**Logo, company name

Description automatically generated**

Data Driven Decision Making

Table of Contents

[**Introduction** 3](#_Toc98684131)

[**Problem 1:** 3](#_Toc98684132)

[**Problem Definition:** 3](#_Toc98684133)

[**Answer 1a:** 5](#_Toc98684134)

[**Answer 1b:** 6](#_Toc98684135)

[**Answer 1c:** 6](#_Toc98684136)

[**R Code:** 7](#_Toc98684137)

[**Problem 2:** 8](#_Toc98684138)

[**Problem Definition:** 8](#_Toc98684139)

[**Answer 2a:** 10](#_Toc98684140)

[**Answer 2b:** 10](#_Toc98684141)

[**Answer 2c:** 10](#_Toc98684142)

[**Part 2(Problem 2):** 11](#_Toc98684143)

[**Production Plan Table (3-years):** 12](#_Toc98684144)

[**R Code:** 13](#_Toc98684145)

[**Problem 3** 14](#_Toc98684146)

[**Problem Definition:** 14](#_Toc98684147)

# **1.0 Abstract**

Multiple linear regression will be the primary focus of this research. The term "multiple linear regression" refers to using more than one variable to predict the response variable. As a result, our primary objective here is to estimate the market value of your car. We'll use a statistical model to figure out the value of your car.

# **2.0 Background**

Driving a car has long been regarded as a necessity for most people worldwide. Most of the world's population own a car to get to high school or college. There is a glut of used vehicles available for second-hand cars. After starting my career as a university graduate this summer, I will need to purchase my car. I intend to build an accurate and realistic prediction model using the "car" dataset that can be used to estimate the future worth of used cars and then select the most cost-effective bargain.

## **Project Questions –**

1. **Is it possible to precisely predict the price of a used car?**

Predicting future car costs can assist me in making more thoughtful purchasing selections depending on my budget, the condition of the car, and my desire for a specific model. My own car's value may be estimated using this prediction model, making it easier to sell used cars for a fair price on the secondary market.

1. **What are the most prominent factors?**

Car model, mileage, trim, make, and so on are all included in this dataset, containing information on several aspects of a vehicle. Using several prediction models, I hope to find the elements that have a substantial impact on the pricing of cars.

1. **What is the effect of those prominent factors?**

After I've identified the most prominent factors, I'd like to see what impact those factors have on the cost of a car. I'd be able to see potential offers in the used-car market if I had these insights.

# **3.0 Methodology**

## **3.1 Data variables**

|  |  |
| --- | --- |
| **“Car”** | |
| **Price** | Retail price of the used car. |
| **Mileage** | Total number of miles the car has driven. |
| **Make** | Brand of the car such as Saturn, Chevrolet, and Pontiac. |
| **Model** | Specific model of each brand such as Ion, Cavalier. |
| **Trim** | Car model such as Quad Coupe 2D etc. |
| **Type** | Car body type such as coupe, sedan, etc. |
| **Cylinder** | Number of cylinders in the engine. |
| **Liter** | Discrete measure of engine size. |
| **Doors** | Number of doors |
| **Cruise** | Whether the car has cruise control or not (1 = cruise) |
| **Sound** | Whether the car has upgraded sound systems or not (1 = upgraded) |
| **Leather** | Whether the car has leather seats or not (1 = Leather seats) |

## **3.2 Data Exploration**

First, we did some basic summaries of the data to enable that there were no quality issues such as extreme outliers or missing values before moving on to further analysis. That's why we went ahead and started digging into the data.

**“The basic summary statistics that were performed can be found in the appendix towards the end of the report.”**

### **3.2.1 Data Visualisation**

I began our exploratory study by creating data visualisations, which I believe is an essential initial step in any research. It is a valuable tool for making sense of the data. The data stats are not clear to us if we only look at the numbers.

1. **Histogram of Car Prices**

Chart, histogram

Description automatically generated

**Fig 1: Histogram of Car Prices**

The above visualisation is the representation of car prices in the data set. This clearly shows us that data in context with prices is rightly skewed and most of the cars are priced towards the lower end of the graph. This shows us most of the cars are priced somewhere around 10000-20000 price range.

1. **Histogram of car prices showing makes of the car**

Chart

Description automatically generated

**Fig 2: Car prices showing the make of the car**

To show us the price distribution of cars, we created a second visualisation. Also, we can observe that the prices of Saturn and SAAB cars are normally distributed, whereas the prices of Cadillac cars tend to be on the higher end of the distribution curve. Because of the wide range of costs, I have realised that this is essential to consider when making price predictions. To investigate this, we created a facet wrap that shows the price distribution for each manufacturer. This facet wrap visualisation helped us see which make had a more normal distribution or a right skew or left skew price distribution.

A picture containing timeline

Description automatically generated

**Fig 3: Facet wrap of car prices showing the make of the car**

1. **Scatterplot of car price vs mileage with different makes**

Scatter chart

Description automatically generated

**Fig 4: Scatterplot of car price vs mileage with different makes**

Car prices and make are shown in relation to the mileage in the above visualisation. There is a distinct downward trend in the price of a Cadillac which means more mileage on a Cadillac has a more significant effect on the price, whereas the blue dots, which represent SAAB, show a flat pattern that indicates that mileage has little to no effect on the price of this make. This demonstrates the impact of mileage on the cost of various automobile models.

**“Few more visualisations were performed before moving ahead you can find them in the appendix towards the end of the report”.**

## **Measures of Association**

### **3.3.1 PCA Analysis**

Chart, scatter chart

Description automatically generated

**Fig 5: PCA Analysis**

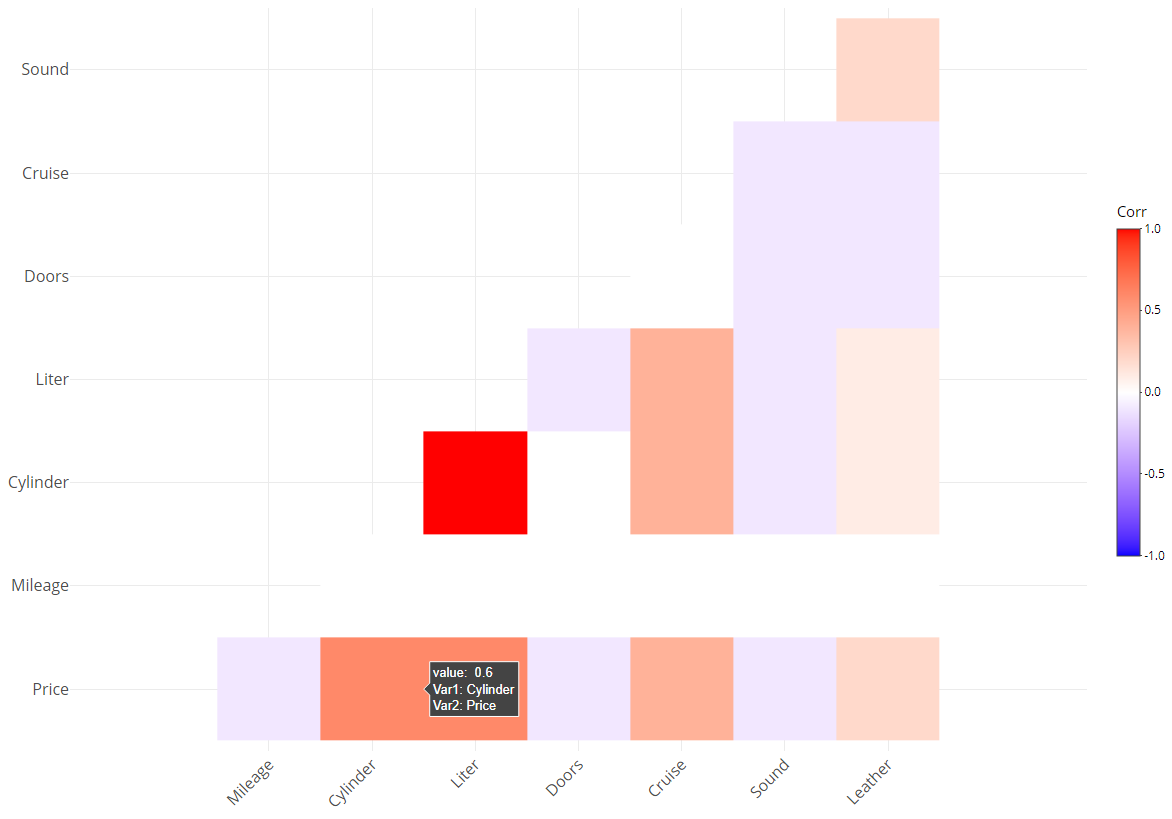
To conduct our study, we first built a dummy dataset using dummy variables. As a first step, we performed a principal component analysis (PCA) to help us decompose the data into more manageable chunks or we can say to reduce the dimensionality of the data. Only one enormous cluster can be seen in the PCA graph, which tells us little to nothing about the data set. We only cared about the interactions between the variables and the resulting clusters.

**“Summary of the PCA analysis can be found in the appendix towards the end of the report”.**

### **3.3.2 Interactive Correlation Plot**

The next step in investigation was to look at the correlation between the different variables. The interactive correlation chart is the consequence of this. To put it another way, when we hover the cursor on a variable, it shows the correlation between two variables. This helped in understanding the correlation between the variables and after having a brief look at the correlation data I wanted to dive in this a little more to get the most out of it. Hence we created a pairwise correlation graph.

**(P.S: Since the graph cannot be displayed interactively in Turnitin, we've provided two examples for you to peruse instead.)**

****

**Fig 6.1: Interactive Correlation Plot**

**Chart

Description automatically generated**

**Fig 6.2: Interactive Correlation Plot**