global Fraud Conference* , Amsterdam, Netherlands , March 24th, 2014.

[14] Lovro Šubelj , Štefan Furlan, Marko Bajec, "An expert system for detecting automobile insurance fraud using Social Network Analysis," *ELSEVIER,* p. 14, 2010.

[15] J. P. Harrison, "Chapter5: Strategic Planning and SWOT Analysis," in *Essentials of Strategic Planning in Healthcare*, Health Administration Press, 2010, 2010, p. 7.

[16] Ruchi Verma, Sathyan Ramakrishina Mani, "Using analytics for Insurance Fraud Detection," p. 10.

[17] S. IQ, "Capitalizing on Big Data Analytics," *white paper,* p. 18.

[18] T. ADANE, "MINING INSURANCE DATA FOR FRAUD DETECTION: THE CASE OF AFRICA INSURANCE SHARE COMPANY," in *MINING INSURANCE DATA FOR FRAUD DETECTION: THE CASE OF AFRICA INSURANCE SHARE COMPANY*, Addis Abaab, Addis Ababa University, 2011 e.c, p. 123.

[19] Ana-Ramona BOLOGA, Razvan BOLOGA, Alexandra FLOREA, "Big Data and Specific Analysis Methods for Insurance Fraud Detection," *Big Data and Specific Analysis Methods for Insurance Fraud Detection,* p. 10.

[20] e. o. S. Company, "https://www.esurance.com/," [Online]. Available: https://www.esurance.com/. [Accessed Dec 2015].

[21] H. Brink, "VALIDITY AND RELIABILITY IN QUALITATIVE RESEARCH," in *Curationis,*, 1993.

[22] Chun-Wei Tsa, Chin-Feng Lai, Han-Chieh Chao and Athanasios V. Vasilakos, "Big data analytics: a survey," *Springer -Journal of Big Datal,* p. 32, 2015.

[23] R. Tutorial, "R Tutorial," 13 July 2016. [Online]. Available: http://www.tutorialspoint.com/r/index.htm.

[24] Andrie de Vries, Joris Meys , "R For Dummies," 13 July 2016. [Online]. Available: http://www.dummies.com/store/product/R-For-Dummies.productCd-1119962846.html.

[25] S. Special Interest Group, "Big Data Analytics," Adama, 2016.

[26] SAS, "http://www.sas.com/en_us/software/sas-hadoop.html," [Online]. Available: http://www.sas.com/en_us/software/sas-hadoop.html. [Accessed August 2016].

[27] H. Tutorial:, "Welcome to Apache™ Hadoop®!," [Online]. Available: www.hadoop/Welcome to Apache™ Hadoop®!.htm. [Accessed 16 september 2016].

[28] ComputerWeekly.com. [Online]. Available: www.hadoop/Big data storage Hadoop storage basics.htm. [Accessed 12 July 2016].

[29] H. Tutorial, "Hadoop Tutorial," [Online]. Available: https://www.tutorialspoint.com/hadoop/hadoop_mapreduce.htm. [Accessed 28 october 2016].

[30] S. IQ, "Capitalizing on Big Data Analytics," *white paper,* p. 18.

[31] U. Regent, Proposal perepartion and submission module 2, UC Regent, 2008-2009.

[32] Nawsher Khan,Ibrar Yaqoob,Ibrahim Abaker Targio Hashem,Zakira Inayat,Waleed Kamaleldin Mahmoud Ali,Muhammad Alam,Muhammad Shiraz,and Abdullah Gani, "Big Data: Survey, Technologies, Opportunities, and Challenges," *Scientific World Journal,* p. 18, 2014.

[33] S. t. p. t. know, "Big Data Meets Big Data Analytics," *white paper,* p. 2, 2012.

[34] S. Whitepaper, "Capitalizing on Big Data Analytics," *StackIQ White pape,* p. 18.

[35] "Wikipedia," Jan 2014. [Online].

[36] D. Kumar, "Introduction to Big Dat Analytics," in *SIG for MSc program*, Nazerte, adama, 2016.

[37] T. &. F. LLC, "ITPerformanceImprovement," 2008—2014 . [Online]. Available: http://www.ittoday.info/ITPerformanceImprovement/Articles/2014-07Raghupathi.html. [Accessed 26 may 2016].

[38] Yuri Demchenko, Canh Ngo, Peter Membrey, "Architecture Framework and Components for the Big Data Ecosystem," *System and Network Engineering,* p. 31, 2013.

[39] J. Ruotolo, "Coalition Against Insurance Fraud," 18 June 2015. [Online]. Available: http://www.Articles on insurance fraud.htm.

[40] J. Zhu, "Data Modeling for Big Data," *Data Modeling for Big Data,* p. 7, 2012.

[41] qlikview.com, "qlikview.com," *qlikview.com,* p. 10, june 2012.

APPENDICES

Appendix-A

Questionnaire

RESEARCH: A BIG DATA ANALYTICS FRAMEWORK FOR FRAUD DETECTION IN VEHICLE INSURANCE:

Objective:

The objective of this questionnaire is to collect data concerning the study of big data analytics based solution to vehicle insurance industry. The study mainly focuses on improving predictive analytics methods on vehicles insurance industry data and helps in finding Fraud claim and suggesting the fair pricing solutions using big data tools and technologies.

*Code:*_____*Date:*_____
Position: _____*Phone:* _____

==========

1. *How many vehicles are registered totally to your insurance company?*
   _____
2. *How many new vehicles are registered to your office per year?*
   _____
3. *How many vehicles are insured per a year?*
   _____
4. *Do you have any database? If yes, what is the SW?*
   _____
   _____
5. *Is your database structured or not?*
   _____
6. *What kinds of claim data are available in your organization?*
   _____
   _____
   _____
7. *What are your mining techniques to investigate your client claim?*
   _____
   _____
   _____
8. *What are your main entities in your claim investigation?*
   _____
   _____
   _____
   _____

9. *Which one is the main fraudster entity?*

_____

10. *What are the procedures you follow to find fraud?*

_____

_____

_____

11. *What testing tools have you used for detecting fraud claim?*

_____

_____

_____

12. *What kind of failures you have experienced?*

_____

_____

13. *What are the main sources of data for your Database?*

_____

_____

14. *What is the most available (repeated) insurance claim came to you?*

_____

15. *Do you have any Incident Investigation form?*

_____

_____

16. *Do you think your system have a drawback? Yes / No can you explain it?*

_____

_____

17. *Does your organization use the concepts of big data?*

_____

_____

18. *Are using Big Data technologies for insurance data analysis? If so, when did you start?*

_____

_____

19.. *Are you using any Predictive System for identifying profit and loss?*

_____

_____

20. *Give details on server and client systems configurations for your insurance system in your organization.*

A. Hardware Specifications:

i. Processor's Model & Speed: _____

ii. Hard Disk's Model & Size: _____

iii. RAM Size: _____

iv. Cache: _____

B. Software Specifications:

i. Operating System: _____

ii. Programming Languages: _____

iii. Application Software: _____

iv. Performance Monitoring Software: _____

21. How often you find that current system hardware should be upgrade to improve performance of insurance industry? (Please tick right option)

Never/Often/Continuous

Thank You for Your Help!!

Abraham Worku

Appendix-B

Big Data:

'Big Data' is similar to 'small data', but bigger in size but having data bigger it requires different approaches Techniques, tools and architecture an aim to solve new problems or old problems in a better way Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

We gather Big Data Everywhere: Lots of data is being collected and warehoused such as Web data, e-commerce purchases at department/grocery stores, Bank/Credit Card transactions, Social Network Many more…

Big Data has many Characteristics and according to different scholars its divided: the 3 V, The 5 V or The 7 V's:

- Volume: large amounts of data Zeta bytes/Massive datasets
- Velocity: data comes in many different forms from diverse sources
- Variety: Data is live streaming
- Value: data alone is not enough; how can value be
- Veracity: can we trust the data? How accurate is it?
- Validity: ensure that the interpreted data is sound
- Visibility: data from diverse sources need to be stitched together

Real-time big data isn't just a process for storing Petabytes or Exabyte of data in a data warehouse, It's about the ability to make better decisions and take meaningful actions at the right time. Fast forward to the present and technologies like **Hadoop** give you the scale and flexibility to store data before you know how you are going to process it. Technologies such as **MapReduce, Hive and Impala** enable you to run queries without changing the data structures underneath. Our newest research finds that organizations are using big data to instantly target customer-centric outcomes, tap into internal data and build a better information ecosystem. Big Data is already an important part of the billion database and data analytics market.[25]

Appendix-C

Hadoop Technology

What is Hadoop is Open source software framework designed for storage and processing of large scale data on clusters of commodity hardware, created by Doug Cutting and Mike Carafella in 2005: Cutting named the program after his son's toy elephant.

As the World Wide Web grew in the late 1900s and early 2000s, search engines and indexes were created to help locate relevant information amid the text-based content. In the early years, search results were returned by humans. But as the web grew from dozens to millions of pages, automation was needed. Web crawlers were created, many as university-led research projects, and search engine start-ups took off (Yahoo, AltaVista, etc.).

Why is Hadoop important?

- **Ability to store and process huge amounts of any kind of data, quickly.** With data volumes and varieties constantly increasing, especially from social media and the Internet of Things (IoT), that's a key consideration.
- **Computing power.** Hadoop's distributed computing model processes big data fast. The more computing nodes you use the more processing power you have.
- **Fault tolerance.** Data and application processing are protected against hardware failure. If a node goes down, jobs are automatically redirected to other nodes to make sure the distributed computing does not fail. Multiple copies of all data are stored automatically.
- **Flexibility.** Unlike traditional relational databases, you don't have to preprocess data before storing it. You can store as much data as you want and decide how to use it later. That includes unstructured data like text, images and videos.
- **Low cost.** The open-source framework is free and uses commodity hardware to store large quantities of data.
- **Scalability.** You can easily grow your system to handle more data simply by adding nodes. Little administration is required.[26]

Apache Hadoop:

Apache Hadoop: - Hadoop Distributed File System (HDFS) and Map Reduce is a fast-growing big-data processing platform defined as an open source software project that enables the distributed processing of large data sets across clusters of commodity servers for analyzing data collected.

The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing. The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to

deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

- **Hadoop Common**: The common utilities that support the other Hadoop modules.
- **Hadoop Distributed File System (HDFS)**: A distributed file system that provides high-throughput access to application data.
- **Hadoop YARN**: A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce**: A YARN-based system for parallel processing of large data sets. [27]

Hadoop is a highly scalable analytics platform for processing large volumes of structured and unstructured data. By large scale, we mean multiple petabytes of data spread across hundreds or thousands of physical storage servers or nodes.

The Map step inputs data and breaks it down for processing across nodes within a Hadoop instance. These “worker” nodes may in turn break the data down further for processing. In the Reduce step, the processed data is then collected back together and assembled into a format based on the original query being performed.

To cope with truly massive-scale data analysis, Hadoop’s developers implemented scale-out architecture, based on many low-cost physical servers with distributed processing of data queries during the Map operation. Their logic was to enable a Hadoop system capable of processing many parts of a query in parallel to reduce execution times as much as possible. This can be contrasted with legacy-structured database design that looks to scale up within a single server by using faster processors, more memory and fast shared storage.

To implement the data storage layer, Hadoop uses a feature known as HDFS or the Hadoop Distributed File System. HDFS is not a file system in the traditional sense and isn’t usually directly mounted for a user to view (although there are some tools available to achieve this), which can sometimes make the concept difficult to understand; it’s perhaps better to think of it simply as a Hadoop data store.

HDFS instances are divided into two components: the name node, which maintains metadata to track the placement of physical data across the Hadoop instance and data nodes, which actually store the data.

You can run multiple logical data nodes on a single server, but a typical implementation will run only one per server across an instance. HDFS supports a single file system name space, which stores data in a traditional hierarchical format of directories and files. Across an instance, data is divided into 64MB chunks that are triple-mirrored across the cluster to provide resiliency. Obviously in very large Hadoop clusters, component or even entire server failure will occur so the duplication of data across many servers is a key design requirement of HDFS.

Supporting Hadoop on shared storage doesn't work in the traditional sense, as workload and storage distribution are inherent to Hadoop. However, we are seeing storage supplier products that support Hadoop natively and they point the way to a likely future direction in storage of large-scale distributed architectures.[28]

HDFS:

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. HDFS relaxes a few POSIX requirements to enable streaming access to file system data. HDFS was originally built as infrastructure for the Apache Nutch web search engine project. HDFS is part of the Apache Hadoop Core project.

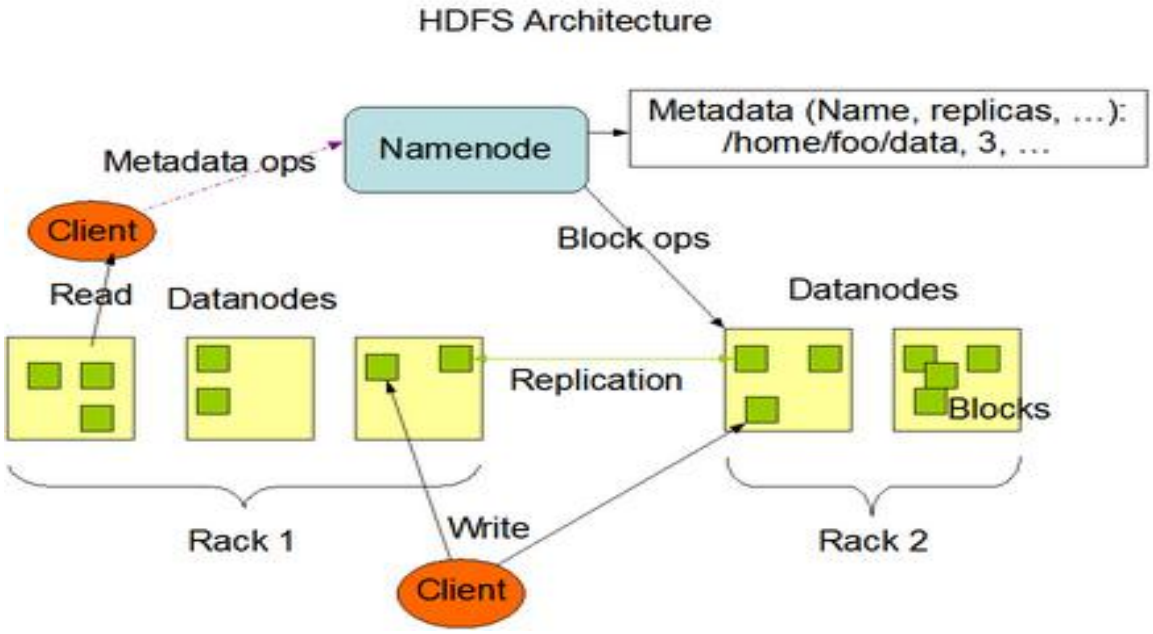


Figure HDFS diagram

MapReduce:

MapReduce is a framework using which we can write applications to process huge amounts of data, in parallel, on large clusters of commodity hardware in a reliable manner.

MapReduce is a processing technique and a program model for distributed computing based on java. The MapReduce algorithm contains two important tasks, namely Map and Reduce. Map takes a set of data and converts it into another set of data, where individual elements are broken down into tupls (key/value pairs). Secondly, reduce task, which takes

the output from a map as an input and combines those data tuples into a smaller set of tuples. As the sequence of the name MapReduce implies, the reduce task is always performed after the map job.

The major advantage of MapReduce is that it is easy to scale data processing over multiple computing nodes. Under the MapReduce model, the data processing primitives are called mappers and reducers. Decomposing a data processing application into mappers and reducers is sometimes nontrivial. But, once we write an application in the MapReduce form, scaling the application to run over hundreds, thousands, or even tens of thousands of machines in a cluster is merely a configuration change. This simple scalability is what has attracted many programmers to use the MapReduce model.

- Generally, MapReduce paradigm is based on sending the computer to where the data resides!
- MapReduce program executes in three stages, namely map stage, shuffle stage, and reduce stage.
    - **Map stage**: The map or mapper's job is to process the input data. Generally, the input data is in the form of file or directory and is stored in the Hadoop file system (HDFS). The input file is passed to the mapper function line by line. The mapper processes the data and creates several small chunks of data.
    - **Reduce stage**: This stage is the combination of the **Shuffle** stage and the **Reduce** stage. The Reducer's job is to process the data that comes from the mapper. After processing, it produces a new set of output, which will be stored in the HDFS.
- During a MapReduce job, Hadoop sends the Map and Reduce tasks to the appropriate servers in the cluster.
- The framework manages all the details of data-passing such as issuing tasks, verifying task completion, and copying data around the cluster between the nodes.
- Most of the computing takes place on nodes with data on local disks that reduces the network traffic.
- After completion of the given tasks, the cluster collects and reduces the data to form an appropriate result, and sends it back to the Hadoop server.[29]

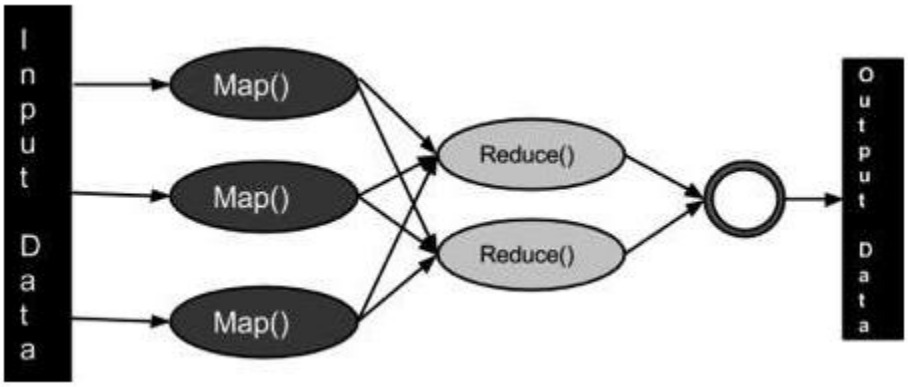


Figure MapReduce diagram

Appendix-D

R-Programming:

R is a programming language and software environment for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand, and is currently developed by the R Development Core Team. The core of R is an interpreted computer language which allows branching and looping as well as modular programming using functions. R allows integration with the procedures written in the C, C++, .Net, Python or FORTRAN languages for efficiency. R is freely available under the GNU General Public License, and precompiled binary versions are provided for various operating systems like Linux, Windows and Mac. R is free software distributed under a GNU style copy left, and an official part of the GNU project called GNU S.

Installing R on Windows from R Studio Website

1.Download R Studio from https://www.rstudio.com/products/rstudio/download/.
2. Run the installation file.
3. Open R Studio

When R starts, it undergoes the following process steps:

- R starts in the working directory, also called the workspace.
- If present, the. Rprofile file's commands are executed.
- If present, the. RData file is loaded.

Some useful codes:

getwd() - return working directory
setwd() - set working directory

Help Command	Function
help.start()	# Load HTML help pages into browser
help(package)	# List help page for "package"
?package	# Display short form for "help(package)"
help.search("keyword")	# Search help pages for "keyword"
?help	# Search for more options
help(package=base)	# List tasks in package "base"

Many data scientist programmers and statisticians use R to design tools for analyzing data and to contribute their codes as pre-assembled collections of functions and objects called packages. Each R package is hosted at http://cran.r-project.org.
Available R packages are listed here:

R Package	Function
library()	# List available packages to load
library("package")	# Load the package
library(help="package")	# List package contents
detach("package:pkg")	# Unload the loaded package "pkg"
install.packages("package")	# Install the package

Explanation of some code used in the chapter four experimentation

>data<-read.CSV("MotorInsurance.csv")

- data is the variable that represent and hold the dataset that have all the unstructured and structured data.
- Read.csv is used to read the data that is MotorInsurance in csv format.

>View(data)

- View function command: provides all the dataset that we want work on it: to visualize what it looks like.

>summary (data)

- Summery function will display the summery of the data set including Min, Median, Max values of all variables in our dataset.

>data $ Claims

- Is code will handle all the data of claim variable. The data is the variable that represents and hold the dataset of the claim.

>cor (data$ claims, data$ Payment)

- Cor is correlation function that used to see the relationship between Claim and Payment variables in the dataset MotorInsurance.

>lreg<-lm (data$Payment~data$Insured+data$Claims+data$Make+data$Bonus+ data$Zone+data$Kilometres)

- linear regression function lm () used to handle and analysis the relationship between Dependent variables (Payment) and Independent variables.

>Plot (data$ Claims, data$ Payment)

- plot function used to graph the relationship between claims and Payment.

Appendix-E

Hive:

Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Hive is a data warehouse infrastructure tool to process structured data in Hadoop. It resides on top of Hadoop to summarize Big Data, and makes querying and analyzing easy. Initially Hive was developed by Facebook, later the Apache Software Foundation took it up and developed it further as an open source under the name Apache Hive. It is used by different companies. For example, Amazon uses it in Amazon Elastic MapReduce.

Hive is not

- A relational database
- A design for OnLine Transaction Processing (OLTP)
- A language for real-time queries and row-level updates

Features of Hive

- It stores schema in a database and processed data into HDFS.
- It is designed for OLAP.
- It provides SQL type language for querying called HiveQL or HQL.
- It is familiar, fast, scalable, and extensible

Hive in the Hadoop Ecosystem the Word Count algorithm, like most that you might implement with Hadoop, is a little involved. When you actually implement such algorithms using the Hadoop Java API, there are even more low-level details you have to manage yourself. It's a job that's only suitable for an experienced Java developer, potentially putting Hadoop out of reach of users who aren't programmers, even when they understand the algorithm they want to use. In fact, many of those low-level details are actually quite repetitive from one job to the next, from low-level chores like wiring together Mappers and Reducers to certain data manipulation constructs, like filtering for just the data you want and performing SQL-like joins on data sets. There's a real opportunity to eliminate reinventing these idioms by letting "higher-level" tools handle them automatically. That's where Hive comes in. It not only provides a familiar programming model for people who know SQL, it also eliminates lots of boilerplate and sometimes-tricky coding you would have to do in Java.