S3774430
Myeonghoon Sun

1. Data Preparation

1-1. Missing Values

Once a dataset can be manipulated in the form of DataFrame, one of the first things one can do for pre-processing is to check whether there are any rows with missing values. Upon inspection, there are four rows with NaN values in all the columns except for *Manufacturer* and *Model,* as well as one row with workable values only in *Engine CC, Male, Model,* and *Transmission.* In both cases, there is no way of recovering missing values in the *Total, Female,* and *Male* columns we need for data exploration. Therefore, it is the most sapient thing to drop these observations. From here on, no row in our dataset contains missing values.

1-2. Redundant Whitespace

After isolating categorical features and subsequently checking the counts of unique values in each of the features in ascending order, found in *Manufacturer* are three values whose counts are 1 each. Somewhere beneath them are also found the same values with higher counts. It can then be inferred that the values with only one count contain trailing whitespace. This issue and potential issues of the same nature have been resolved by removing whitespace from all the values of categorical features.

1-3. Data Entry Error

From the previous check executed in 1-2, it is apparent that *Fuel* contains what seem to be human typos, namely autometic for automatic, diasel for diesel, peatrol for petrol. These typos have been corrected. Furthermore, the *Model* column is now free from this type of error by performing the masking operation to keep only the observation with the values mentioned in data_description.txt.

As with categorical features, numerical features can be extracted to gain more information from their descriptive statistics. Although the descriptive statistics of the features, *Unknown, Female, Male,* and *Total* seem reasonable at first glance, it turns out that there are a few rows where the total number of owners is not equal to the sum of the number of male, female, and unknown owners.

The total values of the above observations seem more like miscalculations as the correct sums near them very closely. Therefore, they have been updated to the correct sums.

1-4. Sanity Checks for Impossible Values

According to data_description.txt, *Engine CC*'s values should range from 0 to 6,500 (inclusive). From its descriptive statistics, we can see that its minimum value and maximum value stay well within the range.

However, the *Price* column has one observation with an impossible value out of its specified range. After other rows of the same model – CityRover – are searched, it becomes obvious that the minus in front of it is a typo. Therefore, it has been replaced by its absolute value.

Given the provided range of *Power*, the masking operation shows that there are three rows that need correcting. The same process seen in dealing with the *Price* column can be applied for the first two rows. It again turns out that the minus signs should be removed. For the last observation, it required searching on the internet for the unprocessed BHPs of the first two rows and its own to accurately conclude that 80120.120000 is meant to be 80.120000; Mini Cooper – Generation 2 in the diesel guise comes with 90 or 110 BHP ("Models of MINI Cooper – Generation 2", 2022), implying its petrol guise is less than 90 and making 80.120000 an ideal candidate to replace the original impossible value with.

We can see a similar trend in the *Transmission* column. Two observations with transmission values greater than 10 have other traceable observations from which the correct values can be obtained. The sole row with a negative value in *Transmission* is once again easily corrected by updating it to its absolute value.

2. Data Exploration

2-1. The Total Number of Owners for Top 10 Vehicles

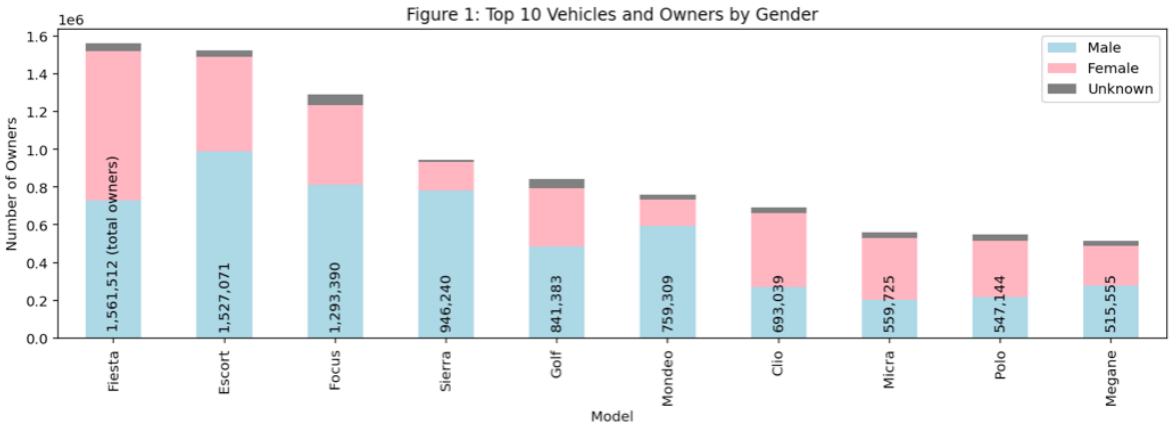
We find multiple observations with the same values all belonging to two of the models: Fiesta and Escort. The values in columns concerning the number of owners among these observations are not completely identical, but very similar as if they were recorded over time. We could certainly argue that they are the same model vehicles produced over the years, having separate data of their own. However, this reasoning does not hold against the counter argument that points out such small differences in the number of owners between the years. Therefore, it is logical to conclude that these are aggregate observations that need to be represented by, per model, one row with the max value in the *Total* column as it is assumed to be the latest data point according to our logic.

Top 10 Vehicles After Cleaning

Model	Male	Female	Unknown
Fiesta	730276.0	789633.0	41603.0
Escort	989746.0	501907.0	35418.0
Focus	814172.0	422731.0	56487.0
Sierra	781210.0	150663.0	14367.0
Golf	483216.0	310604.0	47563.0
Mondeo	594469.0	138468.0	26372.0
Clio	269970.0	390458.0	32611.0
Micra	201479.0	328691.0	29555.0
Polo	216333.0	299110.0	31701.0
Megane	279372.0	208847.0	27336.0

The number of duplicated observations sharing the same *Engine CC, Price, Manufacturer, Power, Model, Transmission,* and *Fuel* is staggering, amounting to 3,757 rows. We could certainly address this problem by creating a row, per vehicle, that contains the highest *Total*, *Unknown*, *Female*, and *Male* values found across its duplicated rows. However, the row with the highest total value does not always contain the maximum value in *Female*, *Male*, or *Unknown* column compared to other rows. We could justify that the drop in the number of male owners from one row to another, which contains the maximum total value, is due to some male owners selling their Fiesta and buying other vehicles. However, if that reason is assumed to be valid, an argument can be made that the row with the maximum total number of owners does not represent the latest data point as there can be cases where the owners sell their vehicles, leading to a non-maximum number in *Total* being the latest data point. Therefore, the best way to address this problem is to take the row with the maximum total value for all the concerned duplicated observations. The resulting total number per vehicle might look very low for a tally for its entire production years, but it is reasonable if it is a record for one particular production year.

After the above cleaning process, we can now plot the composition for the total number of owners by gender for the top ten vehicles with the most owners (Figure 1).



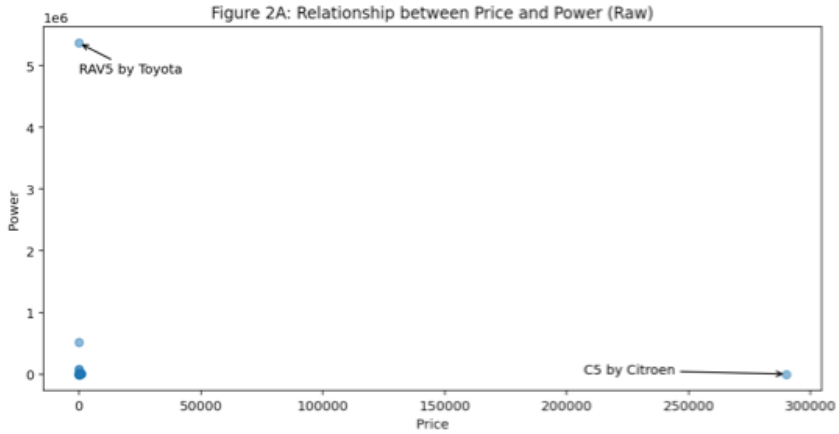
From the graph above, we can see that some models like Sierra and Mondeo are male predominant whereas more female owners own models like Clio and Micra than male owners do. The most popular model Fiesta is close to being gender neutral. The detail statistics can be found in the table in the attached jupyter notebook.

S3774430
Myeonghoon Sun

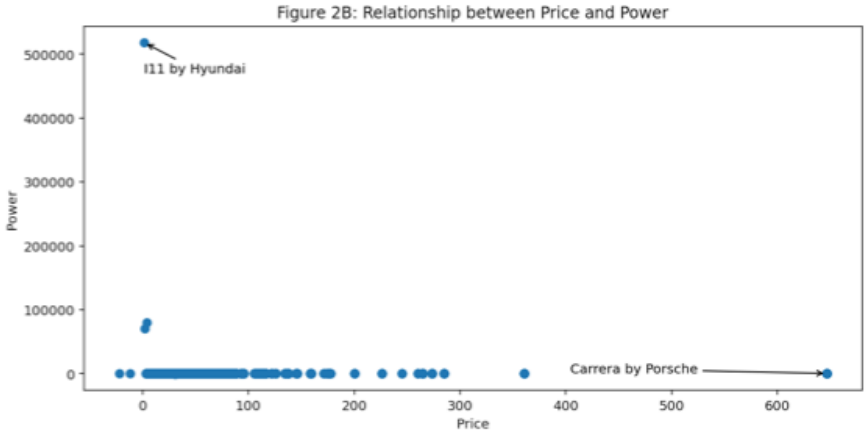
2-2. Exploring Errors in the Price and Power Columns

Since the dataset has been thoroughly cleaned, no anomalies can be observed in the graph below (Figure 2), except for the one data point far in the upper right corner. However, the data point is none other than Carrera by Porsche and its off-the-charts values are reasonable in both axes.

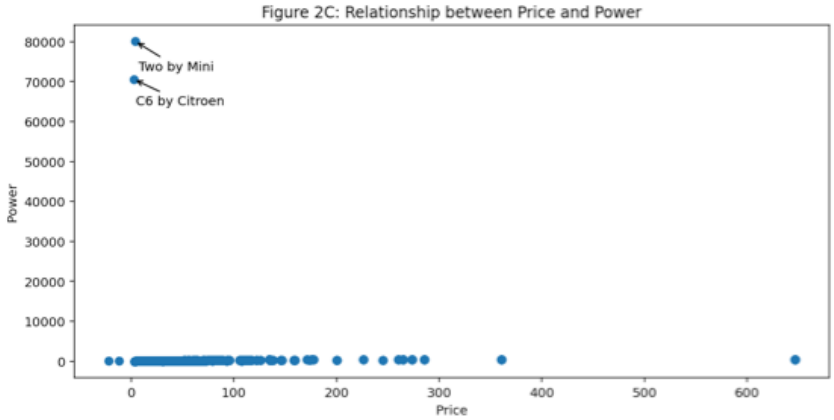
Now let us see how this graph has come about by plotting raw data from two of the columns. As seen in the graph below, two impossible values can be found in both axes in the first plot.



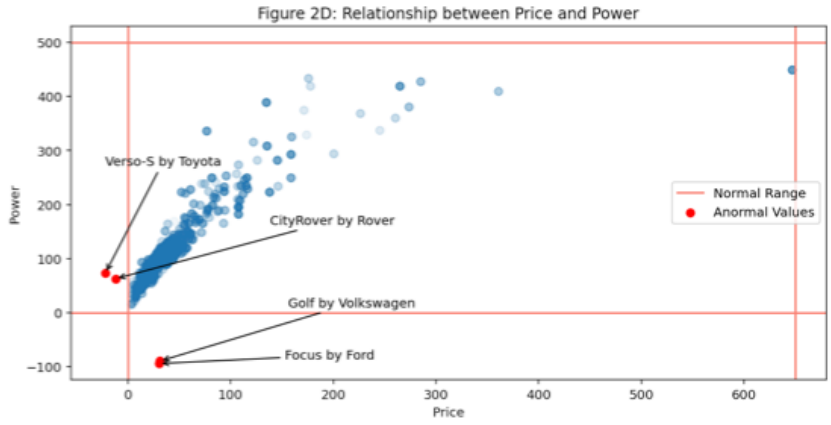
Once they are removed, we can see other values far away from the rest of the data points (Figure 2B below). Here, a data point representing I11 by Hyundai is an impossible value to be removed, whereas the other one is here to stay as mentioned in the beginning.



Once again, we find two data points below to eliminate.



Lastly, we can finish removing the points outside the boundaries in red to come full circle to the graph we saw at the start (Figure 2 above).

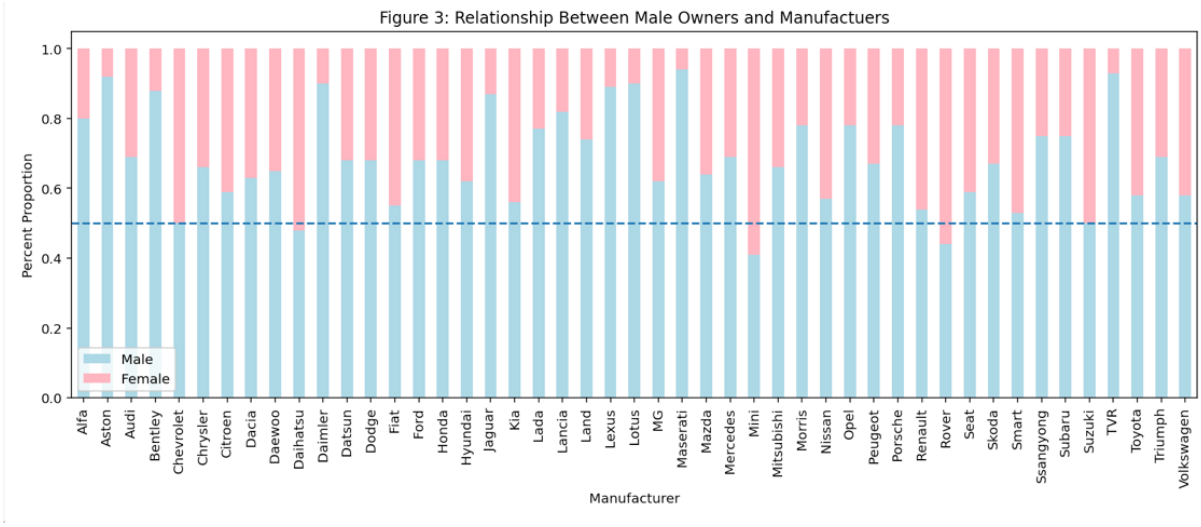


2-3. The Relationships Between Male Owners and Other Columns

2-3-1. The Relationship Between Male Owners and Manufacturers

Even though the task never mentions female owners, we must take them into account to truly understand the relative relationship between male owners and other attributes. Therefore, in all subsequent plots, we will include corresponding data about female owners to give a complete picture.

We can start off with this basic question: what vehicle manufacturers are most popular among men compared to women? To answer this question, we will compare the percentage of male and female owners per each manufacturer.

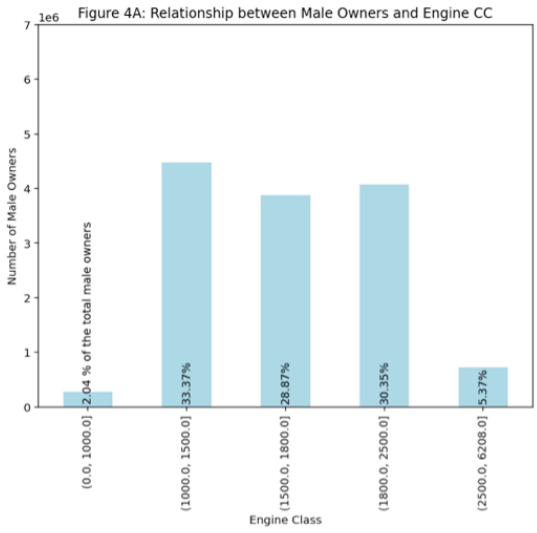


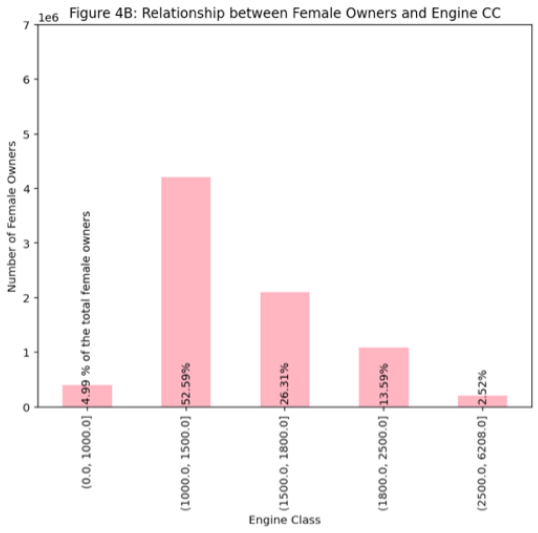
As we can see above, it comes as little surprise that certain manufacturers like Mini, Daihatsu, and Rover have proportionally larger traction from female owners than male counterparts, arguably given the looks of their line of models. On the other hand, we can see the males' overwhelming preference for luxury car brands such as Alfa, Aston, Bentley, Daimler, Jaguar, Lexus, Lotus, Maserati, and TVR.

2-3-2. The Relationship Between Male Owners and Engine CC

Engine capacities can be classified into 5 categories ("Engine Capacity (CC): Engine Volume", 2022):

1. Up to 1000cc (Small Cars and Hatchbacks)
2. 1000cc to 1500cc (Family Cars)
3. 1500cc to 1800cc (Mid-size Cars and Small Wagons)
4. 1800cc to 2500cc (Semi-luxury Cars, Wagons, MPVs and SUVs)
5. Above 2500cc (SUVs, Sports Cars, High-end Luxury Cars)



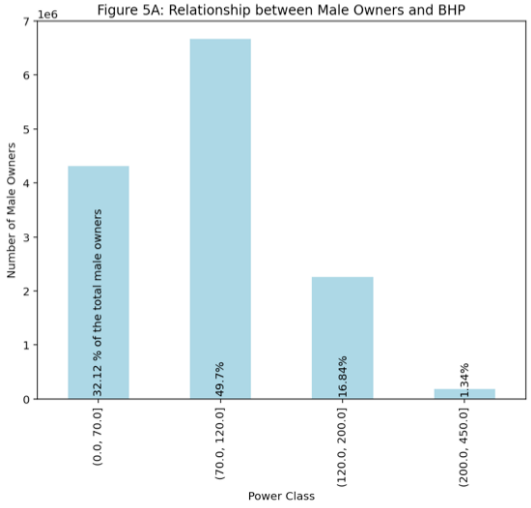


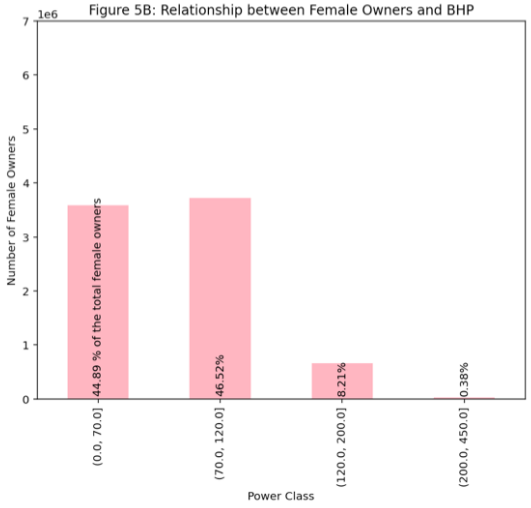
We could see in the graph above that 64.59% of the male owners own vehicles with an engine capacity equal to or greater than 1500cc, whereas 57.58% of the female owners own either small vehicles or family cars. It is reasonable to conclude men lean towards owning vehicles with large engines than women do.

2-3-2. The Relationship Between Male Owners and Power

As with the previous relationship, the *Power* column can be binned into 4 classes ("What is a good BHP for a car?", 2022):

1. Up to 70bhp (Smaller Car)
2. 70bhp to 120bhp (Average Car)
3. 120bhp to 200bhp (Larger SUV)
4. Above 200bhp (Sports and Luxury Car)





The graphs above support the finding in the previous relationship. Men who own vehicles with a powerful engine outnumber female counterparts not only in absolute numbers, but in proportion within their respective populations. In summary, being male has a higher correlation with owning greater-bhp vehicles.

References

- *Models of MINI Cooper – Generation 2*. The Mini Specialist - Mini Sales and Servicing. (2022). Retrieved 3 April 2022, from https://www.theminispecialist.com/mini-knowledge/models-of-mini-cooper-generation-2/.
- *Engine Capacity (CC): Engine Volume*. CarBikeTech. (2022). Retrieved 4 April 2022, from https://carbiketech.com/engine-capacity-cc/.
- *What is a good BHP for a car?*. Homex.com. (2022). Retrieved 4 April 2022, from https://homex.com/ask/what-is-a-good-bhp-for-a-car.