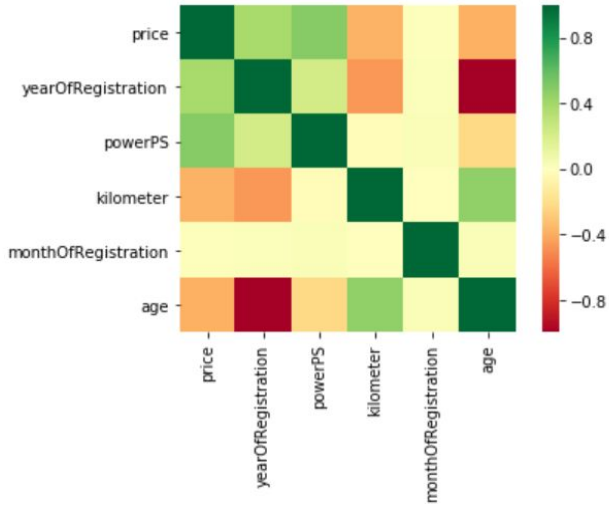
CAPSTONE PROJECT - EDA

After cleaning the large dataset and discarding all unwanted and unnecessary values, the data frame is ready for analysis. Since the dataset variables are mostly binary such as the “salvaged” value or mostly strings, such as “brand” and “model”, but not continuous numeric variables, the size of the covariance matrix where we see the variable correlations will be limited. The heat map generated to display the covariance can be seen on the left. Please note that the algorithm processes only continuous numeric variables, leaving out transmission type, fuel type etc. Binary or discrete variables will be handled as a part of the classification problem in the machine learning phase.



Once we look at the price row of the heatmap, we can see that there is a significant positive correlation between a car’s value and year of registration and power of the engine, with the latter being more prominent. The pearson correlation coefficient between price and year of the vehicle is 0.39 while the same coefficient for power and price is as high as 0.5, which signifies a very strong correlation, whereas the correlation coefficient between kilometer and price is -0.37. The variable “age”, which was created to include the month data to calculate how old the vehicle is, since using only the year of registration would simply ignore that data. When we increase the decimal part of the pearson coefficient to 5 digits, the absolute values of the correlation coefficients for year of registration and age will be 0.38862 and 0.38794. This can be explained looking into how new models are released: if Honda releases the new generation of Civic in 2014, the gap between 2014 and 2013 models will be high, which cannot be caught in the age variable with the same sensitivity.

Two-sided T tests were carried out to check if the null hypothesis that claims 2 independent samples have identical average values are true. Similar tests were also carried out by bootstrap method to analyse relationships between prices of binary values, i.e. prices of diesel vs manual vehicles. We can use this test, if we observe two independent samples from the same or different population, e.i. horsepower value for automatic vehicles and manual vehicles. The test measures whether the average value differs significantly across samples. If we observe a large p-value, for example larger than 0.05 or 0.1, then we cannot reject the null hypothesis of identical average scores. If the p-value is smaller than the threshold, then we reject the **null hypothesis of equal**

averages. The table below depicts the t-test executed to check if different engine and transmission types were similar in changing the price.

PRICE	T Statistic	P Value	Result
DIESEL vs GAS	83.050	0.00000000	Reject H_o
MANUAL VS AUTO	120.509	0.00000000	Reject H_o
SALVAGED VS CLEAN	-64.047	0.00000000	Reject H_o

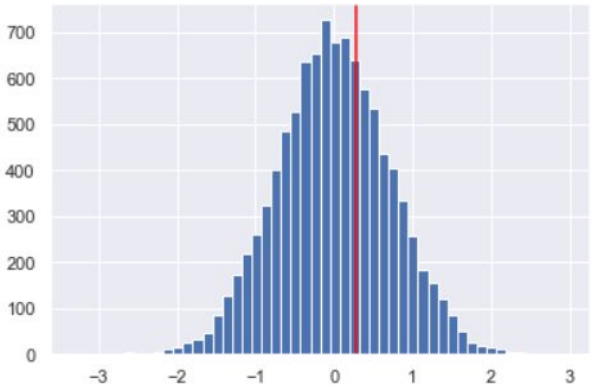
The table below depicts the t-test executed to check if different engine and transmission types were similar in changing the horsepower of the vehicle.

POWER PS	T Statistic	P Value	Result
DIESEL vs GAS	95.365	0.00000000	Reject H_o
MANUAL VS AUTO	249.735	0.00000000	Reject H_o
VW vs SKODA*	-0.385	0.70049115	Accept H_o

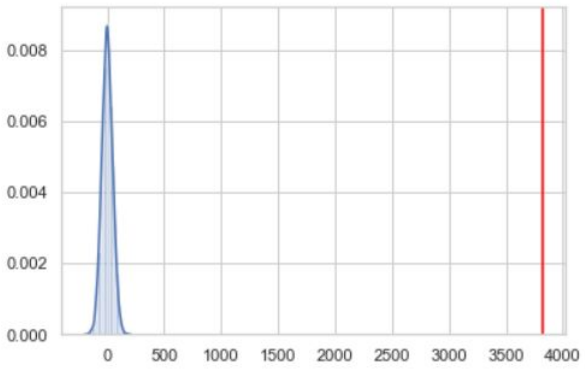
Please note that Skoda, a sister company of Volkswagen has similar values regarding the horsepower, since most vehicles use common engines between these two brands. This was reflected to t-test, and we can clearly see that a random vehicle selected from a pool of VW and Skoda have comparable average horsepowers.

```
two_sample = stats.ttest_ind(df1[df1['brand']=='VOLKSWAGEN'].powerPS,
                             df1[df1['brand']=='SKODA'].powerPS)
print('The t-statistic is %.3f and the p-value is %.8f.' % two_sample)
```

```
The t-statistic is -0.385 and the p-value is 0.70049115.
```

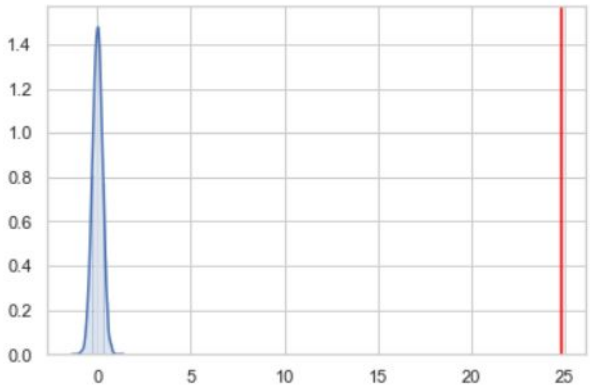


Another approach was using the bootstrap method as discussed above. Looking at the engine horsepower values of both Skoda and Volkswagen, we can see that the empirical difference of means of horsepower of each brand is 0.2796507632004648, and as we carry out



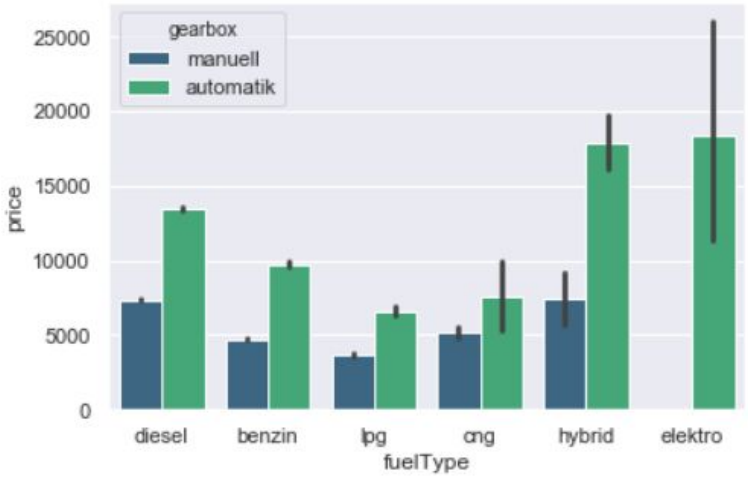
the bootstrap method we can witness the normal distribution of permuted bootstrap replicates and the red line represents the empirical difference of means, with a p value of 0.347.

The graph on the left shows 10000 permuted bootsrap samples drawn, between diesel and gasoline powered vehicles and their mean prices. The blue shaded histogram is the distribution curve of these values. The empirical difference of 3822$ is depicted by the red vertical line and can be observed well right of the distribution curve shaded in blue.



This result suggests us that the mean price of a diesel vehicle and a gasoline vehicle is not similar. We can also carry out the same test for analyzing the power of engines according to their fuel types. The graph is very similar, with the red line at 24.8 Hp, indicating the mean difference of a diesel vs. a gasoline engine. Diesel engines are simply more powerful and more expensive.

Another possible depiction of how car prices are affected by the engine type is the bar plots as seen in the graph below. It is clearly visible that diesel automatic vehicles are more expensive. Diesel engine is very sophisticated and diesel automatic vehicles are much preferred in Europe, since their cost per mile is very low.



The non-numeric discrete variables are a part of the classification problem in machine learning, thus prevalent features were identified.