Question 2

Akhilesh Chauhan(Roll No. 170070)

March 2020

1 PC projection

I first normalised the data due to the difference in variances in the data and then applied PCA on the normalised data. The PC projection comes out to be:

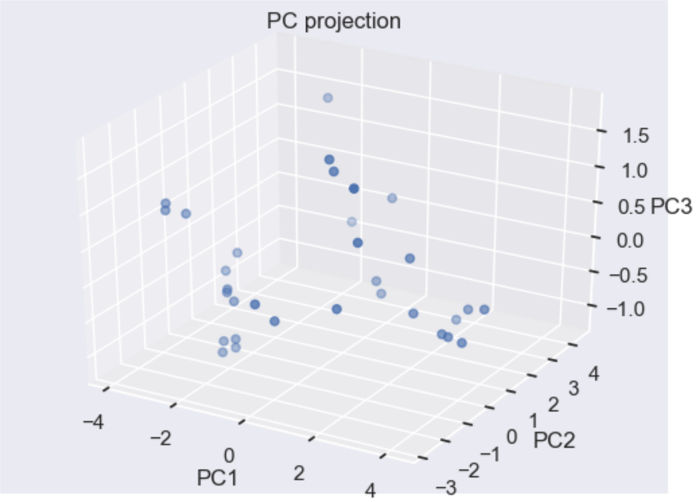

Figure 1: PC projection in the PC plane

2 Percentage of Total Variance

The percentage of total variance explained by 3 PCs is 92.77 percent.

3 Clustering in PC plane

To decide the value of K in Kmeans clustering, I plotted the inertia for all the values of K in the interval [1,10] and looked for the elbow formation in the plot. The elbow is formed at $K = 4$.

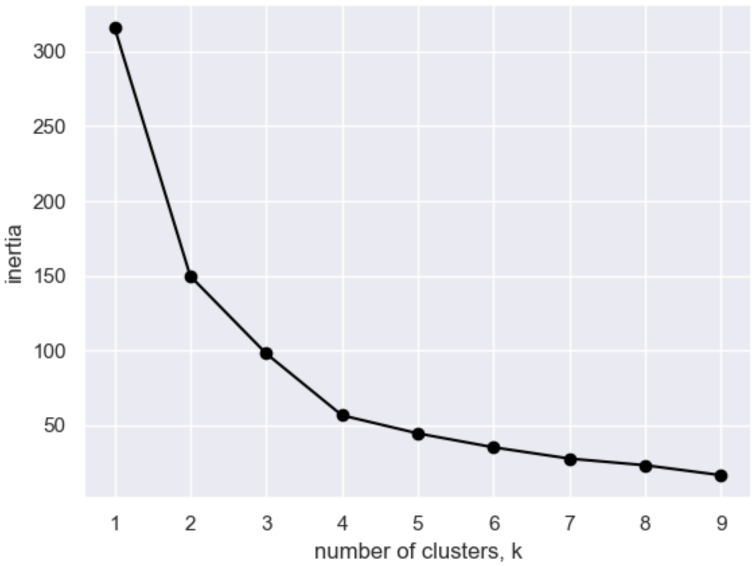

Figure 2: Inertia versus the K value

The Kmeans clusters in the PC projection are:

['Hornet Sportabout', 'Duster 360', 'Merc 450SE', 'Merc 450SL', 'Merc 450SLC', 'Cadillac Fleetwood', 'Lincoln Continental', 'Chrysler Imperial', 'Dodge Challenger', 'AMC Javelin', 'Camaro Z28', 'Pontiac Firebird']

['Mazda RX4', 'Mazda RX4 Wag', 'Ford Pantera L', 'Ferrari Dino', 'Maserati Bora']

['Datsun 710', 'Fiat 128', 'Honda Civic', 'Toyota Corolla', 'Fiat X1-9', 'Porsche 914-2', 'Lotus Europa', 'Volvo 142E']

['Hornet 4 Drive', 'Valiant', 'Merc 240D', 'Merc 230', 'Merc 280', 'Merc 280C', 'Toyota Corona']

The clusters formed almost all have the same number of cylinders, shape of cylinders, transmission and number of forward gears. They have very similar features.

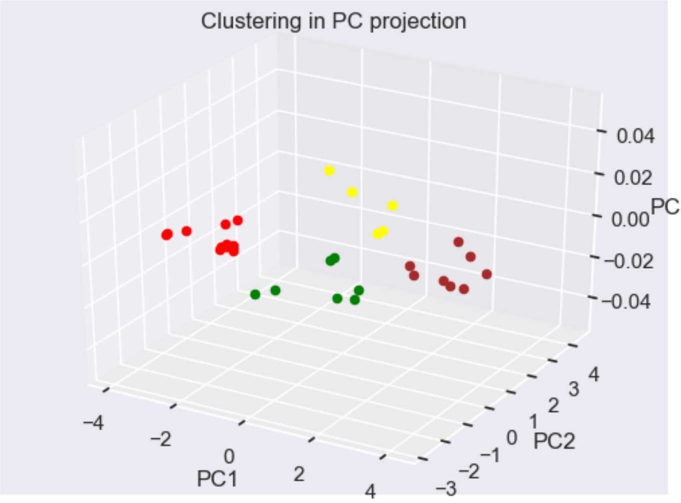

Figure 3: Clusters in the PC plane

4 Outliers

In order to find the outliers, I plotted PC1 vs PC2 for the data as PC1 and PC2 explain a huge amount of variance of the data and I looked for points which are very far away in the plot.

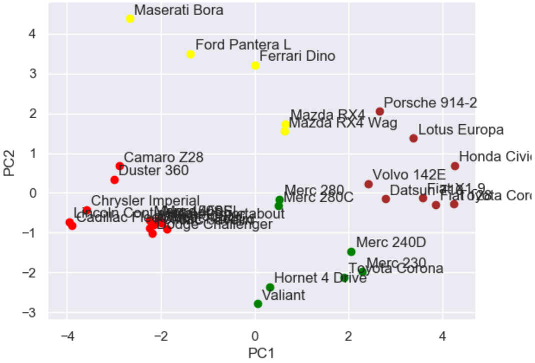

Figure 4: Detection of outliers

Hence as clear from the above graph, the outliers are ”Maseratio Bora” , ”Ford Pantera L” and ”Ferrari Dino”

5 Clustering of data

5.1 Clustering using KMeans

I again formed an inertia vs K graph(for the actual data) for K in the interval [1,10]. I found the elbow to be forming at $K = 4$. So,I chose the value of K to be 4.

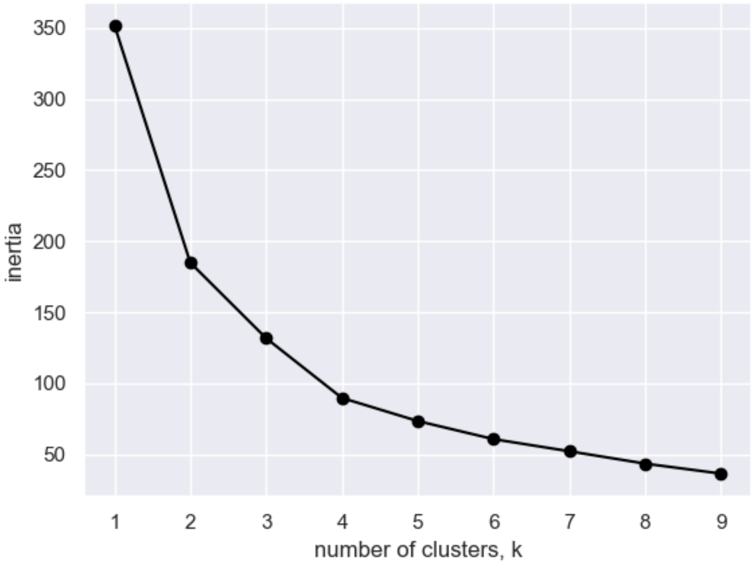

Figure 5: Inertia vs K for the data

I found the clusters to be:

[’Hornet Sportabout’, ’Duster 360’, ’Merc 450SE’, ’Merc 450SL’, ’Merc 450SLC’, ’Cadillac Fleetwood’, ’Lincoln Continental’, ’Chrysler Imperial’, ’Dodge Challenger’, ’AMC Javelin’, ’Camaro Z28’, ’Pontiac Firebird’]

[’Mazda RX4’, ’Mazda RX4 Wag’, ’Ford Pantera L’, ’Ferrari Dino’, ’Maserati Bora’]

[’Hornet 4 Drive’, ’Valiant’, ’Merc 240D’, ’Merc 230’, ’Merc 280’, ’Merc 280C’, ’Toyota Corona’]

[’Datsun 710’, ’Fiat 128’, ’Honda Civic’, ’Toyota Corolla’, ’Fiat X1-9’, ’Porsche 914-2’, ’Lotus Europa’, ’Volvo 142E’]

This is exactly the same clusters that were obtained from the KMeans clustering on the PCs.

5.2 Hierarchical Clustering

Firstly, I plotted the dendogram(using single linkage AHC) of the data and took the resolution to be at level = 7.5.

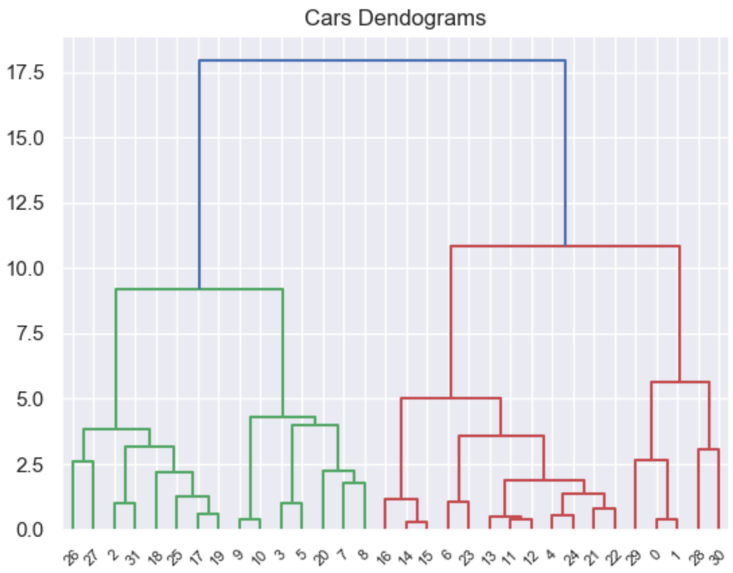

Figure 6: Dendogram for the data(Single linkage AHC)

The clusters I found after that are: ['Mazda RX4', 'Mazda RX4 Wag', 'Ford Pantera L', 'Ferrari Dino', 'Maserati Bora']

['Hornet 4 Drive', 'Valiant', 'Merc 240D', 'Merc 230', 'Merc 280', 'Merc 280C', 'Toyota Corona']

['Hornet Sportabout', 'Duster 360', 'Merc 450SE', 'Merc 450SL', 'Merc 450SLC', 'Cadillac Fleetwood', 'Lincoln Continental', 'Chrysler Imperial', 'Dodge Challenger', 'AMC Javelin', 'Camaro Z28', 'Pontiac Firebird']

['Datsun 710', 'Fiat 128', 'Honda Civic', 'Toyota Corolla', 'Fiat X1-9', 'Porsche 914-2', 'Lotus Europa', 'Volvo 142E']

This is also exactly the same clustering as the one obtained using PCs.