Student Name: Mohammed Usman E Ghani Student ID: S3901999

Data Preparation of Car Buyers Information Dataset

Introduction

Data preparation is a vital step in data science applied in order to yield useful insights from the data. It is the third step of the process, performed after setting the research goal and retrieving the data. It entails transforming, cleansing (detecting and correcting corrupt or inaccurate values), and consolidating data into a dataframe to enrich the data. This will enable the data to be ready for the next data science operations. Data preparation looks after most of the apparent problems, even if an error is generated, these can be addressed relatively quickly as the possible reasons for the errors are narrowed down (Ashraf, 2020). There were a total of 11 features described within the CSV (comma separated values) files. With the help of the ‘data_description.txt’ file, the csv files were briefly explored. These csv files were then imported to a dataframe and merged using the ‘.read_csv()’ and ‘merge()’ functions from pandas and thoroughly studied to reach a better understanding of the dataset.The merged data frame comprised all the information about cars and the car owners and was ready to perform data preparation techniques.

Data preparation

Error 1 : *Dealing with whitespaces and inconsistencies.*

Whitespaces in a dataset are difficult to notice and can easily cause confusion and issues. For example, ‘pandas ‘ is not the same as ‘pandas’ due to the leading whitespace. The dataset contained values with some trailing and leading spaces like ‘Citigo ‘ that needed to be fixed. This was achieved using the str.strip() function that removes space from the front and the back of the string. Moreover, there were inconsistent capitalisations found among the ‘Manufacturer’ and 'Model' columns. Since python is a case-sensitive language, this had to be changed using the str.lower() function to change all the values in both columns to lowercase. For better readability and to avoid additional programming considerations, the Engine CC column was renamed to Engine_CC. A copy of the merged dataframe, ‘car_df’ was created to prevent impact on the original dataframe. The data type of each feature was also taken into consideration and modified (like changing object to int) for better consistency and usability of the dataset. Removing the whitespace and correcting the inconsistencies, ensured a higher quality dataset.

Error 2 : *Dealing with missing values*

Missing data could be a result of people not responding to a survey or even caused by data entry error. These missing values can have significant effects on the results that will be drawn from the dataset in the following processes. When the dataset was checked for missing values, it was found that there were a total of 44 missing values in the entire dataframe. It was then discerned that the missing values found were not very informative and only a minimal amount of observations, 5 observations had missing features. For this reason, the observations that contained these missing values were dropped using the ‘dropna()’ function. After this filtration, the dataset was further condensed down.

Error 3 : *Correcting spelling mistakes*

The ‘Fuel’ column that contains information about the type of fuel compatible with the car had several spelling mistakes like ‘peatrol’, ‘diasel’ and ‘automatic’. These mistakes were detected by creating an array with unique values in the ‘Fuel’ column and compared against the correct spellings. These mistakes were fixed by ustilising the str.replace() function where the mistakes were replaced with the correct spellings.

Error 4 : *Filtering unexpected values*

The ‘Model’ column consisted of additional values not listed in the description.txt file. For example, models ‘forfour’, ‘Jazz’ and ‘forester’ were never listed in the description file. Observations with these values were dropped as they are not expected to be included in the dataframe. This was attained through comparing the ‘Model’ column against the list of the unexpected values and dropping them accordingly. A similar process was applied for the ‘Manufacturer’ column, however, there were no unexpected values found in this column.

Error 5 : *Sanity checks for impossible or out of range values*

Impossible values can contaminate the overall dataset and may lead to biased evaluations which may result in inaccurate conclusions. There were some impossible values present in the columns representing the car attributes. The possible values are defined as follows: values for price are values between 0.0 and 650.0, for transmission between 0.0 and 10.0, for power between 0.0 and 500 and for Engine CC between 0.0 and 6,500. Any values out of these ranges can be considered as impossible values. In addition, a trend was observed amongst values in the ‘Price’, ‘Transmission’ and ’Power’ column, the negative impossible values were exactly identical to other duplicate values indicating a data entry error (data quality issue) in the dataset. These values were dealt by removing the negative sign to convert them to positive numbers using the .abs() function. In order to gain a more visualised context of the ‘Price’ and ‘Power’ column, various graphs were plotted in later tasks before dealing with these errors.

Error 6 : *Dealing with duplicate values*

The enormous size of the dataset is due to duplicate values throughout. Duplicate observations can be detrimental and can lead to incorrect estimations and other data management issues. Therefore it is important to identify observations that are repeated at an early stage. A large proportion of values in the dataset were duplicated and in order to manage this, the male, female, unknown and total population of car owners of the duplicate rows were summed together based on the key features - ‘Manufacturer’, ‘Model, ‘Price’,’Transmission’, ‘Power’, ‘Engine_CC’ and ‘Fuel’. This step resulted in a dataset that was free from any duplicate values.

Discussion

Poor data quality can lead to poor data insights and unreliable results. Therefore, it is crucial to prepare the data before using it for any analytics applications. Although data preparation is a time consuming process, it can be ensured that the data is consistent and in a usable form to enable more informed decisions. After thoroughly performing all the data preparation processes, the data was written into a csv file with the name ‘cleaned car buyers.csv‘ to ensure that it is ready to perform further data science operations.

Data Exploration

Data exploration enables deeper understanding of the data, revealing various characteristics and potential errors in the dataset. It is an important process to improve project efficiency and understand the nature of the data. A range of graphs were plotted to provide further visual context to the car dataset.

Task 2.1

Representation of total number of vehicle owners by gender for the top ten vehicles with most owners.

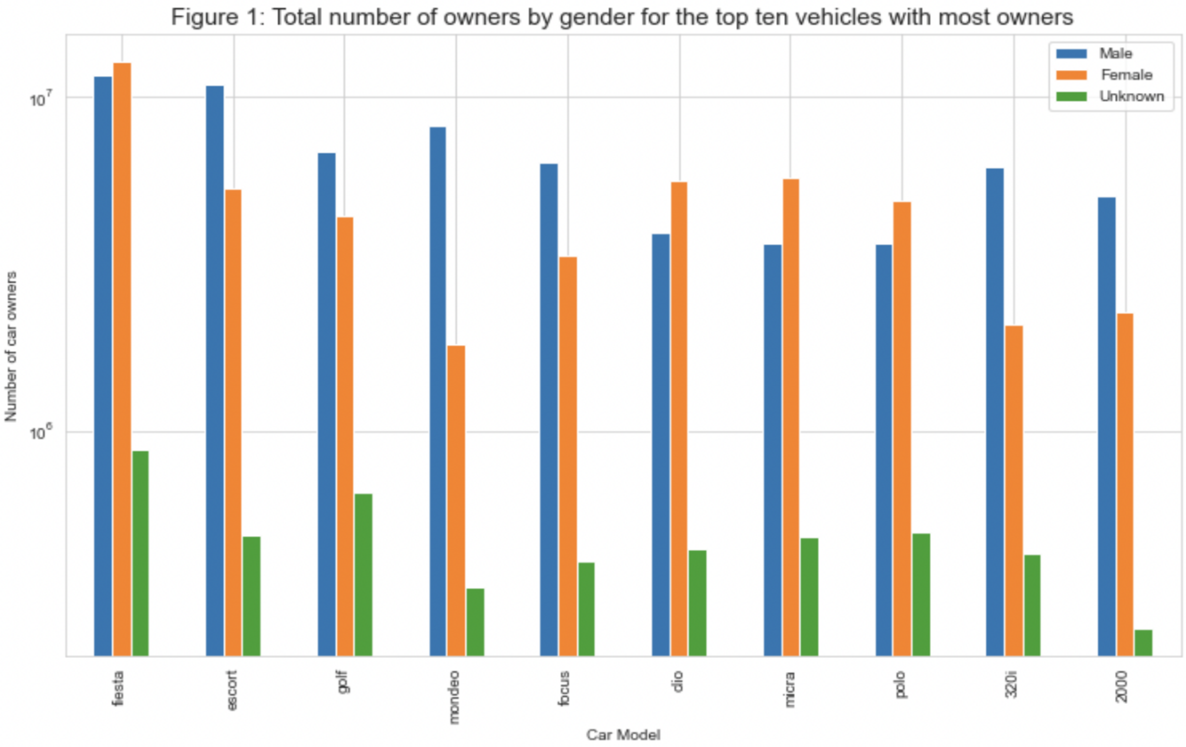


Figure 1 represents the number of car owners by gender for the 10 most owned cars, with the x-axis representing the car models and y-axis representing the number of car owners. It can be inferred from the bar graph that Ford fiesta is the most owned car among the car owners. It is the most owned car across all three genders, male, female and unknown owners. More specifically, Fiesta is the most popular car among female owners as there are more female owners for Fiesta compared to male and unknown genders. Some other female owner dominant cars are Renault clio, Nissan micra and Volkswagen polo. Comparing this to the male owners, there are 6 male owner dominant cars across the plot with Ford escort, Volkswagen golf, Ford focus, Ford mondeo, BMW 320i and Rover 2000 with the highest number of male owners.

Task 2.2

Visualisation of errors in the ‘Price’ and ‘Power’ column

Figure 2a: Distribution of Price column with errors

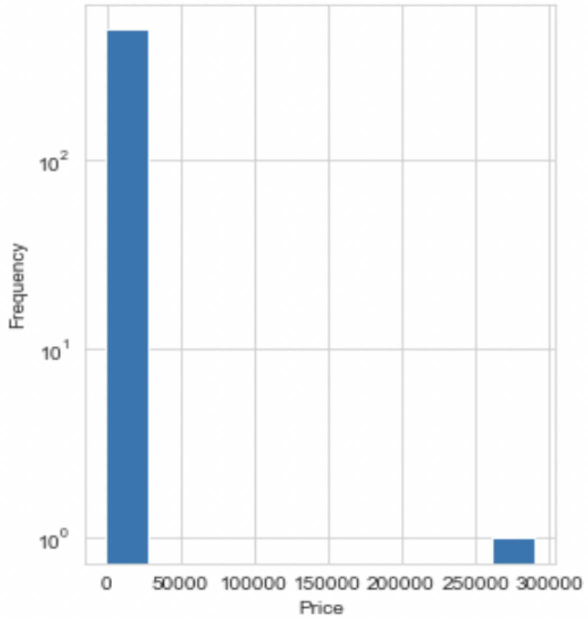
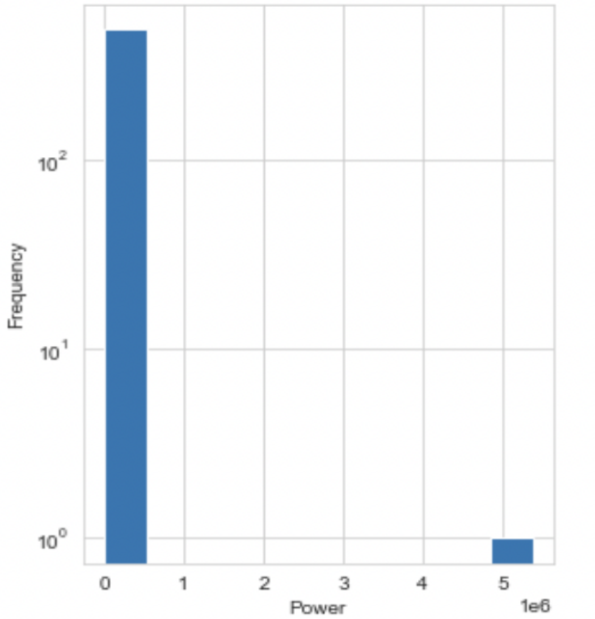


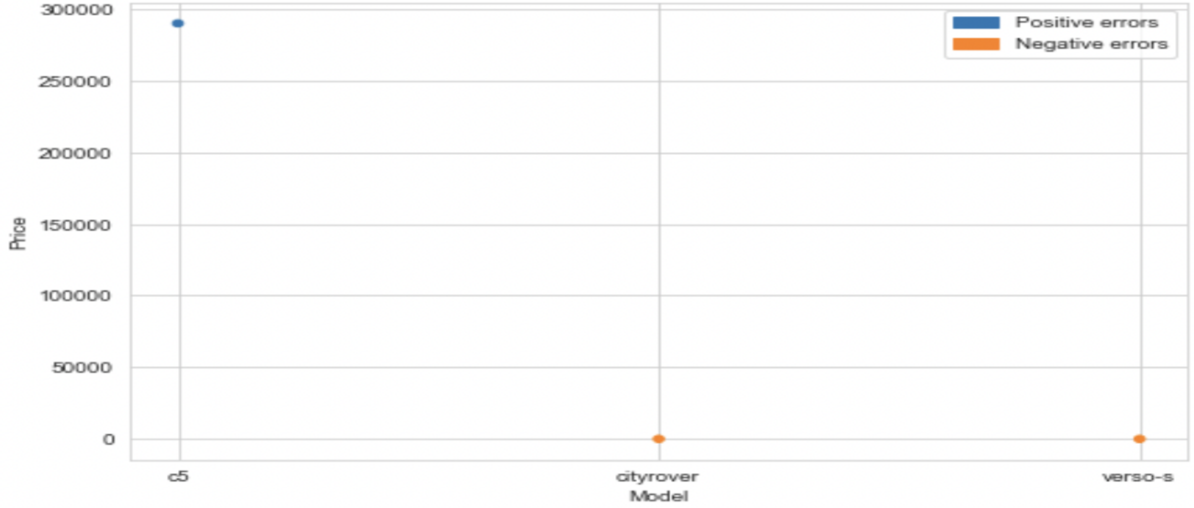
Figure 2b: Distribution of Power column with errors



During the data preparation process, there were errors identified in the ‘Price’ and ‘Power’ column. However, these errors were managed in this task 2 in order to get a visual context of these errors. The price of the car cannot be more than 650 or less than 0 according to the ‘data_description.txt’ file, however, some values that disregard this range were found. It can be observed in figure 2a that there is a small amount of price values above 250000. Similarly, the power of the car which cannot be more than 500 and less than 0 were identified in the ‘Power’ column with 6 values violating this range. This can also be observed in figure 2b where there is a small number of power values more than 500.

In order to determine which car’s information these errors lie on, a scatter plot (Figure 3) with the out of range price values (C5 with the value 290050.26, Cityrover with -11.74 and Verso-s with -22.05) was produced. There were 3 rows in the dataset that disregarded the 0 - 650.0 price range values as presented in the scatterplot. Further, the errors with positive and negative values were differentiated with different colours.

Figure 3: Errors in the Price column



A similar scatterplot (Figure 4) to visualise the power values in the dataset was generated to get a better insight. A total of 6 errors were found in the Focus, Golf, C6, Two and Rav5 car models.

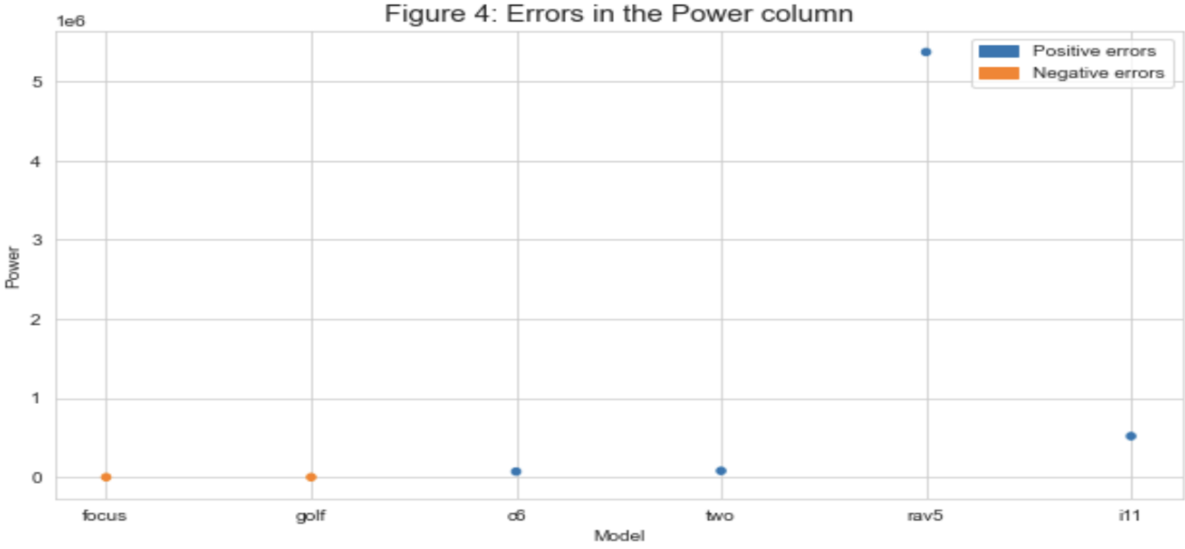


Figure 4: Errors in the Power column

Task 2.3

Relationship between Male car owners and other car attributes

Although men are more likely to spend more money on cars and purchase a car with a higher transmission and engine CC (WrNachbahriter, 2022), the relation in figure 5a, 5b and 5c suggest otherwise. There does not seem to be any correlation between male car owners and the car attributes as it can be shown that as the number of male owner count increases, the price, transmission and engine CC of the car owned does not necessarily increase.

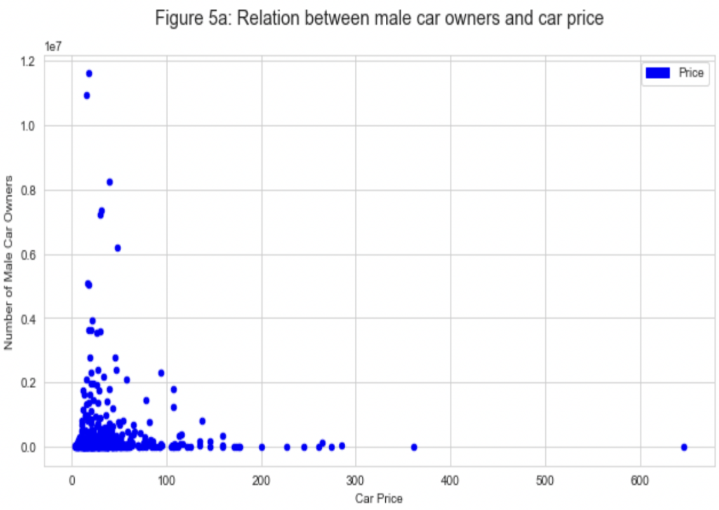


Figure 5a: Relation between male car owners and car price

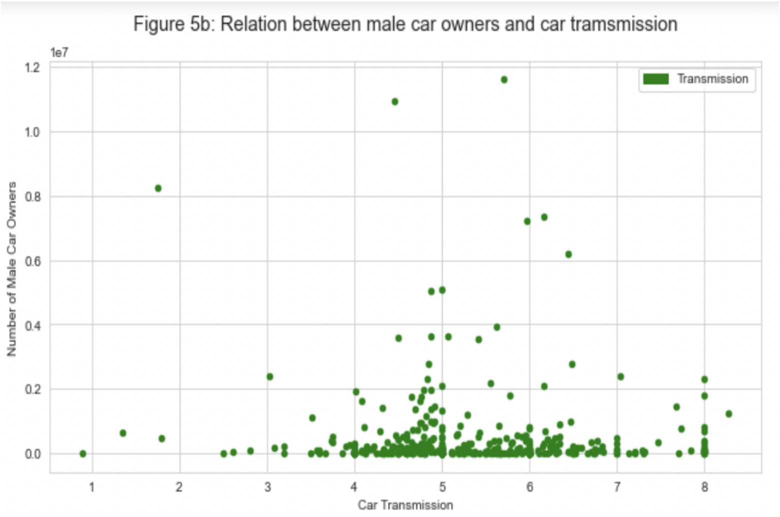
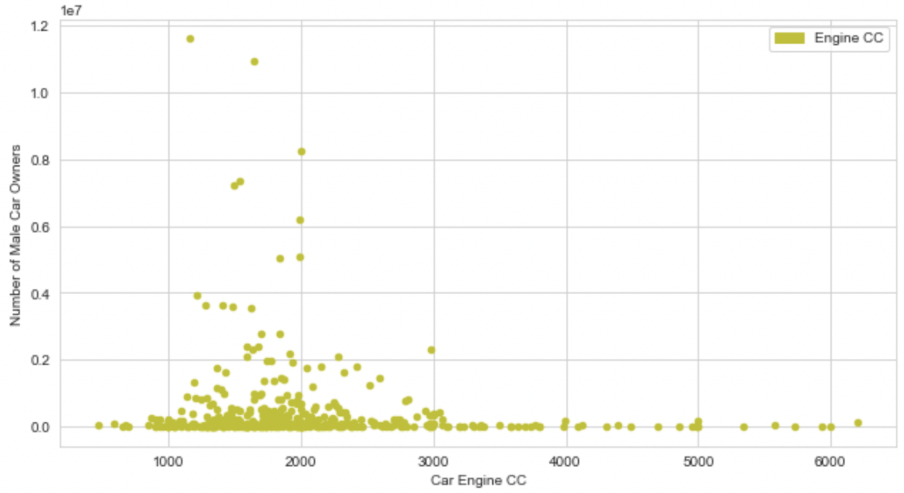


Figure 5b: Relation between male car owners and car tramsmission

Figure 5c: Relation between male car owners and car engine CC



The data points on the scatterplots in Figure 5a, 5b and 5c are scattered randomly with a zero slope without any trends, indicating no linear relationship between the number of male car owners and the car attributes. In addition to this, the data points in the scatterplots are clustered around certain values (4-6 in transmission, 0-100 in price and 1000 to 2500 in engine CC) demonstrating male car owners' preferences for the respective attribute.

Exploring male car owners and other gender columns with respect to the type of fuel used in cars

Figure 6: Total number of car owners by gender for the type of fuel used in the cars

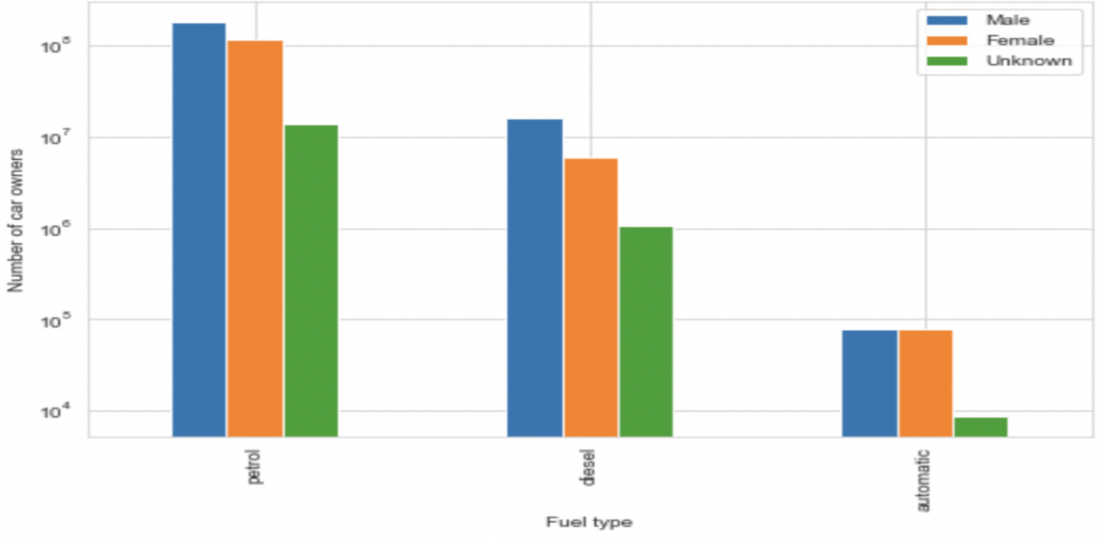


Figure 6 represents the most common fuel types used by the male car owners compared to other genders. Petrol is identified as the most common type of fuel among all car owners followed by diesel and automatic. Male car owners have dominance over all fuel types compared to other car owner genders.

References

ASHRAF, S., 2020. Data Preparation Process Explained: Steps, Benefits, & Tools. [online] Data Preparation Blog. Available at: <https://dataintegrationinfo.com/data-preparation-process/> [Accessed 29 March 2022]

WrNachbahriter, E., 2022. *Study Shows Men Spend More Time and Money Car | Digital Dealer*. [online] Available: <https://digitaldealer.com/dealer-gm/study-shows-men-spend-more-time-and-money-car-shopping-than-women/> [4 April 2022].