Leveraging Clustering Algorithms in Defining Relevant Markets for Competition Policy Analysis

Kurdoglu, Berkay

27 November 2023

Introduction

In antitrust and competition law practices, market definition is the initial stage which is mostly based on actual data regarding the substitution behaviors of both consumers and producers. It is defined by the European Commission as: "*A relevant (product) market comprises all those products and/or services which are regarded as interchangeable or substitutable by the consumer by reason of the products' characteristics, their prices and their intended use*". The concept of the relevant market is crucial from the view of competition law and economics perspective since it defines the landscape of effective competition taking place between products, brands, firms, and even markets. Therefore it must be emphasized that the market share or market power of the one firm or the product entirely depends on how the market is delineated. For instance, one can ask whether the cola and the orange juice can be substitutable and thus might be evaluated in the same relevant market. In the competition economics literature, central to this endeavor is the SSNIP Test (Small but Significant and Non-transitory Increase in Price), a hypothetical analysis to evaluate consumer responses to price changes, thus gauging the substitutability of products or services. Complementing this, Critical Loss Analysis scrutinizes the viability of such price increases by balancing hypothetical losses against actual market reactions. Cross-elasticity of demand further enriches this analysis by quantifying the sensitivity of demand for one product relative to price fluctuations in another, shedding light on inter-product substitutability. Additionally, price correlation analysis, a more supplementary than conclusive approach, reveals interconnected market dynamics through correlated price movements. Therefore, it can be concluded that the backbone of traditional efforts to define relevant markets hinges on demand estimation models through econometric models to dissect demand patterns and elasticity. Nonetheless, due to the problems of endogeneity between price and quantity in the estimation process, these techniques usually require solid and suitable IV variables that cannot be found easily. Moreover, these models' time and data requirements prevent them from being widely used. Owing to these practical limitations and the requirement for a wider range of convincing proof, practitioners commence seeking empirical and more practical tools to help themselves define markets by utilizing cutting-edge AI technologies and a broader spectrum of new-type quantitative instruments.

In light of the foregoing, the identification of relevant markets is a fundamental step in the complex world of antitrust litigation/competition law and a difficult process that often calls for an in-depth grasp of industry dynamics as well as economic data, accurate modelling, and strong foundations. Even while they are reliable, traditional approaches occasionally struggle to fully capture the complex nature of market dynamics, particularly when it comes to quickly changing industries and digital marketplaces. From this motivation, the present study aims to explore the intersection of technology and law, particularly on the use of clustering algorithms to help define relevant markets for antitrust cases. With the help of these algorithms, we can discover insights into competitive landscapes, reveal certain characteristics of ever-changing market structures, and gain a deeper understanding of rivalry relationships between different products and services.

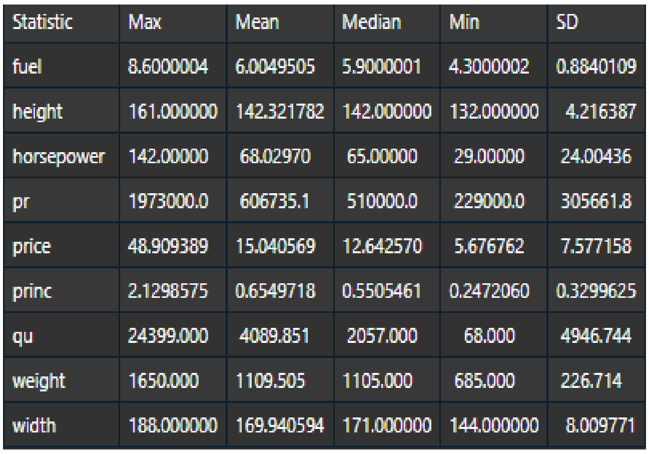
Data

To carry out experimental analysis, I use the dataset on the European car market, collected by Goldberg and Verboven (2001) and Bjornerstedt and Verboven (2014). The total number of observations is 11,483: there are 30 years (1970–1999) and 5 countries (Belgium, France, Germany, Italy, and the United Kingdom), which implies an average of 77 car models per year and country. The car market is divided into five groups according to the segments: subcompact, compact, intermediate, standard, and luxury. The variables I am going to use as a proxy for clustering are sales (qu), price parameters (measured in 1,000 Euro in 1999 purchasing power), the product characteristics such as horsepower (in kilowatts), fuel efficiency (in liter/100 kilometers), width (in centimeters), and height (in centimeters).

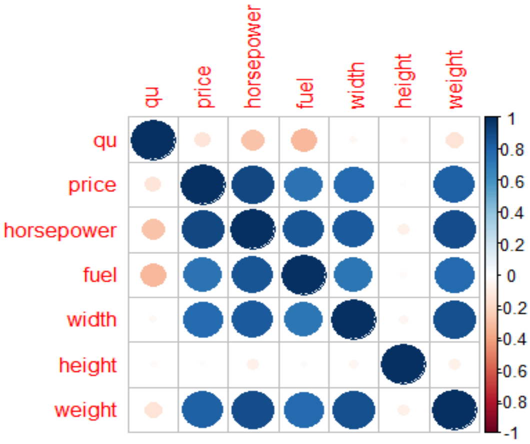
In order to provide uniformity, compatibility, and homogeneity, the year 1999 and the country Belgium which has the maximum observation numbers among the others are chosen. The variables unrelated to studies are dropped. The dataset subject to study comprised 101 observations each signifying a unique car model and its characteristics (price, quantity, horsepower, fuel, width, height, fuel). The segment (character) vector is also transformed into the numerical vector so as to execute comparisons given that the European Commission has a tendency to define the relevant market in the automobile industry considering segmentation[1]. Therefore this variable is going to play a pivotal role and constitute the core benchmark of the analysis.

Descriptive Statistics

Graph 1: Descriptive Statistics for All Variables & Correlation Matrix



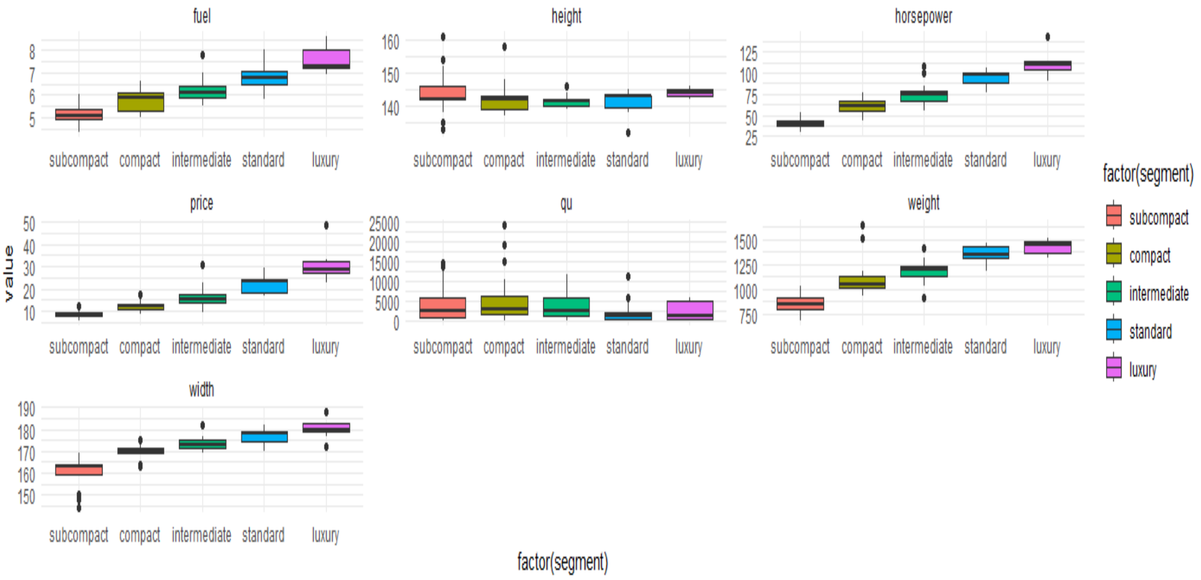
Statistic	Max	Mean	Median	Min	SD
fuel	8.6000004	6.0049505	5.9000001	4.3000002	0.8840109
height	161.000000	142.321782	142.000000	132.000000	4.216387
horsepower	142.00000	68.02970	65.00000	29.00000	24.00436
pr	1973000.0	606735.1	510000.0	229000.0	305661.8
price	48.909389	15.040569	12.642570	5.676762	7.577158
princ	2.1298575	0.6549718	0.5505461	0.2472060	0.3299625
qu	24399.000	4089.851	2057.000	68.000	4946.744
weight	1650.000	1109.505	1105.000	685.000	226.714
width	188.000000	169.940594	171.000000	144.000000	8.009771

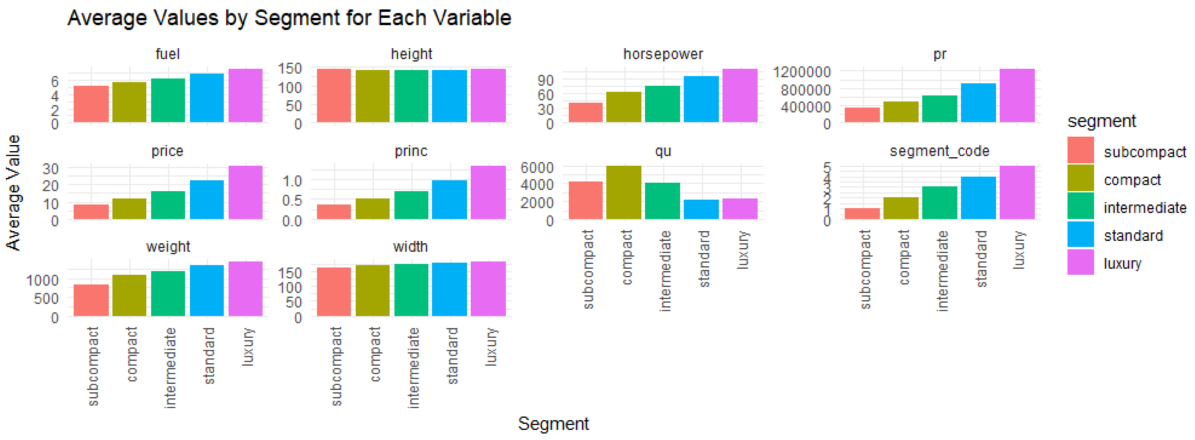


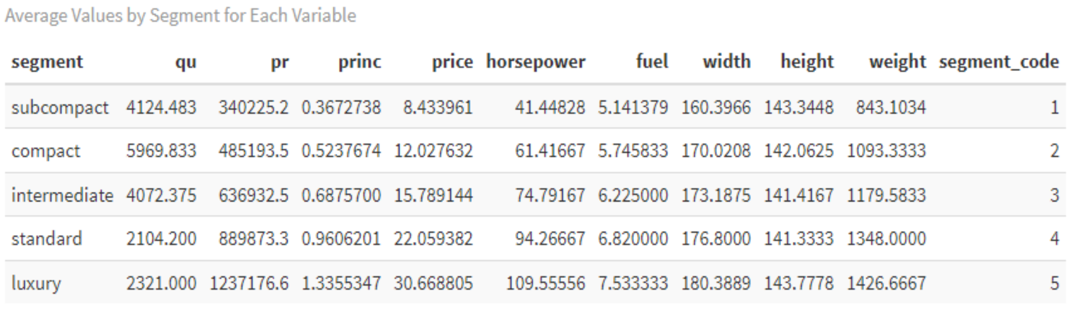
As we can see from the above, the price is strongly correlated with the horsepower, width, weight and fuel respectively, on the other hand, between the price and quantity sold, there is no significant correlation (price and princ are just different presentations of the price therefore not included in the correlation matrix and some of our graph presented later). Accounting that our analysis is directly related to the segmentation of the cars, as of now, all type of characteristics is going to be illustrated by the segmentation classification. In addition, data pre-processing involved normalization to ensure uniformity of scale across all variables

[1] EC/Case M.9730 – FCA/PSA

Graph/Table 2: Descriptive Statistics With Respect to Segmentation



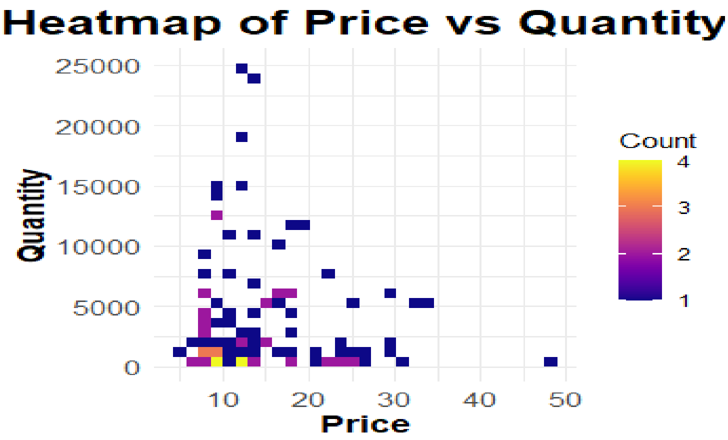



Average Values by Segment for Each Variable

segment	qu	pr	princ	price	horsepower	fuel	width	height	weight	segment_code
subcompact	4124.483	340225.2	0.3672738	8.433961	41.44828	5.141379	160.3966	143.3448	843.1034	1
compact	5969.833	485193.5	0.5237674	12.027632	61.41667	5.745833	170.0208	142.0625	1093.3333	2
intermediate	4072.375	636932.5	0.6875700	15.789144	74.79167	6.225000	173.1875	141.4167	1179.5833	3
standard	2104.200	889873.3	0.9606201	22.059382	94.26667	6.820000	176.8000	141.3333	1348.0000	4
luxury	2321.000	1237176.6	1.3355347	30.668805	109.55556	7.533333	180.3889	143.7778	1426.6667	5

It could be observed from the graph and table presented above that as the segmentation level goes up; price, horsepower, fuel, width, and weight also consistently and gradually increase as well. Nevertheless, there is no prima facie pattern about quantity and height. This result may be problematic from the lens of traditional market definition that mostly relies on bilateral relations between prices and quantity. In conjunction with that clustering algorithms come in quite handy for the situation at hand. This circumstance is demonstrated through the heat map supplied below.

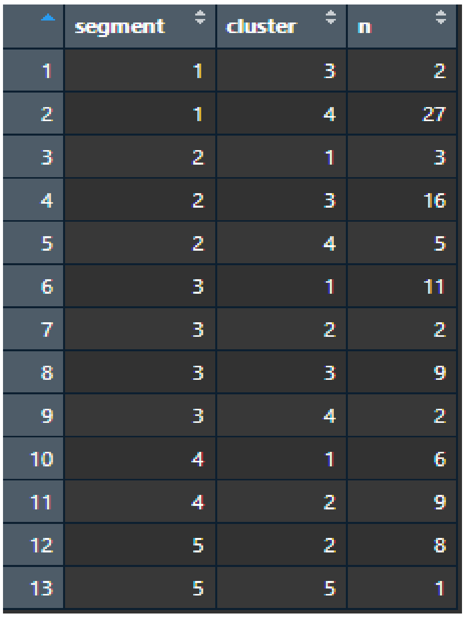
Graph/Table 2: Descriptive Statistics With Respect to Segmentation



Methodology and Results

In my investigation, I deployed multiple approaches to determine which car models can be alternative options (substitution) to each other (constituting a separate sample) by making use of clustering algorithms. In other words, my main objective in this study is to investigate how the market can be defined and how many sub-segments it could have. I also have the premise that the market has 5 segments that can be identified as a benchmark of the ideal cluster number. However, I also assume that the market has more or less than 5 clusters. First of all, the K-Means clustering algorithm is harnessed with the constraint of 5 clusters and without any constraint supplemented by the Elbow Method to ascertain the optimal cluster count. Afterward, hierarchical clustering with Gower distance method can be effective for mixed types of data selected and employed through silhouette score evaluations. Lastly, after reducing dimensionality with Principal Component Analysis (PCA), I applied Gaussian Mixture Models (GMM) to identify clusters. GMM is particularly insightful in the sense that the market segments have the tendency to possess different variances and covariances.

A. K-Means Clustering With Pre-Defined Cluster Number



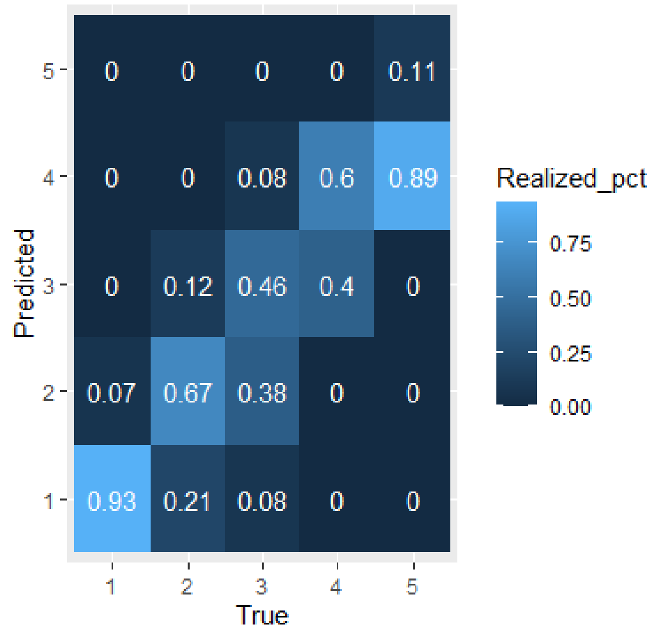
	segment	cluster	n
1	1	3	2
2	1	4	27
3	2	1	3
4	2	3	16
5	2	4	5
6	3	1	11
7	3	2	2
8	3	3	9
9	3	4	2
10	4	1	6
11	4	2	9
12	5	2	8
13	5	5	1

As we assume the market indeed comprises 5 different clusters in the sense of antitrust case law, I particularly specify the number of clusters equals 5 and then match relevant clusters to relevant segments. In the end, I created a confusion matrix and checked the model's performance by looking at the actual segment and predicted segment.

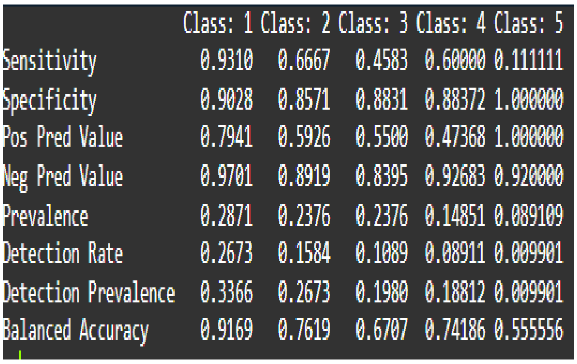
After estimating the clusters, I found the most repeated segment in each cluster group to match every cluster it's closest segment. In the graph located left: segment 1 is mostly reflected by cluster 4; segment 2 is mostly reflected by 3; segment 3 is mostly reflected by cluster 1, segment 4 is mostly reflected by cluster 2, segment 5 is actually mostly reflected by 2 however since we already used that cluster, we are going to go with cluster 5 for segment 5.

This result is indeed truly intuitive and it implies that segment 4 and segment 5 may constitute one relevant market instead of two separate markets.

Graph 3: Confusion Matrix for K-Means Clustering (K=5)



The model actually predicts decently accurate, especially for segment 1 and 2, and it also subtly implies that segment 4 and segment 5 can be substitute to each other.



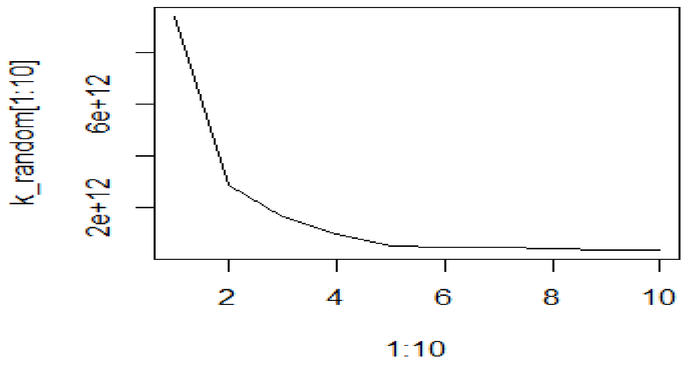
```
                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5
Sensitivity            0.9310   0.6667   0.4583  0.60000 0.111111
Specificity            0.9028   0.8571   0.8831  0.88372 1.000000
Pos Pred Value         0.7941   0.5926   0.5500  0.47368 1.000000
Neg Pred Value         0.9701   0.8919   0.8395  0.92683 0.920000
Prevalence             0.2871   0.2376   0.2376  0.14851 0.089109
Detection Rate         0.2673   0.1584   0.1089  0.08911 0.009901
Detection Prevalence   0.3366   0.2673   0.1980  0.18812 0.009901
Balanced Accuracy      0.9169   0.7619   0.6707  0.74186 0.555556
```

Overall the model offers valuable supplementary and novel information about how accurately a pre-defined relevant market is drawn from the framework of clustering methods.

B. K-Means Clustering (K=The Elbow Methods)

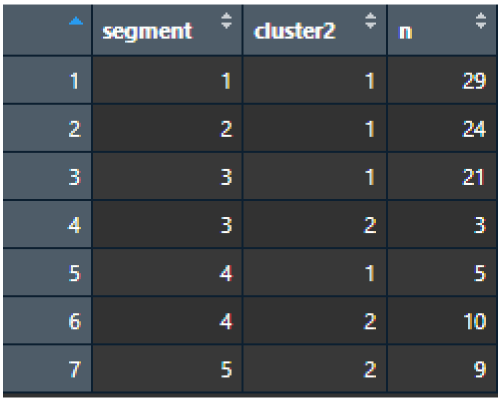
In this case, I am in an effort to define relevant markets by the optimal clustering number obtained by the Elbow Method without any prior information about the market (segments). Once I estimate and define the clusters, I thoroughly examine the control variables with respect to the relevant cluster and try to deduce what are the main characteristics of these identified markets. Hence, the results subtracted from the method are supposed to shed light on possible market segmentation regarding the level of price, horsepower, and so on.

Graph 4: The Elbow Method



The graph plots the within-cluster sum of squares (WSS) against the number of clusters (k) enabling us to look for a k at which the rate of decrease sharply shifts ("the elbow"), indicating that adding more clusters than 2 does not significantly improve the fit. So we picked our cluster number 2 which also can be interpreted as a number of relevant markets.

Graph 5: Most Repetitive Segments by Clusters

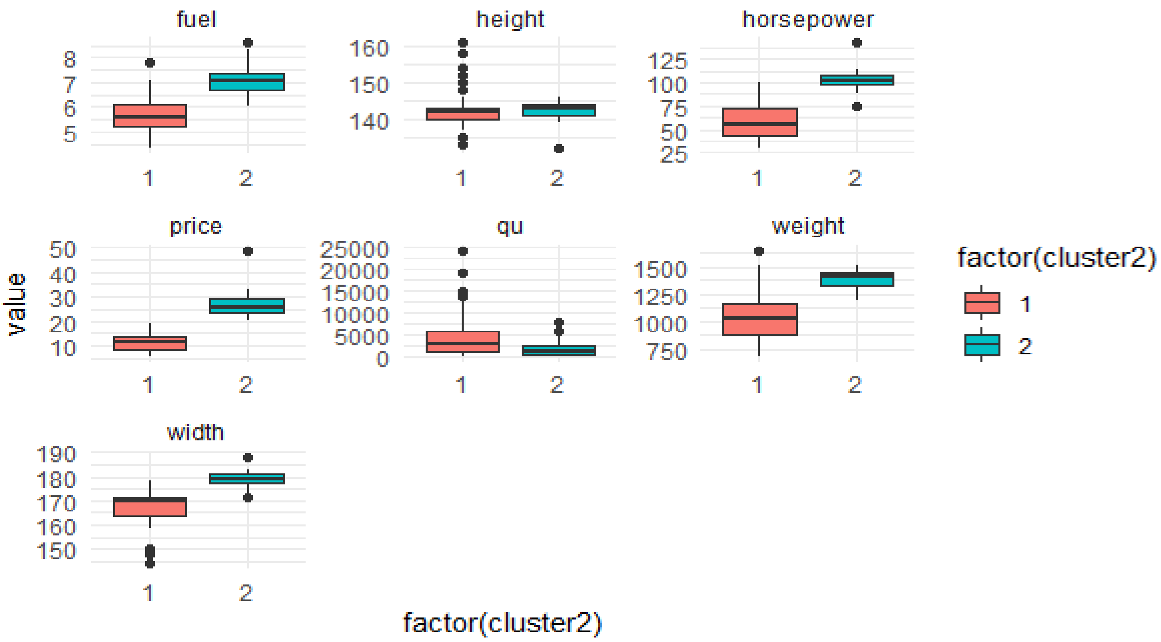


	segment	cluster2	n
1	1	1	29
2	2	1	24
3	3	1	21
4	3	2	3
5	4	1	5
6	4	2	10
7	5	2	9

From the outputs we reached, the market could be actually separated into two parts. Segment 1 (subcompact), Segment 2 (compact), and Segment 3 (intermediate) constitute a single relevant market which means that the brands in these segments are substitutes for each other, whereas Segment 4 (standard) and Segment 5 (luxury) constitute another relevant market that same conditions apply.

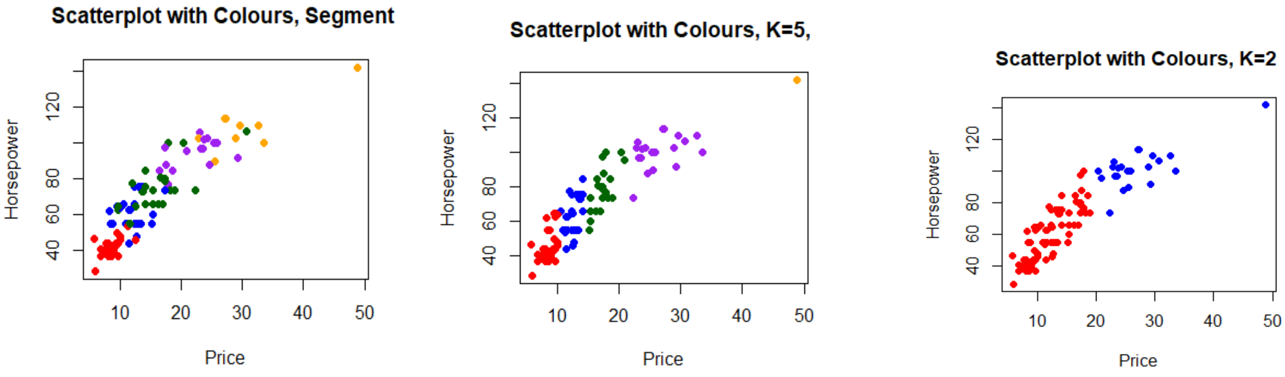
Profoundly, I detected noticeable traits for each the cluster (market) as shown below.

Graph 6: Characteristics of Each Market (By Clusters)- K-Means



Graph 6 articulates remarkable facts that the main aspects of separating markets are consumption of fuel, horsepower, price weight and width and among them, the price and horsepower must be underlined regarding the magnitude of dispersion.

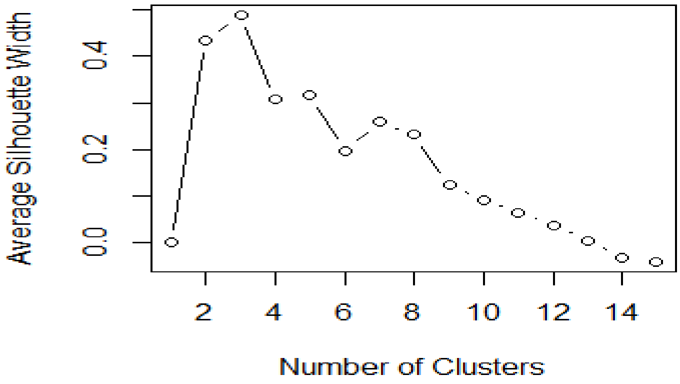
Graph 7: Interlink Between the Price and the Horsepower



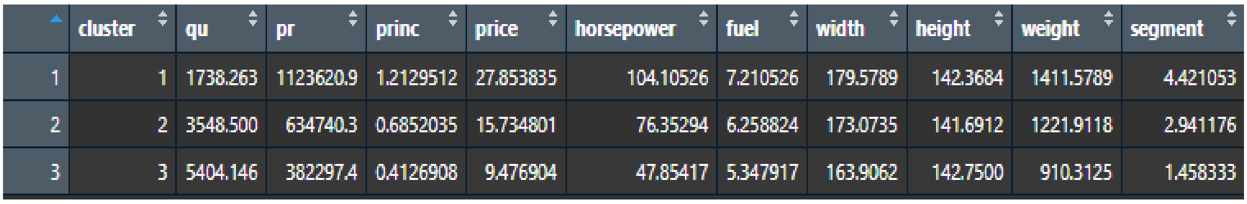
C. Agglomerative Hierarchical Clustering with Gower Distance

This method starts by treating each data point as a single cluster and then successively merges clusters until all points form a single group. Firstly, I calculated the Gower distance matrix that represent the pairwise dissimilarities between each point in my dataset. Then perform agglomerative clustering using with Ward algorithm. To choose optimal cluster number I checked the highest Silhoutte Width, so I can then cut the tree (dendrogram) at a certain height.

Graph 8: Avarage Silhoutte Width

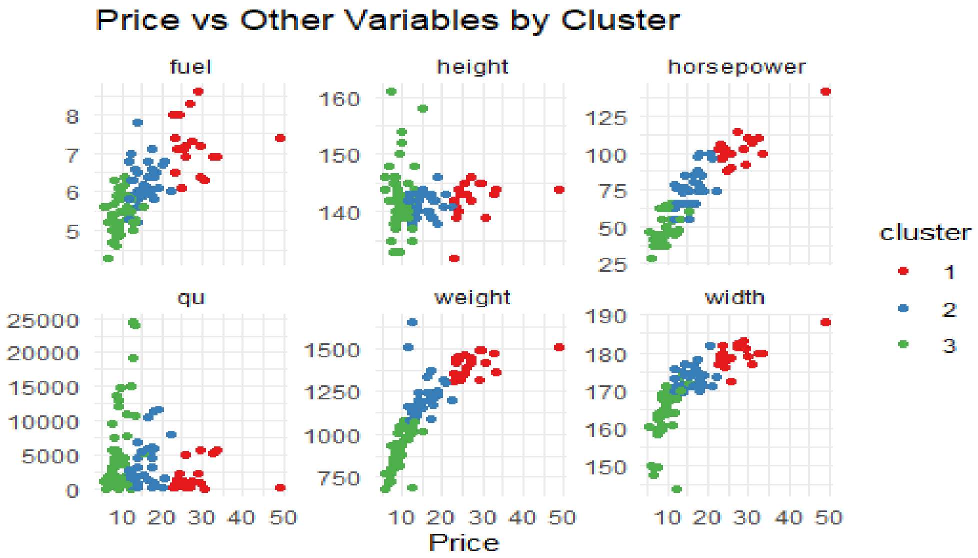


With respect to Graph 8, I selected K=3.



	cluster	qu	pr	princ	price	horsepower	fuel	width	height	weight	segment
1	1	1738.263	1123620.9	1.2129512	27.853835	104.10526	7.210526	179.5789	142.3684	1411.5789	4.421053
2	2	3548.500	634740.3	0.6852035	15.734801	76.35294	6.258824	173.0735	141.6912	1221.9118	2.941176
3	3	5404.146	382297.4	0.4126908	9.476904	47.85417	5.347917	163.9062	142.7500	910.3125	1.458333

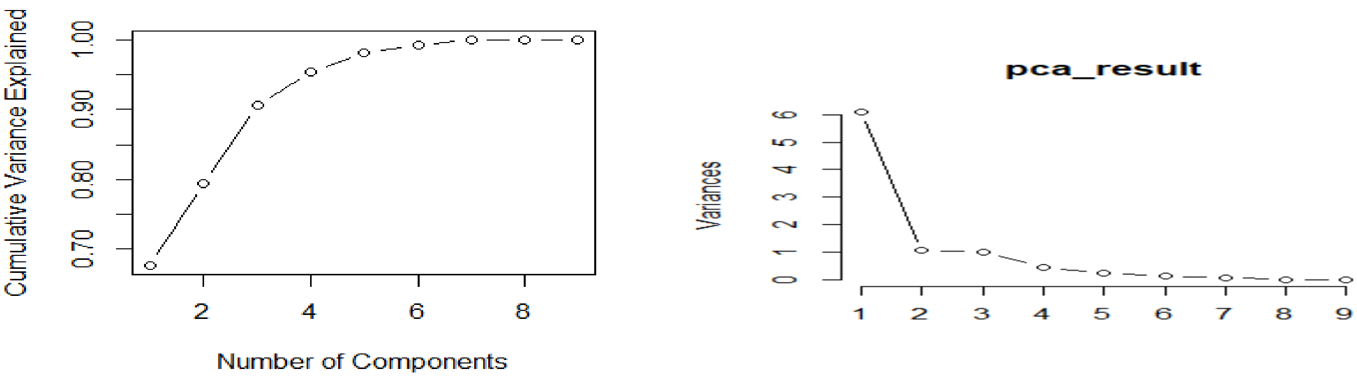
According to the results of the model, one can infer that Segment 1 and 2, Segment 3 and 4, Segment 4-5 can be evaluated as a separate relevant markets and except for height every parameters tends to change as the cluster at hand differ.



D. PCA-GMM

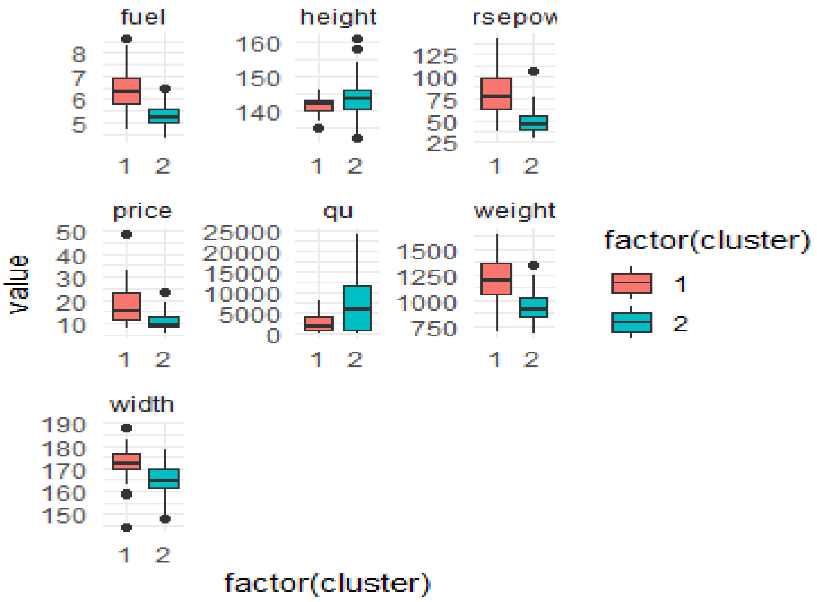
To reduce noise and improve clustering performance I adopted PCA approach for my GMM model. By reducing dimensions through PCA may lead to better clustering performance and by simplifying the data structure, make it easier for GMM to model the data with Gaussian distributions. The number of principal components is the key to capture to maximum variation (explained variance). I went with the benchmark that based on variance explained criterion, but with a more specific target, 95% variance explained.

Graph 9: Number of Component Selection

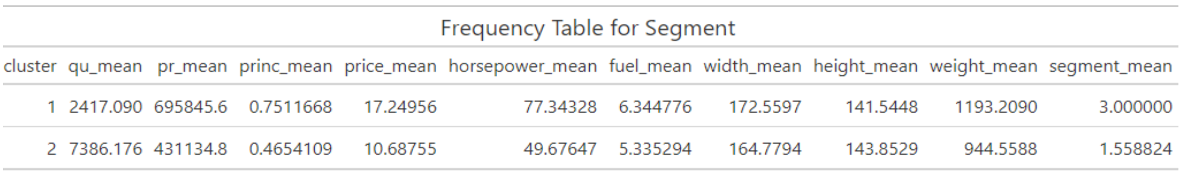


After choosing optimal component number as 3, the GMM define, like K-Means, two separate clusters.

Graph 10: Characteristics of Each Market (By Clusters)-GMM



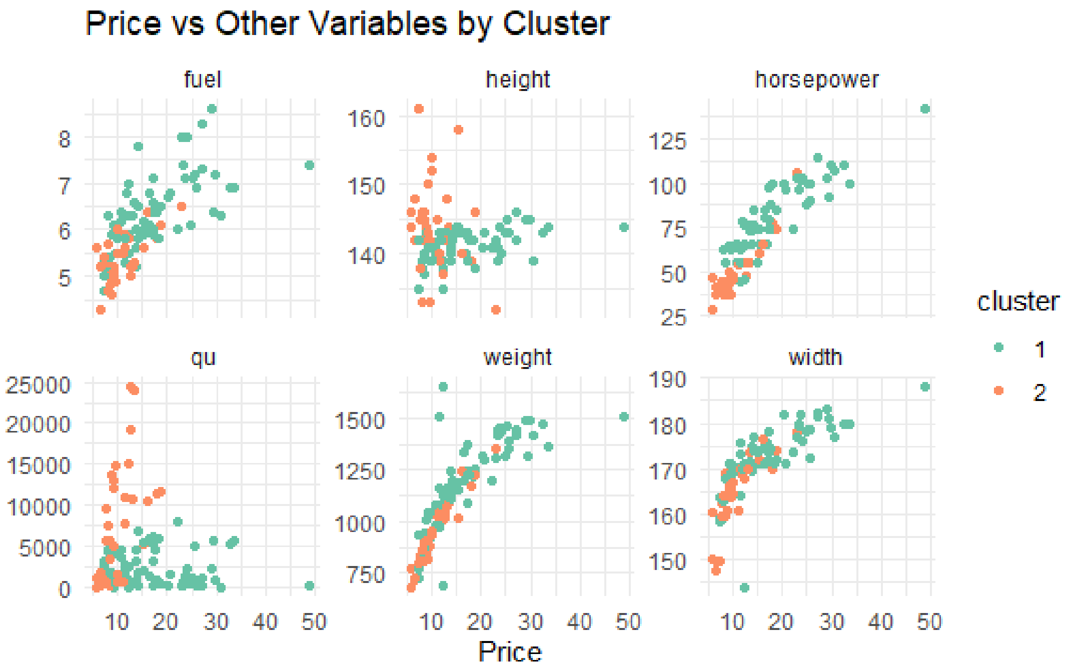
,



Frequency Table for Segment

cluster	qu_mean	pr_mean	princ_mean	price_mean	horsepower_mean	fuel_mean	width_mean	height_mean	weight_mean	segment_mean
1	2417.090	695845.6	0.7511668	17.24956	77.34328	6.344776	172.5597	141.5448	1193.2090	3.000000
2	7386.176	431134.8	0.4654109	10.68755	49.67647	5.335294	164.7794	143.8529	944.5588	1.558824

Graph 11: Price vs Other Variables (By Clusters)



Unlike the others, patterns are not easy to detect in Graph 11 with respect to variable's relations with price.

In conclusion, the clustering algorithms can be useful and practical tools to identify relevant markets, test claimed relevant markets, or fortify or rebut established relevant market boundaries in competition law cases. However, it must be boldly underlined that these methods are more supplementary and auxiliary and certainly less conclusive and decisive. As a last mark, it will not be unfolded if one presumes that in the ever-changing digital market atmosphere, the potential added value stems from ML-based quantitative techniques that keep skyrocketing.