Data Mining for Improving Decision-Making Facility in Vehicle Maintenance Management

By

L. A. M Perera

179473L

Faculty of Information Technology.
University of Moratuwa
October 2020

Data Mining for Improving Decision-Making Facility in Vehicle Maintenance Management

L. A. M Perera

MSC/IT/11/179473L

Master of Science in Information Technology
Faculty of Information Technology
University of Moratuwa
October 2020

Data Mining for Improving Decision-Making Facility in Vehicle Maintenance Management

L. A. M Perera

MSC/IT/11/179473L

Dissertation submitted to the Faculty of Information Technology,
University of Moratuwa, Sri Lanka for the fulfillment of the
requirements of Degree of Master of Science
in Information Technology
October 2020

Declaration

We declare that is our own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from the published or unpublished work of others has been acknowledged in the text and a list of references is given.

Name of Student Signature of Student

L.A.M Perera ……………………….…

Date: ………………….…

Supervised by

Name of Supervisor Signature of Supervisor

Mr. S.C. Premarathne ……………………….…

Date: ………………….…

Acknowledgement

I am sincerely grateful to Senior lecturer Mr. Saminda Premarathne for his guidance, supervision, motivation, advice, and for dedicating his precious time to our project works.

He endorsed my project idea and guided me to identify the right path in the project. Hence I would like to express my prestigious gratitude to Mr. Saminda Premarathne, Senior Lecturer of the University of Moratuwa for his invaluable support and commitment.

Secondly, I would also like to special thank Mohamed Ferdiehouse for providing the basic knowledge for writing the dissertation as well as I express my sincere gratitude to all the lecturers in the MSc degree program.

Finally, I would like to acknowledge with gratitude, to my loving parents, Liyanage Gunapala and Malini; my brothers, Sudarsha and Malith; my sisters, Muditha Ayeshani, Bhagya Gurusinghe, and Thaveesha Gamage, and all my besties of MSC degree program and my Boss, Director Pradeep Kannangara. They kept me going, and this thesis would not have been possible without them.

Abstract

One of the major problems associated with companies that manage large vehicle fleets is how to manage and operates their vehicle fleet more efficiently. However, the company can’t do it properly without a systematic procedure. Having an automated system for updating vehicle data and related operations of the vehicle's maintenance process, is a very effective way to manage the operations of a vehicle fleet. There are several key factors to consider when managing a fleet of vehicles. Decision-makers must consider Vehicle acquisition, Human, Fuel management, Maintenance, Health and Safety, Compliance. Accordingly, the factors of vehicle type, model, fuel usage, driving efficiency, vehicle condition, spare parts, breakdowns and accidents, vehicle repairs, and services should be considered. All of the above factors should be monitored and managed in the best possible manner. Therefore, using an automated vehicle management system to manage fleets, is very essential for a company that manages a large vehicle fleet. Therefore, this research will emphasize how data mining techniques can be used to analyze data related to vehicle maintenance and finding factors that affect the vehicle maintenance cost using the model identified.

Keywords— Vehicle maintenance, Vehicle fleet management, Decision making, Data Mining

Table of Contents

List of Tables

List of Figures

Chapter 01

Introduction

1.1 Prolegomena

Vehicle maintenance is the process of repairing and maintaining a vehicle. The vehicle maintenance process involved several complex activities such as fuel management, spare parts inventory management, vehicle service management, license renewal, testing, accident claims, vehicle break down, and repair management. These operations should be well administered and managed to use efficiently.

Fleet Management is a very significant process specially for companies that use large-scale vehicle fleets. It keeps track of vehicle allocation for daily operations, driver allocation, spare parts allocation, vehicle breakdown management, vehicle repairs, vehicle servicing, and cost management, etc.

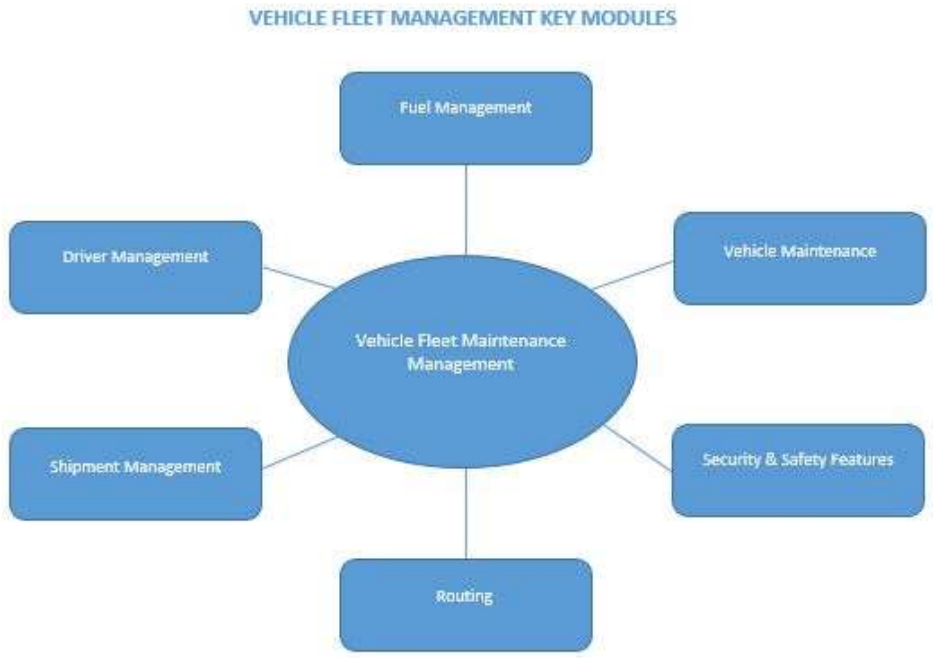

VEHICLE FLEET MANAGEMENT KEY MODULES
Fuel Management
Driver Management
Vehicle Maintenance
Vehicle Fleet Maintenance Management
Shipment Management
Security & Safety Features
Routing

Figure 1.1: Vehicle fleet management key modules

In the past few decades, vehicle maintenance management decisions have been limited to a few areas of vehicle maintenance management. At the present, the combination of the following areas of maintenance (Figure 1.2) integrated with the past role of fleet maintenance management and generated an integrated management process due to the changing role of vehicle fleet maintenance management.

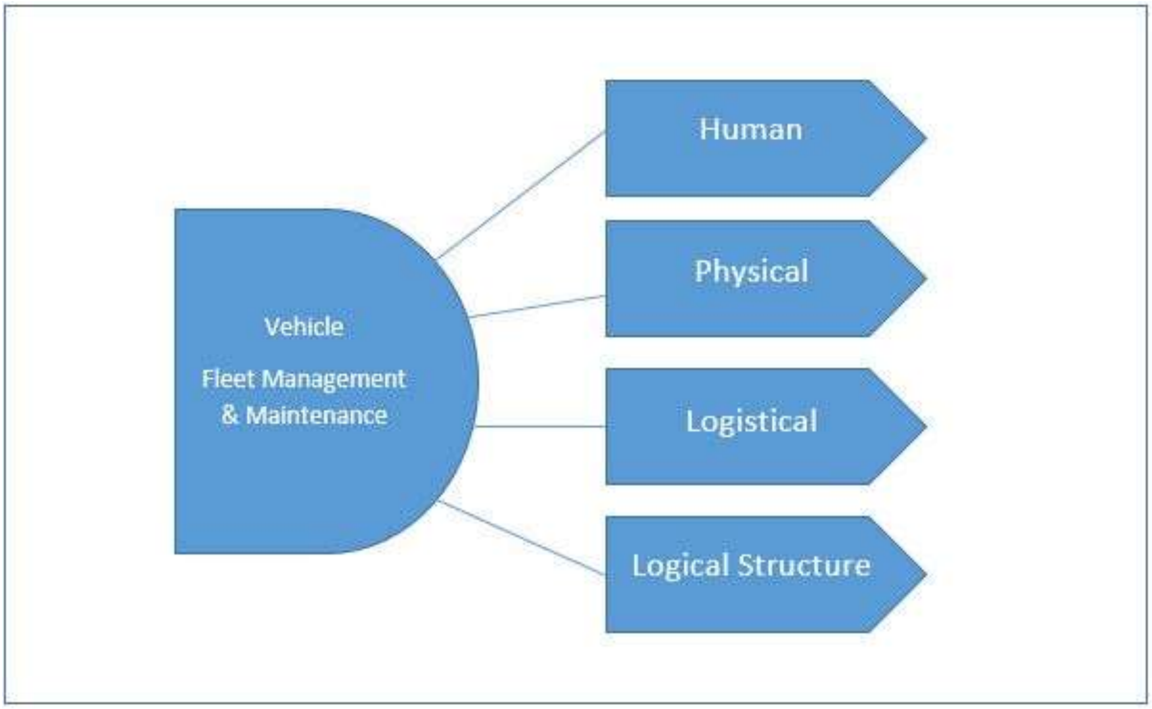


Figure 1.2: Current role of fleet management and maintenance

Although an individual can manage a personal vehicle or several vehicles easily, it is a difficult and responsible task for a company that manages a fleet of vehicles. Studied transport companies most often owned several types of vehicle models based on their operational requirements. The vehicle fleet of the companies which provide the security services is the observed sample in this study. Therefore, due to this complex operation, they have appointed a Fleet Manager to manage the vehicle fleet more efficiently and cost-effectively in organizations that maintain a large fleet of vehicles. An effective fleet management process can support the management to decrease the total cost of vehicle maintenance, improving fleet safety, optimize driver performance, and efficiently manage the vehicle fleet at the operations.

According to the considered sample of this study, the transport process is the core process that makes a profit to the company. Hence it should be conducted to manage associated safety and environmental protection from the maintenance impact. Therefore, maintenance should be managed in a manner that does not impact safety and environmental protection. In order to manage a large group of vehicles, information on the day-to-day operation of the vehicle and details of fuel consumption and repairs must be collected. Most of the companies collect and insert the same data into their database systems, but due to a lack of knowledge on data management, they may not have the data needed to make decisions. Some companies maintain a well-functioning database, but they do not know how to use that data to make decisions. Hence it is more important to conduct a study to identify the factors that influence vehicle maintenance expenses to provide some solutions for the companies that maintain a large vehicle fleet.

The problem to be considered in this study is to determine the relationship between the variables and level of significance and the weight of impact the attributes considered. By conducting this study, the results of this analysis are expected to be used for the fleet managers to increase the profitability of the fleet management and to enhance the efficiency of the fleet. The resulting model and the equation can be used to the managers which factor to focus more on when taking maintenance management decisions on the operations.

1.2 Background of the Study

Most fleet management companies do not have a systematic approach to vehicle maintenance process management. This research focuses mainly to develop an analysis model to provide support in the decision-making process for AB Securitas Privet Limited, one of the leading companies providing security services in Sri Lanka.

There are over 200 vehicles in the fleet of this company and they have no systematic vehicle maintenance system to monitor the vehicles and related operations. A lot of issues could be identified in the current vehicle management process because there is no proper data analysis process. The research aims to support company management in decision making by providing a feasible solution for the vehicle management process by generating accurate reports and it will be a beneficial solution for a company that maintains a huge vehicle fleet

1.3 Problem Statement

When we considering about large vehicle fleet of a company there are so many expenses related to vehicle maintenance. But sometimes some vehicle expenses are different from each other. It is difficult to identify why some of the vehicles have lower maintenance costs while some other are very high. There are hundreds of vehicles on the vehicle fleet and they are running on several types of conditions as well as different manufactured years and various models. As well as the vehicles are assigned for various routes and running on various distances daily.

Because of this, a company that maintains a large number of vehicles should get a clear idea of the factors that affect the maintenance cost of these vehicles. If they can identify the factors which are affecting this high cost, they can be reduced the unnecessary expenses of the process. The main part of this project is identifying the main factors affecting the maintenance cost of a vehicle and how it can be organized a vehicle fleet more profitably.

1.4 Aim and Objectives

Aim

Develop a data analysis model to identify the factors affecting vehicle maintenance expenses and visualizing it using a business intelligence tool for improving decisions making facility in the vehicle maintenance management process is the main purpose of this project.

Objectives

- Finding variables (running distance, manufactured date, model, brand, driver, no of accidents) that are most affected to increase the vehicle maintenance cost of vehicles
- To predict the vehicle maintenance cost of each vehicle based on factors identified by the model
- To manage the vehicle fleet according to the factors identified by the model
- To suggest a solution for manage the vehicle maintenance costs based on research results

1.5 The Proposed Solution

We propose a simulation that uses a historical vehicle maintenance expenses database to predict the most influential variables for vehicle maintenance cost. Accordingly, a service-providing company has been selected and over 20,000 daily records with the fields required for the analysis. A summary is prepared using all daily data and the data is preprocessed using a few techniques. The regression analysis method is used for data analysis and modeling.

1.6 Structure of the Thesis

The overall structure of the project is presented here, the first chapter gives an overview of the project, the objectives, the background of the problem, and the proposed solution. The second chapter will critically review the literature in the data mining technology in factors affecting Vehicle maintenance expenses with specially companies that contain a large vehicle fleet. The third chapter is about data mining technology and the technologies used for the analysis. Forth chapter will present the methodology for finding the factors that affect the vehicle maintenance expenses and the process while the Fifth chapter gives a detailed description of the design and analysis of the research. In the sixth chapter is presented the implementation of the research and an evaluation and conclusion for the project problem at the seventh chapter.

Chapter 02

Review of literature

2.1 Introduction

Critically reviews the existing literature for evaluating the use of data mining techniques to identify the factors that affect vehicle maintenance expenses in this chapter. To identify the factors that affect vehicle maintenance cost, research articles, web pages, and SOPs, and company operation guidance books were reviewed.

2.2 Vehicle Maintenance Procedure in Fleet Management

The process of monitoring whole vehicle fleet performance and maintenance is called fleet management. That is the process of controlling maintenance, efficient vehicle routing with low cost, vehicle tracking, and navigating financing, and replacement [8].

The best vehicle fleet management system helps to achieve the following,

- Reduce the operational costs
- To Increase the productivity
- Guarantee compliance

There are many functions and responsibilities in the vehicle management system [5],

- Vehicle maintenance
- Monitoring fuel consumption
- Driver management
- Asset utilization
- Fuel cost analysis
- Route planning and management
- Productivity optimization
- Waste reduction and much more

Moreover, these are even hard to manage. Companies usually hire a vehicle fleet manager to oversee all these processes. But this is a difficult task and takes a lot of time.

When considering the vehicle maintenance process relevant to the selected company, there is a certain procedure to that operation. After studying and investigating its functionality, the vehicle maintenance expenses can be categorized into the following main fragments [4].

- Emission test and license renewal
- Vehicle Break down repair
- Vehicle Accident repair
- Vehicle regular Services parts replacements

2.3 Issues with Existing Vehicle Fleet Management Procedure in the Transport Services Providing Companies

Most of the companies have not been maintained systematic procedures to follow up the procedure of vehicle maintenance. Companies face challenges in managing fleet management processes.

In this project, a survey was conducted on the fleet management system based on 5 companies providing security services. Based on the details of that study, the general procedure of the companies and the following issues were identified in the vehicle fleet management process.

Normally all type of vehicle services is done through outdoor workshops. If the company has repair facilities, they will be repaired at the company workshop. Accordingly, only the cost of the replacement parts and the other materials are added to maintenance cost and if repairs are done through outdoor workshops, a certain billing process will add to the vehicle maintenance costs.

There is a separate warehouse for maintaining auto parts. A person is issued parts and lubricants for the vehicles from the stock based on the approved request of the transport manager. Those records of the issuing items are entered into a separate database of the company. Hence, details of the stock items issuing for each vehicle and daily outdoor repairing expenses are entered to separate databases. The structure of the databases is different and does not use standard data formats as well as the two systems are not interconnected. Therefore it is hard to generate complete details of the vehicle maintenance expenses of company vehicles.

All vehicles are maintained by the company's transportation department and each vehicle is inspected by a technician at the end of every day. There is no separate driver assigned to a separate vehicle, there is a non-systematic rotating mode of driver reservation for vehicles.

The non-structured systems and lack of data evaluations methods incur additional costs for the company's vehicle maintenance process. All the companies should maintain the best database management system for the vehicle maintenance process and find the factors that most influence the increase in vehicle maintenance costs.

2.4 Review of the Methods Used in Finding Influential Factors for Vehicle Maintenance Cost

Various Researches conducted shows us the factors that affect the lifetime cost of vehicle ownership and operating costs. Lifetime vehicle operating costs can be divided into fixed and variable costs according to the Victoria transport policy institute [2]. Variable costs are costs that vary with the mileage of the vehicle. For example – maintenance, fuel, repair, can be shown. Fixed costs do not vary with the vehicle mileage. For example – vehicle purchase cost or lease cost, insurance taxes can be shown.

Specifically, the author suggests that for a five-year-old vehicle the capital depreciation would be $0.055 per mile and $0.10 per mile for excess mileage charge for vehicle leases. Also, it is suggested by this research that an average of $0.15 to $0.20 per mile for vehicle operating costs (for example fuel, oil tire wear). For distance-based costs (Maintenance, depreciation, etc.) $0.10 to $0.20 per mile [2].

A comparative analysis on “factors influencing fuel consumption and CO2 emissions of passenger cars in real-world operating conditions” by [7]. The author has used identified and investigated the factors of Auxiliary systems, Aerodynamics, Weather conditions, Driving, Vehicle condition, Operating mass, Occupancy rates, Road condition, Fuel characteristics, Certification tests as the attributes which affect the CO2 emissions and fuel consumption. In this case study, the impact of the factors for CO2 emissions and fuel consumption has been analyzed deeply with a technical aspect using external and internal factors. The effect of the factors for the dependent variable has been analyzed one by one using descriptive analysis techniques and equations. [7]

The report by wheels.com [6] has been published a chart including the most significant factors that ultimately affect fleet costs. Fuel, depreciation, maintenance, lease funding, and fees, tax and registration, safety, and collision have been addressed as the major attributes of the study. The report has been provided a general assessment of the potential impact each factor has on drivers. The most significant factor is depreciation, accounting for 45% of the total cost of vehicle ownership. The second most important factor is fuel, accounting for 33% [Figure 2.1]. Although this analysis does not reflect the characteristics of the vehicle and other external variables, as well as the methodology used, has not been included in the report [6].

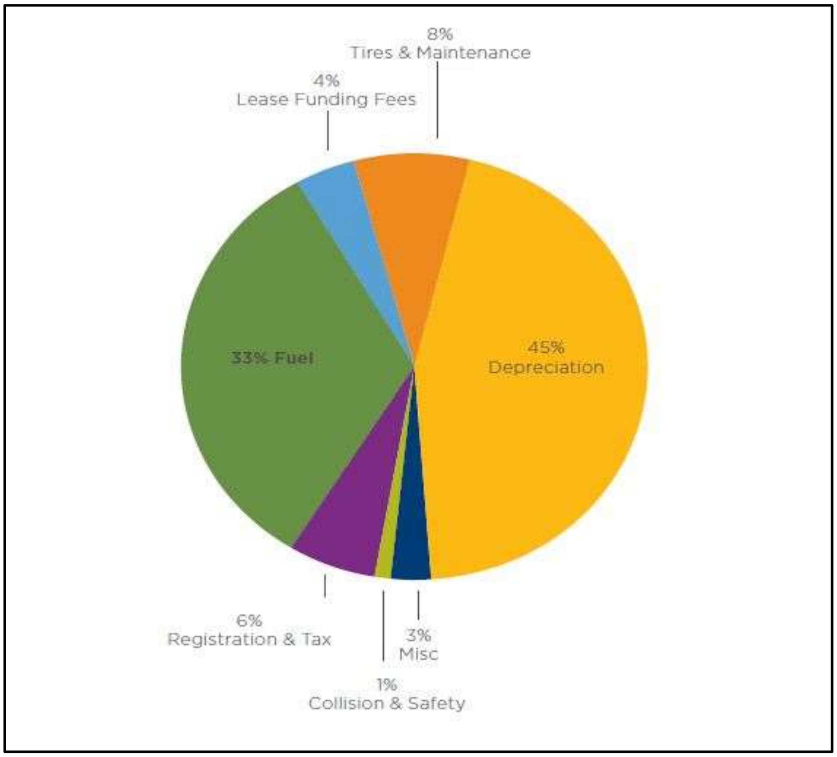


Figure 2.1: Factors affecting the vehicles maintenance cost

Author [1] has been developed a tool for operating a vehicle from buying to removal. The purchased price, financing, fuel, maintenance, insurance cost have been considered as the main attributes. The tool has been developed to calculate to cost of each factor and data for use in individualized scenarios. The output of the research is a developed tool and it only considered the factors which contain direct costing [1].

2.5 Research Problem Identification

Most security services companies do not use a systematic procedure or automation system to manage the company's fleet. They have separate data sources and are not significantly linked. Therefore, many companies are unaware of the major factors that have contributed to the increase in vehicle maintenance costs. Many security companies maintain a large number of vehicles and perform various day-to-day operations at the request of customers. Hence it is difficult to manage a large vehicle fleet with these complex operations. They must first maintain a clear and structured database of vehicle maintenance in their day-to-day operations. Then it helps to easily find the factors that increase the cost of vehicle maintenance.

In considering the literature review, many authors have conducted surveys based on both fixed and variable costs of vehicle maintenance. But in this study, only the variable cost of vehicles was considered. Some studies have been conducted based on specific segments or external factors. Some research has been done based on the industrial side of the vehicle. Hence according to this gap, this research is being done based on all the variables that affect vehicle maintenance expenses and analyzes to find the key factors affecting vehicle maintenance cost.

2.6 Summary

In considering the overall summary of the literature review, the all research of [7], [6], and [1] authors have been considered the effect of the factors for vehicle maintenance cost. But all are considered a special area for each study and research [6] has been shown the output of the research without describing the methodology used. As well as it has not been considered the attributes of characteristics of the vehicle. Because of that in this study expect to do a comprehensive analysis for finding the most appropriate data mining technique that can be used to identify the factors affect for vehicle maintenance cost.

Chapter 03

Technology Adapted

3.1 Introduction

This chapter discusses which kind of technologies have been used for this project and how they have been applied to the work plan. While the analysis has been identified Regression in data mining is the best technological approach to address the problem. This chapter emphasizes how the selected technology can achieve the targeted goal by evaluating different regression models in the contextual factors affecting vehicle maintenance expense.

3.2 What is Data Mining?

The process of analyzing large amounts of data to find patterns, correlations, and insights is data mining. Association rule mining, classification, clustering, regression, and sequential patterns are the main analytical methods of data mining. The data mining method can be used to examine disordered data and repetitive noise, clean and data preprocess using a variety of data mining tools and techniques. As well analytical methods can be selected based on the objectives of each project and the results of the analysis, therefore, help users make better decisions.

3.3 Reasons for Choosing Regression Analysis to Identify the Factors Affecting Vehicle Maintenance Costs

Data mining can be classified into major five areas including association rule mining, classification, clustering, regression, and sequential patterns.

The simple definition of regression analysis is an analytical method of identifying factors affecting a selected variable or variables. The regression analysis leads to identifying which factors are most influential, which variables can be removed and what is the relationship of each variable, and the influence of the variables.

There are many independent variables in this analysis and need to look at what factors influence the dependent variable on the analysis. There are many relationships among the variables. Variables of different data types. Therefore, those relationships should be considered and used for analysis so as not to affect the accuracy of the results. According to these reasons, regression analysis is the best analytical methodology to use this analysis.

3.4 Supervised Learning Algorithms

Supervised learning algorithms provide a very effective set of functions to process the history data and classify the processed data using learning the data patterns.

3.4.1 Regression Analysis

Regression analysis is used to identify the variables that affect a selected variable and their relationship with each other. When considering the problem of this analysis, it is based on the relationships of each attribute. Find out the factors that affect the cost of vehicle maintenance is the major purpose of this project. Therefore, factors affecting vehicle maintenance costs must first be identified. Then, to identify the correlation of each variable, each of the variables must be considered with the dependent variable. According to this scenario, regression analysis is the ideal analytical technique that can be used to achieve the results of this project. Accordingly, regression analysis has been used for this study.

Following advantages will be provided in the use of regression analysis [12].

- The relationship between the dependent variable and the independent variables can be identified.
- The strength of the effect of the independent variables on the dependent variable can be identified in regression analysis.

Multiple Linear Regression

Multiple regression analysis is one of the most popular modeling techniques. This is the most commonly used method of linear regression analysis. In this technique, one dependent variable and is the continuous data type, one or more independent variables can be used for the analysis and can be categorical or continuous variables, and the line of the regression is linear. The relationship between the dependent variable and independent variable (s) will be indicated by the multiple linear regression analysis using a best fit straight line.

The Equation as follow [11],

$$\boxed{Y= a + b \ast X + e}$$

Y	**-** Value of the dependent variable
a	- Constant value
b	- The slope of the regression line
X	- Value of the independent variable
e	- The error term

Using this equation it is possible to predict the value of the dependent variable when each independent variable takes on different values.

Dependent Variable – Vehicle maintenance expenses (Continues)
Independent Variables - Age (continuous), Branch & Model (Categorical variable)

According to the type of variables and the linear regression line, the simple linear regression analysis has been used as the analytical method of this project.

3.5 MS SQL Server

SQL is a standard language used to store data and to manage data more efficiently by creating tables in a database. SQL helps save space and stops data duplication and helps to retrieve data by running quarries based on the data requester's intent. SQL is supported to insert data, update, delete, and efficiently stored data and it is easier to generate data according to the user requirement. Data security is also high and it is supported to store data efficiently and to save the memory space of the hard disk. Then has been used SQL database for the project to stored data to get data by executing the queries for the analysis [14]. The SQL database for the project is used to store the data and retrieve it by executing queries for analysis.

3.6 IBM SPSS Statistical Package

SPSS is the most commonly used statistical package for data analysis. It is easy to process and enter data, pre-process data and execute the various analysis. Therefore, researchers, research firms as well as dissertation writers, managers, and students widely use this SPSS program for their analysis. SPSS comprised the descriptive statistics, text analysis, regression analysis, bivariate statistics, cluster analysis, factor analysis, R extension, etc.

Data preprocessing also can be done with the same software. Because of the above features and specifications, have been using the SPSS statistical package for the analysis of the project. Due to the advantages of the above features and specifications, the SPSS Statistical Package has been used for project analysis.

3.7 Microsoft Excel

Due to some system problems of the respective company, some data had to be collected manually. Have been collected them using Microsoft Excel and stored them in the SQL server database.

3.8 Summary

According to the compressed of the chapter, it has interpreted the used data mining technic for the analysis and the software used for the model creation and databases used for data storing. The regression analysis has been chosen as the data mining technic for this analysis using the SPSS software. The SQL database management server has been selected to store the data in this research.

Chapter 04

Methodology

4.1 Introduction

This chapter discusses the procedures and applications used to conduct research and the input, data collection, data preprocessing process, data analysis methods and techniques, and how they are used to obtain the final results for the research.

4.2 Hypothesis

Most of the data analysts use regression analysis to identify the factors which affect something in quantitative. When studying the analytical methods of the subject in statistics, the most suitable approach to answering the question of this research is regression analysis. Hence that method has been chosen for this analysis.

H_0 - There is no relationship between the X variables and the Y variable

H_1 -There is a relationship between the X variables and the Y variable

The vehicle maintenance cost is the dependent variable in this research analysis. Finding factors that affect the above dependent variable is the aim of the project. According the above hypothesis can be used to identify the impact of independent variables for the dependent variable.

4.3 Input

The study mainly focused on the vehicle maintenance field and one of the main objectives of the study is finding variables that are mostly affected by increasing the vehicle maintenance cost of a vehicle. Accordingly, the indirect approach of this study is to identify how to reduce the maintenance cost of a vehicle.

The selected sample dataset includes maintenance details of 84 vehicles in the van category for the year 2018. It was prepared by SQL server database by querying the 20,000 data records of the year 2018.

Based on the main objective of the project, the factors which can be affected by vehicle maintenance cost has been identified as follow,

Response variable

Vehicle maintenance cost

Independent Variables

Vehicle Number

Vehicle category

Vehicle model

Engine capacity

Gear mode

Manufacturer of the vehicle

Duration of vehicle used

Running distance

No of Accidents

Main running area of the vehicle

Fuel usage

When considering the variables which are going to use for this study, both quantitative and qualitative variables are available. Although out of them most of the variables can be measuring, ranking, categorizing. Hence those are best for identifying patterns and making generalizations using the quantitative methods.

4.4 Output

In this research, the appropriate regression analysis will be identified based on the number of variables and their data type. After that by the regression analysis, the most accurate model will be selected using all the all variables in the database. The model will then keep the very important variables and remove the unwanted variables from the final model. An equation is finally defined as the output to identify the attributes that most affect vehicle maintenance costs.

4.5 Process

This topic describes the outline of the applications, process of data collecting and the data selecting methods, data preprocessing, and evaluation.

4.5.1 Data Selection

The existing data of the vehicle fleet of the selected company has been considered as the data source for this study. The study is conducted by associating a selected company that is maintaining a large vehicle fleet for its operations. Accordingly, this study uses a secondary data source of existing vehicle maintenance data of daily operations of the fleet. Due to the limitations in the data collecting process, there is a limited number of vehicle maintenance data as we could not use the whole dataset of the company for the analysis. The selected sample dataset includes maintenance details of 84 vehicles in the van category for the year 2018.

Identifying the Required Data

Even though there are several vehicle types such as van, lorry, and motorbike are available at the vehicle fleet, only the van category has been considered for the selected sample. Since there are very few vehicles are available in other categories only the vehicle category was selected because otherwise, it can lead to an incorrect output of the analysis.

When considering the main aim of the project, the factors should be identified which can be affected by vehicle maintenance cost, accordingly first of all it should have to identify the dependent variable and independent variables separately.

After observation of the structure of the process, daily operations of the vehicles, and vehicle maintenance process the related variables have been identified as follows. Accordingly, it supported selecting the exact data needed and sort the attributes required.

- Dependent Variable

Vehicle maintenance cost

- Independent Variables

Internal variables	External variables
Vehicle category	Road traffic
Vehicle model	Road condition
Engine capacity	
Gear mode	
Manufacturer of the vehicle	
Duration of vehicle used	
Running distance	
No of Accidents	
Main running area of the vehicle	
Fuel usage	

Table 4.1 variable selections

The above-identified variables lead to select the most required attributes.

4.5.2 Data Collection

The selected company does not have a strong and well-structured database management system. The daily data of each operation store them in the excel sheets and SQL database. Most of the data sheets have not been linked to each other and those are entered separately. Due to this matter, the data entry officers were guided by the researcher to identify which kind of data was needed for this research, and data was gathered under the requirements of the researcher. The company has provided the data for the years 2018 and 2019 from their existing vehicle maintenance datasheets.

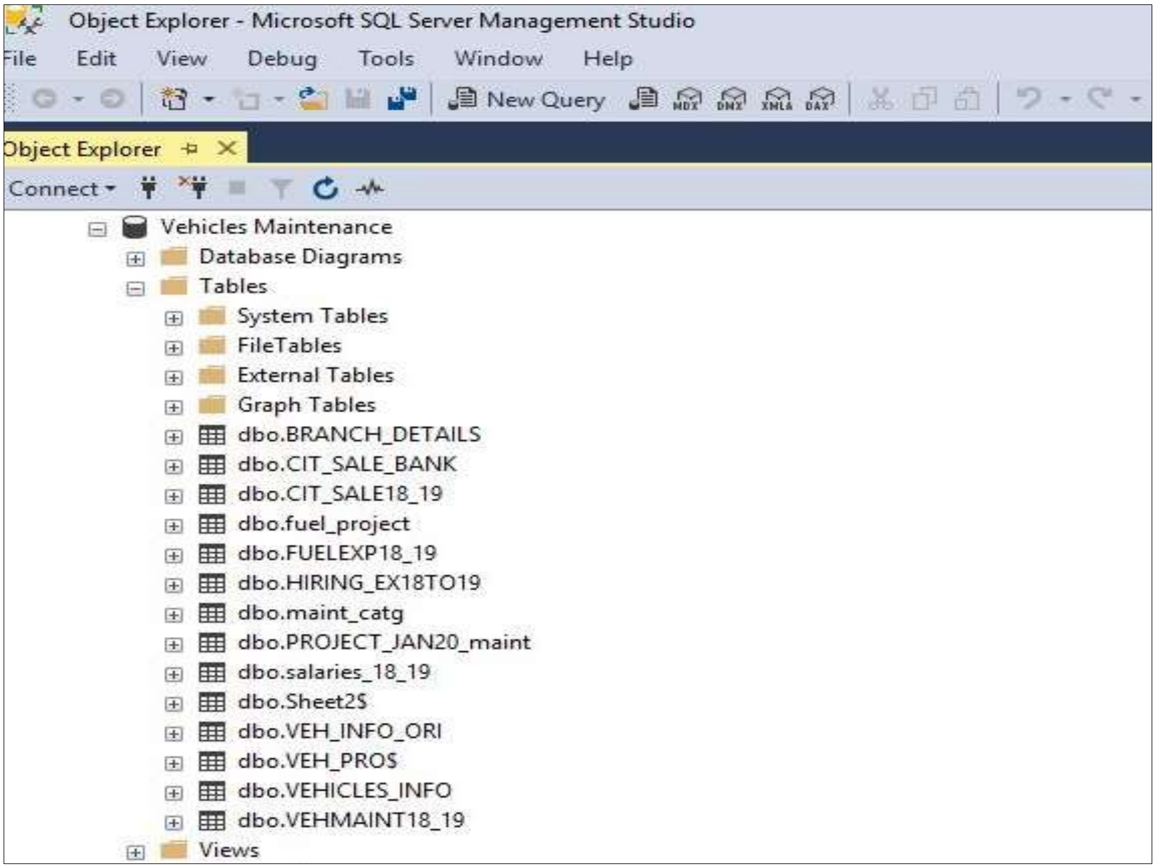
Data sources were gathered by the following individuals

Vehicle details	– Provided by the transport manager
Fuel usage of vehicles	– Data entry officer (transport division)
Vehicle maintenance expenses	– Data entry officer (transport division)
Vehicle Accidents	– CCTV monitoring division (mail details)

Data were available on daily basis and provided by excel sheets. There were 25447 data records are available for the last two years. These data were entered into a SQL database system which contains tables created after analyzing the relationships between the collected data.

4.5.3 Creating Database and Tables

A new database was created to store data collected from the data source. It is contained 7 tables including relevant data.



```
⊟ Vehicles Maintenance
   ⊞ Database Diagrams
   ⊟ Tables
      ⊞ System Tables
      ⊞ FileTables
      ⊞ External Tables
      ⊞ Graph Tables
      ⊞ dbo.BRANCH_DETAILS
      ⊞ dbo.CIT_SALE_BANK
      ⊞ dbo.CIT_SALE18_19
      ⊞ dbo.fuel_project
      ⊞ dbo.FUELEXP18_19
      ⊞ dbo.HIRING_EX18TO19
      ⊞ dbo.maint_catg
      ⊞ dbo.PROJECT_JAN20_maint
      ⊞ dbo.salaries_18_19
      ⊞ dbo.Sheet2$
      ⊞ dbo.VEH_INFO_ORI
      ⊞ dbo.VEH_PRO$
      ⊞ dbo.VEHICLES_INFO
      ⊞ dbo.VEHMAINT18_19
   ⊞ Views
```

Figure 4.1: Table list of vehicle maintenance database

The tables are organized as follow,

Vehicle information

Vehicle_Number - **Primary Key**

Registered_Date

Manufactured_Date

V_type

V_Model

Purchased_Date

Engine_Capacity

Branch

Vehicle_Cost

Vehicle fuel expenses

Vehicle_Number - **Primary Key**

Fuel_Type

Fuel_Liter

Fuel_Cost

Rate

Supplier

Filling_Meter

Distance_Run

Remarks

Vehicle maintenance expenses

Vehicle_Number - **Primary Key**

Date

Bill_Date

Bill_Number

Meter_Reading

Maintenance_Category

Part_Code

Description

Rate

Quantity

Cost

Supplier

Remarks

Vehicle parts master

Part_Code - **Primary Key**

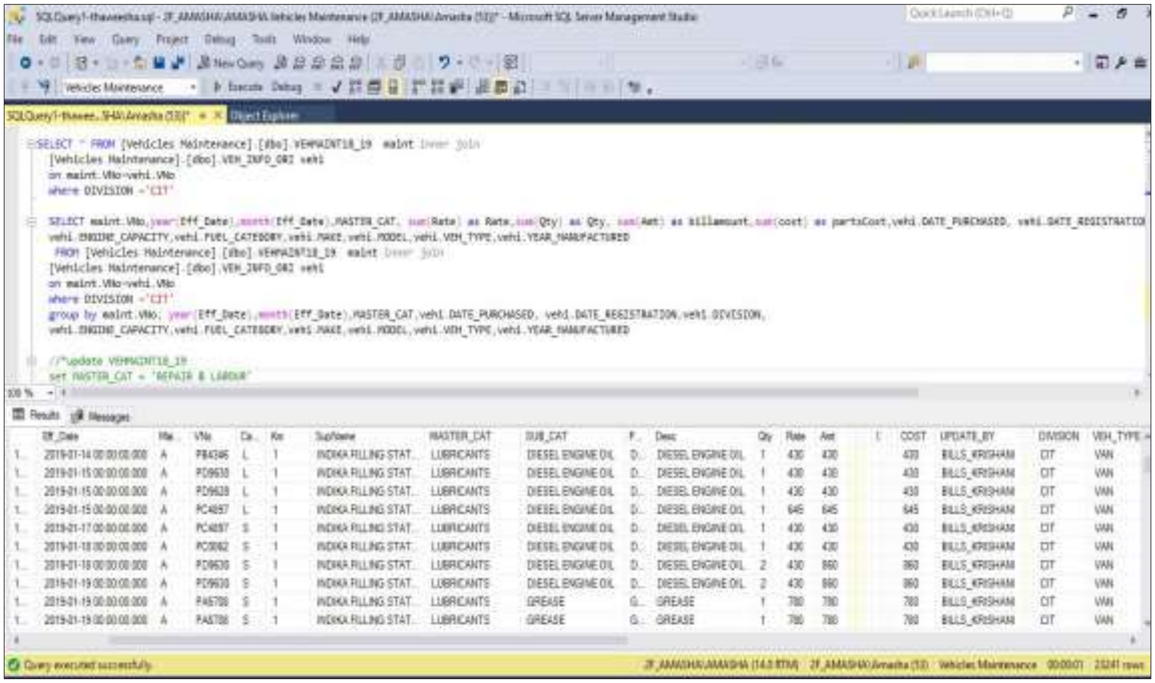
Part_Name

Part_Category

SQL Query for selecting a dataset

```
SELECT *
FROM [Vehicles Maintenance].[dbo].[VEH_PRO$] maint inner join
[Vehicles Maintenance].[dbo].[FUELEXP18_19] vehi on maint.V_No=vehi.V_No and
maint.YEAR=vehi.YEAR and maint.MONTH=vehi.MONTH where DIVISION ='CIT'
```

```
SELECT maint.VNo,year(Eff_Date),month(Eff_Date),MASTER_CAT, sum(Rate) as Rate,sum(Qty) as Qty,
sum(Amt) as billamount,sum(cost) as partsCost,vehi.DATE_PURCHASED,
vehi.DATE_REGISTRATION,vehi.DIVISION,
vehi.ENGINE_CAPACITY,vehi.FUEL_CATEGORY,vehi.MAKE,vehi.MODEL,vehi.VEH_TYPE,vehi.YEAR
_MANUFACTURED FROM [Vehicles Maintenance].[dbo].VEHMAINT18_19  maint inner join
[Vehicles Maintenance].[dbo].VEH_INFO_ORI vehi on maint.VNo=vehi.VNo where DIVISION ='CIT'
group by maint.VNo, year(Eff_Date),month(Eff_Date),MASTER_CAT,vehi.DATE_PURCHASED,
vehi.DATE_REGISTRATION,vehi.DIVISION,
vehi.ENGINE_CAPACITY,vehi.FUEL_CATEGORY,vehi.MAKE,vehi.MODEL,vehi.VEH_TYPE,vehi.YEAR
_MANUFACTURED
```



```sql
SELECT * FROM [Vehicles Maintenance].[dbo].VEHMAINT18_19  maint inner join
  [Vehicles Maintenance].[dbo].VEH_INFO_ORI vehi
  on maint.VNo=vehi.VNo
  where DIVISION ='CIT'

SELECT maint.VNo,year(Eff_Date),month(Eff_Date),MASTER_CAT, sum(Rate) as Rate,sum(Qty) as Qty, sum(Amt) as billamount,sum(cost) as partsCost,vehi.DATE_PURCHASED, vehi.DATE_REGISTRATION
vehi.ENGINE_CAPACITY,vehi.FUEL_CATEGORY,vehi.MAKE,vehi.MODEL,vehi.VEH_TYPE,vehi.YEAR_MANUFACTURED
 FROM [Vehicles Maintenance].[dbo].VEHMAINT18_19  maint inner join
[Vehicles Maintenance].[dbo].VEH_INFO_ORI vehi
on maint.VNo=vehi.VNo
where DIVISION ='CIT'
group by maint.VNo, year(Eff_Date),month(Eff_Date),MASTER_CAT,vehi.DATE_PURCHASED, vehi.DATE_REGISTRATION,vehi.DIVISION,
vehi.ENGINE_CAPACITY,vehi.FUEL_CATEGORY,vehi.MAKE,vehi.MODEL,vehi.VEH_TYPE,vehi.YEAR_MANUFACTURED

//*update VEHMAINT18_19
set MASTER_CAT = 'REPAIR & LABOUR'
```

	Eff_Date	Ma...	VNo	Ca...	Km	SupName	MASTER_CAT	SUB_CAT	F...	Desc	Qty	Rate	Amt	I	COST	UPDATE_BY	DIVISION	VEH_TYPE
1...	2019-01-14 00:00:00.000	A	PB4346	L	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	1	430	430		430	BILLS_KRISHAM	CIT	VAN
1...	2019-01-15 00:00:00.000	A	PD9630	L	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	1	430	430		430	BILLS_KRISHAM	CIT	VAN
1...	2019-01-15 00:00:00.000	A	PD9628	L	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	1	430	430		430	BILLS_KRISHAM	CIT	VAN
1...	2019-01-15 00:00:00.000	A	PC4897	L	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	1	645	645		645	BILLS_KRISHAM	CIT	VAN
1...	2019-01-17 00:00:00.000	A	PC4897	S	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	1	430	430		430	BILLS_KRISHAM	CIT	VAN
1...	2019-01-18 00:00:00.000	A	PC0062	S	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	1	430	430		430	BILLS_KRISHAM	CIT	VAN
1...	2019-01-18 00:00:00.000	A	PD9630	S	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	2	430	860		860	BILLS_KRISHAM	CIT	VAN
1...	2019-01-19 00:00:00.000	A	PD9630	S	1	INDIKA FILLING STAT...	LUBRICANTS	DIESEL ENGINE OIL	D...	DIESEL ENGINE OIL	2	430	860		860	BILLS_KRISHAM	CIT	VAN
1...	2019-01-19 00:00:00.000	A	PA5708	S	1	INDIKA FILLING STAT...	LUBRICANTS	GREASE	G...	GREASE	1	780	780		780	BILLS_KRISHAM	CIT	VAN
1...	2019-01-19 00:00:00.000	A	PA5708	S	1	INDIKA FILLING STAT...	LUBRICANTS	GREASE	G...	GREASE	1	780	780		780	BILLS_KRISHAM	CIT	VAN

Query executed successfully.   2F_AMASHA\AMASHA (14.0 RTM)   2F_AMASHA\Amasha (53)   Vehicles Maintenance   00:00:01   23241 rows

SQLQuery1-thaweesha.sql - 2F_AMASHA\AMASHA.Vehicles Maintenance (2F_AMASHA\Amasha (53))* - Microsoft SQL Server Management Studio
Quick Launch (Ctrl+Q)
File Edit View Query Project Debug Tools Window Help
New Query
Vehicles Maintenance
Execute Debug
SQLQuery1-thawee...SHA\Amasha (53))*
Object Explorer
100 %
Results
Messages

Figure 4.2: Data quarrying in the SQL database

4.5.4 Data Preprocessing

Data preprocessing is a technique of data mining that contains converting raw data into a meaningful and efficient format. Most of the raw data is frequently inconsistent and incomplete, it contains many errors. Raw data is usually incomplete and that data cannot be sent to analysis and cannot fit a model. That causes some errors. Because of that, the data should be needed to preprocess before sending through a model [9]. Therefore, it can be controlled to some extent or totally by using the data preprocessing methods.

Data Pre-Processing Steps

1. Data Cleaning
2. Data Transformation
3. Data Reduction

4.5.4.1 Data Cleaning

When some data is missing in the dataset or there is meaningless data that does not contain an understandable format those Missing data and Noisy data can be handle by different methods of data preprocessing techniques in data mining.

4.5.4.2 Missing Data

At the data collection stage of this study, the requested data provided by the mentioned company by separate data sets and by different data sources. Hence when examining the dataset there was a lot of missing data in the data sets.

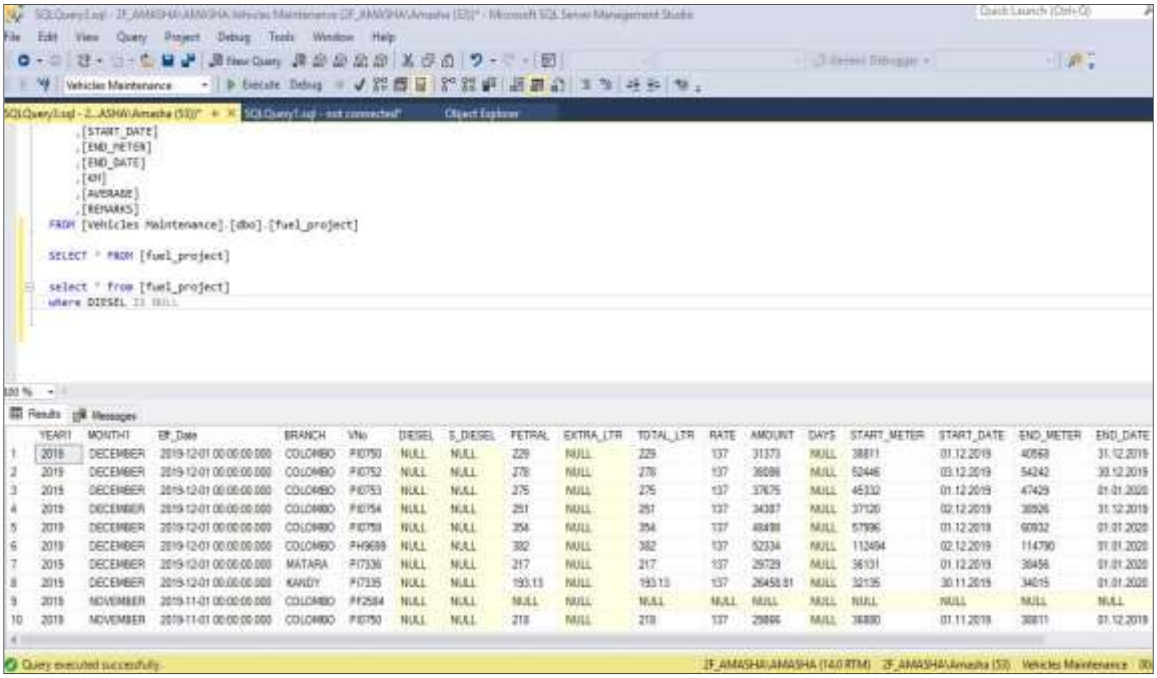
Ignore the Tuples

This method is only used with a large-size data set and when multiple missing values are present in a tuple. When considered the dataset of the study around 20,000 records existed in the database. Hence when the dataset was cleaning removed the records which contained more than one missing value within a tuple.

Checking for Missing Values

Identified the no of missing values of each variable by quarrying the datasets of the SQL database.

```
select * from [fuel_project]
where DIESEL IS NULL
```


SQLQuery1.sql - 2F_AMASHA\AMASHA.Vehicles Maintenance (2F_AMASHA\Amasha (53))* - Microsoft SQL Server Management Studio
File Edit View Query Project Debug Tools Window Help
New Query
Vehicles Maintenance
Execute Debug
SQLQuery1.sql - 2...ASHA\Amasha (53))*
SQLQuery1.sql - not connected*
Object Explorer
100 %
Results
Messages

```
      ,[START_DATE]
      ,[END_METER]
      ,[END_DATE]
      ,[KM]
      ,[AVERAGE]
      ,[REMARKS]
  FROM [Vehicles Maintenance].[dbo].[fuel_project]

  SELECT * FROM [fuel_project]

  select * from [fuel_project]
  where DIESEL IS NULL
```

	YEAR1	MONTH1	EF_Date	BRANCH	VNo	DIESEL	S_DIESEL	PETRAL	EXTRA_LTR	TOTAL_LTR	RATE	AMOUNT	DAYS	START_METER	START_DATE	END_METER	END_DATE
1	2019	DECEMBER	2019-12-01 00:00:00.000	COLOMBO	PI0750	NULL	NULL	229	NULL	229	137	31373	NULL	38811	01.12.2019	40568	31.12.2019
2	2019	DECEMBER	2019-12-01 00:00:00.000	COLOMBO	PI0752	NULL	NULL	278	NULL	278	137	38086	NULL	52446	03.12.2019	54242	30.12.2019
3	2019	DECEMBER	2019-12-01 00:00:00.000	COLOMBO	PI0753	NULL	NULL	275	NULL	275	137	37675	NULL	45332	01.12.2019	47429	01.01.2020
4	2019	DECEMBER	2019-12-01 00:00:00.000	COLOMBO	PI0754	NULL	NULL	251	NULL	251	137	34387	NULL	37120	02.12.2019	38926	31.12.2019
5	2019	DECEMBER	2019-12-01 00:00:00.000	COLOMBO	PI0758	NULL	NULL	354	NULL	354	137	48498	NULL	57996	01.12.2019	60932	01.01.2020
6	2019	DECEMBER	2019-12-01 00:00:00.000	COLOMBO	PH9698	NULL	NULL	382	NULL	382	137	52334	NULL	112494	02.12.2019	114790	01.01.2020
7	2019	DECEMBER	2019-12-01 00:00:00.000	MATARA	PI7336	NULL	NULL	217	NULL	217	137	29729	NULL	36131	01.12.2019	38456	01.01.2020
8	2019	DECEMBER	2019-12-01 00:00:00.000	KANDY	PI7335	NULL	NULL	193.13	NULL	193.13	137	26458.81	NULL	32135	30.11.2019	34015	01.01.2020
9	2019	NOVEMBER	2019-11-01 00:00:00.000	COLOMBO	PF2584	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL	NULL
10	2019	NOVEMBER	2019-11-01 00:00:00.000	COLOMBO	PI0750	NULL	NULL	218	NULL	218	137	29866	NULL	36890	01.11.2019	38811	01.12.2019

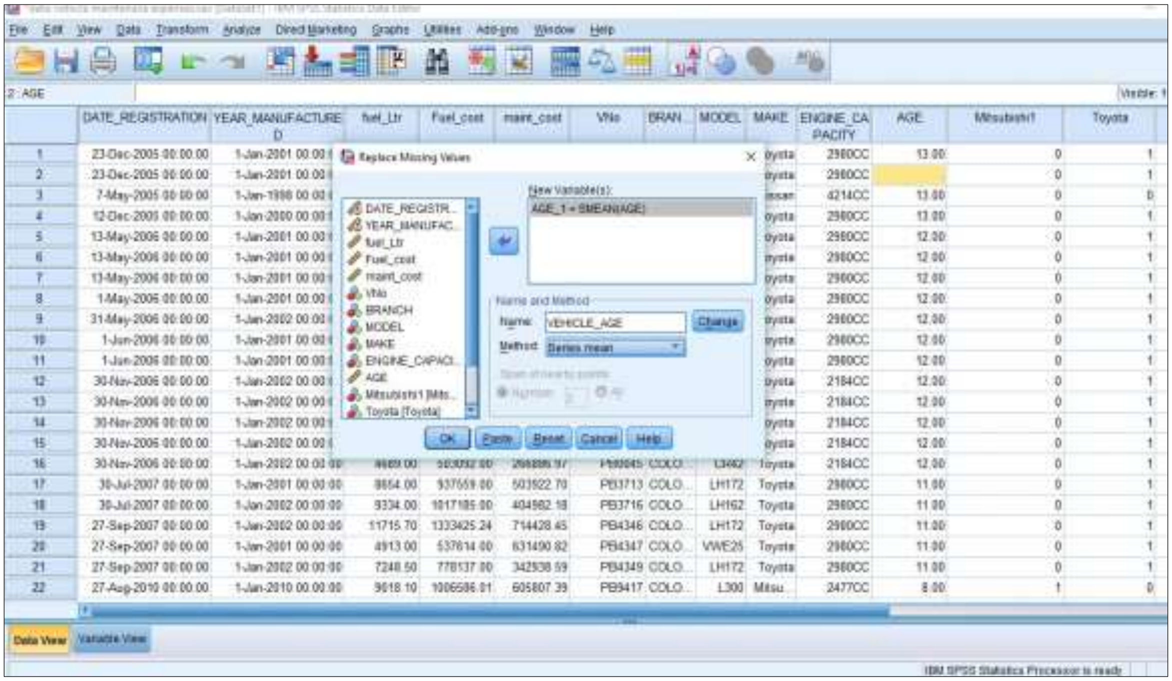
Query executed successfully.

2F_AMASHA\AMASHA (14.0 RTM) 2F_AMASHA\Amasha (53) Vehicles Maintenance

Figure 4.3: Identifying the missing values of the dataset

Replacing Missing Values

By using the dataset of the project, the missing values were replaced by using the mean and mode of the values of the variable.



Replace Missing Values

New Variable(s): AGE_1 = SMEAN(AGE)

Name: VEHICLE_AGE Change

Method: Series mean

	DATE_REGISTRATION	YEAR_MANUFACTURED	fuel_Ltr	Fuel_cost	maint_cost	VNo	BRAN...	MODEL	MAKE	ENGINE_CAPACITY	AGE	Mitsubishi1	Toyota
1	23-Dec-2005 00:00:00	1-Jan-2001 00:00							oyota	2980CC	13.00	0	1
2	23-Dec-2005 00:00:00	1-Jan-2001 00:00							oyota	2980CC		0	1
3	7-May-2005 00:00:00	1-Jan-1998 00:00							ssan	4214CC	13.00	0	0
4	12-Dec-2005 00:00:00	1-Jan-2000 00:00							oyota	2980CC	13.00	0	1
5	13-May-2006 00:00:00	1-Jan-2001 00:00							oyota	2980CC	12.00	0	1
6	13-May-2006 00:00:00	1-Jan-2001 00:00							oyota	2980CC	12.00	0	1
7	13-May-2006 00:00:00	1-Jan-2001 00:00							oyota	2980CC	12.00	0	1
8	1-May-2006 00:00:00	1-Jan-2001 00:00							oyota	2980CC	12.00	0	1
9	31-May-2006 00:00:00	1-Jan-2002 00:00							oyota	2980CC	12.00	0	1
10	1-Jun-2006 00:00:00	1-Jan-2001 00:00							oyota	2980CC	12.00	0	1
11	1-Jun-2006 00:00:00	1-Jan-2001 00:00							oyota	2980CC	12.00	0	1
12	30-Nov-2006 00:00:00	1-Jan-2002 00:00							oyota	2184CC	12.00	0	1
13	30-Nov-2006 00:00:00	1-Jan-2002 00:00							oyota	2184CC	12.00	0	1
14	30-Nov-2006 00:00:00	1-Jan-2002 00:00							oyota	2184CC	12.00	0	1
15	30-Nov-2006 00:00:00	1-Jan-2002 00:00							oyota	2184CC	12.00	0	1
16	30-Nov-2006 00:00:00	1-Jan-2002 00:00:00	4669.00	503092.00	268886.97	PB3045	COLO...	L3442	Toyota	2184CC	12.00	0	1
17	30-Jul-2007 00:00:00	1-Jan-2001 00:00:00	8654.00	937659.00	503922.70	PB3713	COLO...	LH172	Toyota	2980CC	11.00	0	1
18	30-Jul-2007 00:00:00	1-Jan-2002 00:00:00	9334.00	1017185.00	404982.18	PB3716	COLO...	LH162	Toyota	2980CC	11.00	0	1
19	27-Sep-2007 00:00:00	1-Jan-2002 00:00:00	11715.70	1333425.24	714428.45	PB4346	COLO...	LH172	Toyota	2980CC	11.00	0	1
20	27-Sep-2007 00:00:00	1-Jan-2001 00:00:00	4913.00	537614.00	631490.82	PB4347	COLO...	VWE25	Toyota	2980CC	11.00	0	1
21	27-Sep-2007 00:00:00	1-Jan-2002 00:00:00	7248.50	778137.00	342938.59	PB4349	COLO...	LH172	Toyota	2980CC	11.00	0	1
22	27-Aug-2010 00:00:00	1-Jan-2010 00:00:00	9018.10	1006586.01	605807.39	PB9417	COLO...	L300	Mitsu...	2477CC	8.00	1	0

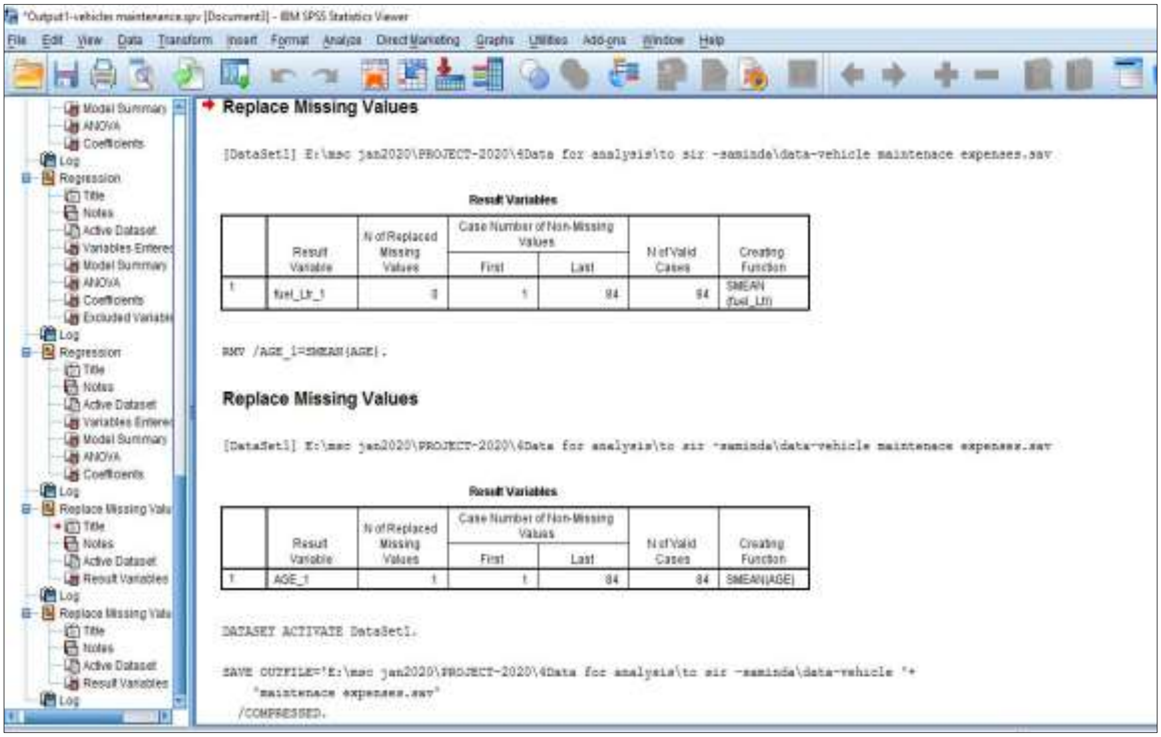
Figure 4.4: Replacing missing values

4.5.4.3 Noisy Data

Noisy data is meaningless data and does not help to give an accurate result of an analysis. This data cannot be defined by data mining software. These nosy data can be occurred due to errors in the data collection process, errors in the data entry process, etc [9]. But, this type of error can be identified and controlled by using some data mining techniques. In this study, the clustering method has been applied. This method shows the separation of similar data into one group as a cluster and the other data are shown separately.

4.5.4.4 Data Transformation

These steps have been taken to convert the data into data suitable for the mining process. Following are the methods of the data transformation,



Replace Missing Values

```
[DataSet1] E:\msc jan2020\PROJECT-2020\4Data for analysis\to sir -saminda\data-vehicle maintenace expenses.sav
```

Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
			First	Last		
1	fuel_Ltr_1	3	1	84	84	SMEAN (fuel_Ltr)

```
RMV /AGE_1=SMEAN(AGE).
```

Replace Missing Values

```
[DataSet1] E:\msc jan2020\PROJECT-2020\4Data for analysis\to sir -saminda\data-vehicle maintenace expenses.sav
```

Result Variables

	Result Variable	N of Replaced Missing Values	Case Number of Non-Missing Values		N of Valid Cases	Creating Function
			First	Last		
1	AGE_1	1	1	84	84	SMEAN(AGE)

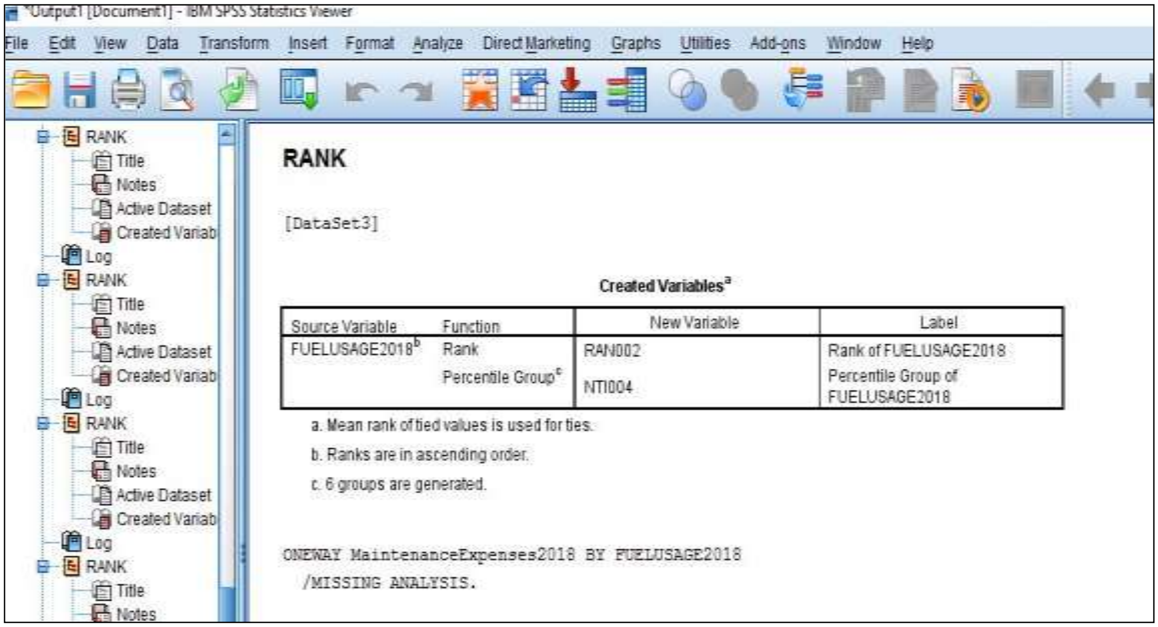
```
DATASET ACTIVATE DataSet1.

SAVE OUTFILE='E:\msc jan2020\PROJECT-2020\4Data for analysis\to sir -saminda\data-vehicle '+
    'maintenace expenses.sav'
  /COMPRESSED.
```

Figure 4.5: Replacing the missing values using average

Discretization

The discretization method is used for the numerical attributes and the conceptual levels or interval levels data are replaced for the raw data of those attributes [9]. When considering the dataset of this study, some attributes contain an extensive range of values and because of that, it is hard to get accurate output from this dataset. Accordingly, the data ranking method has been used on this dataset.

The attribute of 'Fuel usage' of vehicles has been ranked into 6 groups using the mean rank of tied values and ranks are in ascending order.



RANK

```
[DataSet3]
```

Created Variables[a]

Source Variable	Function	New Variable	Label
FUELUSAGE2018[b]	Rank	RAN002	Rank of FUELUSAGE2018
	Percentile Group[c]	NTI004	Percentile Group of FUELUSAGE2018

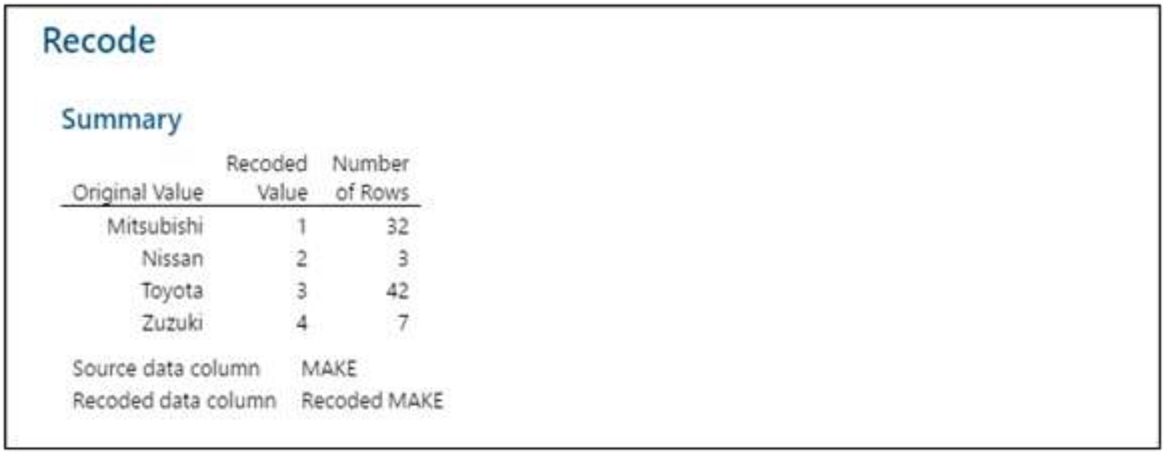
a. Mean rank of tied values is used for ties.

b. Ranks are in ascending order.

c. 6 groups are generated.

```
ONEWAY MaintenanceExpenses2018 BY FUELUSAGE2018
  /MISSING ANALYSIS.
```

Figure 4.6: Ranking of variables (spss)



Recode

Summary

Original Value	Recoded Value	Number of Rows
Mitsubishi	1	32
Nissan	2	3
Toyota	3	42
Zuzuki	4	7

Source data column MAKE
Recoded data column Recoded MAKE

Figure 4.7: Recording of variable (spss)

4.5.4.5 Data Reduction

Data mining involves manipulating a large amount of data and using it for meaningful analysis. Dealing with a large amount of data makes it difficult to perform an analysis. Data reduction methods can be used as a solution to these difficulties. Hence these techniques support saving storage space and increase efficiency and reduce time wastage.

Numerosity Reduction

The numerosity reduction has been applied to this study by using a regression model for the dataset. This reduction method allows the data model to be stored instead of storing the entire data, which is an advantage.

4.5.5 Data Mining

In data mining, regressions are useful for analyzing and estimating the relationships between two or more variables. When considering the aim of this study it is focused on finding relationships among the variables. Hence regression analysis could be used as the data mining technique of this study. But before using that kind of data mining technique for a study the nature of variables and data should be recognized and categorized [9].

Most of the variables of this data set are quantitative and some are qualitative. The dependent variable is continuous and independent variables of ‘Fuel usage’ and ‘Age’ also continues variables. The variables of Branch, Model, and Brand are qualitative and they can be used as categorical variables. Accordingly when considering the type of variables and the aim of the project a logistic regression analysis is the most suitable data mining and analytical technique for this study.

Regression Analysis

“In statistical modeling, regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome variable') and one or more independent variables (often called 'predictors', 'covariates', or 'features'). The most common form of regression analysis is linear regression, in which a researcher finds the line (or a more complex linear combination) that most closely fits the data according to a specific mathematical criterion” [12].

4.6 Software Used for the Analysis

The Minitab which statistical package for data mining has been used in this analysis and It is very user-friendly software and easy to use with a large amount of data and to build suitable models and get accurate results. Allows the application to easily store data, pre-process data, and execute various analytics based on processed data. Even though the SPSS data analyzing software was used for the analysis before this, it was low-efficiency work with the categorical data. The application of Minitab is mostly easy to use with the categorical data when recoding them into another format. Hence in this scenario, the Minitab has been used as the analytical software.

4.7 Statistical Method Used

The ‘problem’ of the study and objectives should be considered to decide which statistical method can be used for this analysis. Finding factors that affect the vehicle maintenance cost is the main purpose of this project. But it is not enough to consider only the purpose of the study and the characters of the variables and attributes should be considered.

4.8 Users

The selected model of the regression analysis can be used for the selected category of security companies as well as companies that maintain a large number of vehicles. Also, this analysis can be used to determine factors affecting vehicle maintenance costs year on year basis.

4.9 Summary

The methodology and procedure used for this research have been discussed in this section. According to that, it presents how this methodology was used to identify factors affecting vehicle maintenance costs using data mining technology.

Chapter 05

Research Design and Analysis

5.1. Introduction

The analytical approach to the use of data mining methods to improve vehicle maintenance management decision-making facilities by determining factors affecting vehicle maintenance costs is described in this section. Accordingly, several variables were considered for the analysis to determine the factors that affect vehicle maintenance cost. The steps to achieve the goal are described in this chapter.

5.2. Research Design

This research mainly focuses to find out the factors which affect vehicle maintenance expenses. Accordingly, data mining techniques have been used for analyzing the maintenance expenses to find out hidden patterns among the variables. The steps of scientific research methodology have been followed to achieve the goal successfully. The main steps in the scientific research approach are observation, background study, constructing a hypothesis, testing hypothesis draws conclusions, reporting, and evaluating.

A literature review was conducted to find out more about previous studies related to the selected topic and the nature of those studies. In that stage search about that kind of researches earliest done and study deeply the main objectives of those. As well as identifying the research gaps and proposed solutions of them. The literary review helps lead research in the right direction and builds a meaningful purpose for research.

Accordingly, the information of the factors that influence the vehicle maintenance expenses was studied by reviewing the literature, and the data mining techniques were studied by similar research. Based on the research questions, the hypothesis was buildup. Using the SPSS analytical software the hypothesis was tested and results were generated using the regression analysis. That method of analysis helps to decide the extent to which independent variables affect dependent variables. The conclusion will be offered based on the final model of the result.

5.3 Detailed Research Design

Many companies do not consider which factors affect the most to increase the maintenance cost of a vehicle. That negligence leads to unnecessary expenses. Hence identifying the reasons for the increase in vehicle maintenance costs of a company that maintains a large number of vehicles is the primary research question of the study.

The overall design of the analysis included the following steps.

- Data Gathering – Gathered data from the security company of Sri Lanka for 12 months of data.
- Data Storing – Gathered data store in suitable tables of a SQL database.
- Data Preprocessing – The process of cleaning and processing data as required for analysis.
- Data Analyze and Model creation – Data analysis and selection of the most suitable regression model for the research.
- Evaluation – Evaluate the suitability of the regression model and the final output.
- Conclusion – Presenting the conclusion based on the outcome of the model.

5.3.1 Sub Research Question One

What are the existing mechanisms and procedures in the security companies to identify the influential factors associated with vehicle maintenance expenses?

5.3.2 Sub Research Question Two

What are the variables that most affect vehicle maintenance cost?

5.3.3 Sub Research Question Three

How to use the resulting output to maintain a vehicle fleet more cost-effectively

5.4 Summary

The process and the design of the research and primary and sub-research questions have been presented step by step in this section.

Chapter 06

Implementation

6.1 Introduction

The results will be visualized using a business intelligence tool to support the decisions making process of the company management.

6.2 Data Pre-Processing

Data has been pre-processed using the following steps for the analysis as discussed in chapter four.

1. Data Cleaning
2. Data Transformation
3. Data Reduction

6.3 Attribute Selection

8 independent variables have been used for this analysis and 84 records of the data set. These data were compiled based on 20,000 records for the year 2018.

- **The Dependent Variable**

Vehicle maintenance expenses

- **Independent Variables are as Follow,**

Average miles per liter (AVG_MILES) – Continuous variable

Total running distance (TOTAL_KM) - Continuous variable

No of accidents - Continuous variable

Age (total years used) - Continuous variable

Fuel usage Total liter – Categorical variable

Assigned Branch – Categorical variable

Make – Categorical variable

Vehicle Model - Categorical variable

Engine Capacity - Continuous variable

6.3.1 Attributes Selection for the Model

Stepwise Selection of Terms

Candidate terms: AVG_MILES, TOTAL KM, ACCIDENTS, BRANCH, MODEL, MAKE, FUEL USAGE-category-2018, age_category

	-----Step 1----		-----Step 2----		-----Step 3----	
	Coef	P	Coef	P	Coef	P
Constant	42702		42702		174178	
FUEL USAGE-categry-2018	529043	0.000	366750	0.000	357833	0.000
ACCIDENTS			65931	0.000	62890	0.000
AVG_MILES					-15905	0.016
BRANCH						
S		96120.8		72729.8		70312.7
R-sq		71.29%		83.79%		85.05%
R-sq(adj)		67.80%		81.56%		82.77%
R-sq(pred)		63.98%		78.23%		79.65%

	-----Step 4-----	
	Coef	P
Constant	394533	
FUEL USAGE-categry-2018	335465	0.000
ACCIDENTS	55313	0.000
AVG_MILES	-32728	0.000
BRANCH	-157735	0.003

S	64090.4
R-sq	88.44%
R-sq(adj)	85.68%
R-sq(pred)	*

α to enter = 0.15, α to remove = 0.15

If a term has more than one coefficient, the largest in magnitude is shown.

Even though considered 8 independent variables were for the analysis, only attributes of Fuel usage, Number of accidents, Average Mills per litter, and Branch have been selected for the model using the stepwise selection method. The model accuracy is 88.44% and it was the highest accuracy level compared to previous models (Appendix 1, 2, 4, 5). Accordingly, this regression model was selected.

6.3.2 Descriptive Statistics

Statistics

	Total				
Variable	Count	N	N*	Mean	Sum
AVG_MILES	84	84	0	7.2	606.3
TOTAL KM	84	84	0	65059.0	5464915.0
ACCIDENTS	84	84	0	1.1	91.0
FUEL USAGE-categry-2018	84	84	0	6.2	519.0
Age_Category	84	84	0	3.0	251.0
Recoded BRANCH	84	84	0	2.8	232.0
Recoded VEH_TYPE	84	84	0	3.0	252.0
Recoded MODEL	84	84	0	7.0	591.0

Table 6.1 Descriptive statistics

6.4 Model Creation

Using the identified attributes of the study regression analysis were conducted depend on the characteristic of independent variables and dependant variable. The analysis has been done by more times to find the most suitable model for the study.

6.4.1 Hypothesis

For the analysis of this project, the Regression analysis has been applied to the dataset.

The hypothesis of the analysis

H0: there are no statistically significant factors between the variables that influence Vehicle maintenance expenses.

H1: there is at least one statistically significant factor between the variables that influence Vehicle maintenance expenses.

6.4.2 Choosing the Best Model

The selected variables are continuous, Categorical variables, and dummy variables used for the regression model creation. The best model is fitted based on the output of each trial.

The normal probability plot presents the spreading of the data set of the analysis. According to the following chart, the data set has been spread linearly.

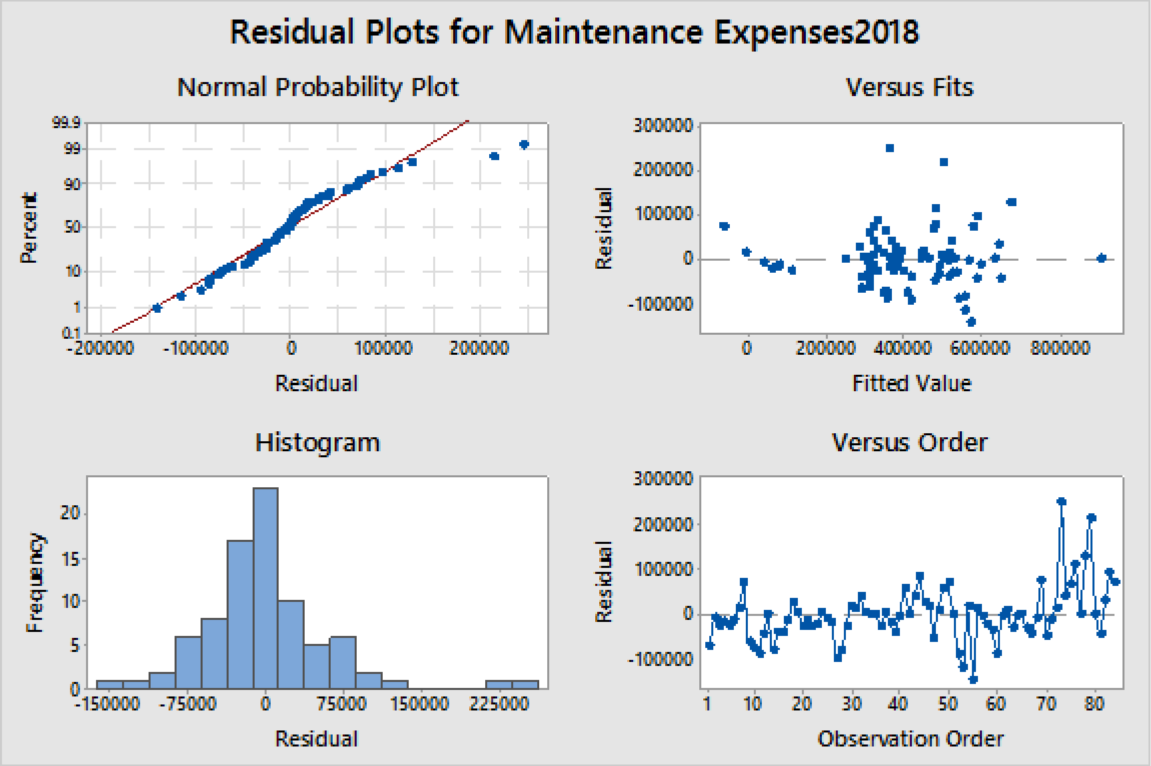


Figure 6.1: Residual plots for Vehicle maintenance expenses

6.5 Summary

This chapter presents the process of model creation and analysis. Interpreted how can attributes selecting, data preprocessing and model creating using the SPSS software. In the next section, the evaluation and the conclusion will be presented.

Chapter 07

Evaluation

7.1 Introduction

The output of the analysis is discussed in this section. Using the regression analysis the model summary, equation, and identify factors affecting the vehicle maintenance costs are presented. The suitability of the analytical techniques and model used in the analysis are also discussed.

7.2 Evaluation for Regression Analysis

The multiple linear regression analysis was selected as the analytical method of the research. There is a continuous dependent variable and continuous and categorical and dummy variables as the independent variables. According to that, the linear regression analysis was selected for this analysis to model creating.

When selecting a model for a regression analysis it should be selected the most accurate and suitable model for the relevant analysis. Otherwise, it cannot be trusted about the reliability of the output of the analysis. Hence using all attributes of the database tested models to select the most accurate one for this research and selected the highest rate of R-squared and adjusted R-squared model for the final analysis (Appendix A, B, D, E).

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
64090.4	88.44%	85.68%	*

When considering the model summary, R-squared and adjusted R-squared are very high and it indicates the goodness and accuracy level of the model. In this analysis, the R-squared is 88.44% and the R-squared of the most recently rejected model is 87.09% (Appendix 5). Accordingly when compare the model summaries of the previous analysis and present one is the most suitable and accurate model for this analysis as well as it can be used as the data mining technique in the future to identify the factors affecting the vehicle maintenance cost.

7.3 Evaluation Results

The project is based on the aim of the study to determine the factors that affect the cost of vehicle maintenance. To solve the research problem, a regression analysis was performed using a sample dataset.

Accordingly, the dependent variable is the vehicle maintenance costs, and eight independent variables were used to determine their relationship in this study. After fitting the regression model, the following variables have been identified as factors affecting vehicle maintenance costs.

- Accidents
- Average miles
- Fuel usage
- Branch

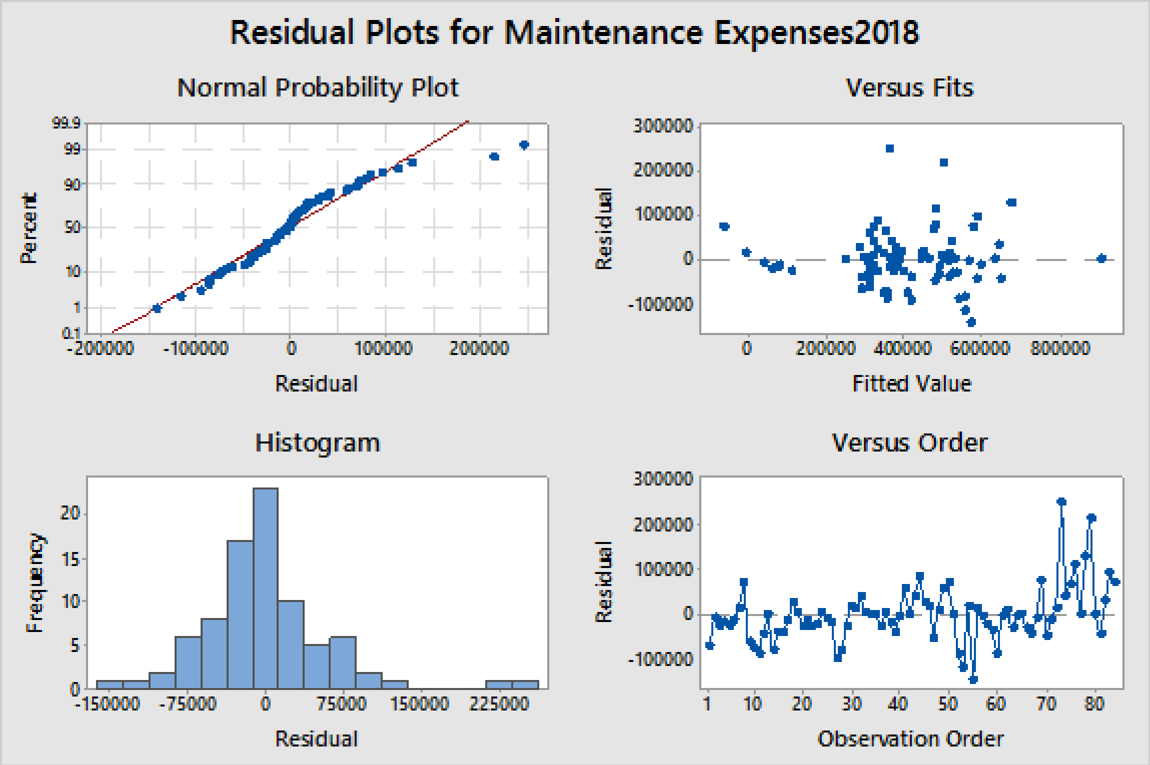


Figure 7.1: Residual plots for Vehicle maintenance expenses

The residual plot is expresses the linearity of the dataset.

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	16	2.10619E+12	1.31637E+11	32.05	0.000
AVG_MILES	1	74944586011	74944586011	18.25	0.000
ACCIDENTS	1	1.70758E+11	1.70758E+11	41.57	0.000
BRANCH	5	80750670726	16150134145	3.93	0.003
FUEL USAGE-categry-2018	9	4.74364E+11	52707149009	12.83	0.000
Error	67	2.75208E+11	4107582017		
Total	83	2.38140E+12			

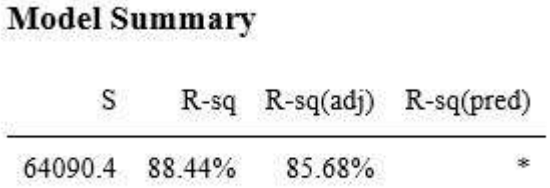
Table 7.1 Analysis of Variance

According to the created model of the final analysis, the variables of Make, Model, Total KM, and AGE category have been removed from the selected list of variables and the other variables have been selected for the model getting less than 0.05 P values for each.

The p-value indicates the relationship between two variables and the nature of that relationship. When the value of 'P' is less than 0.05, it means that the variable under consideration is significantly related to the dependent variable. Otherwise, if the 'P' is greater than 0.05, there is no relationship between the two variables.

When considering the analysis of the variance table, the p-value of selected variables which are AVG Milles, Accidents, Branch, Fuel Usage less than the significant level of 0.05. It expresses the statistical significance of these variables. This means that the above variables affect the vehicle maintenance costs and those factors are appropriate to add to the regression model.

The other variables are not statistically significant and that means the variable does not correlate with the dependant variable. Because of that, they have been removed from the model.



Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
64090.4	88.44%	85.68%	*

When considering the model summary, R R-squared and adjusted R R-squared are very high and it indicates the goodness and accuracy level of the model. In this analysis, the R-squared is 88.44% and the R-squared of the most recently rejected model is 87.09% (Appendix 5). Accordingly when compare the model summaries of the previous analysis and present one is the most suitable and accurate model for this analysis as well as it can be used as the data mining technique in the future to identify the factors affecting the vehicle maintenance cost.

Coefficient Table

The coefficient table provides the details of the relationships between dependent and independent variables. The size and the direction of the relationship of the variables can be defined by the coefficient table. The values given for the independent variables can be used for the equation and the Y variables can be predicted accordingly [13].

When the P-value of the predicted variable approaches zero, the correlation with the dependent variable becomes stronger. Accordingly variables of ‘number of ACCIDENTS’, ‘average MILES per liter’, ‘BRANCH’ and ‘FUEL USAGE” are respectively the highest correlated variables of the model. Large coefficient values indicate the order of effect for the dependent variable

As well as when considering the categorical variables of the analysis, the ‘Fuel usage’ is significantly correlated with the dependant variable and the P-value is 0.000 (Table 6.2). It indicates a higher relationship between fuel usage with the vehicle maintenance cost. The variable of fuel usage has been divided into 10 categories and all the categories are statistically significant. That means Fuel usage is mainly affected by the maintenance cost of a vehicle.

When considering the categorical variable of ‘Branch’ it has 5 categories and All groups except the MATARA branch are statistically significant. That meant the variable of Branch is affected by vehicle maintenance cost excluding the MATARA Branch category.

The p-value of the coefficient defines the variables to be included and the variables to be removed for the final model. According to the result of the analysis of this study, ‘MODEL’, ‘AGE’, ‘VEHICLE TYPE’, and ‘MAKE’ are specified as variables to be removed. The accuracy of the model decreases due to not keeping non-statistically significant attributes.

The variable of ‘AGE’ was not statistically significant and it has been removed from the fitted model. Generally, Age of a vehicle meant, that vehicle is used for a long time. In practice, if consider a vehicle which age is higher but is not run long time the maintenance costs can be reduced. That can be caused to removal from the model.

But sometimes the affection of the external variables and variables which couldn’t found the data (the traffic condition, driving Speed, road conditions) Can be affected to create wrong relationships among the variables. That was a disadvantage of a regression analysis.

Coefficients Table of the Analysis

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	394533	81018	4.87	0.000	
AVG_MILES	-32728	7662	-4.27	0.000	2.10
ACCIDENTS	55313	8579	6.45	0.000	2.44
BRANCH					
C_NORTH	-123848	30416	-4.07	0.000	1.98
COLOMBO	-85213	23696	-3.60	0.001	2.86
JAFFNA	-157735	77386	-2.04	0.045	1.44
KANDY	-93668	46997	-1.99	0.050	1.56

MATARA	-49271	33103	-1.49	0.141	1.71
FUEL USAGE-categories					
2	240023	46125	5.20	0.000	1.97
3	162349	38620	4.20	0.000	2.02
4	217641	32040	6.79	0.000	2.39
5	255422	40373	6.33	0.000	1.87
6	250394	37276	6.72	0.000	2.45
7	335465	35393	9.48	0.000	2.45
8	285100	37815	7.54	0.000	3.83
9	295353	38868	7.60	0.000	2.66
10	326701	37320	8.75	0.000	3.73

Table 7.2 Coefficients table of the final model

Since there is at least one variable statistically significant, the alternative hypothesis is accepted and the null hypothesis is rejected. When considering the result of the analysis, the ‘H1’ is accepted. According to that it indicates, there is at least one significant variable among the variables that affect vehicle maintenance expenses.

Regression Equation

Maintenance Expenses2018	=	394533 - 32728 AVG_MILES + 55313 ACCIDENTS
		+ 0.0 BRANCH_A_PURA
		- 123848 BRANCH_C_NORTH - 85213 BRANCH_COLOMBO
		- 157735 BRANCH_JAFFNA - 93668 BRANCH_KANDY
		- 49271 BRANCH_MATARA
		+ 0.0 FUEL USAGE-categry1
		+ 240023 FUEL USAGE-categry2
		+ 162349 FUEL USAGE-category-2018_3
		+ 217641 FUEL USAGE-category-2018_4
		+ 255422 FUEL USAGE-category-2018_5
		+ 250394 FUEL USAGE-category-2018_6
		+ 335465 FUEL USAGE-category-2018_7
		+ 285100 FUEL USAGE-category-2018_8
		+ 295353 FUEL USAGE-category-2018_9
		+ 326701 FUEL USAGE-category-2018_10

Table 7.3 Regression Equation

For reaction analysis, there is a reaction line to present the relationship between the variables. It is also known as the ‘best-fitting line’. The algebraical expression of this best-fitting line is the equation of the regression analysis. The regression equation is used to predict the values of the dependent variable based on the values of the independent variables.

The regression equation is as follows,

$$\mathbf{Y_e = a + bX}$$

Where ‘Y’ indicates the value of the dependent variable and ‘X’ indicates the value of the independent variables. In this study ‘a’ is the constant value and it is 394,533. The variables of AVG MILLS, the number of ACCIDENTS, BRANCH, and FUEL USAGE are the independent variables. The dependant variable which is vehicle maintenance cost can be predicted based on the values of independent variables. The predicted values of each variable are included in the above regression equation (Table 7.2).

Fits and Diagnostics for Unusual Observations

Obs	Maintenance Expenses2018	Fit	Resid	Std Resid		
13	247722	247722	-0	*		X
55	430406	555104	-124698	-2.12	R	
73	604216	399941	204274	3.45	R	
77	902194	729824	172371	3.06	R	
79	714428	547448	166981	2.87	R	

Table 7.4 Fits and Diagnostics for Unusual Observations

7.4 Summary

The results of the analysis and the evaluation of the measures have been presented in this chapter. Accordingly based on the output of the regression model, removed attributes and the selected factors, the ratio of the effect of the selected attributes have been defined. The next chapter will be presented the conclusion and future works of the research.

Chapter 08

Conclusion and Future works

8.1 Introduction

The overall summary of the analysis and the conclusion, limitations of the study, and further work of the research are discussed in this chapter. While conducting an analysis the impacted or influenced the interpretation of the findings from the research can describe as the limitations. The future aspects of the research were also discussed in this research.

8.2 Overview of the Research

The regression analysis was selected as the analytical method of this research. The aim is to develop a more efficient and cost-effective vehicle management system by identifying the factors that most affect vehicle maintenance costs. Hence studied the data mining technics to use for this analysis and selected the multiple linear regression analysis as the analytical technic for the research. Accordingly, the results of the analysis can be used to assist a company in maintaining a large number of vehicles to make the best decision. The results of the research could be used to reduce vehicle maintenance costs and manage the fleet more efficiently.

According to the output of the study finally identified the variables which are Fuel usage, no of accidents, average mills per liter, and branch are the most influential factors in determining the maintenance cost of a vehicle. The model output can be used by an organization to control the vehicle fleet more efficiently based on the impact of each factor. When considering the aim and objectives of the study, different data mining techniques have been used for the data preprocessing and data analyzing phase. The multiple regression analysis was used for the main analysis of the project and it is the best data analyzing tool for this study and can be used in the future for this kind of study.

8.3 Limitations

- Limited values for the class variable
- Some of the important variables could not be collected and some data are uncompleted

  Example – Distance run by the vehicles, driving behavior (acceleration and speed), and Vehicle condition. Those variables are hard to measure and because of that they couldn't be considered as attributes for this analysis
- It is hard to measure the external variables.

8.4 Future Works of the Project

For companies that maintain a large fleet of vehicles, plan to design a system to implement an automated data visualization dashboard that is updated daily using this analytics model. Expect to be able to visualize the daily analysis with an automated system and reminders of the notification and maintenance stages.

8.5 Summary

The overview of the project and the conclusion of the result are discussed in this chapter. As well as the suitability of the data mining technics used and the limitations of the project and future expectations and how it can be improved in the future are also presented as the final chapter of the thesis.

References

[1] Daejin Kim, (2018), Personal vehicle ownership and operating cost calculator, Georgia Institute of Technology

[2] Litman, A. (2009), Transportation Cost and Benefit Analysis, Victoria Transport Policy Institute, PP 7-13, Jordan

[3] Riis, J. O, Luxhoj, J. T, Thorsteinsson U. A, (1997), Situational Maintenance Model. International Journal of Quality & Reliability Management v.4

[4] Rogic K, Sutic B, Kolaric G, (2008), Methodology of Introducing Fleet Management System

[5] Venezia, F. (2000), Transit Fleet Maintenance, Transportation Research Board publications

[6] Wheels.com, (2017), Factors that influence fleet operating costs, Wheels Inc. All Rights Reserved

[7] Zacharof, N. G. and Fontaras, G. (2016), "Review of in-use factors affecting the fuel consumption and CO2 emissions of passenger cars", European Union

[8] "Benefits of fleet management-systems" [Online] Available: https://www.verizonconnect.com/au/resources/article/benefits-fleet management systems [Accessed 28-Dec-2020]

[9] "Data Preprocessing in Data Mining" [Online] Available: https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/ [Accessed 20-Apr-2020]

[10] “How to write a research methodology” [Online] Available:
https://www.scribbr.com/dissertation/methodology/
[Accessed 28-Apr-2020]

[11] “Multiple-linear-regression analysis” [Online] Available:
https://corporatefinanceinstitute.com/resources/knowledge/other/multiple-linear-regression
[Accessed 12-Dec-2020]

[12] “Regression Analysis Tutorial and Examples”
https://blog.minitab.com/blog/adventures-in-statistics-2/regression-analysis-tutorial-and-examples/
https://en.wikipedia.org/wiki/Regression_analysis
[Accessed 29-Apr-2020]

[13] “Sampling and data analysis” [Online] Available:
https://people.umass.edu/~mcclemen/581Sampling.html
[Accessed 20-Mar-2020]

[14] “SQL tutorials” [Online] Available: https://www.w3schools.com/sql/
[Accessed 20-Mar-2020]

Appendixes

Appendix A - Regression analysis (Model 1)

Dependent variable – Vehicle maintenance cost

Regression Analysis: Maintenance Expenses2019 versus ..., Age, MAKE

The following terms cannot be estimated and were removed:

MODEL, MAKE

Method

Categorical predictor coding	(1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	18	2.42665E+12	1.34814E+11	11.61	0.000
FUEL USAGE2019	1	45400314466	45400314466	3.91	0.052
Age	12	1.00504E+12	83753286977	7.21	0.000
BRANCH	5	2.38190E+11	47638076453	4.10	0.003
Error	65	7.54646E+11	11609939224		
Total	83	3.18130E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
107749	76.28%	69.71%	*

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	257200	103951	2.47	0.016	
FUEL USAGE2019	16.32	8.25	1.98	0.052	3.93
Age					
3	-72615	109490	-0.66	0.510	4.86
4	12848	122428	0.10	0.917	7.19
5	1042	120628	0.01	0.993	15.44
6	164334	105356	1.56	0.124	8.42
7	288034	128677	2.24	0.029	2.78
8	209160	112354	1.86	0.067	8.74
9	257146	115958	2.22	0.030	8.38
10	241466	116659	2.07	0.042	4.47
13	273095	122181	2.24	0.029	6.05
14	227878	116624	1.95	0.055	12.05
15	107417	120792	0.89	0.377	5.91
16	113354	151010	0.75	0.456	1.94
BRANCH					
C_NORTH	-150033	48932	-3.07	0.003	1.82
COLOMBO	-161822	45100	-3.59	0.001	3.67
JAFFNA	-255558	134953	-1.89	0.063	1.55
KANDY	-232307	85315	-2.72	0.008	1.81
MATARA	-222679	61176	-3.64	0.001	2.07

Regression Equation

Maintenance Expenses2019 = 257200 + 16.32 FUEL USAGE2019 + 0.0 Age_2 - 72615 Age_3
+ 12848 Age_4 + 1042 Age_5 + 164334 Age_6 + 288034 Age_7
+ 209160 Age_8 + 257146 Age_9 + 241466 Age_10
+ 273095 Age_13
+ 227878 Age_14 + 107417 Age_15 + 113354 Age_16
+ 0.0 BRANCH_A_PURA - 150033 BRANCH_C_NORTH
- 161822 BRANCH_COLOMBO - 255558 BRANCH_JAFFNA
- 232307 BRANCH_KANDY - 222679 BRANCH_MATARA

Fits and Diagnostics for Unusual Observations

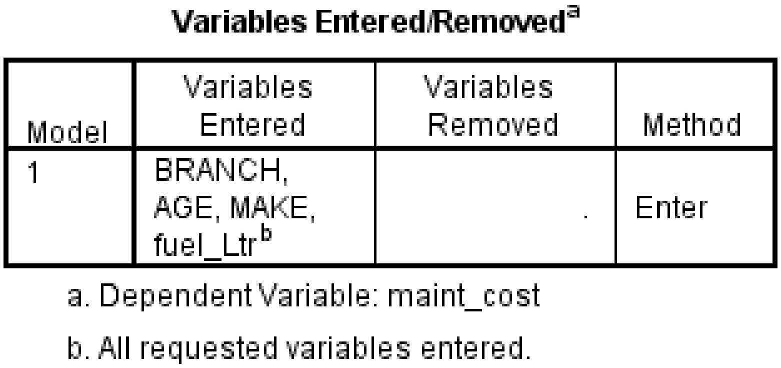
Obs	Maintenance Expenses2019	Fit	Resid	Std Resid		
9	311141	311141	-0	*		X
40	313302	313302	-0	*		X
60	978799	679058	299741	3.03	R	
74	836847	642572	194275	2.10	R	
76	881254	522268	358987	3.78	R	
82	937603	753205	184398	2.04	R	

R Large residual
X Unusual X

Residual Plots for Maintenance Expenses2019

P-value is lower than the above variables and according to that can be said the variables above significantly match.

Appendix B - Regression analysis (Model 2)

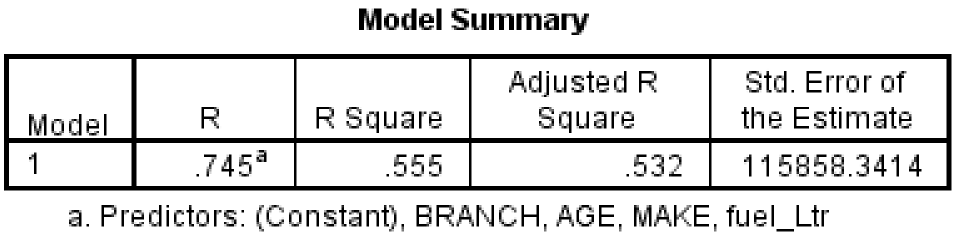


Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	BRANCH, AGE, MAKE, fuel_Ltr[b]	.	Enter

a. Dependent Variable: maint_cost

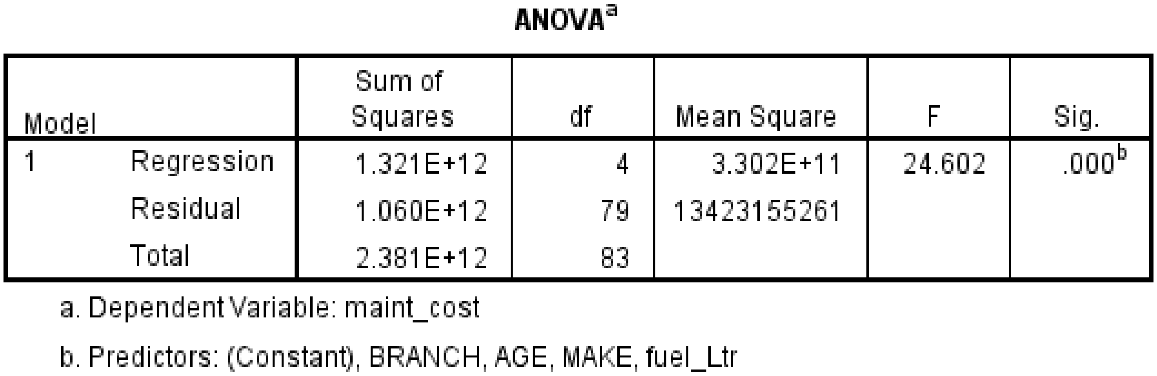
b. All requested variables entered.



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.745[a]	.555	.532	115858.3414

a. Predictors: (Constant), BRANCH, AGE, MAKE, fuel_Ltr

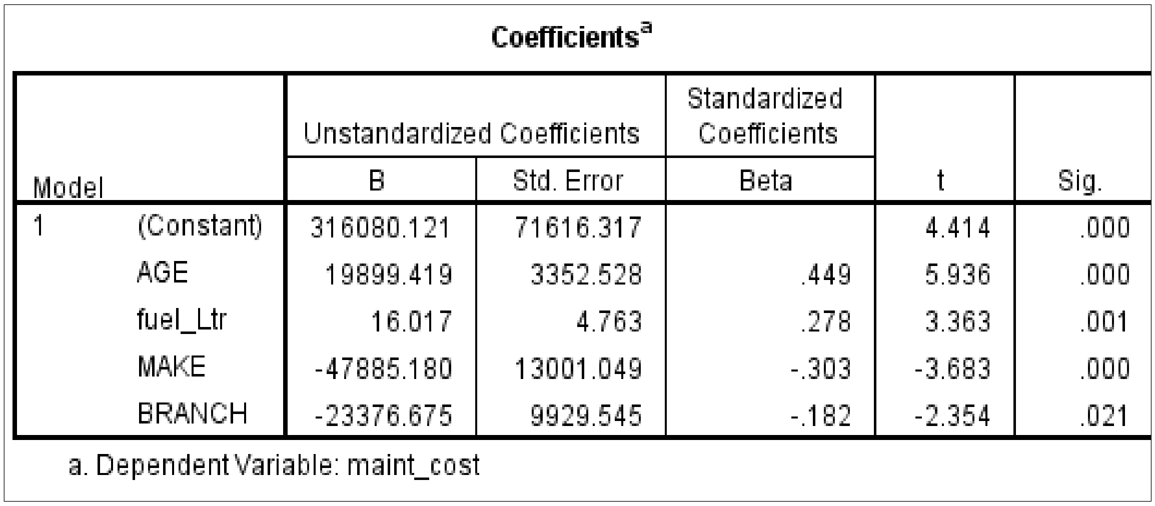


ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1.321E+12	4	3.302E+11	24.602	.000[b]
	Residual	1.060E+12	79	13423155261		
	Total	2.381E+12	83			

a. Dependent Variable: maint_cost

b. Predictors: (Constant), BRANCH, AGE, MAKE, fuel_Ltr

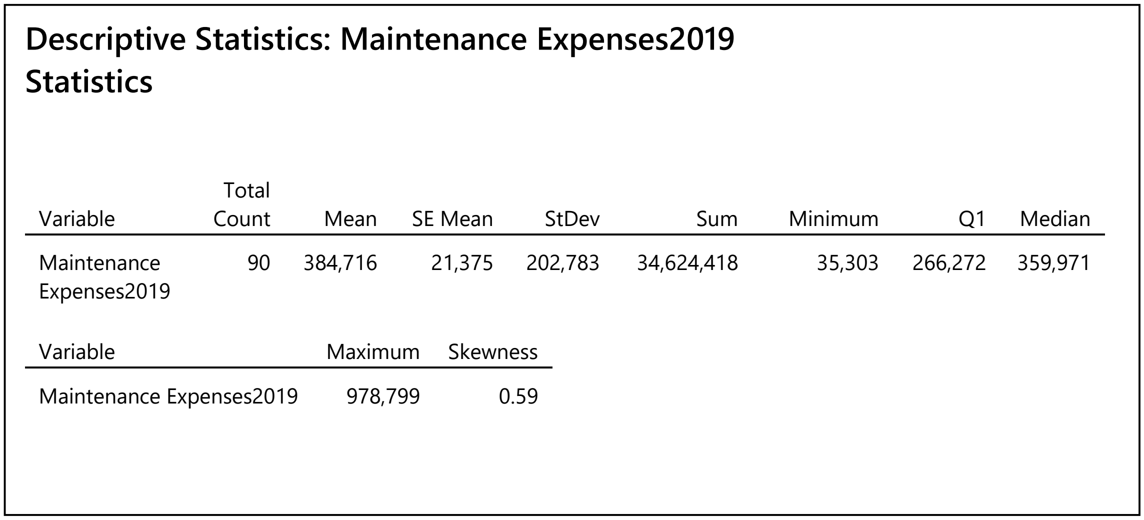


Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	316080.121	71616.317		4.414	.000
	AGE	19899.419	3352.528	.449	5.936	.000
	fuel_Ltr	16.017	4.763	.278	3.363	.001
	MAKE	-47885.180	13001.049	-.303	-3.683	.000
	BRANCH	-23376.675	9929.545	-.182	-2.354	.021

a. Dependent Variable: maint_cost

Appendix C - Descriptive statistics



Descriptive Statistics: Maintenance Expenses2019

Statistics

Variable	Total Count	Mean	SE Mean	StDev	Sum	Minimum	Q1	Median
Maintenance Expenses2019	90	384,716	21,375	202,783	34,624,418	35,303	266,272	359,971

Variable	Maximum	Skewness
Maintenance Expenses2019	978,799	0.59

Appendix D - Regression Analysis- Model 3

Regression Analysis: Maintenance Expenses2018 versus ... KE, BRANCH

The following terms cannot be estimated and were removed:

MAKE

Method

Categorical predictor coding (1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	24	1.85926E+12	77469044885	8.75	0.000
AGE_2018	1	24858112	24858112	0.00	0.958
FUEL_cat_18	9	1.54666E+11	17185056246	1.94	0.063
MODEL	9	2.98387E+11	33154095489	3.75	0.001
BRANCH	5	1.66842E+11	33368368496	3.77	0.005
Error	59	5.22141E+11	8849852799		
Lack-of-Fit	35	4.12967E+11	11799062862	2.59	0.009
Pure Error	24	1.09174E+11	4548921458		
Total	83	2.38140E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
94073.7	78.07%	69.16%	*

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	290406	198930	1.46	0.150	
AGE_2018	679	12814	0.05	0.958	23.36
FUEL_cat_18					

2	221997	118013	1.88	0.065	11.39
3	175365	106584	1.65	0.105	10.32
4	167602	105522	1.59	0.118	9.11
5	282750	106401	2.66	0.010	10.28
6	254864	103131	2.47	0.016	8.70
7	253913	105689	2.40	0.019	10.14
8	155223	104864	1.48	0.144	8.99
9	148683	102203	1.45	0.151	9.48
10	187778	100973	1.86	0.068	8.34
MODEL					
DA17A	-139842	179186	-0.78	0.438	23.28
E24	-13738	120843	-0.11	0.910	3.22
L300	74457	97421	0.76	0.448	21.24
LH162	-15883	67875	-0.23	0.816	1.98
LH171	427595	122806	3.48	0.001	1.68
LH172	53925	66900	0.81	0.423	4.06
LH200	-86266	131042	-0.66	0.513	31.51
QD32	39709	106638	0.37	0.711	1.27
VWE25	209623	106860	1.96	0.055	1.27
BRANCH					
C_NORTH	-155676	46185	-3.37	0.001	2.12
COLOMBO	-98685	38571	-2.56	0.013	3.52
JAFFNA	-344571	117779	-2.93	0.005	1.55
KANDY	-153351	68668	-2.23	0.029	1.54
MATARA	-116409	53310	-2.18	0.033	2.06

Regression Equation

Maintenance Expenses2018 = 290406 + 679 AGE_2018 + 0.0 FUEL_cat_18_1 + 221997 FUEL_cat_18_2
+ 175365 FUEL_cat_18_3 + 167602 FUEL_cat_18_4
+ 282750 FUEL_cat_18_5 + 254864 FUEL_cat_18_6
+ 253913 FUEL_cat_18_7 + 155223 FUEL_cat_18_8
+ 148683 FUEL_cat_18_9 + 187778 FUEL_cat_18_10 + 0.0 MODEL_CR42
- 139842 MODEL_DA17A - 13738 MODEL_E24 + 74457 MODEL_L300
- 15883 MODEL_LH162 + 427595 MODEL_LH171 + 53925 MODEL_LH172
- 86266 MODEL_LH200 + 39709 MODEL_QD32 + 209623 MODEL_VWE25
+ 0.0 BRANCH_A_PURA - 155676 BRANCH_C_NORTH
- 98685 BRANCH_COLOMBO
- 344571 BRANCH_JAFFNA - 153351 BRANCH_KANDY
- 116409 BRANCH_MATARA

Regression Analysis: Maintenance Expenses2018 versus ... GE-AT 2018

The following terms cannot be estimated and were removed:

MAKE, FUEL USAGE-categry-2018, AGE-AT 2018

Method

Categorical predictor coding	(1, 0)

Analysis of Variance

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Regression	17	2.07395E+12	1.21997E+11	26.19	0.000
AVG_MILES	1	61451848568	61451848568	13.19	0.001
TOTAL KM	1	43680511248	43680511248	9.38	0.003
ACCIDENTS	1	1.64456E+11	1.64456E+11	35.30	0.000
BRANCH	5	24905583335	4981116667	1.07	0.385
MODEL	9	1.43339E+11	15926609208	3.42	0.002
Error	66	3.07448E+11	4658301912		
Total	83	2.38140E+12			

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
68251.8	87.09%	83.76%	*

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	557960	80299	6.95	0.000	
AVG_MILES	-37273	10262	-3.63	0.001	3.32
TOTAL KM	1.967	0.642	3.06	0.003	4.74
ACCIDENTS	51777	8714	5.94	0.000	2.22
BRANCH					
C_NORTH	-62558	31801	-1.97	0.053	1.91
COLOMBO	-53092	28418	-1.87	0.066	3.63
JAFFNA	-99498	77446	-1.28	0.203	1.27
KANDY	-70674	48702	-1.45	0.151	1.47
MATARA	-37459	32904	-1.14	0.259	1.49
MODEL					
DA17A	-189477	51818	-3.66	0.001	3.70
E24	2411	58077	0.04	0.967	1.41
L300	-7208	37360	-0.19	0.848	5.94
LH162	-69041	46851	-1.47	0.145	1.80
LH171	224462	83653	2.68	0.009	1.48
LH172	-9278	38891	-0.24	0.812	2.61
LH200	-49814	38020	-1.31	0.195	5.04
QD32	-6823	76101	-0.09	0.929	1.23
VWE25	99354	76415	1.30	0.198	1.24

Regression Equation

Maintenance Expenses2018 = 557960 - 37273 AVG_MILES + 1.967 TOTAL KM + 51777 ACCIDENTS
+ 0.0 BRANCH_A_PURA - 62558 BRANCH_C_NORTH
- 53092 BRANCH_COLOMBO
- 99498 BRANCH_JAFFNA - 70674 BRANCH_KANDY
- 37459 BRANCH_MATARA
+ 0.0 MODEL_CR42 - 189477 MODEL_DA17A + 2411 MODEL_E24
- 7208 MODEL_L300 - 69041 MODEL_LH162
+ 224462 MODEL_LH171
- 9278 MODEL_LH172 - 49814 MODEL_LH200 - 6823 MODEL_QD32
+ 99354 MODEL_VWE25

Fits and Diagnostics for Unusual Observations

	Maintenance					
Obs	Expenses2018	Fit	Resid	Std Resid		
13	247722	247722	0	*		X
51	462935	462935	-0	*		X
55	430406	572069	-141663	-2.18	R	
73	604216	359720	244496	3.89	R	
77	902194	902194	-0	*		X
78	800809	674185	126624	2.20	R	
79	714428	500837	213591	3.52	R	
80	631491	631491	-0	*		X

R Large residual

X Unusual X