Course Name- PRACTICAL DATA SCIENCE WITH PYTHON

Title – **Report of assignment 1**

Name – Sasha Nazareth Student number- s3912063

## Data Preparation: -

The approach I went ahead with for data preparation was to impute a value from an estimated or theoretical distribution.

1. Converting the file into a csv file

To convert the data set into a csv file firstly, I merged the columns one by one using the merge function (.merge). As there was an extra index column while I was merging it was first showing an error to merge. Therefore, I dropped the index column and then converted the file into a csv file.

2. How I dealt with missing values in the data-

There were two ways for me one was to the drop the values which were null and second was to compare with other columns and fill the values.

Hence, firstly I started to check how many missing values are present in the column. I used the function, {cleaned_car_buyers ["row name"].isnull().sum()} to check all the missing values for each column. Once I was aware of the number of missing values, I started to compare with the column that had a strong relationship with the other column. For example, for the model missing values I compared it with the manufacturer column as a specific model will only be from one manufacturer. I used this approach for columns engine CC, manufacturer, power, price and transmission.

- Error 1- of not having enough data in a particular row

Once I filled the missing value for model I realised, in that specific row (65) all the values were missing. Therefore, it did not make sense for me to keep that row. Hence, I dropped the row 65. Using the [.drop] function.

Manufacturer's missing value, I compared it with model and filled the values. Using [.at].

For the missing values of power, transmission, price, engine CC and fuel I have compared all the missing values with the model of the row and then filled the null values, using the same function mentioned above.

In fuel there were a few values which had diesel and petrol in the same model, but I took the value which was the most prevalent in that model. Just to check once again, I also compared it with the manufacturer column. And then only filled in with the appropriate value.

- Error 2- of not having data for four columns in the same rows

Going ahead with my data cleaning for male, female, total and unknown at first, I thought I could use the same approach. But I realised there were four rows which were null altogether therefore, could not compare also hence, I could not use the same approach.

I assumed how to fill such values, hence filled in “gender not specified”. But as I moved forward with my data cleaning, I realised I could not use this approach as total is dependent on the number of males and females. Hence, I could not leave it as “gender not specified”.

- Error 3- of ‘,’ in integer values

In the end, I noticed in the data the numbers have ‘,’ and hence would not be considered as an integer. I wanted to change the data type of males, females, total and unknown as integer values (I checked the data type using the [.dtypes] function) and a ‘,’ would not allow me to change it to an integer, hence I had to remove the ‘,’. (to remove the ‘,’ I used the replace function [ .str.replace(',' ,' ')])

- Error 4- of data types

Once, I sorted the ‘,’ I noticed the data types were first defined as an “object”. Therefore, I replaced the commas and changed the type as a float at first. (using the [.dtypes] function) Then I compared the null values with the similar model and filled in the values. I rounded up the figures too. In the end after replacing and filling the values I changed the data type as an integer.

One noteworthy thing that I noticed was when I tried to change the ‘object’ type into an ‘integer’ first it did not work, an error always came. I had to change the ‘object’ type into ‘float’ first and then only I could change it into an ‘integer’ in the end.

- Error 5- of exponential values when taking the mean for many columns

The most interesting find for me was when I was taking the mean values for all the ten columns, exponential values were coming up as the mean. At first, I was going to drop those rows as an exponential value could affect my graph, but what I learnt was if I round up the values the error goes away. Hence, then I started to round up the values so that I do not see exponential values. I used [.mean().round()] to solve my problem.

Error 3,4 and 5 were present in Male, female and unknown. I dealt with each column the same way.

While for the total values I added the male, female and unknown data together as it was mentioned in the data description.

3. How I cleaned all the inappropriate values according to the description given.

- Error 6- of negative values

Firstly, for the price there were two negative values and when I compared it with the same model, I observed it was just a negative value which was filled. It could have been an error made by the person building the model. Therefore, I just filled in the positive value. This was

for the first condition of price. In transmission and power there were a few negative values too and I used the same approach to fill the values too.

- Error 7- of wrong data added in the dataset

While checking for the inappropriate there were a few values which were wrongly put in. for example in price for the model C5 in one of the rows the price was given as 290050.260250 while all the other C5 models had the price of 34.909057. Hence, it was just a mistake of filling in wrong data. Therefore, after comparing with the same model I filled in the right price. This error was in the transmission data set too and I handled it in the same way.

- Error 8- of zero values in the dataset

In transmission, there were 127 rows which had the value as 0. At first, I did not want to drop all the values as it would affect the data set. But as I checked all the models, none of them were mentioned in the data description which was provided to us. Hence, in the end I dropped all the values as they were not significant for our data.

When I tested for the values for power which were more than 500 there were four values which cropped up. Unfortunately, these models had only one row, and these models had nothing for me to compare. Hence, then I tried to compare it with the manufacturer. But only for the manufacturer 'Citroen' the mean value was less than 500. I checked for the other three manufacturers, but all their mean values were more than 500 which did not meet our data description requirements. Therefore, I could only fill the value for the model C6. While for the others, I dropped the values as there was no other way in my opinion.

Engine CC met all the data requirements, hence, did not need to change anything.

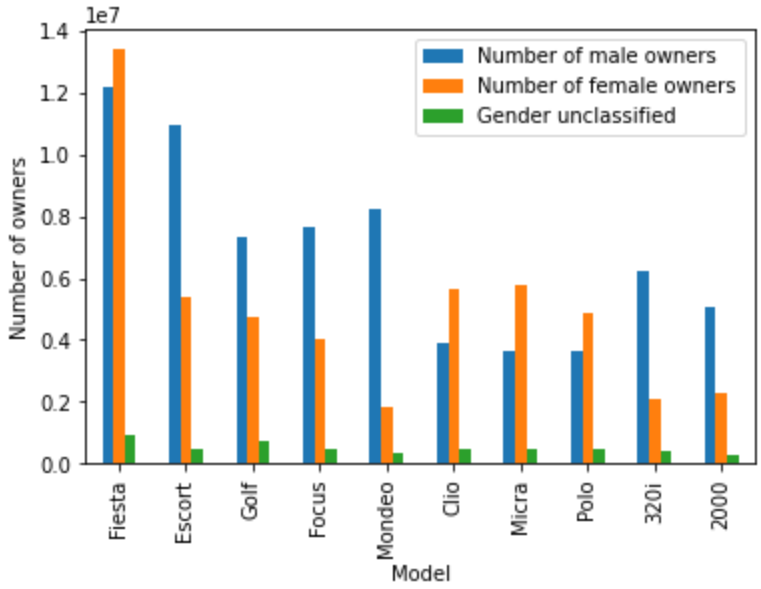
- Error 9- Wrong spelling mistakes

In fuel, there were spelling mistakes made in petrol, diesel and automatic. I used the replace function to change all the wrong spelling mistakes into the correct ones.

In the end, after cleaning all the data I changed the column names which needed to be changed by using the rename function (.rename). And stored the data in a csv file.

- Error 10- in graphs

For visualisation, I faced a few errors, hence, would like to mention it. While I wanted to create a bar graph, until and unless I changed the model index into a list there was error coming as to it could not create my graph due to a key error in 'model'. The model column was an index in the first data frame as it was grouped by the model column. Therefore, I created a new data frame with all the lists in it. That's when my bar graph was created with no errors. For the other graphs I did not face any problem.




**Data exploration:-**

**Task 2.1**

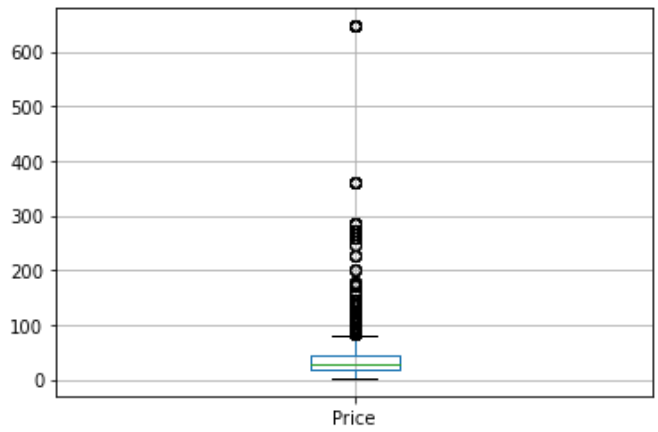
In this task I chose to use a bar graph as we needed to analyse the composition.

Firstly, I grouped the data with respect to 'Model'. And sorted the values by 'Total number of owners'. Lastly, I created a new data frame with the models which had the greatest number of owners.

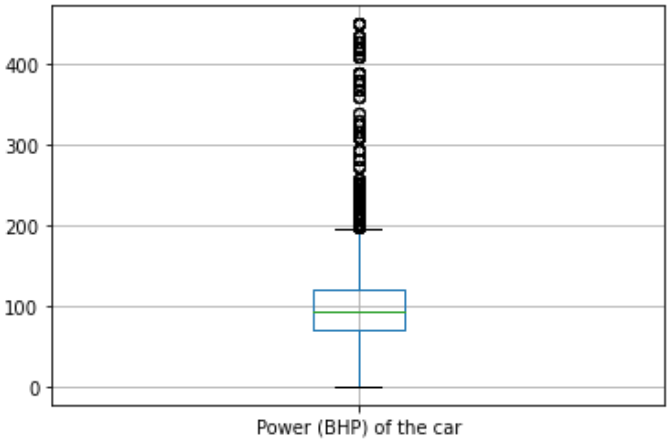
As we can observe, Fiesta is the topmost model. And an interesting finding is that it is bought the most by females. It is also the model which is bought the most by the group of 'gender unclassified' and 'males' if we compare it with all the models.

**Task 2.2: -**

I chose to create a box plot for these questions, as according to me errors can be seen evidently in a box plot.




As we can observe in the above diagram there are no errors seen because in task 1, I have cleaned the data and corrected the inappropriate values according to the data description. Hence, in this graph there are no errors seen. If there were errors in my data after cleaning, they would have been below 0 value, if there were any negative values. And they would have also been more than 650. But as mentioned in the data description I cleaned the negative values and whichever value was above 650 (there was only 1 model C5) which I checked and filled in the right value. Most of the values are above the third quartile.



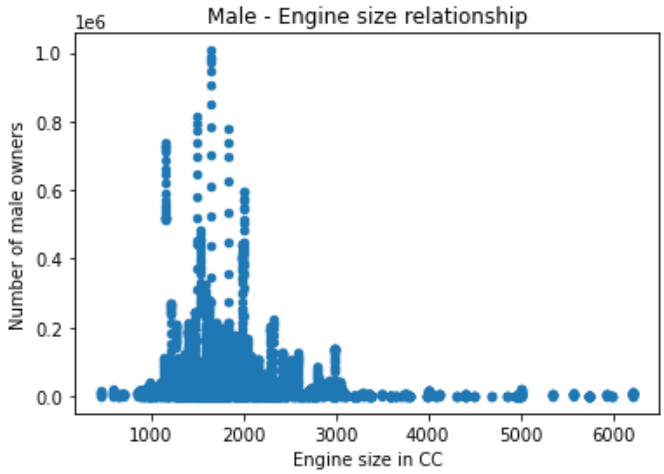

While in the power graph, there is no errors too. It is the same scenario as explained in the price graph. As the errors were cleaned by me in the data cleaning process, we cannot notice any errors. There were two negative values in the original data, hence I changed the values to a positive integer. In the data description the highest value for power was 500 and there were four values which were above 500. I could only change one value to the correct value, but the others I dropped because I could not make any comparisons. In the power graph too, most of the values are in the third quartile.

Therefore, in the two graphs above we cannot see any error as these errors were sorted in the data cleaning process itself. If I had not cleaned the data well, I would've noticed errors in my graph at the end.

**Task 2.3: -**

**First graph-**

I used a scatter plot to analyse the relationships because to show any type of a "relationship" between two entities it is the best visualized by a scatter plot.



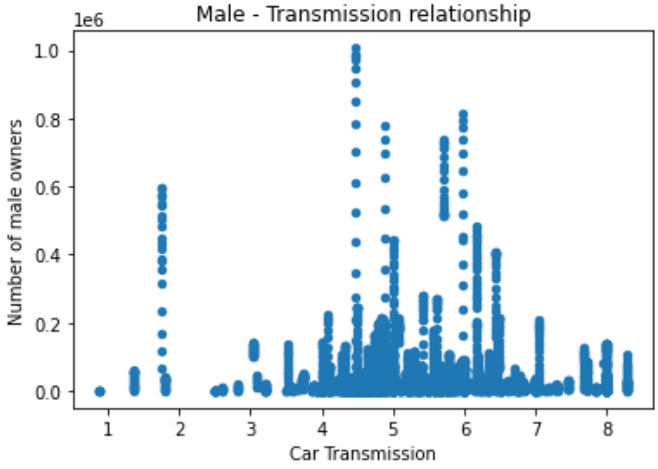

Hypothesis- After 1000 engine size, the number of male owners increase up till 3000 CC hence a positive relationship, but after 3000 CC there is a negative relationship as, the number of male owners decrease as the engine size increases.

In the above graph, I observed a relationship between male and the engine size of a car. As we can notice, the highest number of male owners prefer to buy a car whose engine size is

between 1000 engine CC to 3000 engine CC approximately. As the engine size increases the number of male owners is decreasing after 3000 engine CC.

Insights- For male owners a higher engine size does not matter much.

**Second graph-**

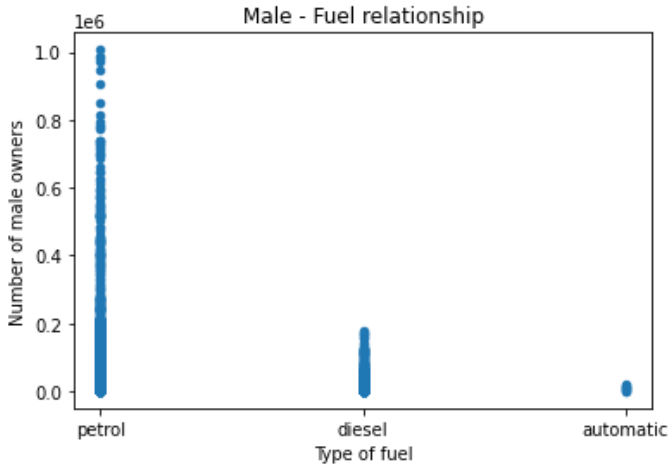



Hypothesis- Number of male owners are increasing when car transmission is in between 4 to 7.

In this graph as we can notice, the number of male owners is mainly between 4 to 7. From 1 to 3 the graph does have many owners, but not as much as 4 to 7. We cannot assume a positive or negative relationship here.

Insights- Male owners buy cars with the transmission value between 4 to 7.

**Third graph-**




Hypothesis- Number of male owners prefer petrol cars than diesel or automatic.

In the above graph it is quite evident that male owners prefer a car's fuel type which is petrol. The number of diesel and automatic owners is less. It can also be assumed that the second choice of a male owner would be diesel and the last choice would be automatic. Hence, we can assume that male owners and petrol have a positive relationship.

Insights- a male owner will choose a petrol, fuel type car.

Self-analysis- By doing this assignment, I have learnt a lot. I also do realise that I could have done better with making my code more precise. I did take a lot of time in fixing the errors. But as I am learning, I will apply these insights to my future assignments.