UNIVERSITÀ DEGLI STUDI DI MILANO

Data Science for Economics

Statistical learning project

CAR PRICES IN THE ARABIC PENINSULA

Professor: Silvia Salini

Author: Guglielmo Berzano

Student id: 13532A

e-mail: guglielmo.berzano@studenti.unimi.it

October 2023

Abstract

The analysis contained in this report aims to find the main factors behind the determination of car prices in the markets of the Arabic peninsula. The analysis will be conducted thanks to the implementation of both supervised and unsupervised statistical learning methods through R. Principal Component Analysis will be used for the unsupervised part to discover which are the variables that mostly determine price while Robust Linear Regression, Regression Trees and Random Forest are the methods of choice for the supervised analysis. In this case the objective is to create the best model according to the R^2 metric and to understand, as before, the most influent variables for the determination of price. The analysis continues by designing general models that have *Area* as a variable and not as a discriminant. The goal is to see if selling cars in a market rather than in another is a clue fact or if the selling market – meaning the selling country – is not influent.

Furthermore, each analysis will be filled with comparisons between countries of the Arabic peninsula to see how preferences change across countries of the same geographical area.

Table of Contents

Table of Figures

List of Tables

1. INTRODUCTORY ANALYSIS

The dataset is called *Cars in the middle east* and it is taken from the website Kaggle. The dataset contains information about cars sold in the Arabic Peninsula's markets, their selling price along with the currency of the country in which the cars are sold in. The ultimate goal of this analysis is to be able to predict with high accuracy the price of a car based on its technical characteristics and to understand which are the most influent variables that determine price. The dataset was obtained thanks to web scraping and has not been updated for about two years.

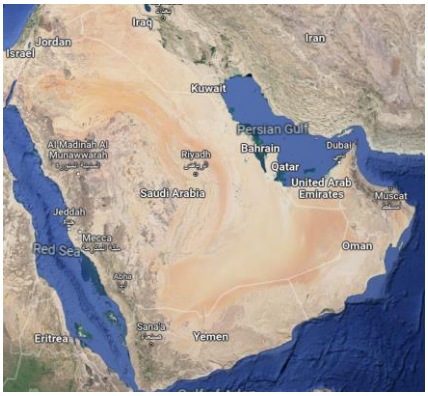

Figure 1 – Arabic peninsula, source: Google Earth

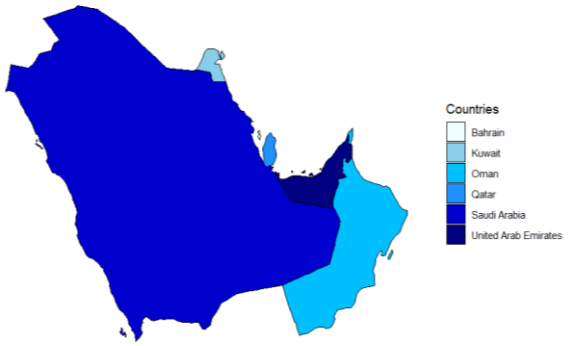

Figure 2 – Analysed countries

In this case the data is retrieved from the following countries of the Arabic peninsula:

- Bahrain
- Kuwait
- Oman
- Qatar
- Saudi Arabia
- United Arab Emirates

1.1 Descriptive analysis

After having imported the dataset in R Studio, I started the analysis by taking a look at the values at my disposal. The original dataset consisted in 5668 rows, 20 columns and had the following structure:

Table 1 - Dataset structure

Engine_Capacity	Cylinders	Driving	Fuel_Capacity	Liters_For_100km	Fuel_Type	Horsepower	Torque	Transmission	Top_Speed
1,2	3	Front Wheel Drive	42	4,9	Petrol	76	100	Automatic	170
1,6	4	Front Wheel Drive	50	6,4	Petrol	102	145	Automatic	180
1,5	4	Front Wheel Drive	48	5,8	Petrol	112	150	Automatic	170
1,4	4	Front Wheel Drive	35	5,1	Petrol	98	127	Automatic	170
1,6	4	Front Wheel Drive	50	6,4	Petrol	102	145	Automatic	180
1,4	4	Front Wheel Drive	45	6,3	Petrol	100	132	Automatic	183

Seating_Capacity	Acceleration_0100	Length	Width	Height	Wheelbase	Trunk_Capacity	Name	Price	Currency
5	14	4,24	1,67	1,51	2,55	450	Mitsubishi Attrage 2021 1.2 GLX (Base)	34099	SAR
5	11	4,35	1,99	1,53	2,63	510	Renault Symbol 2021 1.6L PE	44930	SAR
5	10,9	4,31	1,81	1,62	2,58	448	MG ZS 2021 1.5L STD	57787	SAR
5	12	3,64	1,6	1,48	2,38	314	Chevrolet Spark 2021 1.4L LS	53790	SAR
5	11	4,35	1,99	1,53	2,63	510	Renault Symbol 2021 1.6L LE	54780	SAR
5	13,4	4,44	1,73	1,48	2,6	396	Hyundai Accent 2021 1.4L Base	53460	SAR

Table 2 – Detail of each variable

Variable	Unit of measure or levels	R type
Engine_Capacity	Liters	double
Cylinders	-	double
Driving	All, front, rear wheels	factor
Fuel_Capacity	Liters	double
Liters_For_100km	Liters	double
Fuel_Type	Petrol, diesel, hybrid	factor
Horsepower	Break horsepower (bhp)	double
Torque	Nanometers	double
Transmission	Auto, manual, CVT	factor
Top_Speed	Km/h	double
Seating_Capacity	-	integer
Acceleration_0100[1]	Seconds	double
Length	Meters	double
Width	Meters	double
Height	Meters	double
Wheelbase	Meters	double
Trunk_Capacity	Liters	double
Name	-	char
Price	Based on the currency	double
Currency	(see the following page)	factor

[1] Acceleration_0100 refers to the amount of time it takes for a car to reach 100 km/h. The higher this number, the slower the car.

In order to make comparisons between countries possible, I created two more variables: *PriceEURO* and *Area* which are respectively the converted price of cars in EURO[2] and the Area a particular car was sold in.

The distributions of the numeric variables can be summarised by the following boxplots:

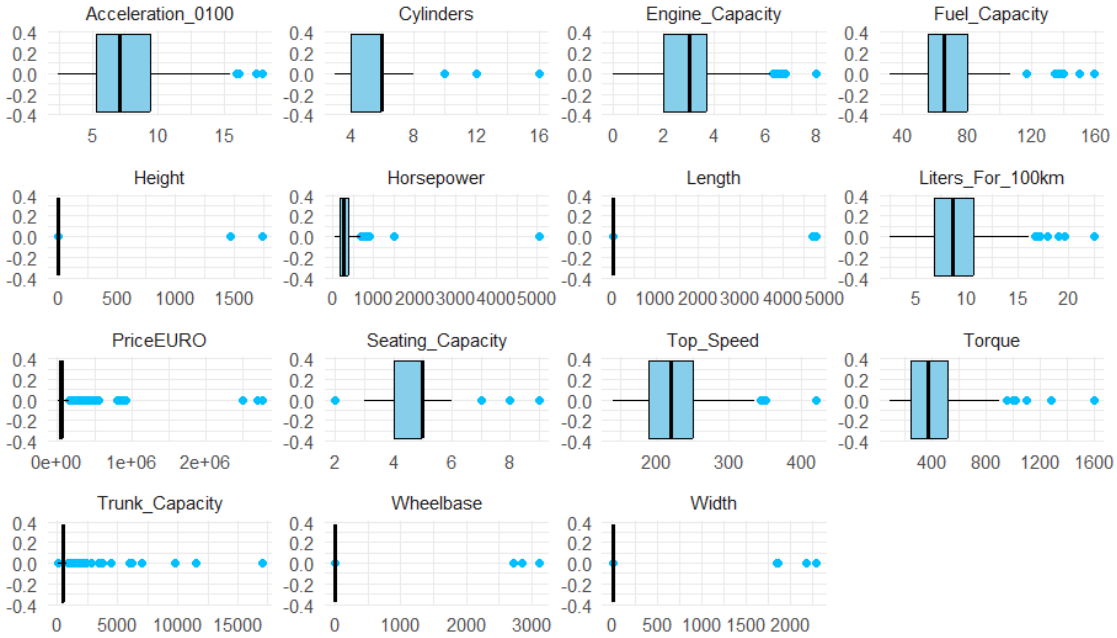

Figure 3 – Boxplots for the numeric variables

Clearly there are some extreme outliers which need to be removed. Not all the outliers will be removed, like the ones in the *Acceleration_0100* since those are just the tail of the distribution.

After the outlier removal the boxplots became:

[2] Exchange rates were retrieved on the 23th of June with the following values:
1 BHD = 2.44 EURO, Bahrain currency
1 KWD = 2.99 EURO, Kuwait currency
1 OMR = 2.39 EURO, Oman currency
1 QAR = 0.25 EURO, Qatar currency
1 SAR = 0.25 EURO, Saudi Arabia currency
1 AED = 0.25 EURO, United Arab Emirates currency

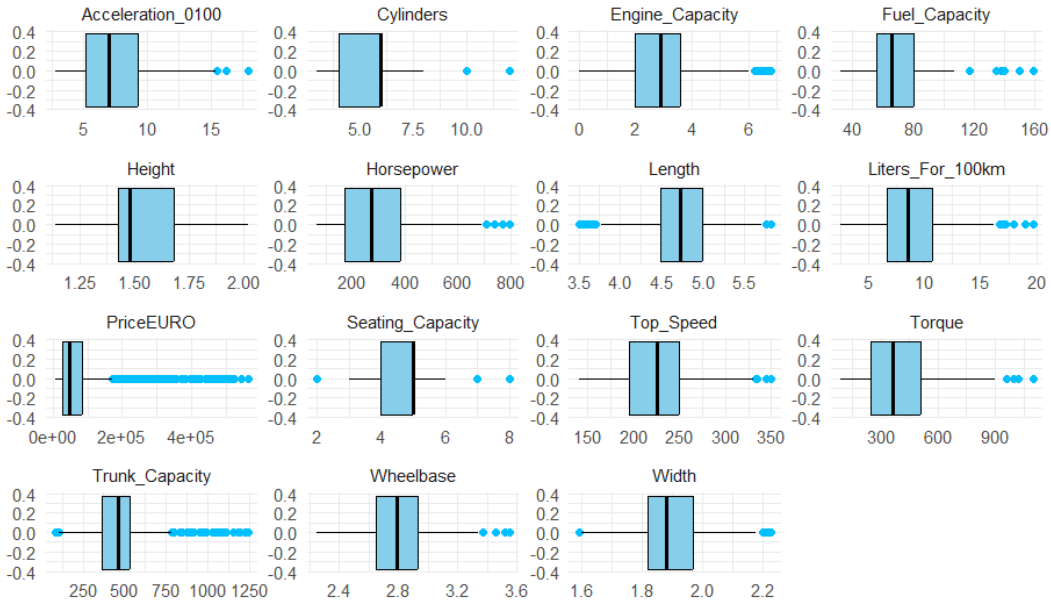

Figure 4 – Boxplots for the numeric variables after outlier removal

Still, some outliers are present but they do not seem that problematic. This can be acknowledged also by looking at the histograms:

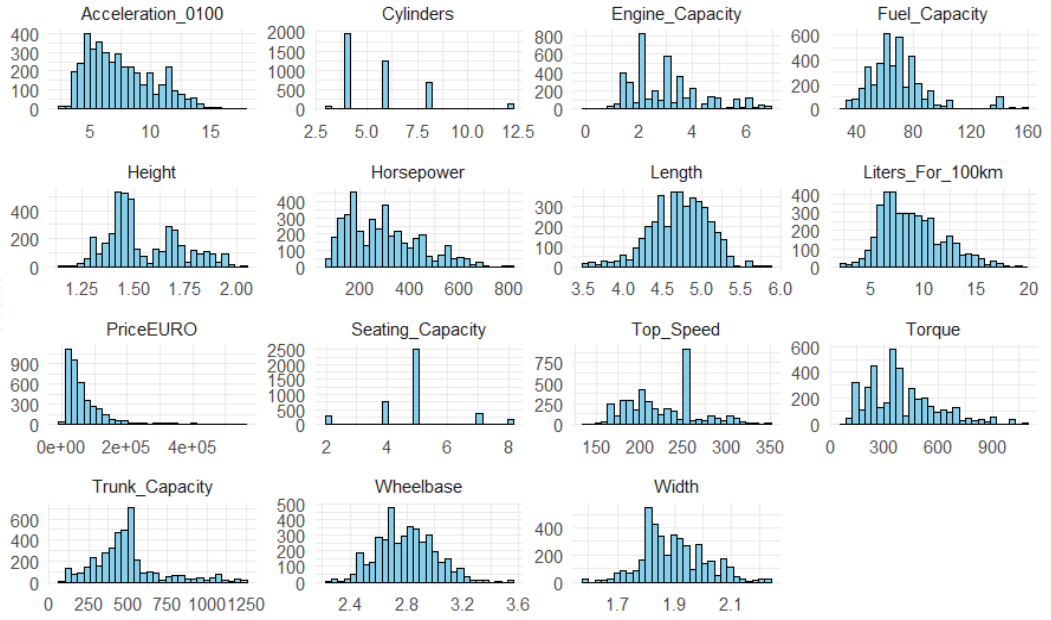

Figure 5 – Histograms for data distributions

Now, by looking at the correlation plot, it is possible to see whether there exists any close relation between the variables:

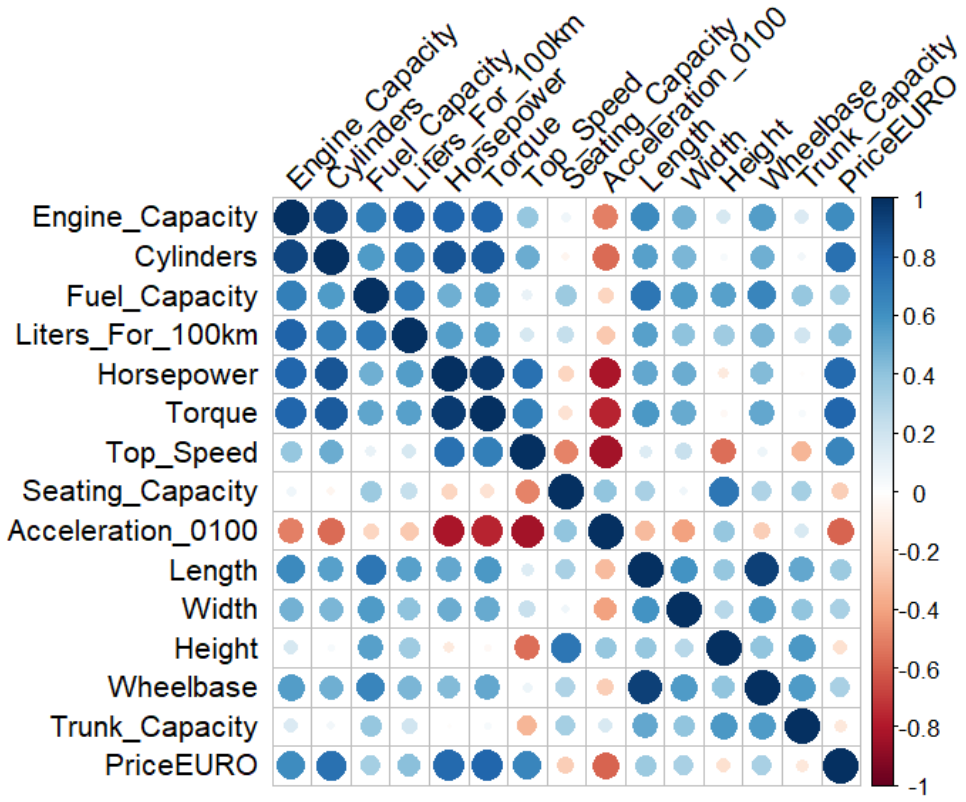

Figure 6 – Correlation between the variables

It seems that many variables are positively correlated with each other, especially in the upper-left corner while some others are negatively correlated. In particular, you can notice the extremely high correlation that exists between *Horsepower, Torque* and, even if it has lower magnitude, *Top_Speed* or between *Wheelbase* and *Length*. For sure some deeper analysis is needed.

In order to continue, I created one dataset for each level of the factor variable *Area*, resulting in six total datasets which will open the possibility of comparisons across countries. Each dataset has the same columns as the original ones but *Name, Price, Currency* and *Area*. By doing this we lose the cars' name but, since this is not the focus of the analysis, it is not a problem.

2. UNSUPERVISED LEARNING

Given the correlation plot at Figure 6, we may be interested in understanding better how variables are correlated between one another, thus we can perform a *Principal Component Analysis* (PCA). The main objective of the implementation of PCA is to have a preliminary idea of the existing relations within data and see whether a dimensionality decrease could improve the result of the analysis.

Before implementing the PCA, I split the factor variables *Driving*, *Fuel_Type* and *Transmission* into dummies, since PCA is not affected by multicollinearity, and removed the y column *PriceEURO*.

The variable *Driving*, whose levels were "All wheel drive", "Front wheel drive" and "Rear wheel drive", after the split became *DrivingAWD*, *DrivingFWD* and *DrivingRWD*. I used the same structure to convert also the other categorical variables: first the original column name, maybe shortened – *FT* for *Fuel_Type* and *Trans* for *Transmission* – and then a shortened version of the factor level.

2.1 Principal Component Analysis (PCA)

The table below indicates the percentage of variance explained by the first three principal components and their cumulative variance grouped by *Area*.

Table 3 – Proportion of explained variance of the first three principal components and cumulative variance, grouped by Area

P.C.	Bahrain	Kuwait	Oman	Qatar	Saudi Ar.	UAE
1st	33.17%	32.7%	32.8%	32.9%	32.5%	33%
2nd	17%	17.2%	16.4%	16.5%	18%	16.3%
3rd	9%	9.6%	9.4%	9.9%	10.5%	9.5%
Tot.	59.17%	59.5%	58.6%	59.3%	61%	58.8%

Since we ended up with more than twenty variables, each principal component will be able to explain a relatively small percentage of variance. Indeed when we get to the third component we are only able to explain around sixty percent of the total variance. You can find, in Appendix A, the effect that each variable has on the first – and most important – PC and the

detailed correlation each variable has up to the 6th principal component. This dimension was choosen since, on average, after the 6th principal component the eigenvalues were less than one, then no more important in the PC interpretation.

2.1.1 Interpretation of the PCA

The charts called "Contribution of variables to 1st Principal component - ..." just show the magnitude, in absolute value, that each variable has in the determination of the first PC. In order to see the sign of that effect, you must look at the relative chart on the right, especially at the first column.

It is immediate to see a pattern among all of these charts, especially if they are compared with the correlation chart in Figure 6. In particular, the variables *Engine_Capacity, Cylinders, Horsepower* and *Torque* are always the most influential variables across every country with respect to the first principal component. Also, by looking at Figure 6, it is possible to notice that these are also the most correlated variables with *PriceEURO*, thus we can conclude that the first principal component is a nice summary of how variables influence price. This thesis is confirmed also if we look at *Acceleration_0100* which has a negative correlation with *PriceEURO* that is perfectly captured by the first principal component.

One particular case is the one of the variable *Top_Speed*. Indeed, if we look at Figure 6, we see a positive correlation with *PriceEURO* but this is not very well captured by the principal components. For every country, in the charts on the left, *Top_Speed* barely remains above the red dashed line – which intercepts the y-axis at point 1/(nr of variables) – which denotes a lack of relevance of this variable. Furthermore, in the cases of Bahrain, Kuwait and Qatar, *Top_Speed* has a negative effect on the second PC.

Then we could look at the variable *Trunk_Capacity*. If we look at the impact it has on the first PC, we immediately see that it is almost zero which supports the idea that the 1st PC summarises well the correlation with *PriceEURO* since, according to Figure 6, these two variables are complitely uncorrelated with each other. On the other hand, if we look at the 2nd

PC, we see that *Trunk_Capacity* has many controversial effects which let us think that this principal component does not have any linkage with *PriceEURO*.

For this reason, even if the 1st PC showed us the most relevant features that determine car prices, the 2nd PC is much harder to interpret since it lacks a precise explanation. This leads us to abandon the PCA for dimensionality reduction and to seek for alternative methods that could also deal with problematic situations like multicollinearity, non-scaled variables or non-gaussian distributions.

2.1.2 Linear PCs model

However, mathematically speaking, PCs are significant if their eigenvalues are greater or equal than one. Thus if we analyse the first n PCs such that their eigenvalues satisfy the property I have just mentioned, we can get the most explicative model possible with the PCs, also reducing the dimensionality of the dataset. Furthermore, since the PCs are orthogonal variables, we do not need to worry about multicollinearity.

So, first we find the first n PCs such that their eigenvalues are equal or greater than one. The computations shows that for every analysed country, the value n is equal to 7. For the following analysis we will use a dataset in which the columns are first 7 PCs. Now that we have obtained a reducted dataset, one for each country, by implementing a simple linear regression model, we get:

Table 4 - Significance of first 7 PCs by country

Area	Principal Components						
	1st	**2nd**	**3rd**	**4th**	**5th**	**6th**	**7th**
Bahrain	***	***		***	***		***
Kuwait	***	***	*	***	***	**	.
Oman	***	***		***	***	*	***
Qatar	***	***	***		***	**	
SA	***	***	***	***		***	***
UAE	***	***	***	***	***		***

As you can see from the table, the first two principal components are very statistically significant for every analysed country and, from the third PC on, we lose some of that singnificance for some countries.

For analysing the goodness of the model we have just created we can look at the R^2 metric which formula is summarized as follows:

$$R^2 = 1 - \frac{SSR}{TSS} = 1 - \frac{\sum_i[(y_i - \hat{y}_i)^2]}{\sum_i[(y_i - \bar{y})^2]} \quad (1)$$

The results are summarized in the following table:

Table 5 - R^2 for a regression run on PC

Area	R^2
Bahrain	0.8709772
Kuwait	0.8297523
Oman	0.8581698
Qatar	0.8348107
Saudi Arabia	0.8206330
United Arab Emirates	0.8653526

We can clearly see that the R^2 is quite high especially in Bahrain, in Oman and in the UAE.

2.1.3 PCs regression trees model

After having analysed the simple linear model we can try to use some more advanced algorithms like the *Classification And Regression Trees (CART)* to check whether it fits better the data. The results, in terms of R^2 are summarised in the following table and the trees can be seen in Appendix B.

Table 6 – R^2 for regression tree made by PCs

Area	R^2
Bahrain	0.798516
Kuwait	0.8100026
Oman	0.7812357
Qatar	0.7608527
Saudi Arabia	0.7687297
United Arab Emirates	0.7488691

In this case, the R^2 is not excellent anywhere except for Kuwait in which it has a value slightly higher than 0.80. We can consider these results acceptable but, in general we must say that the tree algorithm did not perform great with principal components.

Furthermore, the principal component analysis is among the hardest methods to interpret so, even though the overall results are acceptable, we can probably obtain something better and easyer by implementing supervised learning methods. Later in the second to last chapter, page 18, we will discuss how these results compare to the ones obtained in the supervised learning models, trying to understand which are the best methods in terms of interpretability and model performance.

3. SUPERVISED LEARNING

In this section I am trying to predict, using supervised learning models, the outcome of the response variable *PriceEURO* given information about the cars. Since the y variable is numerical, the problem will be referred as *regression*.

First, I thought about implementing the linear regression model but, due to problems that will be further discussed later, I decided to change model to implement regression trees and random forest.

3.1 Linear regression model

To begin, I looked at the distribution of the y variable *PriceEURO* in all the dataset which, as can be seen also from Figure 5, does not follow a normal distribution at all. To try to fix this, I applied the logarithm transformation to the response variable and succeeded in obtaining a quite good approximation of a normal distribution. From now on, all the analysis will be conducted by using the logarithmic transformation of the variable *PriceEURO*.

In order to apply the linear model, one of the assumptions is that the residuals must be normally distributed. For this reason, I performed a Shapiro-Wilk normality test on the residuals and these were the results:

Table 7 – P-Value of Shapiro-Wilk normality test on ln(PriceEURO), grouped by Area

Area	P-Value
Bahrain	6.974e-08
Kuwait	0.007798
Oman	0.003011
Qatar	0.001321
Saudi Arabia	5.581e-12
United Arab Emirates	0.1725

The Shapiro-Wilk normality test compares the null hypothesis H_0 that the data is normally distributed against the alternative H_1 in which the data are not normally distributed. If we

select an $\alpha = 0.95$, where 1 - $\alpha = 0.05$, we reject null hypothesis for every country except for UAE, thus concluding that, on average, the distributions of the errors are not gaussian.

The next analysis will be about overcoming this problem along with dealing with multicollinearity.

3.2 Best subset selection

There is the necessity of modify the data so that models can be implemented more easily. To do so, I first performed the *Variance Inflation Factor* (Vif) analysis for each *Area* and found out that there were two main groups of multicollinear variables, namely the ones that had a Vif ≥ 10: *Horsepower/ Torque* and *Engine Capacity/Cylinders.* So, in order to maintain as much information as possible, I removed from all the subsets the variables *Horsepower* and *Engine Capacity.* After this, I did a best subset selection according to the *Mallows' Cp* criterion, removed the non-important variables and, in some cases, created dummies for those variables in which just one level out of two or three of a factor variable was considered important by the algorithm.

In Appendix C you can find the Vif charts after the removal of the two multicollinear variables and the charts for the best subset.

Now that we have subsetted the dataset we can implement some models, namely robust linear regression, the regression trees and random forest.

3.3 Robust Linear Regression

Despite the removal of multicollinear variables, distributions of the residuals obtained by the linear models on the y variable *PriceEURO* of course remained non-gaussian. So, since multicollinearity is not a problem anymore, we may be interested in running a robust linear regression, also to neutralise the effect that outliers and extreme leverage points have on the regression. below you can find a graphical representation of the strong presence of both high leverage points and outliers.

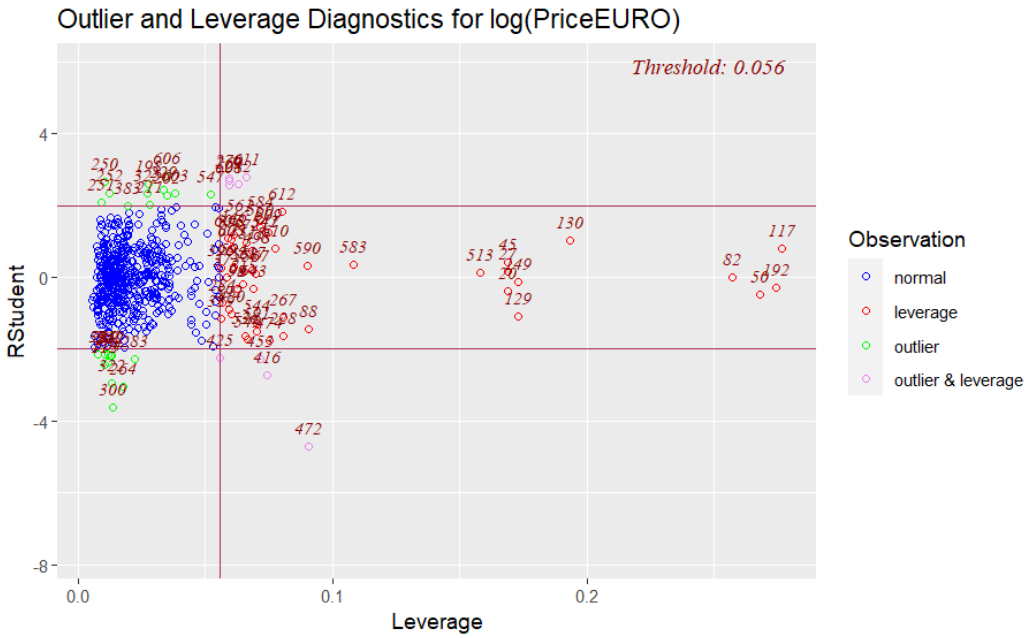

Figure 7 - Residuals by characteristics, Kuwait linear regression taken as an example

Thus, to implement the robust regression I created the function *rob_reg* which first standardises the numeric variables in order not to be influenced by different unit of measures and outputs coefficients computed on a casual training set obtained by selecting randomly 80% of the observations, their significance level and the R^2 of the regression computed with respect to the remaining 20% of the observations. The coefficients of the robust regression were computed with MM-type estimates, by using the package *robustbase*. The R^2 for the robust regression is summarised in the table below:

Table 8 – R^2 for robust linear regression

Area	R^2
Bahrain	0.8886621
Kuwait	0.8531936
Oman	0.8713075
Qatar	0.8709802
Saudi Arabia	0.8458653
United Arab Emirates	0.8914368

We can see that the robust regression has a very good R^2 for every *Area* we consider, reaching a peak of almost 0.9 for UAE.

In the following table are reported the most relevant variables, therefore the ones with the highest coefficient estimates in absolute value among the significant variables.

Table 9 - Most important variables according to robust linear regression

	Bahrain	**Kuwait**	**Oman**	**Qatar**	**Saudi Ar.**	**UAE**
1st	Max Speed	Max Speed	Torque	Torque	Max Speed	Torque
2nd	Torque	Torque	Max Speed	Max Speed	Torque	Max Speed
3rd	Fuel Cap.	Fuel Cap.	Cylinders	Wheelbase	Height	Cylinders

The regression finds that the most influential variables in the analysis clearly are *Torque* and *Max_Speed* but since it is not clear which is the most impactful one, further analyses are needed.

3.4 Regression Trees

In this case, to implement the regression trees, I created the function *tree_construct* which returns the best regression tree pruned according to the minimization of the *cost of pruning* (cp), which consists in selecting the cp which minimizes the column *xerror* of the *cptable*, meaning that we select the cp such that the error is the smallest possible. The pruned trees are shown in the Appendix D.

Since we are making a regression over a logarithmic variable, the interpretation of the prediction may not be that straightforward. Let's take, for instance, the first leaf at the right in Figure 43, page 44. We see that 3.26% of observation have a price such that:

$$ln(PriceEURO) = 12.71$$

Thus the predicted price is equal to:

$$PriceEURO = e^{12.71} = €\ 331{,}041.82$$

The R^2, computed with the formula (1), on the test is summarized as follows:

Table 10 – R^2 for regression trees

Area	R^2
Bahrain	0.8180443
Kuwait	0.7998163
Oman	0.835828
Qatar	0.8542003
Saudi Arabia	0.7114432
United Arab Emirates	0.8046047

Which can also be seen as:

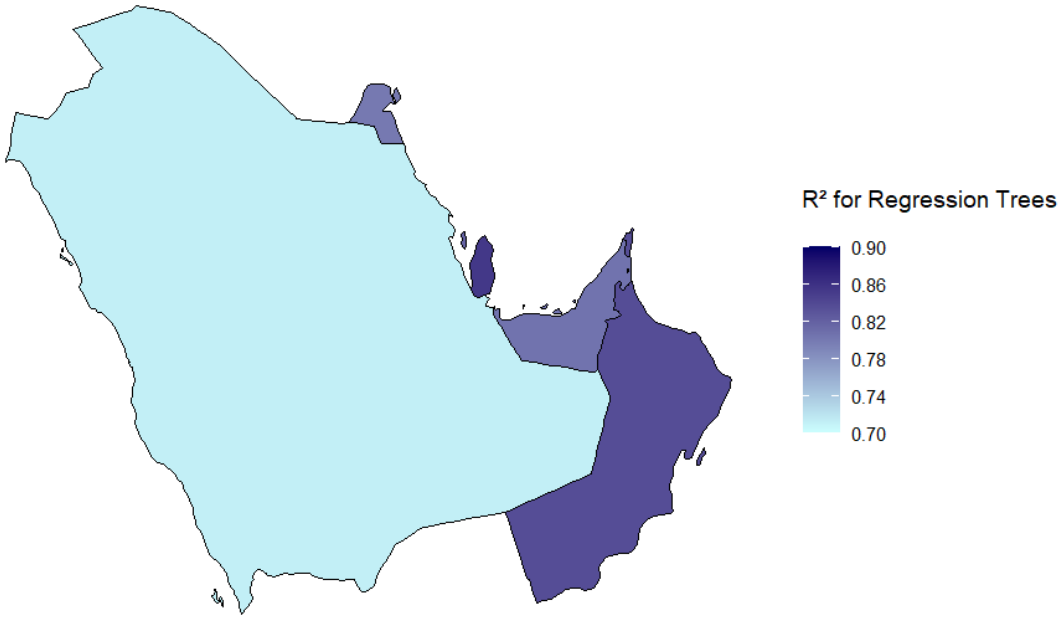

Figure 8 – R^2 for regression trees, map

Now we can look at the variables that, according to the regression trees, are the most important in the determination of the price. Results are summarised in the table below:

Table 11 – Most influential variables according to regression trees

	Bahrain	Kuwait	Oman	Qatar	Saudi Ar.	UAE
1st	Torque	Torque	Torque	Torque	Torque	Torque
2nd	Acceleration	Max Speed	Acceleration	Acceleration	Acceleration	Acceleration
3rd	Max Speed	Cylinders	Max Speed	Max Speed	Max Speed	Max Speed

In this case, clearly *Torque* is the most influential variable while the second most impactful is *Acceleration* followed by *Max Speed*. Here, a pattern is much clearer with respect to the robust regression which was a bit more confusing.

3.5 Random Forest

After the implementation of robust regression and regression trees, I decided to implement the random forest to hopefully obtain a higher prediction power. The algorithm needs two hyperparameters to operate at its maximum potential, namely *mtry* and *ntrees* which must be numbers belonging to $\mathbb{N}$. The first parameter picks *mtry* variables randomly selected as candidates at each split, while the second is simply the number of trees that minimizes the mean squared error (MSE). These results were achieved thanks to cross validation.

In the following table you can find the R^2 values for the random forest algorithm grouped by *Area* optimized with the two hyperparameters.

Table 12 – R^2 for random forest

Area	R^2
Bahrain	0.9458579
Kuwait	0.9260215
Oman	0.9384747
Qatar	0.9434955
Saudi Arabia	0.9077597
United Arab Emirates	0.9359294

The results can also be visualized with the following map:

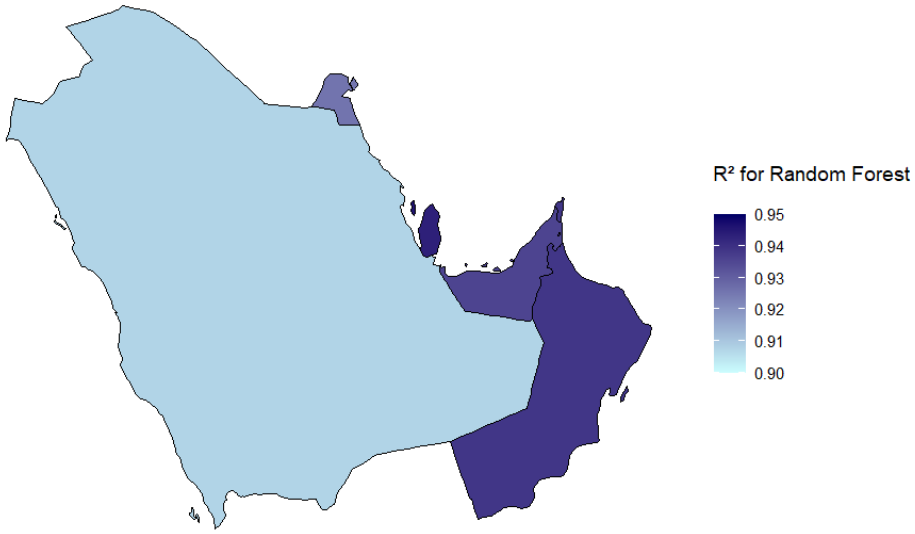

Figure 9 – R^2 for random forest, map

In the following table you can find the most influent variables according to the random forest algorithm:

Table 13 – Most influencial variables according to random forest

	Bahrain	Kuwait	Oman	Qatar	Saudi Ar.	UAE
1st	Torque	Max speed	Torque	Torque	Torque	Trunk Cap.
2nd	Fuel Cap.	Torque	Acceleration	Acceleration	Acceleration	Torque
3rd	Height	Height	Max speed	Fuel Cap.	Max speed	Acceleration

As in regression trees, *Torque* seems the most influential variable, *Acceleration* comes next while the last one varies a lot based on the countries.

4 MODEL COMPARISON

Now it would be interesting to take a look at how these models have performed and how they compare. I will first analyse the prediction power and then I will compare the most important variables.

4.1 Accuracy and R^2

Let's first look at the accuracy of the models, comparing them with the metric used since the beginning of the analysis, the R^2. First, we compare the PCs' model with its normal-data counterpart, for both the regression trees and the linear regression.

Table 14 – Best model Regression trees with PCs vs Regression Trees with normal data

Area	Regression trees with PCs	Regression trees normal data	Best model
Bahrain	0.798516	0.8180443	R.T.
Kuwait	0.8100026	0.7998163	R.T. PCs
Oman	0.7812357	0.835828	R.T.
Qatar	0.7608527	0.8542003	R.T.
Saudi Arabia	0.7687297	0.7114432	R.T. PCs
UAE	0.7488691	0.8046047	R.T.

The principal components' model is better than the normal regression trees for only two countries: Kuwait and Saudi Arabia. This fact, combined with the one that the normal tree is much more interpretable than the one made by using the PCs is making me think that the best model is the one made thanks to normal data.

Instead, results for the comparing of the linear PCA to the robust regression are summarised in the table below.

Table 15 - Best models, Linear PCA vs Robust Regression with normal data

Area	Linear PCA	Robust regression	Best model
Bahrain	0.8709772	0.8886621	Robust regression
Kuwait	0.8297523	0.8531936	Robust regression
Oman	0.8581698	0.8713075	Robust regression
Qatar	0.8348107	0.8709802	Robust regression
Saudi Arabia	0.8206330	0.8458653	Robust regression
UAE	0.8653526	0.8914368	Robust regression

We clearly see that the robust regression performs better in every single case. Furthermore, if we compare interpretability, for sure the robust regression is easier than the linear PCA.

Having said that, the supervised learning models are the best ones in terms of both interpretability and prediction power thus using the PCs models is, in this case, a bad choice. The most relevant advantage of using the principal components models is that by decreasing the dimensionality of the dataset, you can speed up the analysis by a quite good amount. But since in this case we only have 23 columns in each dataset – considering the ones that have been modified to fit the PCA requirements, i.e. only numerical values, and an average of 600 rows per dataset, the computations are all quite fast.

Furthermore, if we check which model is the best between the supervised learning framework, namely robust regression and regression trees, then we find that the robust regression is the best model for every analysed country. Actually we can also notice that regression trees have the massive advantage of being easier than normal regressions and of having a prediction power which is not that bad. Thus, if interpretability is the most important thing, regression trees could be used over robust regression even if a bit of prediction power is lost.

On the other hand, if we also consider the random forest, we find that it is the best model for every country according to R^2 since it always generates a value greater than 0.90.

We can thus conclude that the random forest is the best model for prediction, even if it takes a lot of time to load and is not very easy to interpret.

4.2 Most influential variables

The table below summarises data coming from Table 9, Table 11 and Table 13.

Table 16 – Most influent variables per Area, grouped by model

Country		Robust Regression	Regression Trees	Random Forest
Bahrain	1st	Max Speed	Torque	Torque
	2nd	Torque	Acceleration	Fuel capacity
	3rd	Fuel Cap.	Max Speed	Height
Kuwait	1st	Max Speed	Torque	Max Speed
	2nd	Torque	Acceleration	Torque
	3rd	Fuel Cap.	Cylinders	Height
Oman	1st	Torque	Torque	Torque
	2nd	Max Speed	Acceleration	Acceleration
	3rd	Cylinders	Max Speed	Max Speed
Qatar	1st	Max Speed	Torque	Acceleration
	2nd	Torque	Acceleration	Fuel capacity
	3rd	Height	Max Speed	Width
Saudi Arabia	1st	Torque	Torque	Torque
	2nd	Max Speed	Acceleration	Acceleration
	3rd	Wheelbase	Max Speed	Max Speed
United Arab Emirates	1st	Torque	Torque	Trunk cap.
	2nd	Max Speed	Acceleration	Torque
	3rd	Cylinders	Max Speed	Acceleration

We clearly see that the most influential variable tends to be the same across different models for the same country. Actually, there does not exist a case in which models were able to predict the same variables for the same country but one thing that we can conclude is that, for sure, the most impactful variable for price determination is *Torque*, followed by *Max Speed* and *Acceleration*. The case of Qatar is probably the strangest since the models predicted different variables in each position, suggesting that, for Qatar, does not exist one single variable which can be considered the most important but rather, there could be many variables that influence price with seminal magnitude.

What about the PCA? As we analysed in the chapter 2.1, the first principal component was a good summary of what may influence price. Actually, the first PC found that the most impactful variables were – if we exclude the multicollinear ones – *Torque, Cylinders* and *Length* for every country. These results are not amazing since they are not able to capture the peculiarities that every country has.

Overall, thanks to these models and these comparisons, I think we may have reached a quite good approximation of reality which also let us see how very similar countries may have different preferences.

5 GENERAL MODELS

Now, after having analysed the situation for each and every country separately, it would be interesting to create a general model to see whether the variable *Area* is actually an important factor in the determination of the price. Essentially, the variable that, since the first chapter was used as a discriminant to create many different datasets, now will be maintained inside the only dataset under the name of *Area*. As in the other chapters, I will analyse the robust regression, the regression tree and the random forest for both normal data and for the PCs. I will directly compare the results obtained, since all the aspects related to why things are the way they are, have been already vastly covered.

Using information gained from chapter 3.2, I removed the columns of *Horsepower* and *Engine_Capacity* which were found to be multicollinear with, respectively, *Torque* and *Cylinders*. Moreover, the principal component analysis will be effectuated using the first 7 PCs. The number 7, exactly as in *chapter 2.1.2*, is the number obtained by counting how many PCs have an eigenvalue greater or equal than one.

Let's now take a look at the results of these models in both their "PC version" and in their "normal-data" one.

Table 17 – R^2 for general models

Model	Normal data	Principal components
Robust Regression	0.8744916	0.8605758
Regression Tree	0.8426183	0.8008268
Random Forest	0.9700614	0.959543

As in the previous chapter, we see that normal data outperform the principal component model by quite a great amount. In this case, since thanks to PCs, data have been decreased in dimensionality, allowing computations to be faster. Still, results are worse than the normal data ones in every single case.

As we observed in chapter 4, one single regression tree, even if it is optimized with the *Mallow's cp*, it is found to be the worst model; this is true for both normal data and principal component. By the way, the trees have the following structure:

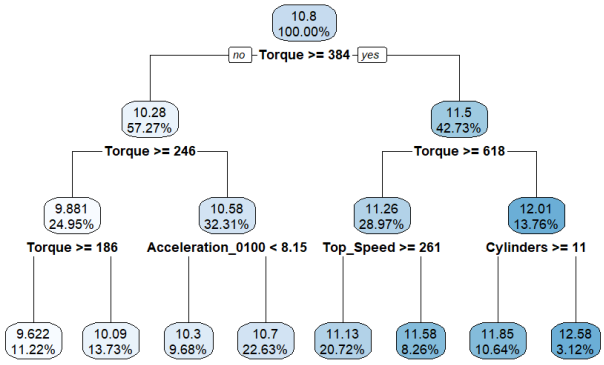

Figure 10 - Tree representation for normal data

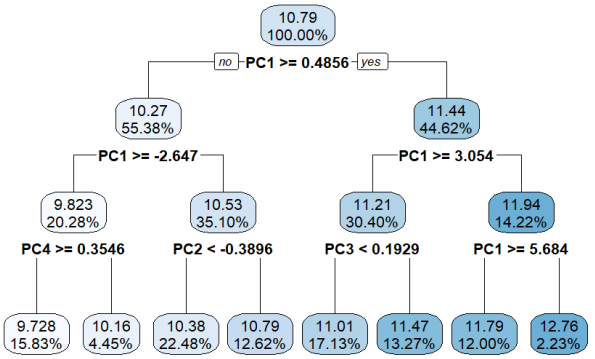

Figure 11 - Tree representation for PCs

Robust regression, as before, is the second-to-best model while random forest is by far the best one, having an incredible R^2 of 0.97 – computed on test data – for both the normal data and the PC. These results show that it is possible to create and optimize models that explain almost perfectly the situation of the car market in the Arabic peninsula.

Probably the reason why in this situation we get so good results while in the *six-dataset way* we do not is that in this case we have many more rows, compared to the other situation. Indeed, in the previous chapters we analysed datasets with 500-800 rows while now we are dealing with a dataset of almost 5,000 rows.

What about most important variables?

Table 18 - Most important variables, general models

Model		Robust Regression	Regression Trees	Random Forest
Normal data	1st	Torque	Torque	Trunk capacity
	2nd	Max Speed	Acceleration	Liters for 100km
	3rd	Area Saudi Arabia	Max Speed	Fuel capacity
Principal components	1st	PC 1	PC 1	PC 1
	2nd	PC 2	PC 3	PC 4
	3rd	PC 5	PC 4	PC 2

Results seem quite odd, especially if we compare them with the ones obtained before and summarised in Table 16. The variable *Torque* only appears twice as most important variable

and it does not show up at all in the top three variables for the random forest being, surprisingly, 7th. The *Area* appears, under the observation *Saudi Arabia*, only in the regression model meaning that, besides being statistically significant, it has also a high magnitude.

Since *Area* is a categorical variable, in the summary of the robust regression, it was split into its observations Bahrain, Kuwait, ..., United Arab Emirates. Of course, the algorithm will remove one of them to avoid dummy variable trap. The result that we got is that not all the countries are significant and, even for those which are, the magnitude is quite low. In particular, just the Saudi Arabia and the UAE are significant while all the others are not. We cannot find the importance that the *Cart* algorithm gives to the variable *Area* since it computes an importance value only for numerical variables. But we see that, for the random forest algorithm, *Area* is 7th, after variables like *Height* and *Width*, making us think that the value of the car is mostly due to its characteristics rather than the market in which it is sold.

What about the models created with the PCs? Every algorithm says that the first principal component is – by far – the most influent. We can now check how variables influence the principal component.

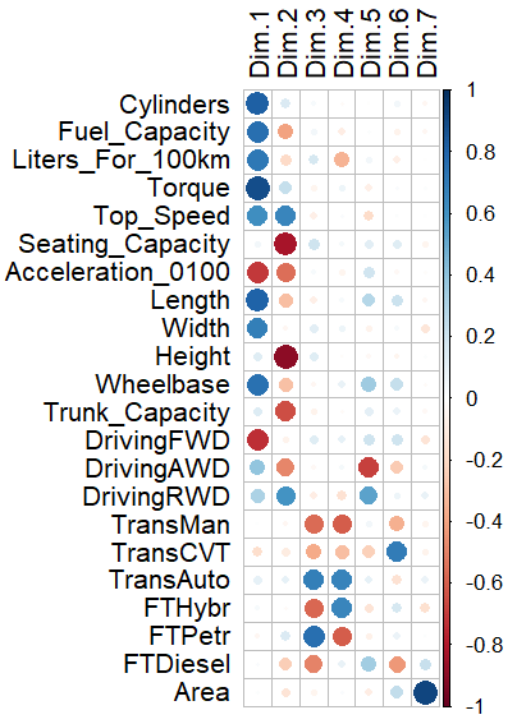

Figure 12 - Variable correlations up to the 7th Principal components

The variable *Area* – the very last one in the picture – does not contribute at all to the first PC but it actually has a correlation coefficient of 1 with the 7th PC, which, honestly seems just a summary of the variable without much other interpretation. This make us think that *Area* is not, indeed, a very important factor in the determination of price. Moreover, as *Area* is essentially the only variable contributing to the 7th principal component, from the following scree plot we can clearly see that it only explains 4.5% of all variability which, to be honest, is a quite low impact.

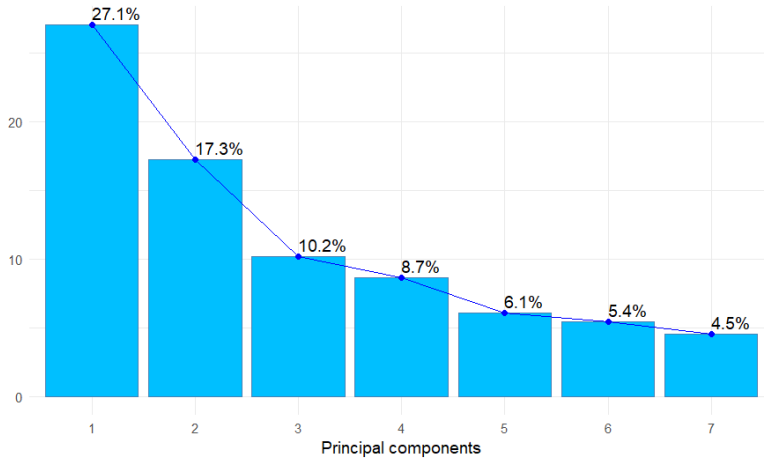

Figure 13 - Percentage of variance explained by each PC

6 FINAL CONSIDERATIONS

In this report we analysed the car's markets of countries of the Arabic peninsula and tried to obtain interesting insights about them. The main objectives were to create models as accurate and as optimized as possible to obtain good predictions of cars' prices according to their technical features and to understand which were the main factors that determined the price.

Unsupervised and supervised learning techniques were implemented after the first round of data exploration and data cleaning in which the original dataset was subsetted according to the analysed countries. To begin, in subchapter 2.1, I performed a Principal Component Analysis (PCA) to see whether there existed any relation between the y variable *PriceEURO* and the features. It was found that the first principal component was able to summarise quite well the effect each variable had on *PriceEURO* according to Figure 6. Instead, the second PC was much more difficult to interpret since it had no apparent relation with anything else. Before continuing the analysis, I applied some supervised learning models to the PC-datasets in order to measure their performance. The results were not amazing and were vastly discussed in this report.

In the following chapter, I started the supervised learning analysis by looking at the distributions of the residuals of the logarithmic transformation of the response variable *PriceEURO*. None of those distributions, according to the Shapiro-Wilk normality test, were gaussian except for the one of United Arab Emirates. Thus I decided not to use the linear regression model for predictive purposes but, instead, I exploited it to solve some linear problems the data had, namely multicollinearity, thanks to the Variance Inflation Factor (Vif) analysis. After having removed variables that suffered from multicollinearity, I used the *Mallows' Cp* criterion to select the best subset.

Since multicollinearity was not a problem anymore, but outliers were, I implemented the robust regression – subchapter 3.3 – to neutralize the effect that those observation had on the regression. Results were great but probably could be improved by other models.

After the robust regression, I implemented regression trees – subchapter 3.4 – to make things more interpretable. I optimized the trees by selecting the *cost of pruning* (cp) that minimized the error and resulted in regression trees easy to interpret but that had a worse performance than the robust regression.

Lastly, I decided to implement the random forest – subchapter 3.5 – to see if I was able to obtain a model with better performances. After the tuning of the hyperparameters *mtry*, which decides how many features must be analysed in each node, and the *ntree*, which controls how many trees is the random forest made of, I was able to obtain a very good model with an R^2 always higher than 0.91 (see Table 12).

So, we can conclude that, according to accuracy and prediction power, the random forest is, for sure, the best model but it is very expensive in terms of computational time and not that simple to interpret, at second place there is the robust regression model and lastly the regression tree. This is true for both normal-data and principal component datasets meaning that, in our case, models better interpreted normal-data rather than PCs. This also meant that there was little to no point in implementing the principal component analysis since there was literally no improvement in prediction power but a little improvement in time taken to do the computations.

Furthermore, if we look at the most important variables – Table 16 – we see that most of the times models agree on which are the most influential features in the determination of price, concluding that, indeed, the analysis generated some quite good results.

Finally, I tried to reassemble the dataset into one to create three general models, one for each method already seen, to see whether the variable *Area* was a determinant factor in the determination of price. Results, in terms of prediction power were amazing since the random

forest model was able to reach an R^2 of 0.97 on the test. Also the robust regression and the tree predictor performed nicely but, as before, they were far from being potent as the random forest. Still, they are incredibly faster than the random forest to load and they return more or less similar results.

We obtained, as Table 18 show, slightly different results regarding the most influential variables compared to the ones observed in Table 16. Here *Torque* is still the most influential variable but with less margin. Furthermore, according to the robust regression, the variable *Area* is in most cases – meaning for most countries – not even statistically significant and in the case of the random forest, the *Area* is only the 7^{th} most important variable. It seems that the most influential factors, rather than the *Area*, are the physical characteristics of the car. Again, the implementation of principal components did not improve the computations nor the interpretability but decreased by a bit the loading times.

After all these analyses we can finally say that a car does not cost differently just because it is sold in a different country but, different countries may have different preferences and thus may value more or less cars with certain characteristics.

APPENDICES

APPENDIX A

Contribution of variables to 1st PC and their correlation up to the 6th principal component, divided by country.

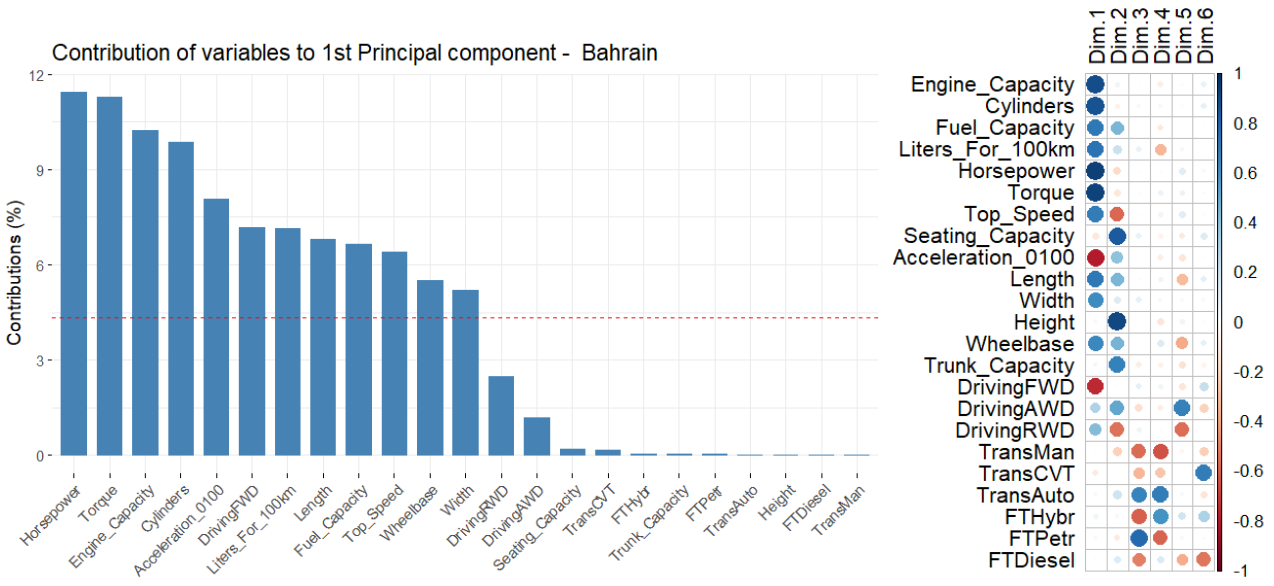

Figure 14 - Correlation with the PCs, Bahrain

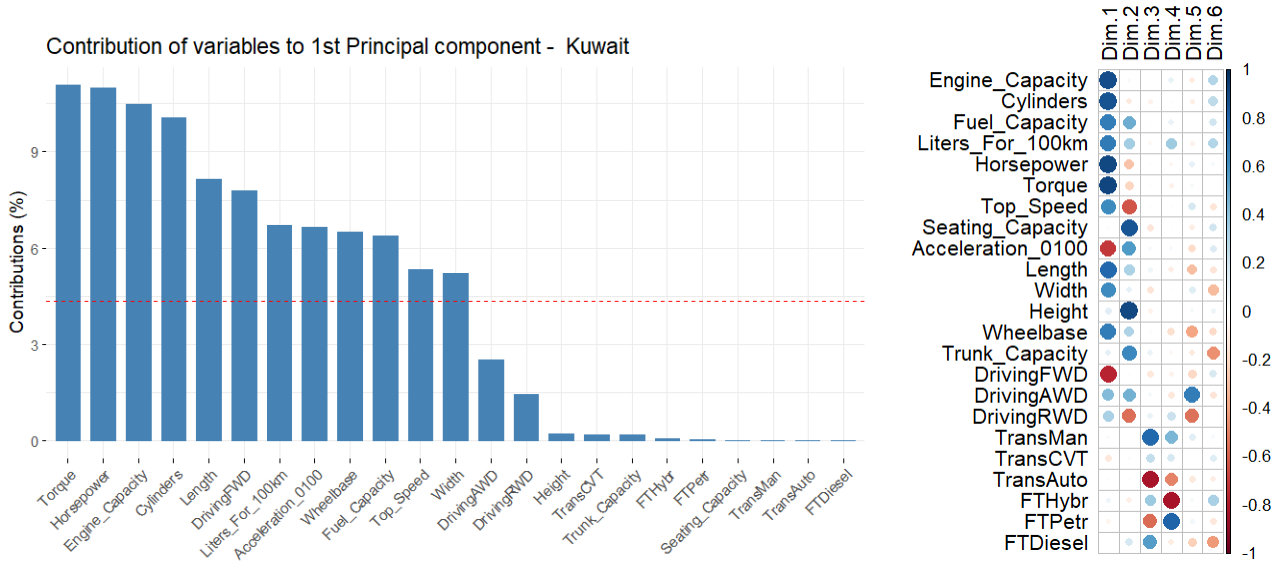

Figure 15 - Correlation with the PCs, Kuwait

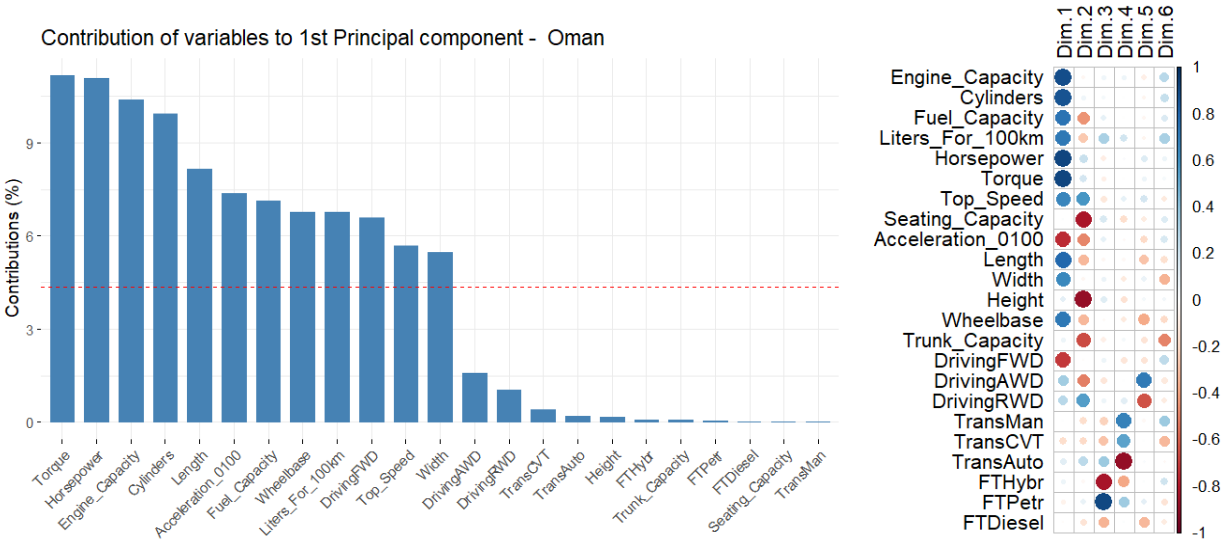

Figure 16 - Correlation with the PCs, Oman

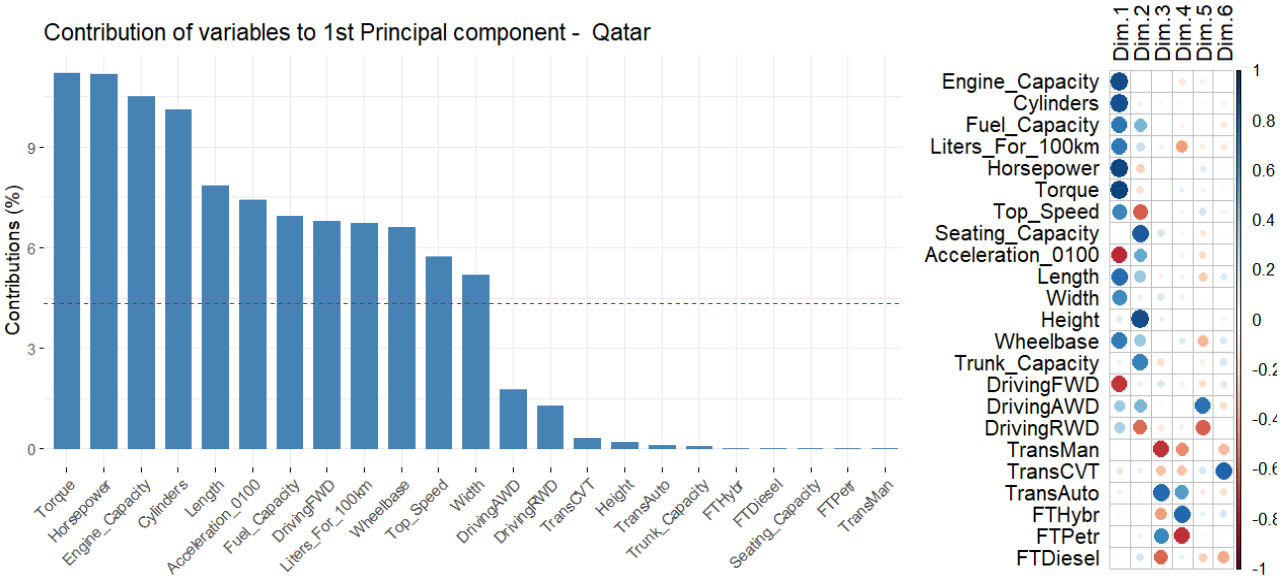

Figure 17 - Correlation with the PCs, Qatar

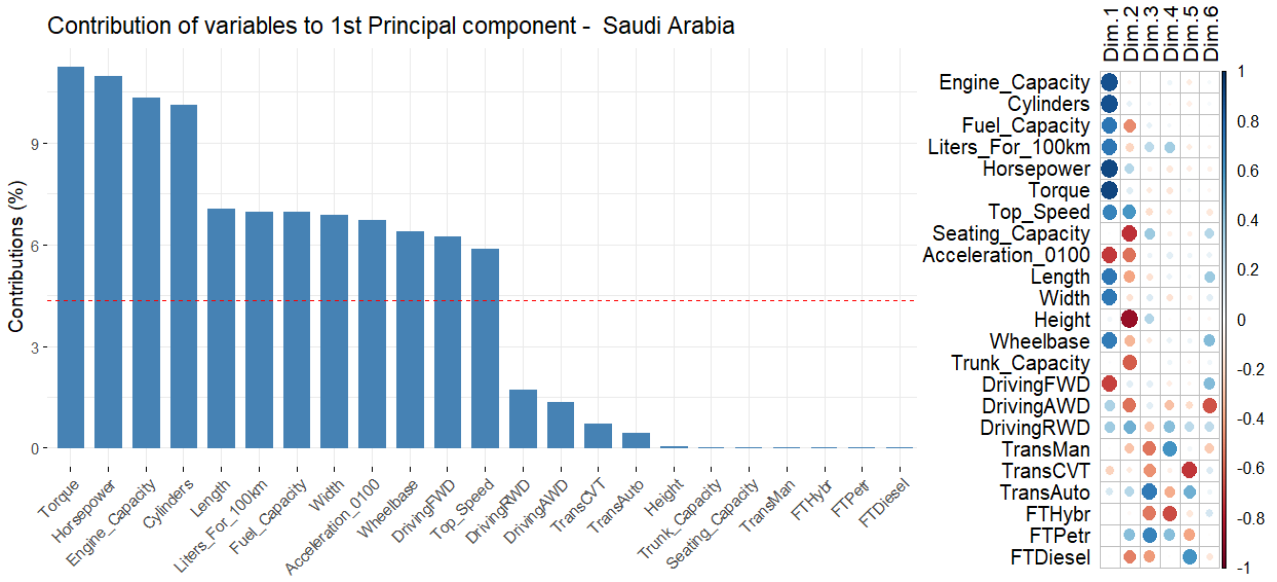

Figure 18 - Correlation with the PCs, Saudi Arabia

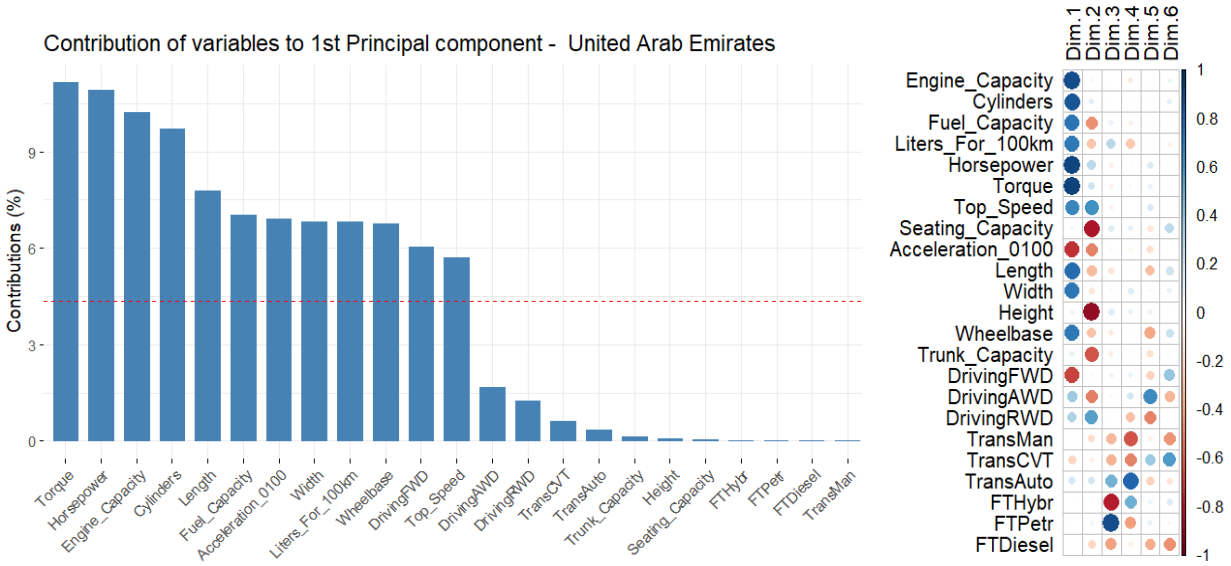

Figure 19 - Correlation with the PCs, United Arab Emirates

APPENDIX B

Regression tree representations of *Cart* algorithm run on PC datasets, divided by country.

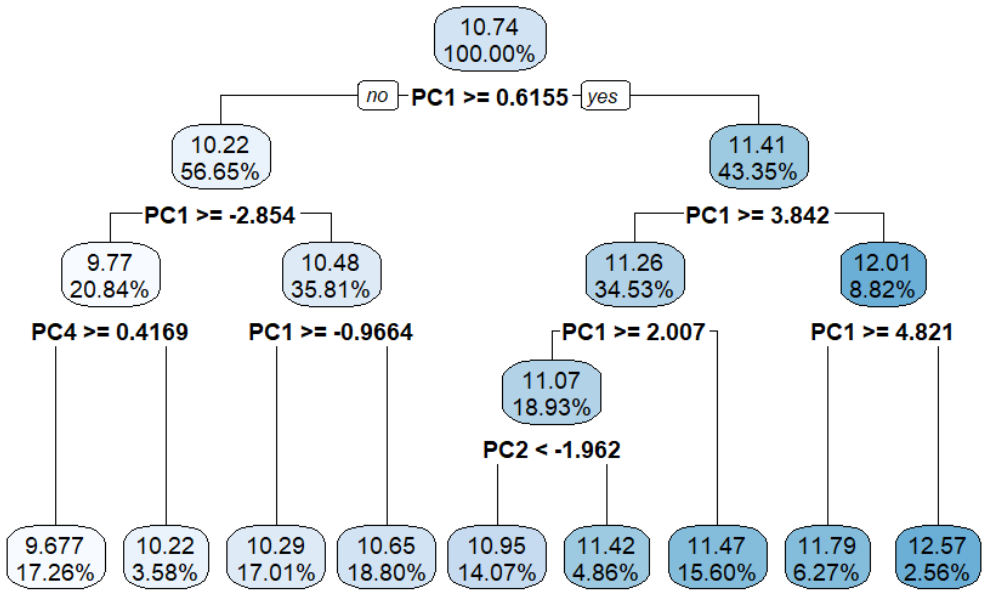

Figure 20 - PCA tree for Bahrain

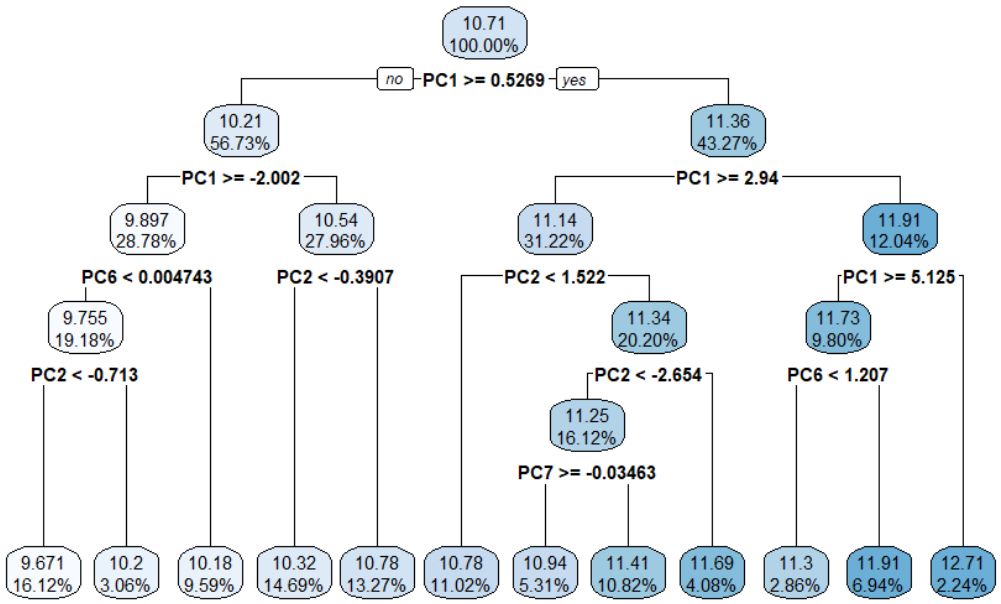

Figure 21 - PCA tree for Kuwait

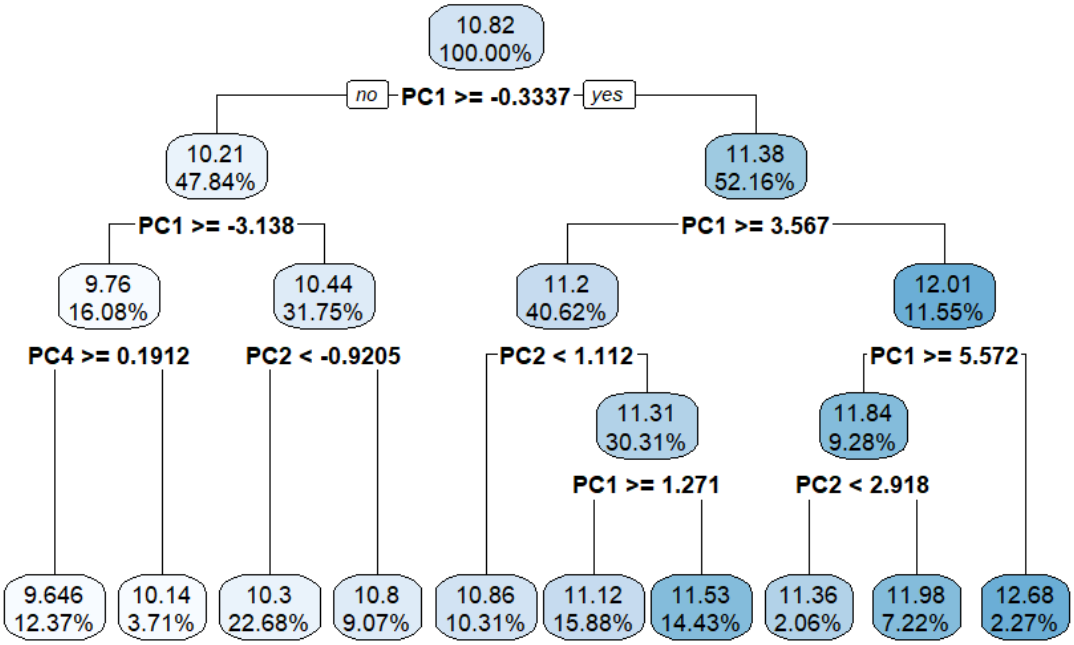

Figure 22 - PCA tree for Qatar

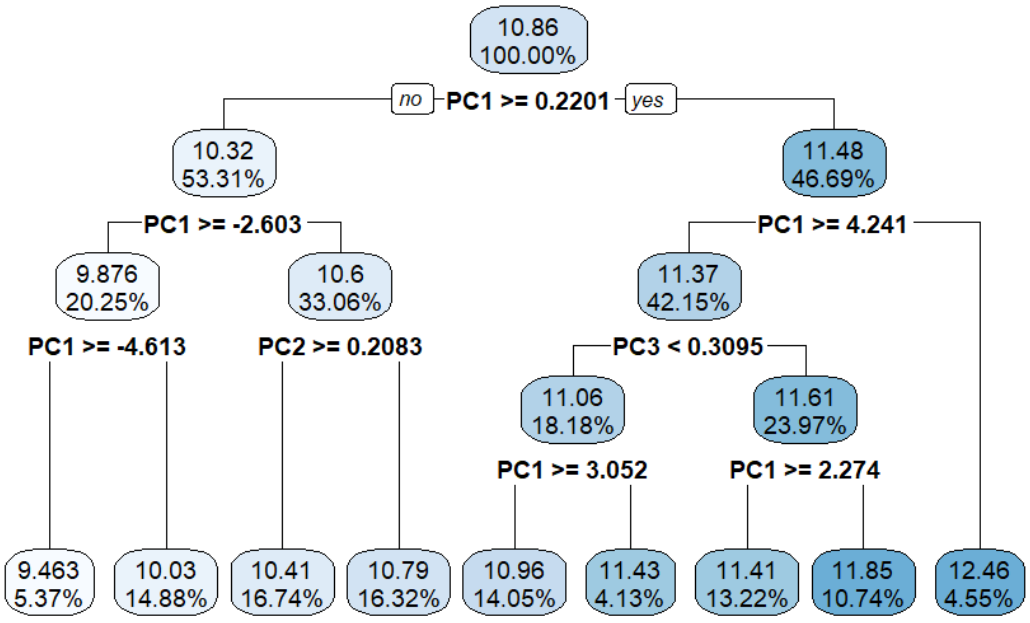

Figure 23 - PCA tree for Oman

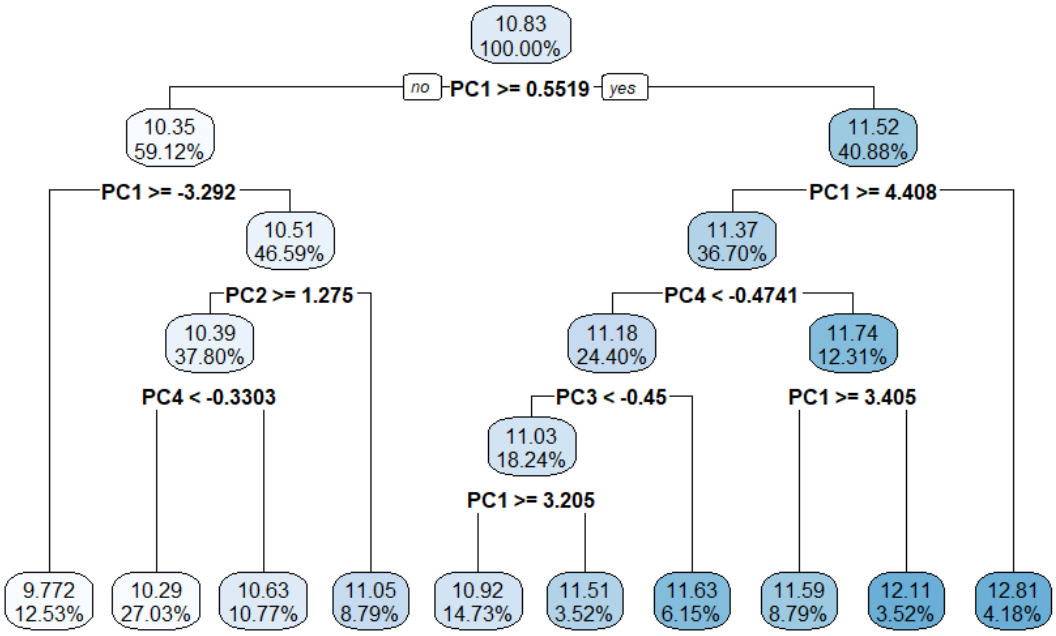

Figure 24 - PCA tree for Saudi Arabia

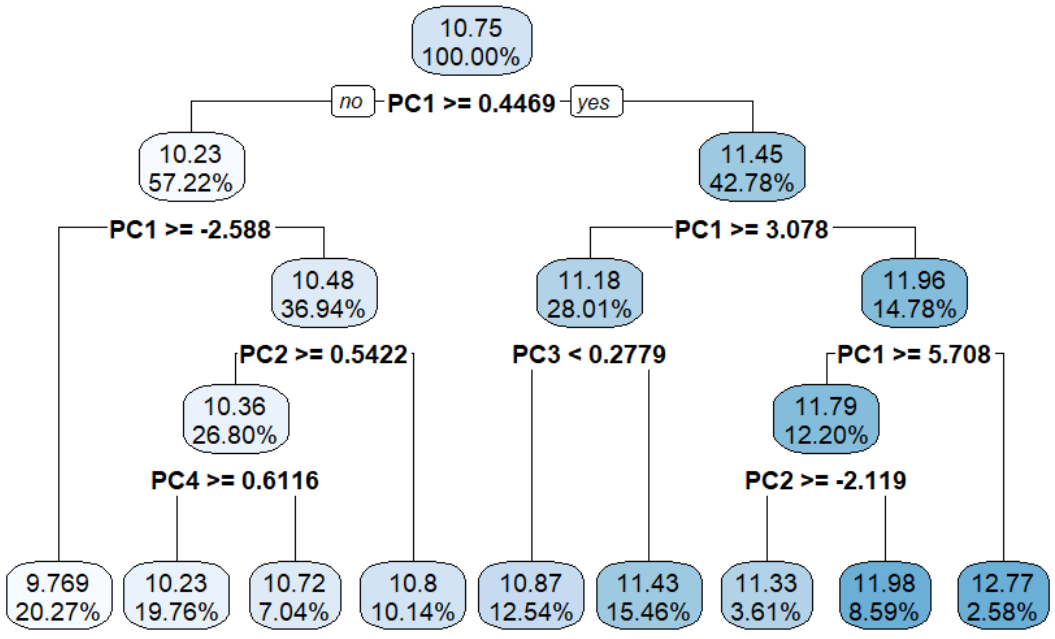

Figure 25 - PCA tree for United Arab Emirates

APPENDIX C

Variance inflation factor (vif) and best subset selection for each country.

- Bahrain

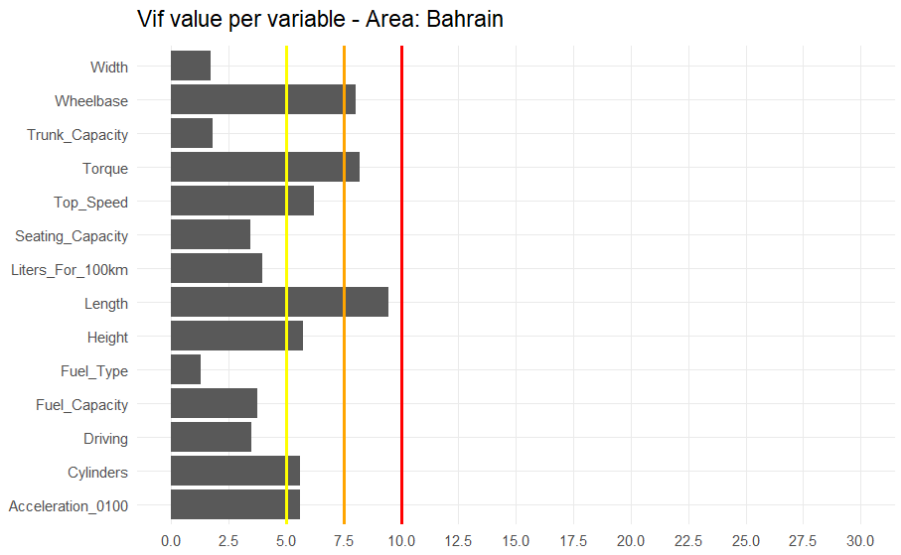

Figure 26 - Vif Bahrain

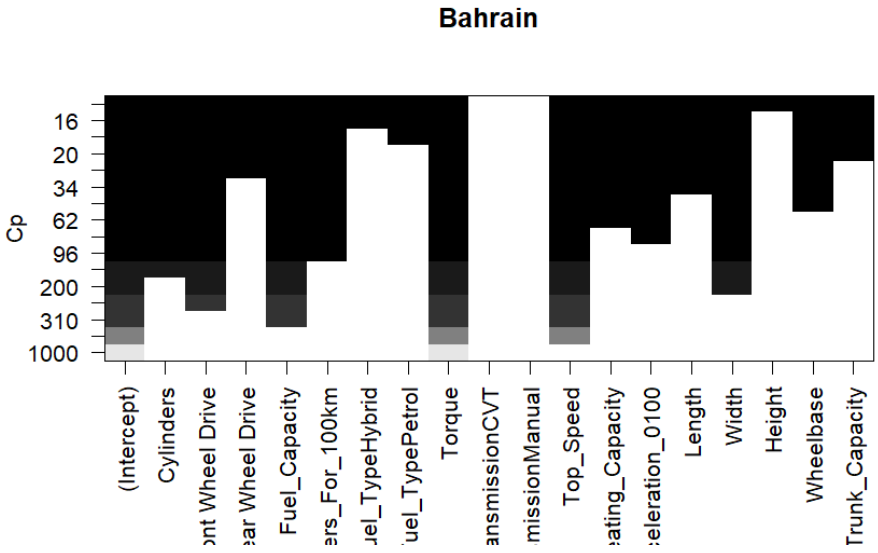

Figure 27 Best subset selection Bahrain

Variables removed from best subset:

- Transmission

- Kuwait

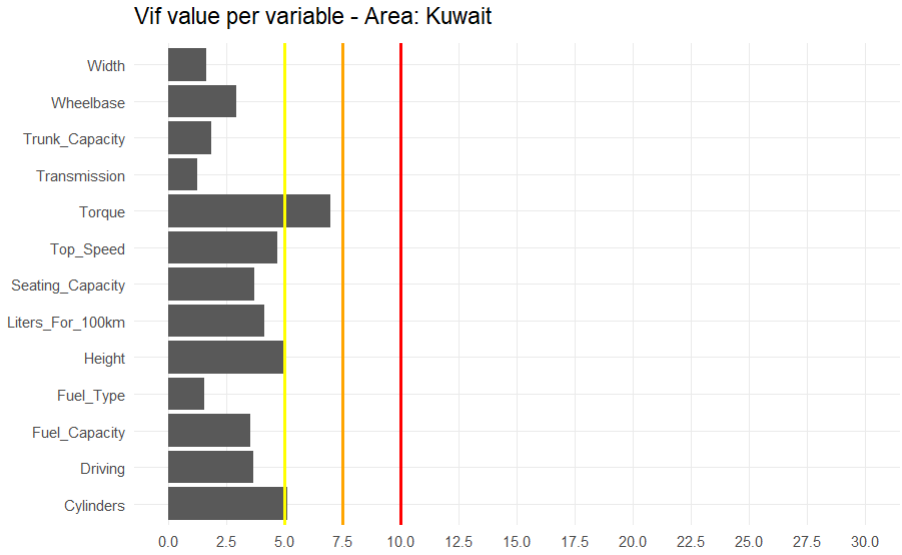

Figure 28 – Vif Kuwait

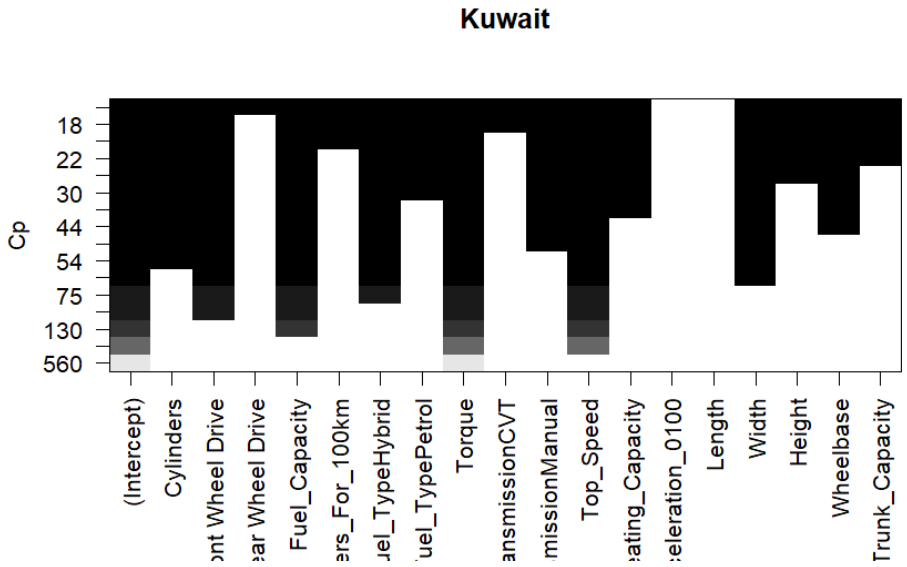

Figure 29 – Best subset selection Kuwait

Variables removed from best subset:

- Acceleration
- Length

- Oman

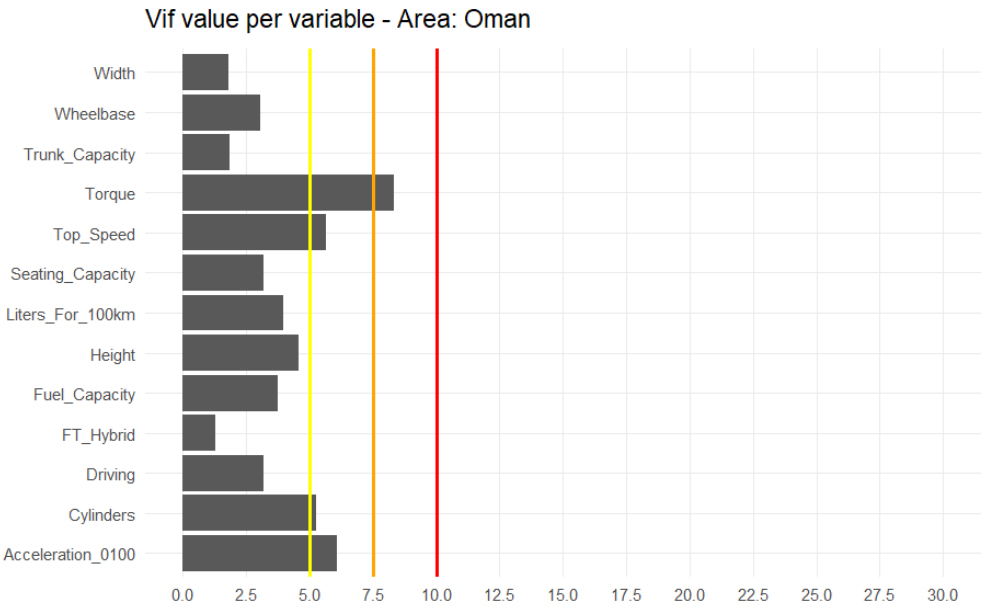

Figure 30 - Vif Oman

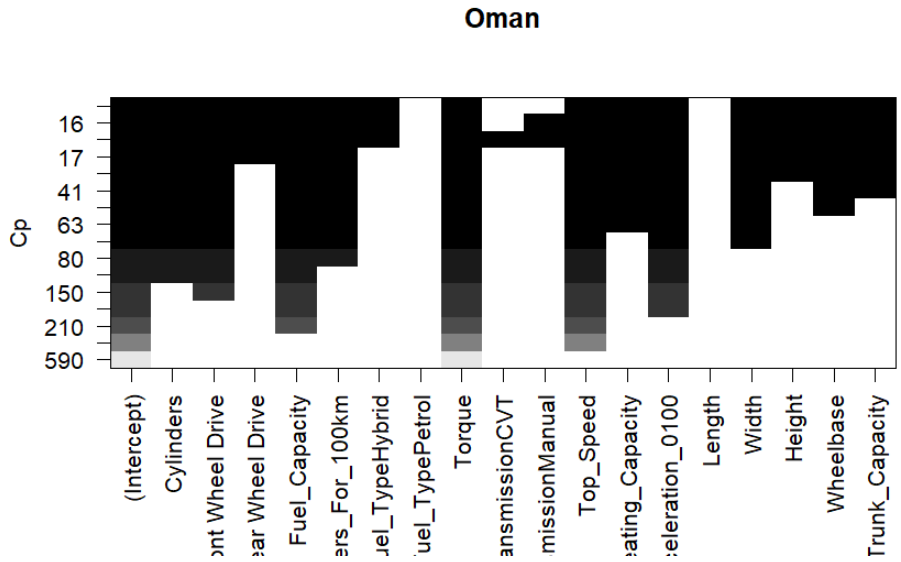

Figure 31 - Best subsect selection Oman

Variables removed from best subset:

- Dummy variable *Petrol* obtained from *Fuel_Type*
- Transmission
- Length

- Qatar

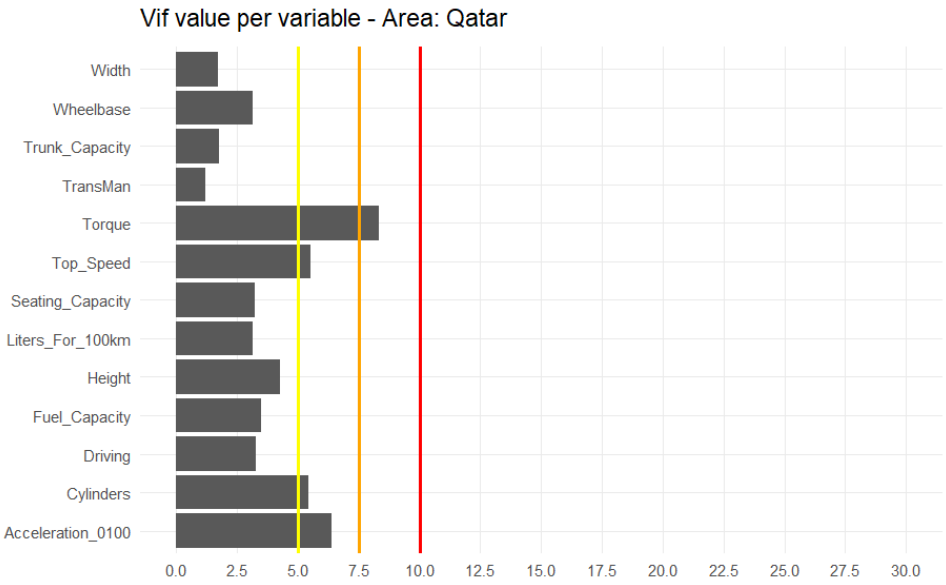

Figure 32 - Vif Qatar

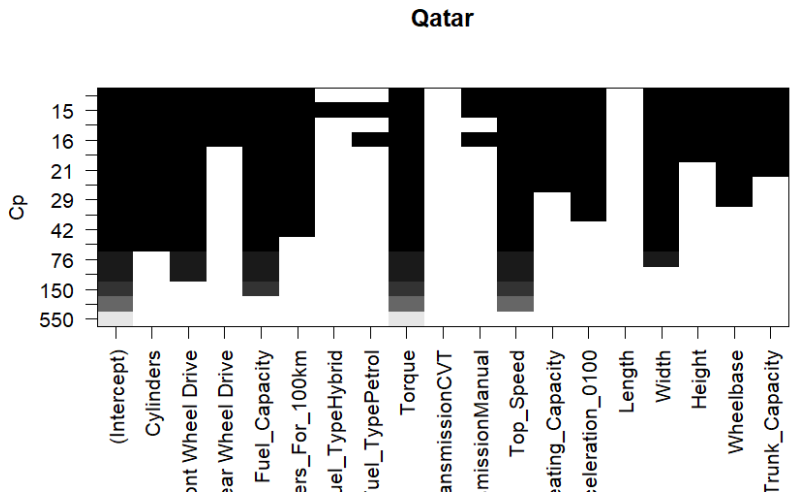

Figure 33 - Best subsect selection Qatar

Variables removed from best subset:

- Fuel Type
- Dummy variable *CVT* obtained from *Transmission*
- Length

- Saudi Arabia

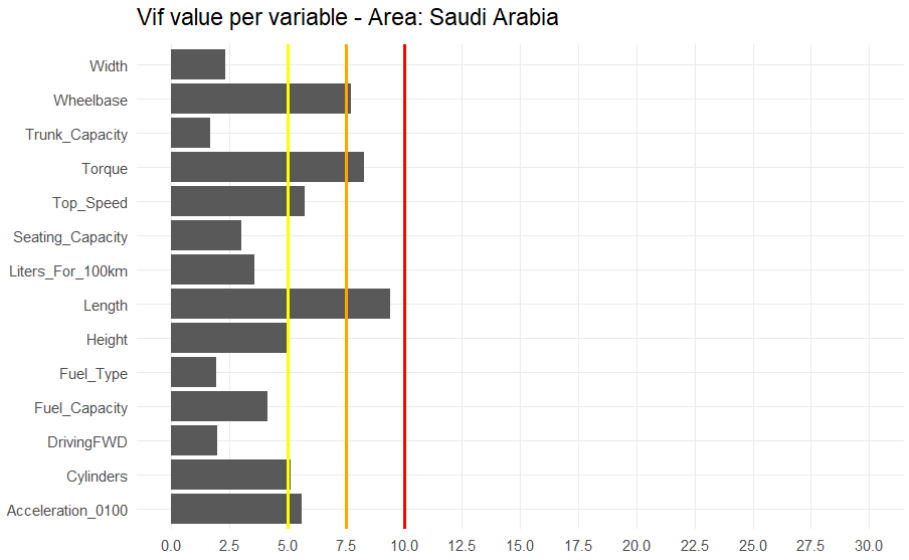

Figure 34 - Vif Saudi Arabia

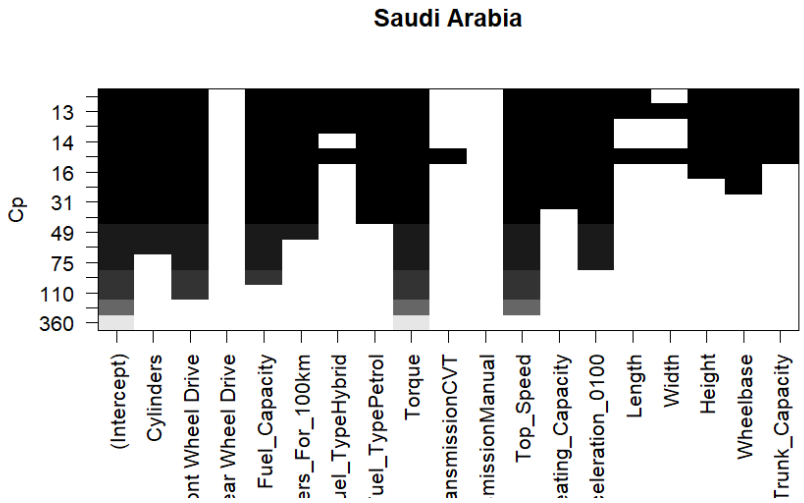

Figure 35 - Best subsect selection Saudi Arabia

Variables removed from best subset:

- Transmission
- Dummy variable *Rear wheel drive* from *Driving*
- Length

- United Arab Emirates

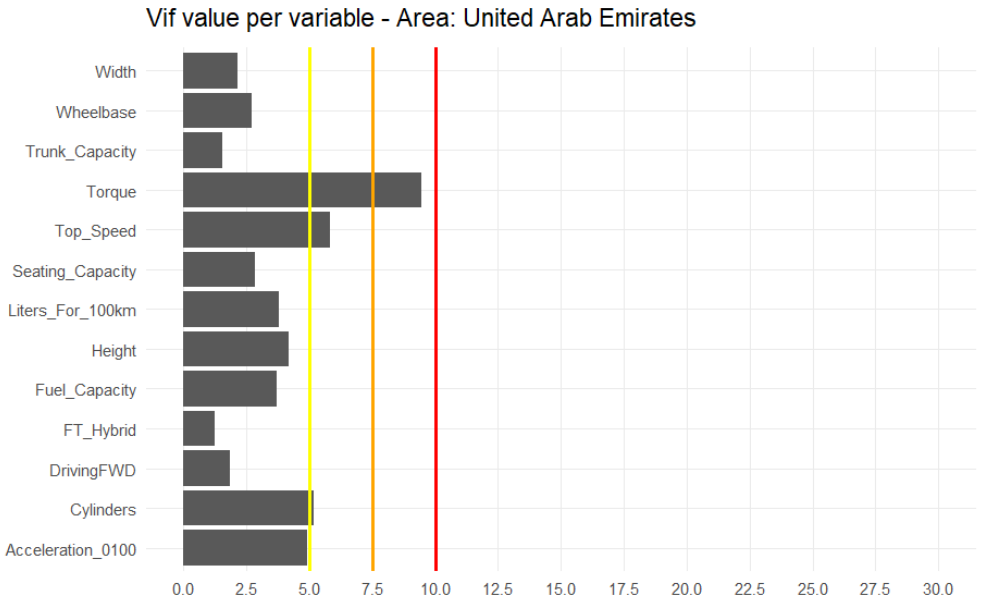

Figure 36 - Vif United Arab Emirates

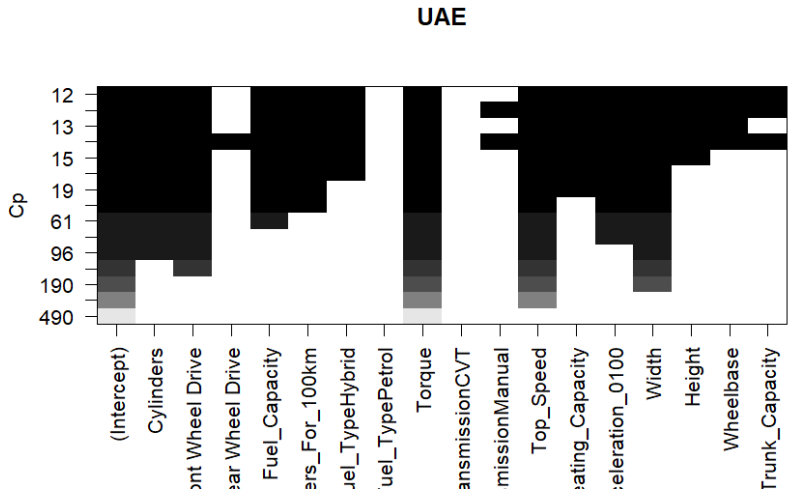

Figure 37 - Best subsect selection United Arab Emirates

I removed also the variable *Length* because it showed a Vif $\geq$ 10

Variables removed from best subset:

- Dummy variable *Rear Wheel Drive* obtained from *Driving*
- Dummy variable *Petrol* obtained from *Fuel_Type*
- Transmission

APPENDIX D

Regression tree representations of the *Cart* algorithm run on normal-data datasets, divided by country.

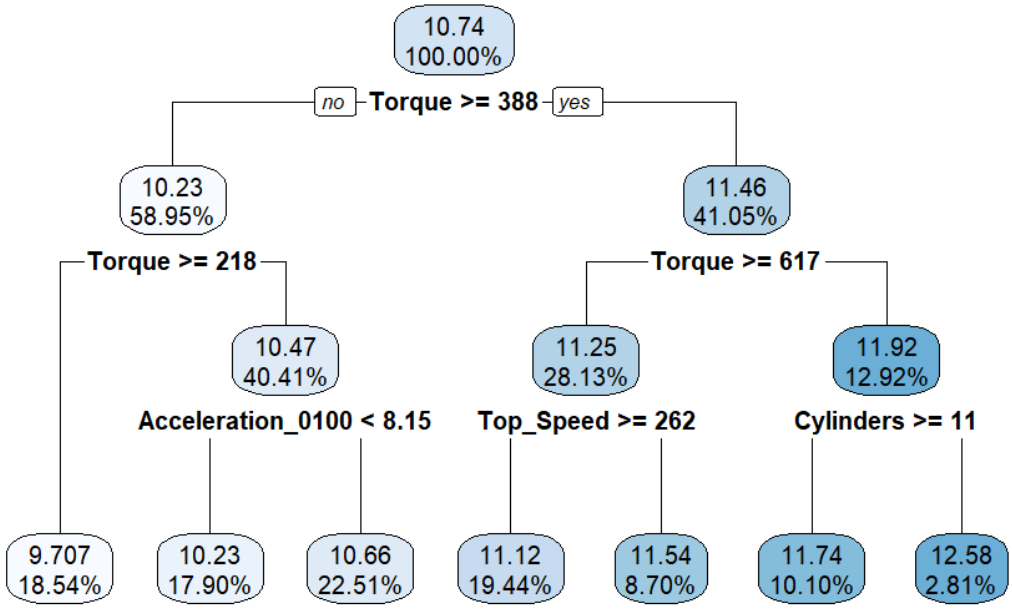

Figure 38 - Regression tree for Bahrain

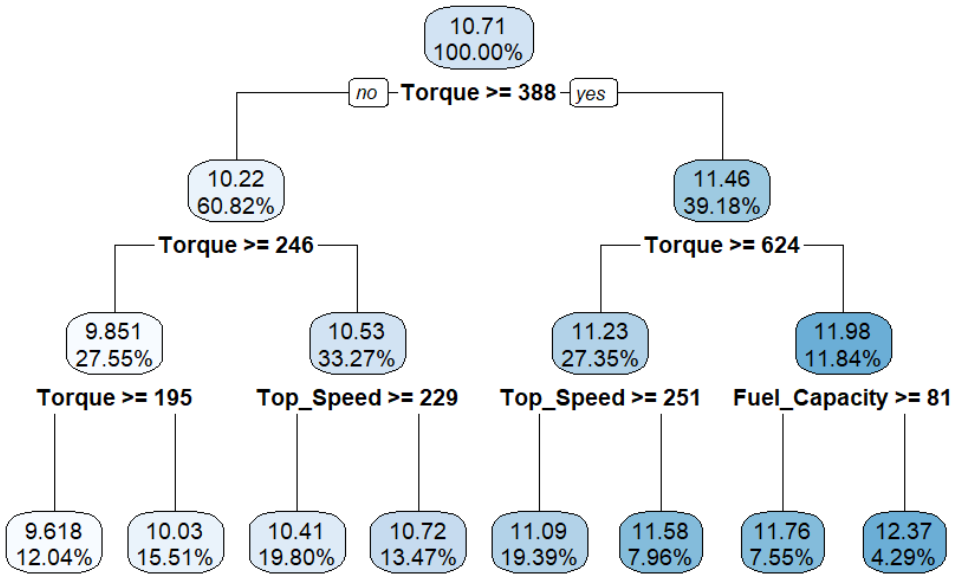

Figure 39 - Regression tree for Kuwait

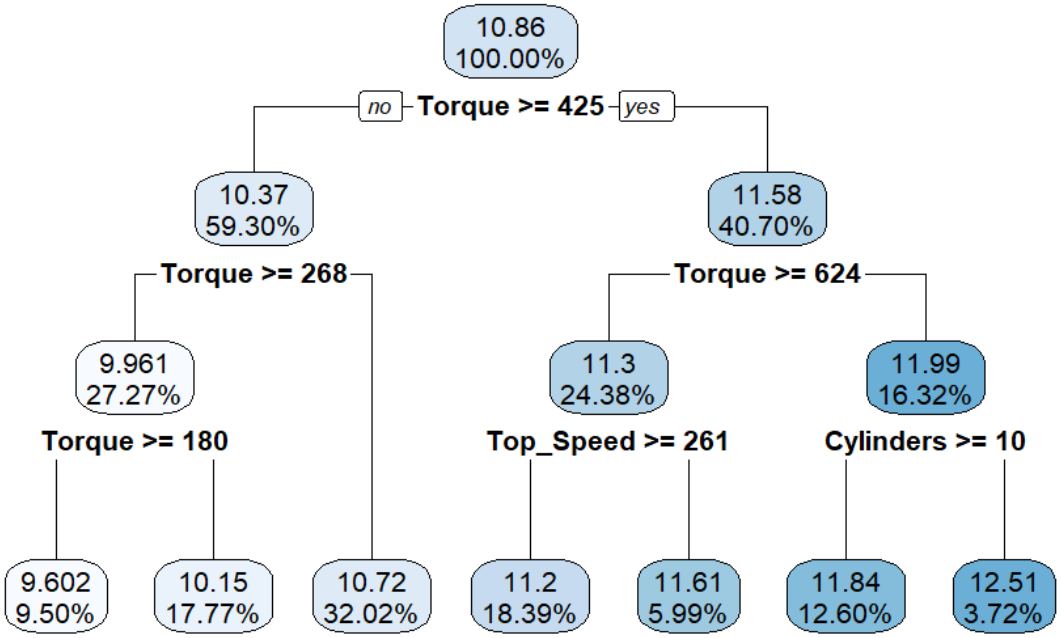

Figure 40 - Regression tree for Qatar

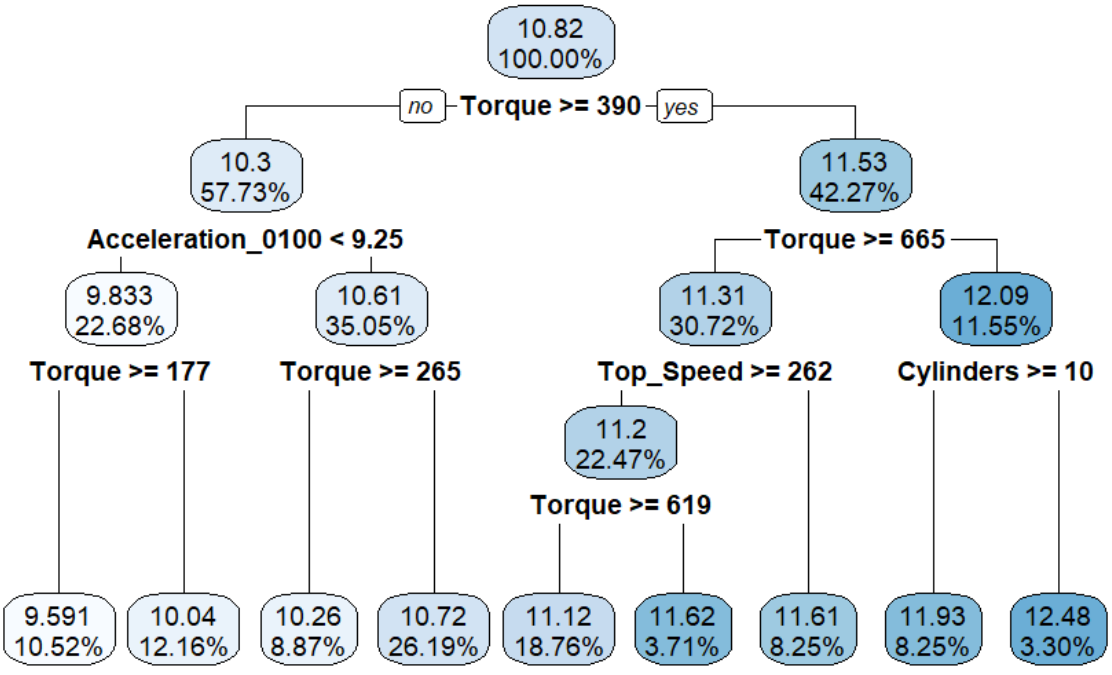

Figure 41 - Regression tree for Oman

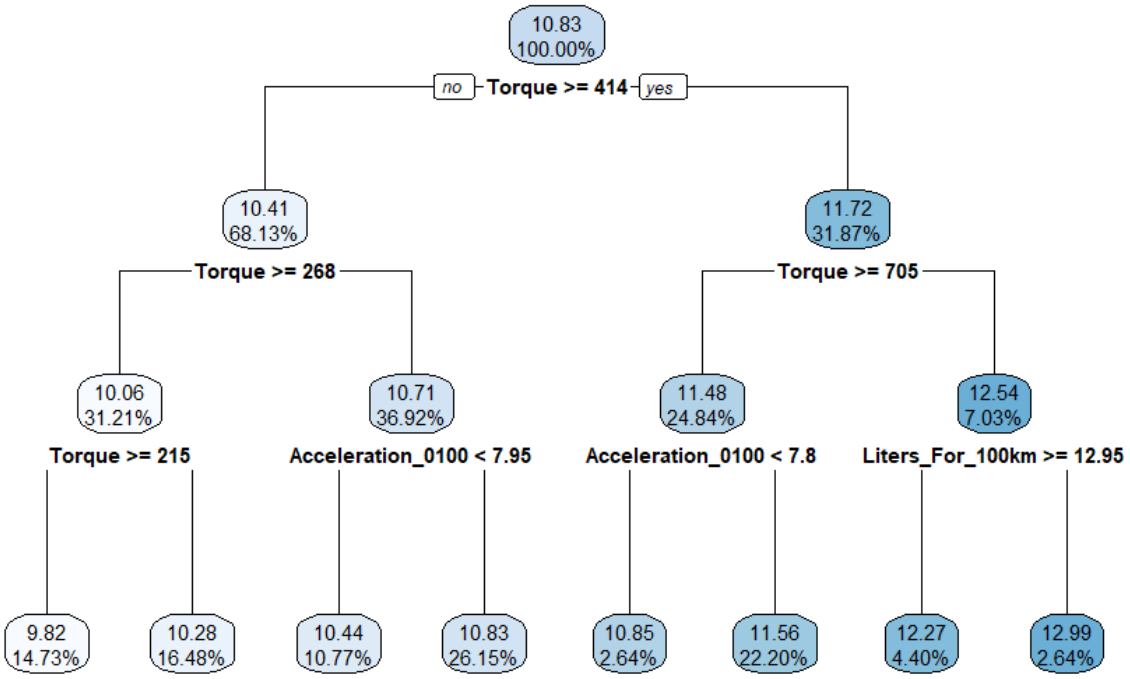

Figure 42 - Regression tree for Saudi Arabia

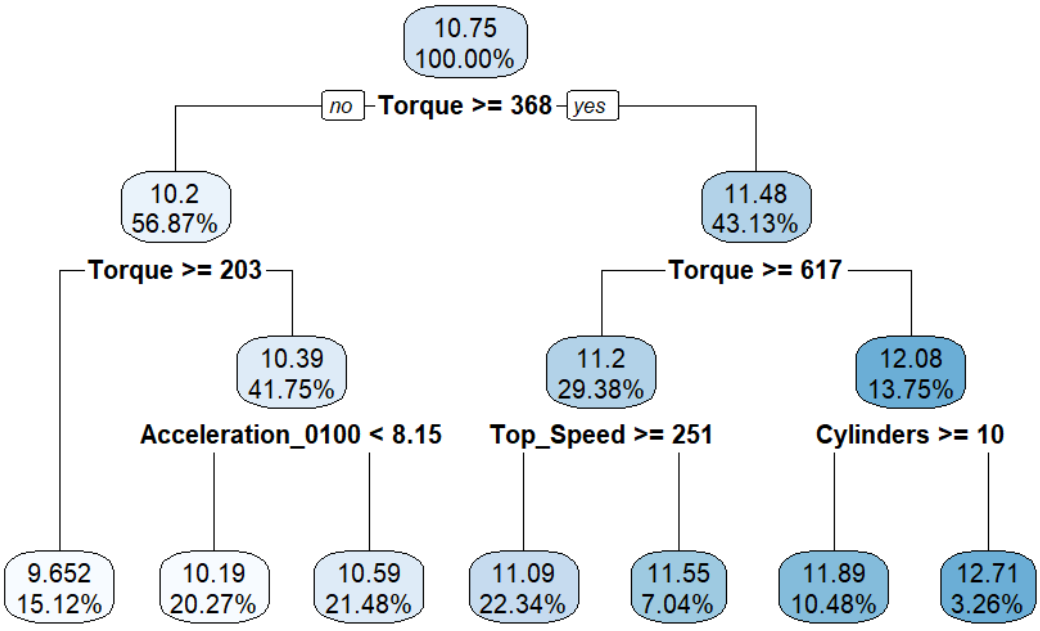

Figure 43 - Regression tree for United Arab Emirates