Assignment 2

Segmentation of Ford Ka Consumers

Paris Ngow

BU425

Salar Ghamat

December 10th, 2022

Table of Contents

Executive Summary

Ford wants to determine the optimal consumer segment to position its new car model, the Ka, towards. They are considering two kinds of sample datasets for this segmentation task, one that contains demographic data and another that contains psychographic data. These datasets were combined after data pre-processing and evaluating the correlation of the features. With this combined dataset, two types of analyses were performed to aid Ford in its segmentation task.

First, a predictive analysis was conducted to identify which consumers would choose the Ka as one of their top three choices within the small car market (deemed Choosers). The results from this model would help the company evaluate the size of their potential market, and based on the features used in the model, develop a profile of who these Choosers are. The logistic regression model used in this analysis indicated that most consumers would choose the Ka than not. It also specified that Choosers of the Ka would likely be older, female consumers who are married and have children. Their income levels would be in the middle to the low end of the spectrum and the Ka would not be their first car purchase.

The company also executed an exploratory analysis to determine the possible segmentations of consumers that make up the small car market. A K-means model was run using four clusters after determining that this number was the point where the within-cluster sum of squares began to diminish. The four clusters created by this model allowed for the identification of four different potential consumer segments in the market, mainly distinguishable by their psychographic characteristics.

However, it would be impractical to try and target all four segments, so Ford should choose one to focus on. The largest and most logical cluster to target is the group of consumers who are environmentally conscious and care about the utilitarian aspects of their cars. Furthermore, the results of the clustering agreed with the supervised model’s profiling of consumers in that they were expected to be 30-40 years old, have a middle-class income, and not be first-time car purchasers. Thus, with this psychographic and demographic understanding of the ideal target consumer, Ford should be able to focus its marketing efforts and create a campaign that appeals to these consumers and what they like.

Problem Definition

Ford wants to use predictive and exploratory models to segment its consumer base and identify key target consumer groups for their newest car model, the Ka. From these classes of groups, the company also wants to determine how to position itself to market to the most promising target group.

Description of Data

Two sample datasets were provided about the Ka model's potential consumers. The first contains demographic information about consumers. There are 10 features and 250 records in this dataset. As seen in the correlation matrix in Appendix A, the Age/AgeCat and Children/ChildCat variables are highly correlated which is not ideal when implementing any analytical model. So, only one of each will be kept. The marital status and income category were converted to factors, and the Preference (target) variable was converted to binary by arbitrarily assigning records with the value 3 (In middle 4) to class 1 or 2.

The second was a dataset with psychographic information derived from consumers' responses to a questionnaire inquiring about their attitudes toward small cars and related topics. The values are the consumer's attitudes towards the topics on a scale of 1-7. There are 62 features and 250 records in this dataset. The dataset was stripped of the highly correlated variables, which can also be seen as questions that ask about similar topics. This left 31 variables in the dataset. Principal component analysis was not conducted because using principal components in the model would make it difficult to interpret the actual features used, which is not ideal as we want to explain the attributes that impact the results of the model.

Then, these datasets were combined to create a comprehensive set with all variables that describe a consumer. This set will be used to train and evaluate the models. Note that both datasets did not contain any missing data. This dataset was split into training and testing sets on an 80:20 basis. Furthermore, a normalized dataset was created for the sake of better performance on the clustering model.

Analysis

For the predictive analysis, a logistic regression model was used to determine the Choosers (those who rank the Ka in their top 3 small cars) and Non-Choosers of the Ka. The goal of this analysis was to predict who would choose the Ka and whether the number of consumers was significant. Using a threshold of 0.5, the model predicted that more consumers would be Choosers of the Ka in the train and test sets. These results are promising regarding the potential response from consumers. But, as seen in the confusion matrix in Appendix B, the model had an accuracy of 64%, which puts the confidence we have in the prediction into question. The AUC for the ROC curve was also 0.634, which indicates that the model was not very good at distinguishing the classes. Still, using the inverse of the coefficients from the model in Appendix C to establish a potential profile of Choosers, they are likely to be older, married, female, have children, and have an upper-middle-class income. But these features were not deemed to be significant in making a prediction. The most significant predictor was whether the Ka would be their first car purchased and the model specified that Choosers would likely not be first-time purchasers.

To conduct the exploratory analysis, the dataset was experimented with to see how it could be segmented using an unsupervised model. In this analysis, the goal was to attempt to create new segments of consumers and describe the characteristics of the groupings. The K-means unsupervised model was used to do this clustering. The model was not exposed to the target labels in the dataset to allow it to segment the data in the way it determined to be optimal. To determine the best number of clusters, the elbow method will be used to evaluate the cluster size where the within-cluster sum of squares (WCSS) begins to diminish. As seen in Appendix D, the elbow indicates that the optimal number of clusters was 4.

Evaluating the clusters, we can see that they are mainly differentiated based on psychographic information. The demographic data averages around the same values for each cluster, indicating that consumers cannot be separated meaningfully by these attributes. By matching the psychographic variables to the actual questions asked, we can derive some insights into the consumers' values and concerns. From the un-normalized model results in Appendix E, consumers are unanimous about the utilitarian aspects of the car such as reliability and performance. But the segments can be differentiated by specific characteristics, as seen in Appendix F. To summarize, the first consumer group sees a car as something meaningful to their external appearance and will care for it through maintenance. The second group is environmentally conscious, and they care about the practical features of their car. The third cluster only sees cars as a means of transportation and has negative opinions about manufacturers. Lastly, the fourth group is indifferent about their cars and will likely choose a vehicle based on what is popular.

However, we should take the model's clustering with a grain of salt. Looking at the WCSS values in Appendix G, we can see that the values are large which means that this optimal clustering was not very succinct and the data points in each cluster were varied. Furthermore, the R^2 value was 29.6%, which signified that the segmentation of the data was only able to explain a small number of the model's errors.

Recommendation & Action Plan

It would not be ideal to try and target all four consumer segments that were identified since Ford only has limited resources to dedicate to marketing. Therefore, the company should target the largest segment made up of consumers who are concerned about the environment and care about the utilitarian features of their vehicles (See Appendices H & I). Thus, Ford should position the Ka as a car boasting practical benefits such as safety, as the Ka would have imported these benefits from the Fiesta chassis. In addition, Ford should advertise efficient fuel consumption, however, they may have to redesign the Ka to have the fuel consumption of the Fiesta 1.8D Navy to do this honestly. Marketing the Ka as an environmentally friendly car would also appeal to consumer concerns related to pollution and fuel consumption that emerged in the early 90s. If we were also to consider the demographic attributes, both models indicated that females around 35 years old with a middle-class income who are not first-time car purchasers would be the ideal consumers to tailor their advertisements towards.

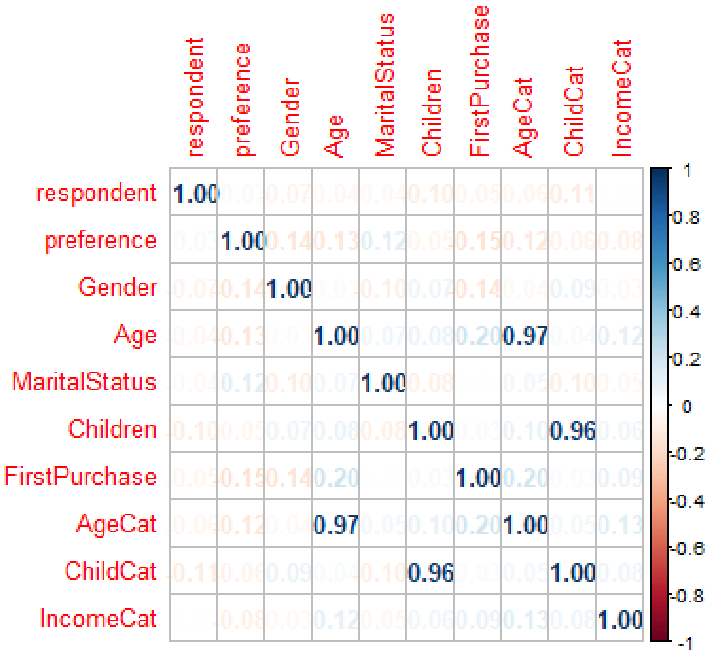
Note that although this segment of consumers (2) comes close to another group (4) in Appendix H, that group is indifferent and does not have strong attitudes about the small car market. They are also easily swayed by what is trending. So, if Ford can create a strong marketing campaign that expresses the benefits of the Ka and it catches on with the public, consumers in cluster 4 may also respond positively.

Conclusion

With the prediction that there will be a market for the Ka, Ford should market the car model to older, experienced car owners who are conscious about the environment and want a capable vehicle.

Appendices

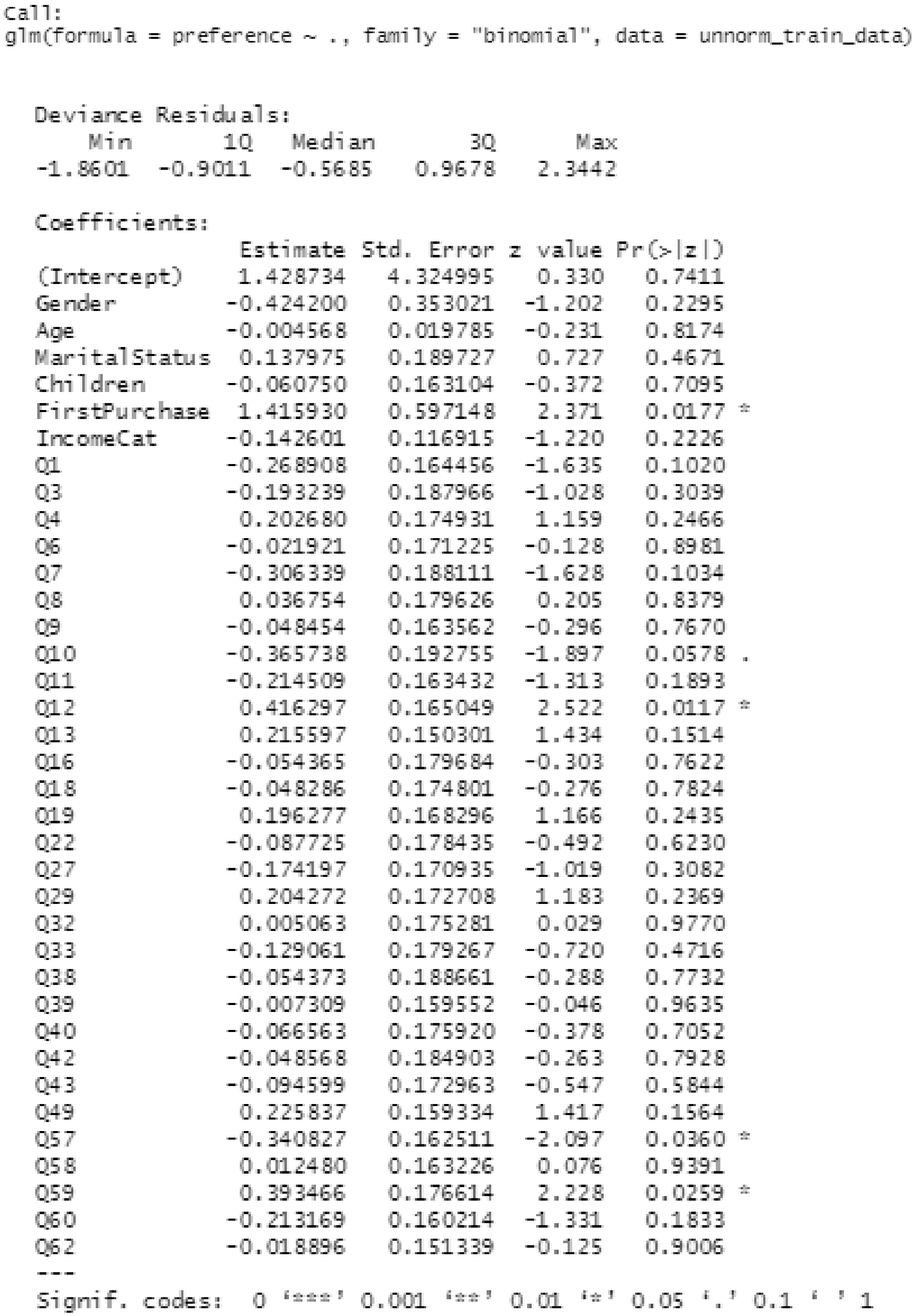
Appendix A – Demographic Data Correlation Matrix



Appendix B – Logistic Regression Model Results

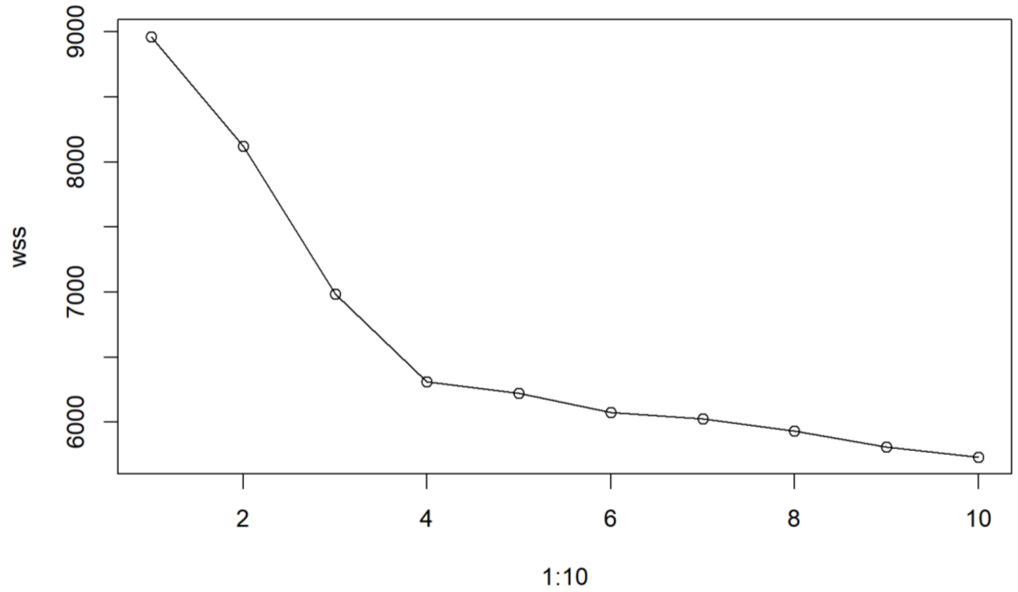
```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 19  8
         1 10 13
                                          
               Accuracy : 0.64            
                 95% CI : (0.4919, 0.7708)
    No Information Rate : 0.58            
    P-Value [Acc > NIR] : 0.2383          
                                          
                  Kappa : 0.2707          
                                          
 Mcnemar's Test P-Value : 0.8137          
                                          
            Sensitivity : 0.6552          
            Specificity : 0.6190          
         Pos Pred Value : 0.7037          
         Neg Pred Value : 0.5652          
             Prevalence : 0.5800          
         Detection Rate : 0.3800          
   Detection Prevalence : 0.5400          
      Balanced Accuracy : 0.6371          
                                          
       'Positive' Class : 0               
```

Appendix C - Logistic Regression Model

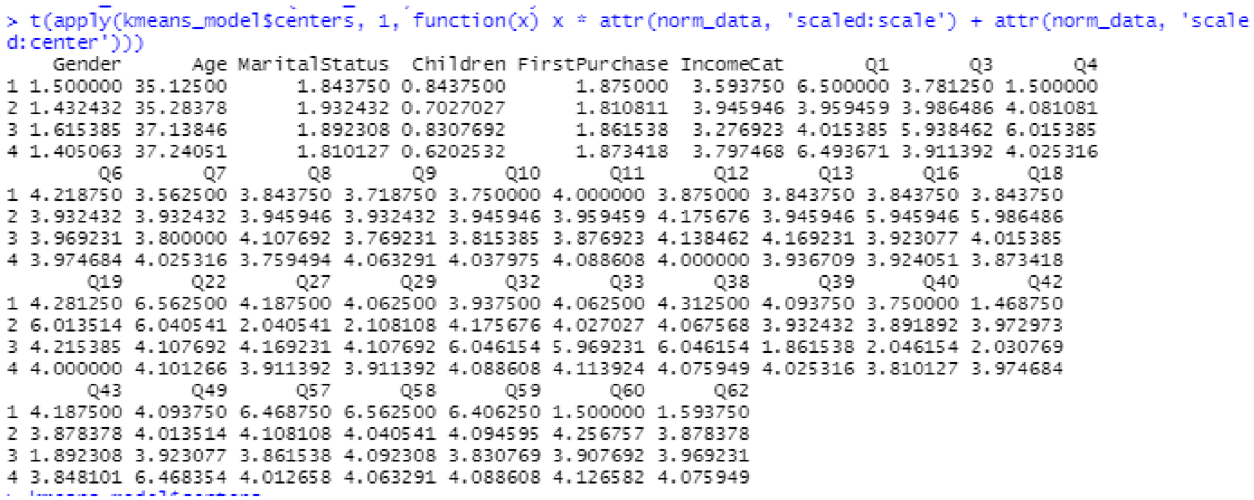


```
Call:
glm(formula = preference ~ ., family = "binomial", data = unnorm_train_data)


  Deviance Residuals:
      Min        1Q    Median        3Q       Max
  -1.8601   -0.9011   -0.5685    0.9678    2.3442

  Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
  (Intercept)    1.428734   4.324995   0.330   0.7411
  Gender        -0.424200   0.353021  -1.202   0.2295
  Age           -0.004568   0.019785  -0.231   0.8174
  MaritalStatus  0.137975   0.189727   0.727   0.4671
  Children      -0.060750   0.163104  -0.372   0.7095
  FirstPurchase  1.415930   0.597148   2.371   0.0177 *
  IncomeCat     -0.142601   0.116915  -1.220   0.2226
  Q1            -0.268908   0.164456  -1.635   0.1020
  Q3            -0.193239   0.187966  -1.028   0.3039
  Q4             0.202680   0.174931   1.159   0.2466
  Q6            -0.021921   0.171225  -0.128   0.8981
  Q7            -0.306339   0.188111  -1.628   0.1034
  Q8             0.036754   0.179626   0.205   0.8379
  Q9            -0.048454   0.163562  -0.296   0.7670
  Q10           -0.365738   0.192755  -1.897   0.0578 .
  Q11           -0.214509   0.163432  -1.313   0.1893
  Q12            0.416297   0.165049   2.522   0.0117 *
  Q13            0.215597   0.150301   1.434   0.1514
  Q16           -0.054365   0.179684  -0.303   0.7622
  Q18           -0.048286   0.174801  -0.276   0.7824
  Q19            0.196277   0.168296   1.166   0.2435
  Q22           -0.087725   0.178435  -0.492   0.6230
  Q27           -0.174197   0.170935  -1.019   0.3082
  Q29            0.204272   0.172708   1.183   0.2369
  Q32            0.005063   0.175281   0.029   0.9770
  Q33           -0.129061   0.179267  -0.720   0.4716
  Q38           -0.054373   0.188661  -0.288   0.7732
  Q39           -0.007309   0.159552  -0.046   0.9635
  Q40           -0.066563   0.175920  -0.378   0.7052
  Q42           -0.048568   0.184903  -0.263   0.7928
  Q43           -0.094599   0.172963  -0.547   0.5844
  Q49            0.225837   0.159334   1.417   0.1564
  Q57           -0.340827   0.162511  -2.097   0.0360 *
  Q58            0.012480   0.163226   0.076   0.9391
  Q59            0.393466   0.176614   2.228   0.0259 *
  Q60           -0.213169   0.160214  -1.331   0.1833
  Q62           -0.018896   0.151339  -0.125   0.9006
  ---
  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Appendix D - Within Cluster Sum of Squares vs. Cluster Size



Appendix E - K-Means Un-Normalized Center Averages



```
> t(apply(kmeans_model$centers, 1, function(x) x * attr(norm_data, 'scaled:scale') + attr(norm_data, 'scale
d:center')))
    Gender      Age MaritalStatus  Children FirstPurchase IncomeCat       Q1       Q3       Q4
1 1.500000 35.12500      1.843750 0.8437500      1.875000  3.593750 6.500000 3.781250 1.500000
2 1.432432 35.28378      1.932432 0.7027027      1.810811  3.945946 3.959459 3.986486 4.081081
3 1.615385 37.13846      1.892308 0.8307692      1.861538  3.276923 4.015385 5.938462 6.015385
4 1.405063 37.24051      1.810127 0.6202532      1.873418  3.797468 6.493671 3.911392 4.025316
        Q6       Q7       Q8       Q9      Q10      Q11      Q12      Q13      Q16      Q18
1 4.218750 3.562500 3.843750 3.718750 3.750000 4.000000 3.875000 3.843750 3.843750 3.843750
2 3.932432 3.932432 3.945946 3.932432 3.945946 3.959459 4.175676 3.945946 5.945946 5.986486
3 3.969231 3.800000 4.107692 3.769231 3.815385 3.876923 4.138462 4.169231 3.923077 4.015385
4 3.974684 4.025316 3.759494 4.063291 4.037975 4.088608 4.000000 3.936709 3.924051 3.873418
       Q19      Q22      Q27      Q29      Q32      Q33      Q38      Q39      Q40      Q42
1 4.281250 6.562500 4.187500 4.062500 3.937500 4.062500 4.312500 4.093750 3.750000 1.468750
2 6.013514 6.040541 2.040541 2.108108 4.175676 4.027027 4.067568 3.932432 3.891892 3.972973
3 4.215385 4.107692 4.169231 4.107692 6.046154 5.969231 6.046154 1.861538 2.046154 2.030769
4 4.000000 4.101266 3.911392 3.911392 4.088608 4.113924 4.075949 4.025316 3.810127 3.974684
       Q43      Q49      Q57      Q58      Q59      Q60      Q62
1 4.187500 4.093750 6.468750 6.562500 6.406250 1.500000 1.593750
2 3.878378 4.013514 4.108108 4.040541 4.094595 4.256757 3.878378
3 1.892308 3.923077 3.861538 4.092308 3.830769 3.907692 3.969231
4 3.848101 6.468354 4.012658 4.063291 4.088608 4.126582 4.075949
```

Appendix F – K-Means Cluster Characteristics

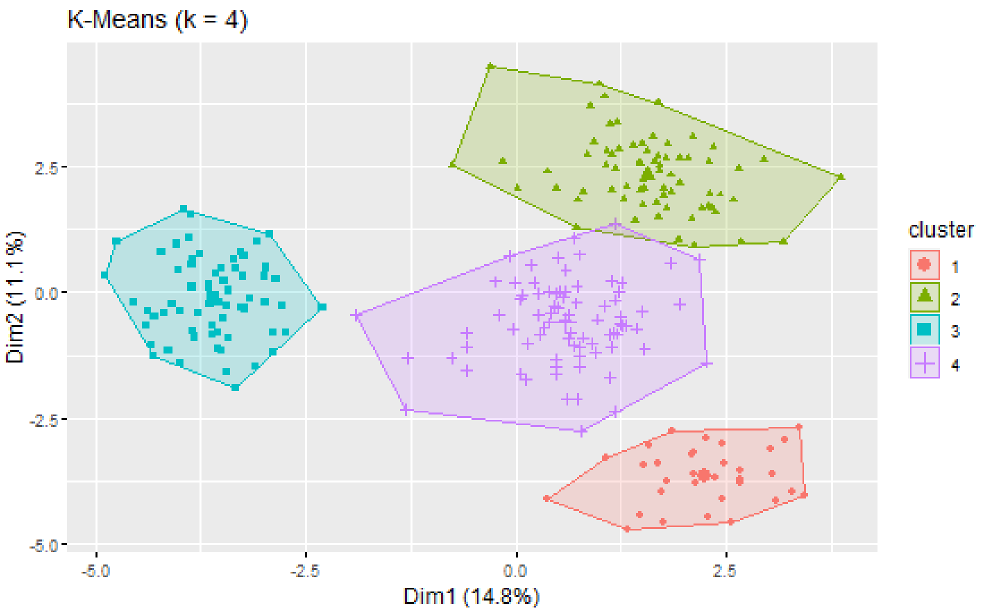
Cluster	Description
Cluster 1	• Cars are extensions of their self-concept • Care about appearance, style of the vehicle • Willing to take care of their cars through frequent maintenance
Cluster 2	• Conscious about environmental sustainability • Concerned about the impacts their vehicle will have on the environment • Value the utilitarian/practical aspects of their car

	- Don’t particularly care for the hedonic attributes of a car
Cluster 3	- Cars are simply a means of transportation - Don’t particularly care for the hedonic attributes of a car - Hold negative opinions about car manufacturers
Cluster 4	- Indifferent about their cars - Don’t hold strong positive or negative opinions about different cars - Likely to choose a car based on what’s popular/trendy

Appendix G – K-Means Within Cluster Sum of Squares (k = 4)

Cluster Within Sum of Squares			
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
737.6808	1906.1605	1535.7498	2130.2639

Appendix H - K-Means Clustering Graph



Appendix I – K-Means Cluster Sizes

Cluster Sizes			
Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**
32	74	65	79