1. Import the Data set

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

amjdata = pd.read\_csv("D:/Users/OtaiAA0B/Desktop/Bootcamp/Project/AmjadDt.csv")

amjdata.head()

|  | **model** | **year** | **price** | **transmission** | **mileage** | **fuelType** | **tax** | **mpg** | **engineSize** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | A1 | 2017 | 12500 | Manual | 15735 | Petrol | 150 | 55.4 | 1.4 |
| 1 | A6 | 2016 | 16500 | Automatic | 36203 | Diesel | 20 | 64.2 | 2.0 |
| 2 | A1 | 2016 | 11000 | Manual | 29946 | Petrol | 30 | 55.4 | 1.4 |
| 3 | A4 | 2017 | 16800 | Automatic | 25952 | Diesel | 145 | 67.3 | 2.0 |
| 4 | A3 | 2019 | 17300 | Manual | 1998 | Petrol | 145 | 49.6 | 1.0 |

amjdata.info()

<class 'pandas.core.frame.DataFrame'>

RangeIndex: 10668 entries, 0 to 10667

Data columns (total 9 columns):

# Column Non-Null Count Dtype

--- ------ -------------- -----

0 model 10668 non-null object

1 year 10668 non-null int64

2 price 10668 non-null int64

3 transmission 10668 non-null object

4 mileage 10668 non-null int64

5 fuelType 10668 non-null object

6 tax 10668 non-null int64

7 mpg 10668 non-null float64

8 engineSize 10668 non-null float64

dtypes: float64(2), int64(4), object(3)

memory usage: 750.2+ KB

amjdata\_clean = amjdata.copy(deep = True)

amjdata.shape

(10668, 9)

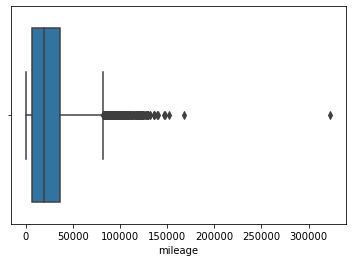
amjdata.isna().sum()

model 0 year 0 price 0 transmission 0 mileage 0 fuelType 0 tax 0 mpg 0 engineSize 0 dtype: int64

As shown above the data set is cleaned with no missing values.

1. Outlier

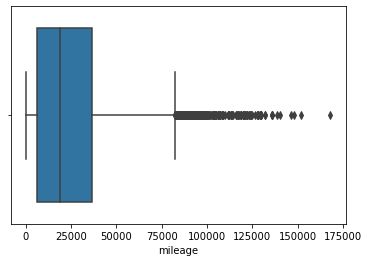
boxa = sns.boxplot(x = 'mileage', data = amjdata\_clean)



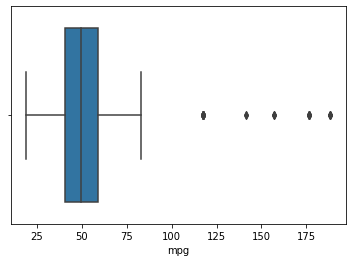
As shown above there is a car above the 300,000 miles which considered an outlier. By remove it the range of data will be from 0 to approxmitly 175000miles.

amjdata\_clean = amjdata\_clean[amjdata\_clean['mileage'] < 175000]

boxa = sns.boxplot(x = 'mileage', data = amjdata\_clean)



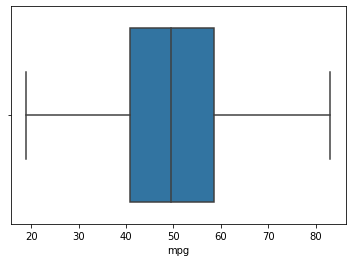
boxa = sns.boxplot(x = 'mpg', data = amjdata)



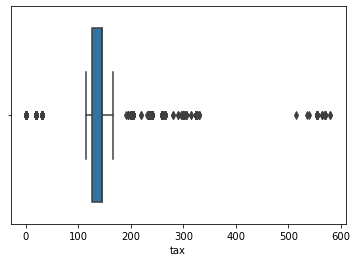
As we can see there are outliers ranging from 120 to 190 mileage. We removed those and make the range from 0 to 110.

amjdata\_clean = amjdata\_clean[amjdata\_clean['mpg'] < 110]

boxa = sns.boxplot(x = 'mpg', data = amjdata\_clean)

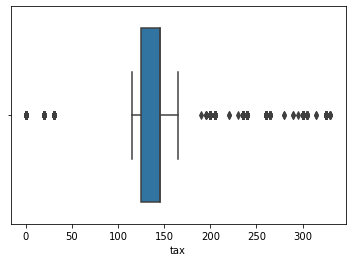


boxa = sns.boxplot(x = 'tax', data = amjdata\_clean)



It is obvious there is outliers in the range from 500 to 600. So, we remove the outliers.

boxa = sns.boxplot(x = 'tax', data = amjdata\_clean)



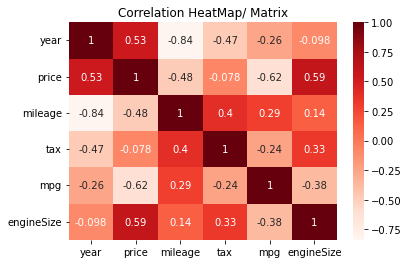
1. Exploratory Data Analysis:

After we removed all the outlier, the data is ready for Exploratory Data Analysis.

sns.heatmap(amjdata\_clean.corr(), cmap ="Reds", annot = True)

plt.title("Correlation HeatMap/ Matrix")

plt.show()



From the Correlation Matrix we get the following information:

* There is a negative correlation between mileage and price. Which means that the car that has more mileage the price is lesser because the car is used more.
* There is a negative correlation between price and mpg. Which means the less mpg the higher price.
* There is positive correlation between engine size and price. It means that having a higher engine size will cost more.
* There is a little positive correlation between price and tax.

fig, axes = plt.subplots(figsize = (12,10), nrows = 2, ncols = 3)

sns.histplot(amjdata\_clean["year"], ax = axes[0,0])

sns.histplot(amjdata\_clean["mileage"], ax = axes[0,1])

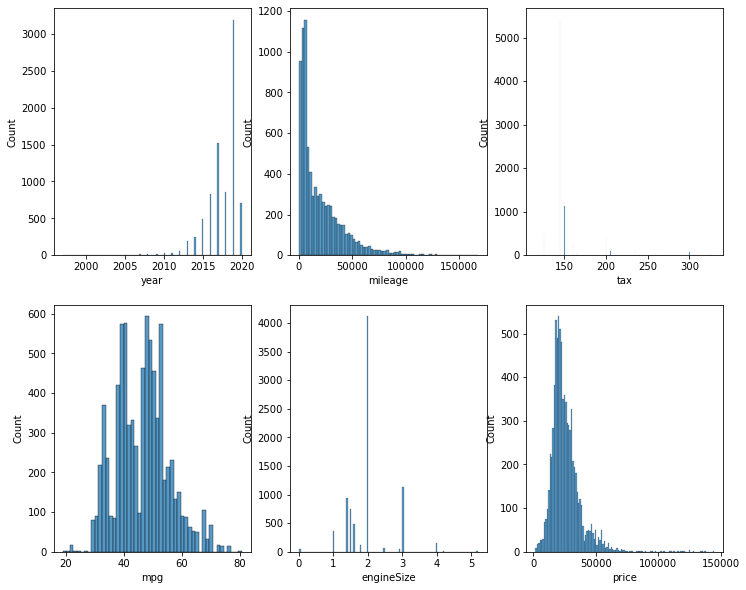
sns.histplot(amjdata\_clean["tax"], ax = axes[0,2])

sns.histplot(amjdata\_clean["mpg"], ax = axes[1,0])

sns.histplot(amjdata\_clean["engineSize"], ax = axes[1,1])

sns.histplot(amjdata\_clean["price"], ax = axes[1,2])

plt.show()

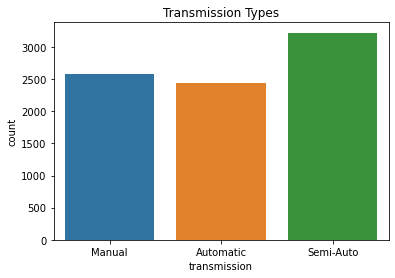


The year column is right-skewed which means that most of the cars are between 2015 to 2020. The mileage is left-skewed which show that most of the cars are driven for more than 5000 miles. The engineSize shows that the most engine size is between 1.5 to 2.

sns.countplot(x = "transmission", data = amjdata\_clean)

plt.title("Transmission Types")

plt.show()



As we can see above, there are 2500+ cars which are Manual and less than 2500 cars are Automatic and 3000+ are Semi-Auto transmission.

print(amjdata\_clean["model"].value\_counts())

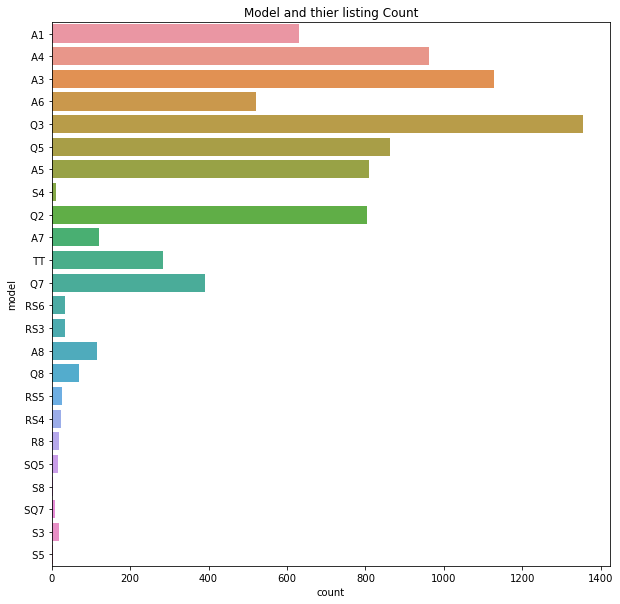
plt.figure(figsize = (10,10))

sns.countplot(y = amjdata\_clean["model"])

plt.title("Model and thier listing Count")

plt.show()

A7 120 A8 116 Q8 69 RS6 34 RS3 33 RS5 26 RS4 23 R8 18 S3 18 SQ5 16 S4 11 SQ7 8 S8 3 S5 3 Name: model, dtype: int64



It is clear that A3's is more than the other models.

plt.subplots(figsize = (18,18))

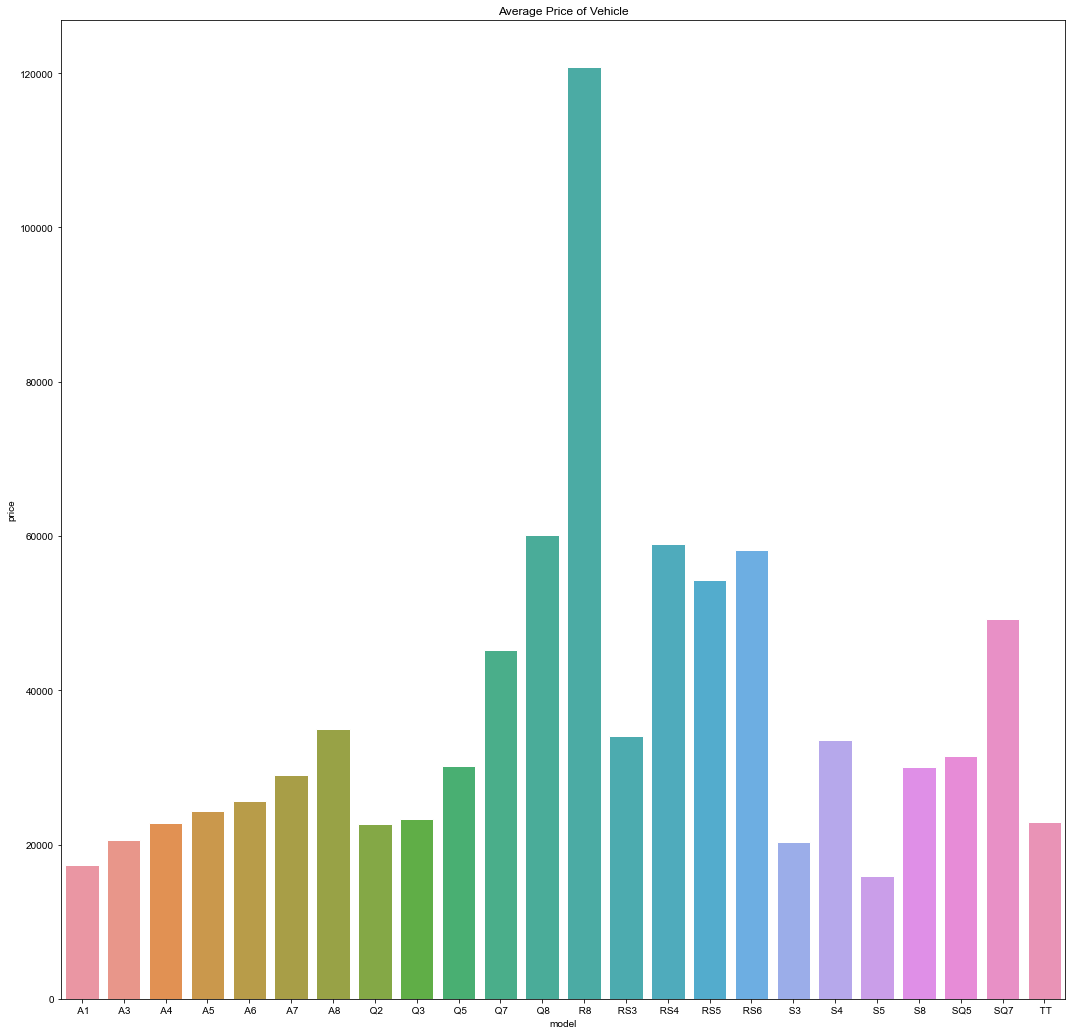
priceByModel = amjdata\_clean.groupby("model")['price'].mean().reset\_index()

plt.title("Average Price of Vehicle")

sns.set()

sns.barplot(x = 'model', y = 'price', data = priceByModel)

plt.show()



The previous chart shows that the price of R8 is higher than the others models.