A picture containing text, nature

Description automatically generated

Nushura Islam (Nush)

**Data Cleaning**

**and**

**Summarising**

**Assignment 1**

Data Preparation

The first cell of this section was dedicated to importing packages and libraries. It was followed by retrieving data files from the computer directory and concatenating those records, creating a single data frame.

**Unnamed index columns**

However, after concatenation, it could be observed that there was an unnamed column before each featured column. Those were acting like an index column for each of the features before concatenation, but on the newly concatenated data frame, there was already an index identifier which meant that the unnamed column was not needed. Hence, to remove the ‘Unnamed: 0’ column, drop( ) method was used.

**Missing values**

There are a few different ways of handling missing values depending on the dataset provided. As the given dataset was a huge dataset consisting of 6102 rows, it was not physically possible to go through all the columns and judge which values were missing. Thus, the dropna( ) method was used, which assisted in dropping all the rows from the table with any missing values. After dropping those rows, it could be observed that the number of rows dropped from 6102 to 6097, which meant that the difference among the dropped rows was not much. Hence, it will not be a problem as more data was present after removing missing values.

**Duplicated rows**

It was checked if there were any duplicated rows using the duplicated( ). This was followed by using drop\_duplicates( ) method for removing the duplicated. However, after trying to drop the duplicated rows, it could be observed that no rows were dropped. Thus, it could be understood that there were no duplicated dataset rows.

**Removing duplicated unique identifier**

The unique identifier for the dataset was the ‘Model’ column. There were few of the same model cars with similar ‘Price’, ‘Transmission’, ‘Power’, ‘Engine CC’ and ‘Fuel’ values, but they differed from the genders and thus, total values. This meant that those rows were not considered when duplicated rows were dropped. This kind of case has to be handled differently. For example, in this case, the highest value is assessed for each of the duplicated models, and the rest are removed. It is implemented by using the nlargest( ) method.

**String to integer conversion**

It could be observed through the data frame that there were integer values written as string values within the ‘Total’, ‘Male’, ‘Female’, and ‘Unknown’ columns. Those values within the columns are initially considered object types that must be specified as string types. This is done using the method called astype( ). It is followed by removing the comma using replace( ) method and converting the string to an integer using the astype( ) method again.

Data Exploration

# 2.1

This section is visualised by using bar charts. There are three bar charts created where each one represents a particular gender. There are a few of the standard implementation techniques that have been used for this section. The head(10) represents the first ten rows of the cleaned dataset. The x and y-axis are pointed out before plotting the chart. In each of the charts, the gender of car owners is on the y-axis, and the car models are on the x-axis. Furthermore, labelling is done to understand which axis represents what resource.

**Male:** The car model called ‘Escort’ is the most popular, and 206 is the least popular among the male car owners.

**Female:** The ‘Fiesta’ car model is the most popular and ‘Sierra’, ‘Mondeo’ and ‘320i’ is the least popular among the female car owners.

**Unknown:** The ‘Focus’ car model is the most popular, and ‘Sierra’ is the least popular among the unknown car owners.

The car model's most significant number of car owners is represented by Fiesta, which could be observed by the ‘Total Car Owners per Model’ bar chart.

# 2.2

This section is visualised by using box-plot diagrams. To explore an entry error, a method that can be used is measuring the outliers present within a boxplot diagram. For both the boxplots of ‘Price’ and ‘Power’, the initial box-plots x-axis’s length was high due to outliers. The x limit for both the features had to be adjusted for a zoomed-in view and better analysis.

**Price**: The price boxplot’s lower quartile value was 1.6, and the upper quartile value was 82. It could be observed through the initial Price boxplots that any integer which was above 82 was an outlier. Thus, to remove the outliers, this method was used.

Any value less than 1.6 and greater than 82 should be removed. After running this code, it could be observed that a new set of outliers are found. This proved that it is not a normal distribution. Hence, the outlier removal process had to be done again, but the limit was 1.6 to 66.6 as 66.6 was the new upper quartile. After plotting the boxplot again, all the outliers got removed.

**Power**: The power boxplot’s lower quartile value was 15, and the upper quartile value was 207.6. It could be observed through the initial Powe boxplots that any value which was above 207.6 was an outlier. Thus, to remove the outliers, this method was used.

# This meant that any value less than 15 and more significant than 207.6 should be removed. After executing this code to make a boxplot, it could be seen that a new set of outliers are found. This proved that it is not a normal distribution. Hence, the outlier removal process had to be done again, but the limit was from 1.5 to 175 as 175 was the new upper quartile. After plotting the boxplot again, all the outliers got removed.

# 2.3

This section is visualised by using scatter plots as it assists in understanding the relationship between two variables. This was concluded by visualising the part of the plot where it is the densest.

**Male car owners versus car price**: It could be observed that the highest number of car owners were between 0 to 200 0000 with car prices between 10 to 50.

**Male car owners versus car transmission**: In this plot, it could be observed that the highest number of car owners were with car transmission between 3.5 to 8.

**Male car owners versus car power**: It could be observed that the highest amount of car owners was with car power between 40 to 160.

Overall, all three plots have no trends or patterns within the plots, which means that there is no particular association.

References

* Wallpaper Access, *Green Patel Wallpapers*, Wallpaper Access, viewed 17 March 2022, < <https://wallpaperaccess.com/green-pastel> >
* **retrieve file:**

Pierson, L 2019, *Python for Data Science Essential Training Part* 1, streaming video, LinkedIn Learning, viewed 26 March 2022, < <https://www.linkedin.com/learning/python-for-data-science-essential-training-part-1/grouping-and-aggregation?autoplay=true&resume=false&u=2104756> >

* **replace()**

W3 Schools 2022, *Python String replace() Method*, W3 Schools, viewed 26 March 2022, <<https://www.w3schools.com/python/matplotlib_labels.asp> >

* **dropna()**

Pierson, L 2019, *Python for Data Science Essential Training Part* 1, streaming video, LinkedIn Learning, viewed 26 March 2022, < <https://www.linkedin.com/learning/python-for-data-science-essential-training-part-1/treating-missing-values?autoplay=true&resume=false&u=2104756> >

* **concat**

Pierson, L 2019, *Python for Data Science Essential Training Part* 1, streaming video, LinkedIn Learning, viewed 26 March 2022, < <https://www.linkedin.com/learning/python-for-data-science-essential-training-part-1/concatenating-and-transforming?autoplay=true&resume=false&u=2104756> >

* **duplicate drop:**

Pierson, L 2019, *Python for Data Science Essential Training Part* 1, streaming video, LinkedIn Learning, viewed 26 March 2022, < <https://www.linkedin.com/learning/python-for-data-science-essential-training-part-1/removing-duplicates?autoplay=true&resume=false&u=2104756> >

* **astype():**

Rajan, S 2019, *Python| Pandas DataFrame.astype()*, GeeksforGeeks, viewed 26 March 2022, < <https://www.geeksforgeeks.org/python-pandas-dataframe-astype/> >

* **nlargest():**

W3 Schools 2022, *Panndas DataFrame nlargest() Method*, W3 Schools, viewed 27March 2022, <<https://www.w3schools.com/python/pandas/ref_df_nlargest.asp> >

* **head():**

Bhutani, K 2018, *Python | Pandas DataFrame/Series.head() method*, GeeksforGeeks, viewed 02 April 2022, <<https://www.geeksforgeeks.org/python-pandas-dataframe-series-head-method/> >

* **Bar chart:**

Pierson, L 2019, *Python for Data Science Essential Training Part* 1, streaming video, LinkedIn Learning, viewed 02 April 2022, < <https://www.linkedin.com/learning/python-for-data-science-essential-training-part-1/creating-standard-data-graphics?autoplay=true&resume=false&u=2104756> >

* **Box plot():**

Chart Explorers 2020, *Boxplots with Python | As fast as possible*, YouTube, viewed 10 April 2022, <<https://www.youtube.com/watch?v=MmApBcxSgMk> >

* **Scatter plots:**

Kalamkar, S 2022, *matplotlib.pyplot.scatter() in Python*, GeeksforGeeks, viewed 10 March 2022, <<https://www.geeksforgeeks.org/matplotlib-pyplot-scatter-in-python/#:~:text=The%20scatter()%20method%20in,in%20one%20affects%20the%20other.&text=The%20scatter()%20method%20takes,array%20containing%20x%2Daxis%20data> >

* **Labelling:**

W3 Schools 2022, *Matplotlib Labels and Title*, W3 Schools, viewed 02April 2022, <<https://www.w3schools.com/python/matplotlib_labels.asp> >