Logo, company name

Description automatically generated Logo, company name

Description automatically generated

**BSc. (Hons.) Data Science Degree.**

**7081 CEM - Business Simulation.**

**Module leader: Mr. Chameera De Silva**

ADDS 21.2P

**15th of January 2023**

**Yapa G.I.D. - COADDS 21.2P-(005)**

**Acknowledgement**

I am highly indebted to Mr. Chameera De Silva for his guidance and supervision. I would like to thank him for his assistance and support for this project. It was a great learning experience.

I have taken a lot of effort into this project. However, completing this project would not have been possible without the support and guidance of a lot of individuals. I would like to extend my sincere thanks to all of them.

I would like to express my gratitude towards my parents & my friends for their kind co-operation and encouragement which help me a lot in completing this project.

My thanks and appreciations also go to my colleague in developing the project. Thank you to all the people who have willingly helped me out with their abilities.

Contents

[**Introduction** 4](#_Toc124724331)

[**Implementation** 5](#_Toc124724332)

[**Conclusion** 18](#_Toc124724333)

## **Introduction**

As a result of recent technical advancements, numerous organizations now use techniques like machine learning extensively. These models typically operate on a collection of specified data points that are supplied as datasets. These models typically operate on a set of predefined data points provided in the form of datasets. These datasets contain historical data on a specific domain. So, it is crucial to organize these data points before applying them to the model. This is where Data Analysis comes in. If the data fed to the machine learning model is not well organized, it produces false or undesirable results. This can result in significant financial losses for the organization. Therefore, it is crucial to employ competent data analysis.

About the Data set

In this report, we will use data about automobiles. Specifically containing various information datapoints about used cars, such as price, color, and so on. We must understand that simply collecting data is insufficient. Raw data is useless. Data analysis is critical in unlocking the information we need and gaining new insights from this raw data.

Let's define some of the issues that we can encounter: For instance, are there statistics on the features and costs of other cars? What attributes of autos influence their costs? Colour? Brand? Does horsepower have a role in price, or is there another factor at play?

## **Implementation**

It is very important to analyze the data set first as in order to find the right solutions. So, we can identify that the given data is available in the form of .csv/. Moreover, the following modules will be needed to analyze the data set.

The modules needed:

* pandas: You can manipulate data in Python using the open-source Pandas package. Pandas offer a simple method for creating, modifying, and handling the data.
* numpy: The foundational Python library for scientific computing is called numpy. An effective multi-dimensional data container for general data is numpy.
* matplotlib: A Python 2D plotting package, matplotlib generates publication-quality graphics in a range of formats.
* seaborn: based on matplotlib, seaborn is a Python data-visualization package. A high-level interface is provided by Seaborn for creating visually appealing and educational statistical visuals.
* Scipy: Scipy is an open-source ecosystem of tools for math, science, and engineering built on the Python programming language.

Steps that are used in the following code

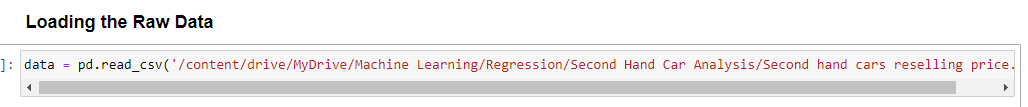
Step 01: Import the modules needed.

Graphical user interface, text, application

Description automatically generated

Step 02: Loading the Raw Data

In this step we expect to connect the data source to python script in order to analyzed.



Check the first five entries of dataset.

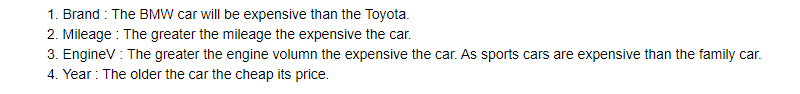
It is important to understand that if your dataset has the right type of data in it. Therefore, the first n rows(head) and last n rows(tail) have been identified as below.

Table

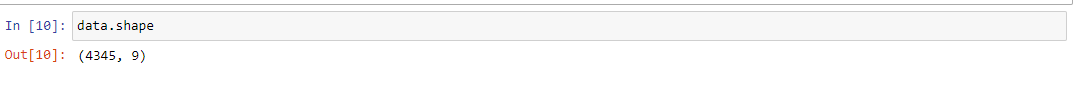
Description automatically generated

Results –

According to the results generated, It can be seen that the data set is quite a big data with almost of 4500 rows. The following variables can be considered when we analyze the data set to find the price of secondhand cars.



The data set shape can be illustrated as below.



Step 03: Preprocessing

The process of converting raw data into an understandable format is known as data preprocessing. We cannot work with raw data, so this is an important step in data mining. Before using machine learning or data mining algorithms, the data quality should be checked. According to the following image, we have identified that there are some missing values in the "Price" and "EngineV” columns.

Graphical user interface, application

Description automatically generated

Step 04: Determining the variables of interest

Dropping data from the table prompts two arguments: the row/column to be dropped and its axis. Axis = 0 represents the row, and axis = 1 represents the column. In this step we are removing the column 'Model' because it is not relevant to our analysis and can cause significant variability due to the large number of unique values.

Graphical user interface, application

Description automatically generated

We now have two columns with missing values that we must deal with carefully. So, we will be drop the rows with missing values in these columns because the number of missing values is less than 5% of the total number of rows.

Graphical user interface, application

Description automatically generated

According to the above image, it can be clearly seen that we have removed all the missing values in our data set.

Step 05: Dealing with outliers

We can infer from the PDF of Prices that the prices have a positive skewness. This may cause issues in our regression. We must eliminate the outliers. We can accomplish this using the quantile method. We will keep the 99 percentile price values from our data according to the below images.

Graphical user interface, application, Word, Teams

Description automatically generated

Graphical user interface

Description automatically generated with medium confidence

Also, we can outlier the data set using Mileage and EngineV columns as to get more smoother the analysis.

Chart, histogram

Description automatically generated Graphical user interface, application, Word

Description automatically generated

According to the above images it can be seen that the 'EngineV' column contains several huge outliers. Because the manual examination of the data reveals that the column contains some values equal to 99.99. This is because we typically fill the null cells with these values. However, in general, the engine volume of a car cannot be greater than 6.5 or less than 0.6. As a result, we must now deal with this blunder as in below illustrations.

Graphical user interface, chart, application, histogram

Description automatically generated

Chart

Description automatically generated with low confidence

Graphical user interface

Description automatically generated

Step 06: Checking the OLS assumptions

According to the below illustrations, the scatter plots show patterns between the dependent and independent variables of our data, but they are not linear. To make our data fit the linear regression model best, we must transform it to obtain a linear relationship between the dependent and independent variables.

Chart, application

Description automatically generated with medium confidence

Therefore, we can further analyze our data using Log Transformation method.

A picture containing table

Description automatically generated

Graphical user interface, application

Description automatically generated

Step 07: Removing Multicollinearity

We can see from the VIF data frame that the variable "Year" has a high degree of multicollinearity. Therefore, we can drop this according to the following steps.

Graphical user interface, text, application

Description automatically generated

So with that it can be seen that we have prepared the data set for analysis.

Step 08: Creating dummies

In order to analyze the data first we need to makesure that our data set is in one format as numerical data. So, to convert that we have used dummy method as in below images.

A picture containing table

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generatedStep 09: Downloading Preprocessed Data

Step 09: Linear Regression Model

Graphical user interface, text, application

Description automatically generatedFuture predictions can now be made scientifically and with high reliability using linear-regression models. The features of linear-regression models are well understood and can be trained extremely quickly since linear regression is a statistical technique that has been around for a very long time. Therefore, we can use Linear Regression Model to predict the second hand cars Re-selling prices.

Step 10: Scaling the data

The process of normalizing the range of features in a dataset is known as scaling and the scaling technique is required for machine learning models to interpret these features on the same scale in the data set.

Application

Description automatically generated with low confidence

Step 11: Test Train Split

Machine learning algorithms that are applicable for prediction-based algorithms and applications are evaluated using the train-test split. We can quickly and easily use this strategy to compare the output of our own machine learning model to output from other machines. Therefore, to analyze the results we can use the test train split technique.

Graphical user interface, text, email

Description automatically generated

Step 12: Creating the Regression

Chart

Description automatically generated

Chart

Description automatically generated

Step 13: Finding the weighs and bias

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

According to the summary table, some dummis are positive while others are negative. This refers to the relationship between these variables and our target variable "Price" or "log price". A positive dummy value means it is directly proportional to the target variable, while a negative dummy value means it is inversely proportional to the target variable.

When we consider the brands, the relationship is a little different. The values of the brands are positive and negative based on the value of the "Audi" brand, which served as the benchmark variable when the dummies were assigned. This, for positive value of a brand means this brand car is expensive than the "Audi" brand car, and if it's negative means it's cheaper than the "Audi".

Graphical user interface, application

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Table

Description automatically generated

According to the descriptive statistics of the df\_pf data frame, we can see that the predictions were generated good results for the majority of the observations, however there were a large number of outliers therefore our model did not perform well there. Therefore, the results can be illustrated as below.

Graphical user interface, text, application

Description automatically generated

## **Conclusion**

Based on the analysis and the linear regression model, it can be concluded that our model fits the data set and predicts the correct values reasonably well, but it is not an exceptional model. Moreover creating a much better fitting model can be done, however this is always a trial and error method; the more the model is tested, the better it gets.