Group Coursework Submission Form

Specialist Masters Programme

Please list all names of group members: (Surname, first name) 1. Bian, Yu 2. Huang, Ruiqi 3. Liu, Ching-Wei	4. Liu, Mengjie 5. Wang, Zhaofei 6. 7. **GROUP NUMBER:** **5**
MSc in: Business analytics	
Module Code: SMM634	
Module Title: Analytics Methods for Business	
Lecturer: Rosalba Radice	**Submission Date:** 11/3/2019

Declaration:

By submitting this work, we declare that this work is entirely our own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the coursework instructions and any other relevant programme and module documentation. In submitting this work we acknowledge that we have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. We also acknowledge that this work will be subject to a variety of checks for academic misconduct.

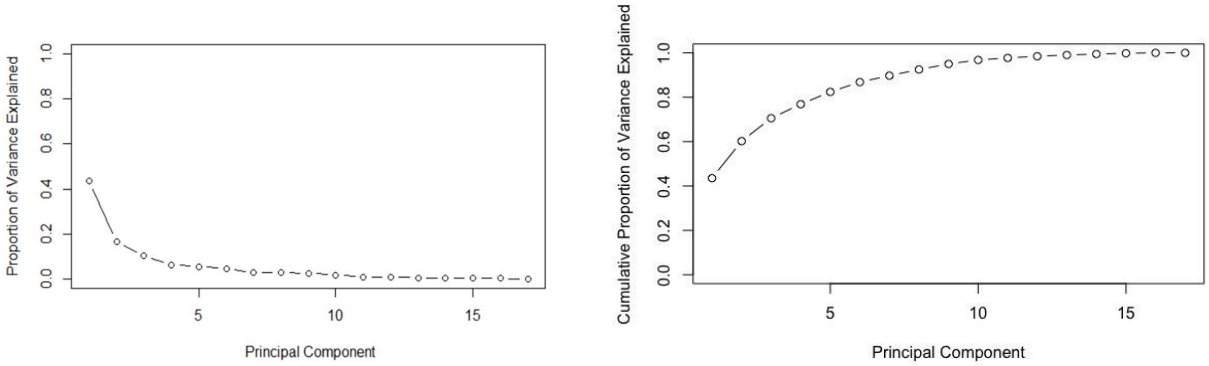
We acknowledge that work submitted late without a granted extension will be subject to penalties, as outlined in the Programme Handbook. Penalties will be applied for a maximum of five days lateness, after which a mark of zero will be awarded.

Marker's Comments (if not being marked on-line):

Deduction for Late Submission: | **Final Mark:** **%**

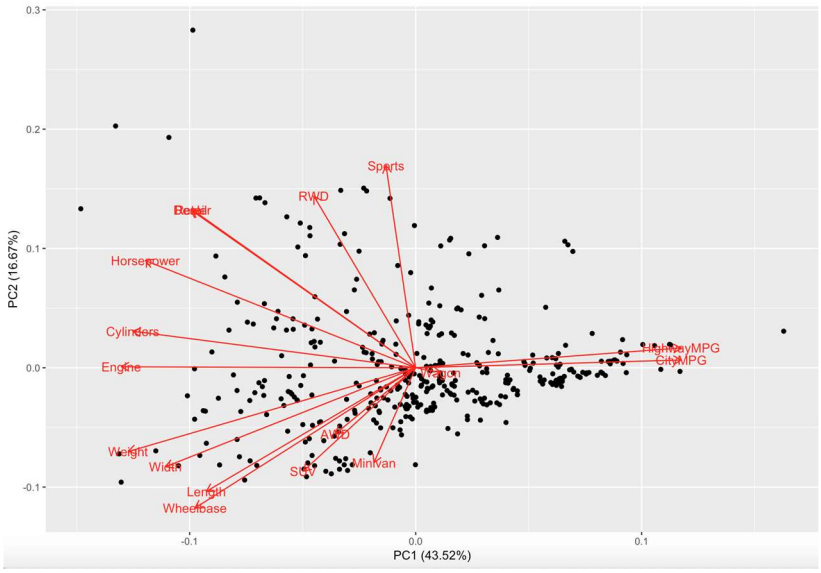
Question 1

```
Importance of components:
                          PC1    PC2    PC3    PC4     PC5     PC6     PC7     PC8     PC9    PC10
Standard deviation     2.7201 1.6836 1.3261 1.0357 0.97181 0.86694 0.70070 0.68760 0.65079 0.55512
Proportion of Variance 0.4352 0.1667 0.1034 0.0631 0.05555 0.04421 0.02888 0.02781 0.02491 0.01813
Cumulative Proportion  0.4352 0.6020 0.7054 0.7685 0.82407 0.86828 0.89716 0.92497 0.94988 0.96801
                          PC11    PC12    PC13    PC14    PC15    PC16    PC17
Standard deviation     0.3911 0.35881 0.31350 0.26767 0.24228 0.18088 0.02771
Proportion of Variance 0.0090 0.00757 0.00578 0.00421 0.00345 0.00192 0.00005
Cumulative Proportion  0.9770 0.98458 0.99036 0.99458 0.99803 0.99995 1.00000
```



To decide how many principles to use, we are interested in knowing the proportion of variance explained (PVE) by each principal component and cumulative proportion of variance explained. Here, we would like to use the smallest number of principal components to get a good understanding of the data.

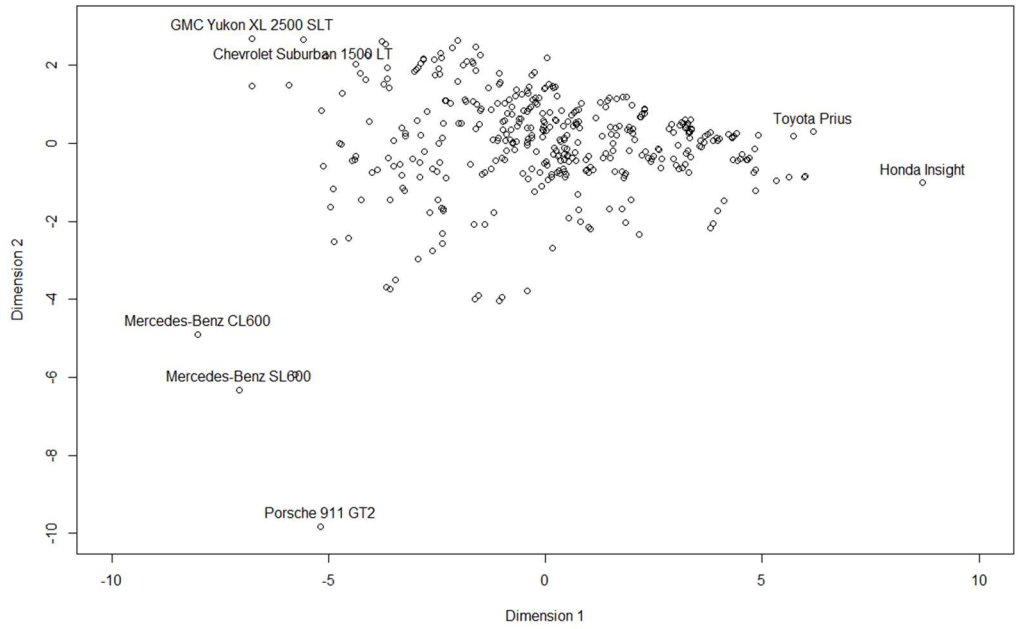
The first principle component explains 43.52% of the variance in the data, and the second principal component explains 16.67% of the variance, etc. The first five PCs together explain 82.41% of the variance in the data, which provide sufficient information about the data.



To interpret the first principle component, all the important features such as MPG, engine and cylinders reflect the fuel efficiency of a vehicle; the second principal component otherwise, reflects that whether the vehicle is a sports car and its outdoor performance contributes massively to the total variance, which is illustrated by features such as sport, RWD and wheelbase.

Question 2

The configuration plot of the 2-dimensions MDS Result:



D1:

	Sports	SUV	Wagon	Minivan	Pickup	AWD	RWD	Retail	Dealer	Engine	Cylinders	Horsepower	CityMPG
Mercedes-Benz CL600	0	0	0	0	0	0	1	4.8258851	4.9806409	2.339127	4.189350	3.964480	-1.389624
Mercedes-Benz SL600	1	0	0	0	0	0	1	4.7371635	4.8831055	2.339127	4.189350	3.964480	-1.389624
Honda Insight	0	0	0	0	0	0	0	-0.7159160	-0.6999345	-1.111480	-1.850181	-2.013077	7.541777
Toyota Prius	0	0	0	0	0	0	0	-0.6449387	-0.6432344	-1.604424	-1.179122	-1.486482	7.351747

	HighwayMPG	Weight	Wheelbase	Length	Width
Mercedes-Benz CL600	-1.466210	1.3322065	0.9578865	0.8338692	0.5116823
Mercedes-Benz SL600	-1.466210	1.2698839	-0.8765737	-0.4503128	0.2147992
Honda Insight	6.873031	-2.3830719	-1.7232477	-2.2632757	-1.2696166
Toyota Prius	4.211571	-0.9099916	-0.1710121	-0.7524733	-0.9727334

D2:

	Sports	SUV	Wagon	Minivan	Pickup	AWD	RWD	Retail	Dealer	Engine	Cylinders	Horsepower
GMC Yukon XL 2500 SLT	0	1	0	0	0	1	0	0.6607889	0.5638369	2.8320709	1.5051142	1.573457
Chevrolet Suburban 1500 LT	0	1	0	0	0	0	0	0.4818249	0.3899937	2.1419495	1.5051142	1.146489
Porsche 911 GT2	1	0	0	0	0	0	1	8.0728400	7.9949675	0.4659403	0.1629962	3.736764
Mercedes-Benz SL600	1	0	0	0	0	0	1	4.7371635	4.8831055	2.3391271	4.1893501	3.964480

	CityMPG	HighwayMPG	Weight	Wheelbase	Length	Width
GMC Yukon XL 2500 SLT	-1.3896236	-1.8210712	3.6834693	3.2156837	2.5712920	2.2929812
Chevrolet Suburban 1500 LT	-1.1995938	-1.6436405	2.0035912	3.2156837	2.5712920	2.2929812
Porsche 911 GT2	-0.6295044	-0.5790566	-0.5686336	-2.0054723	-0.7524733	0.2147992
Mercedes-Benz SL600	-1.3896236	-1.4662099	1.2698839	-0.8765737	-0.4503128	0.2147992

To interpret the configuration, Models such as Mercedes SL600 and CL600 are one extreme of the first dimension while Toyota Prius and Honda Insight are the other extremes. Therefore, the first dimension must measure the most distinctive features between the two extreme groups. The detailed information of all four models is extracted in D1. We could conclude that the two extremes are distinctive in retail price, powers (engine, cylinders, horsepower) and MPGs. Mercedes models generally have higher prices, stronger powers and lower MPGs than the Toyota and Honda models. Therefore, the first dimension is likely to reflect the fuel efficiency, for that, an economic car model such as Toyota would have a lower retail price, lower MPG but will compromise some of the powers as a contrast to luxury brands such as Mercedes.

We extract several extreme models for the second dimension and put their information for comparison as well, displayed as D2. The most distinctive features now are price, horsepower, wheelbase, length and width. GMC Yukon and Chevrolet Suburban have a lower price, lower horsepower, higher wheelbase, larger

length and width than Mercedes and Porsche. Those distinctive features reveal the measure of styles, whereas GMC Yukon and Chevrolet Suburban might be more of a business style and Mercedes SL600 and Porsche 911 are more of a sports car style.

Overall the conclusion is very similar to that of PCA, which is because where squared Euclidean distances have been calculated from a data matrix, classical MDS gives the PCA solutions exactly.

Question3

(a)

In table 3, the first factor represents some overall social mobility of the family members. We can see that the loading values for the third generation's social mobility are the highest, followed by the second generation's and the first generation's. However, since the differences among these values are not very clear, we can apply rotations to find another set of loadings that can be more interpretable. For the second factor, it is obvious that the loading values of the occupation status are relatively higher, which indicates that factor 2 can be interpreted as an "occupation status" factor. In addition, it is noticeable that the second one contrasts further education and qualification variables. In the end, the third factor loads significantly on wife's further education and qualifications and thus can be regarded as a "wife's social mobility" factor. To discuss the third one in more detail, we can notice that it loads positively on the male's social mobility but the values are lower than the female's, while the loadings of the firstborn's social mobility are all negative values.

(b)

In an orthogonal (varimax) rotation, we hold the assumption that variables are uncorrelated.
In table 4, the first factor represents the occupational status as the few large loadings are all occupationally related. The second factor measures the third generation's social mobility as x8, x9 and x10 are clearly the dominant loadings among all. The third factor represents the wife's social mobility as the wife's further education and qualification have the highest loading values.

An oblique rotation leads to correlated factors. In table 5, the first factor loads strongly on the third generation's social mobility and therefore might be interpreted as a "third generation's social mobility" factor. The second factor loads on the occupation status and thus might be regarded as an "occupation status" factor, while for the third factor, wife's qualification and wife's further education have very dominant loadings such that the third factor should be regarded as "wife's social mobility" factor.

When we compare this oblimin result to the varimax rotation above, we can find that the loadings are similar enough that there is no substantial change in how we would interpret the factors, so both of the rotations result in the same factors: "third generation's social mobility", "occupation status", and "wife's social mobility".

Question 4

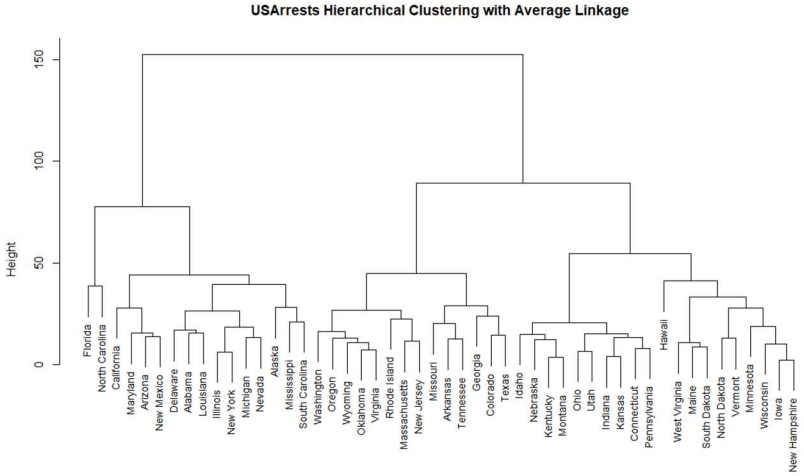
(a)

The Euclidean distance matrix of the four variables:

	Murder	Assault	UrbanPop	Rape
Murder	0.0000	1280.9029	421.3440	109.6001
Assault	1280.9029	0.0000	934.6331	1188.1307
UrbanPop	421.3440	934.6331	0.0000	327.5055
Rape	109.6001	1188.1307	327.5055	0.0000

(b)

The result of the hierarchical clustering analysis with average linkage:



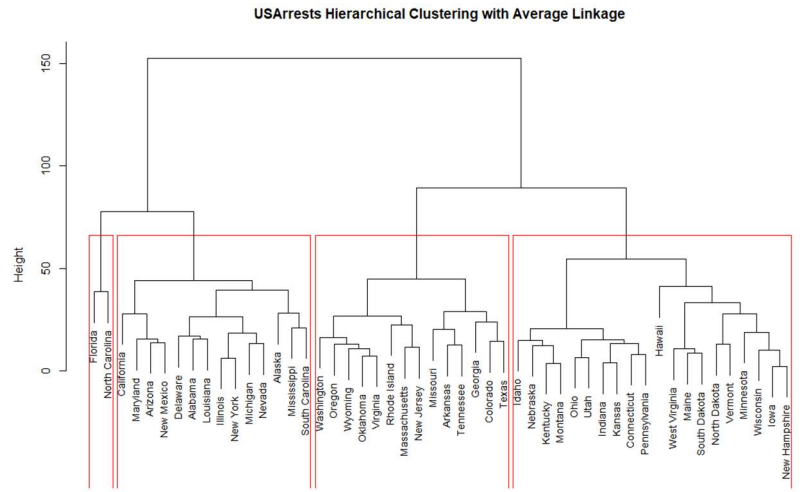
(c)

In this procedure, after we apply the cutree() function, we can get the classification of the 50 US states.

The classes of the US states in 4 clusters:

Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware
1	1	1	2	1	2	3	1
Florida	Georgia	Hawaii	Idaho	Illinois	Indiana	Iowa	Kansas
4	2	3	3	1	3	3	3
Kentucky	Louisiana	Maine	Maryland	Massachusetts	Michigan	Minnesota	Mississippi
3	1	3	1	2	1	3	1
Missouri	Montana	Nebraska	Nevada	New Hampshire	New Jersey	New Mexico	New York
2	3	3	1	3	2	1	1
North Carolina	North Dakota	Ohio	Oklahoma	Oregon	Pennsylvania	Rhode Island	South Carolina
4	3	3	2	2	3	2	1
South Dakota	Tennessee	Texas	Utah	Vermont	Virginia	Washington	West Virginia
3	2	2	3	3	2	2	3
Wisconsin	Wyoming						
3	2						

The result after cutting the dendrogram into 4 clusters:



(d)

The US states grouped in cluster 1:

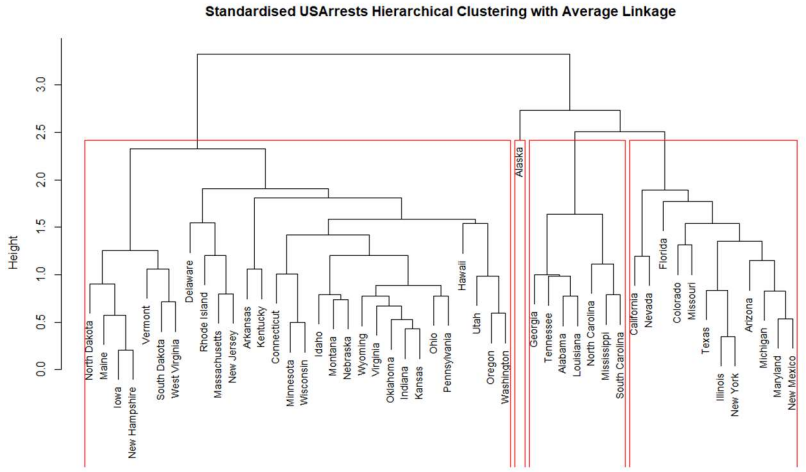
Alabama	Alaska	Arizona	California	Delaware	Illinois	Louisiana	Maryland
1	1	1	1	1	1	1	1
Michigan	Mississippi	Nevada	New Mexico	New York	South Carolina		
1	1	1	1	1	1		

(e)

The classes of the US states in 4 clusters (after scaled):

Alabama	Alaska	Arizona	Arkansas	California	Colorado	Connecticut	Delaware
1	2	3	4	3	3	4	4
Florida	Georgia	Hawaii	Idaho	Illinois	Indiana	Iowa	Kansas
3	1	4	4	3	4	4	4
Kentucky	Louisiana	Maine	Maryland	Massachusetts	Michigan	Minnesota	Mississippi
4	1	4	3	4	3	4	1
Missouri	Montana	Nebraska	Nevada	New Hampshire	New Jersey	New Mexico	New York
3	4	4	3	4	4	3	3
North Carolina	North Dakota	Ohio	Oklahoma	Oregon	Pennsylvania	Rhode Island	South Carolina
1	4	4	4	4	4	4	1
South Dakota	Tennessee	Texas	Utah	Vermont	Virginia	Washington	West Virginia
4	1	3	4	4	4	4	4
Wisconsin	Wyoming						
4	4						

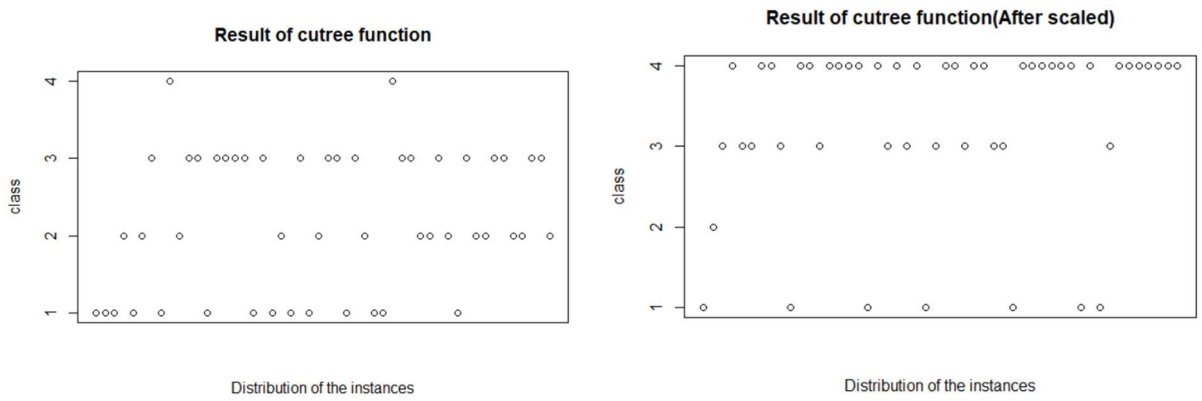
The result after cutting the dendrogram into 4 clusters (after scaled):



The US states grouped in cluster 1 (after scaled):

Alabama	Georgia	Louisiana	Mississippi	North Carolina	South Carolina	Tennessee
1	1	1	1	1	1	1

The distribution of the instances before / after scaled:



After standardising the data, it is obvious that the classifications of the 50 US states are changed. Moreover, when we look the states grouped in cluster 1 in more detail, it is also noticeable that the number of the states in this group decreases after scaling. Finally, if we plot all the clusters, we could see that the distribution of the instances in each class changes dramatically after scaling.