DS 707
Clustering Deliverable: Team 23
Dating Analytics

Anirudh Ravi (IMT2013005)
Nigel Fernandez (IMT2013027)
Revanuru Karthik (IMT2013033)
Rishabh Manoj (IMT2013035)

November 14, 2016

Contents

1 Introduction **2**

2 Clustering into Male and Female **2**

3 Clustering into Attractive and Non attractive Participants **6**

List of Figures

1 Elbow Plot for Male/Female Clustering 3
2 Clus Plot for Male/Female Clustering 4
3 Silhouette Plot for Male/Female Clustering 5
4 Elbow Plot for Attractive/Non-attractive Clustering 6
5 Clus Plot for Attractive/Non-attractive Clustering 7
6 Silhouette Plot for Attractive/Non-attractive Clustering 8

1 Introduction

We chose the two best clustering, namely (1) clustering into males and females and (2) clustering into attractive and non attractive participants. K-means clustering technique worked best in a comparison with other clustering techniques. To determine the optimum number of clusters k we used an elbow plot with within groups sum of squares being the measure. We then picked the k with the maximum average silhouette width. We have provided the elbow plot, the clustering plot using clusplot and the silhouette plot for both clustering.

2 Clustering into Male and Female

We chose two ordinal attributes from our data set namely (1) sports and (2) yoga. These two attributes had the maximum difference between male and female preference in life interests when we analysed our descriptive analysis results. After experimenting with other attributes, these two attributes gave the best external validation results of 65% which is acceptable as males and females have a lot of overlapping interest.

Our elbow plot 1 was used to find the optimum k. The x-axis represents the number of clusters k and the y-axis represents the within groups sum of squares.

From our elbow plot 1 with the x-axis representing the number of clusters k, we chose k to maximize the average silhouette width which corresponded to $k = 4$. We then applied a k-means clustering on our 2 dimensional data with ordinal attributes (1) sports and (2) yoga. We later clustered the obtained 4 clusters into 2 clusters namely male and female. Our clusplot 2 displays the clustering obtained with the two axis representing sports and yoga. The black and red colours were used to externally label the data according to the real gender of the participant. Black refers to male and red refers to female. Our clustering when grouped in this manner, clusters *pink* and *green* as cluster *female* and clusters *blue* and *red* as cluster *male* produced an external validation of 65%.

We had chosen k from our elbow plot to maximize the average silhouette width. This corresponded to $k = 4$ with an average silhouette width equal to 0.59. The silhouette plot with our result is included 3

Figure 1: Elbow Plot for Male/Female Clustering

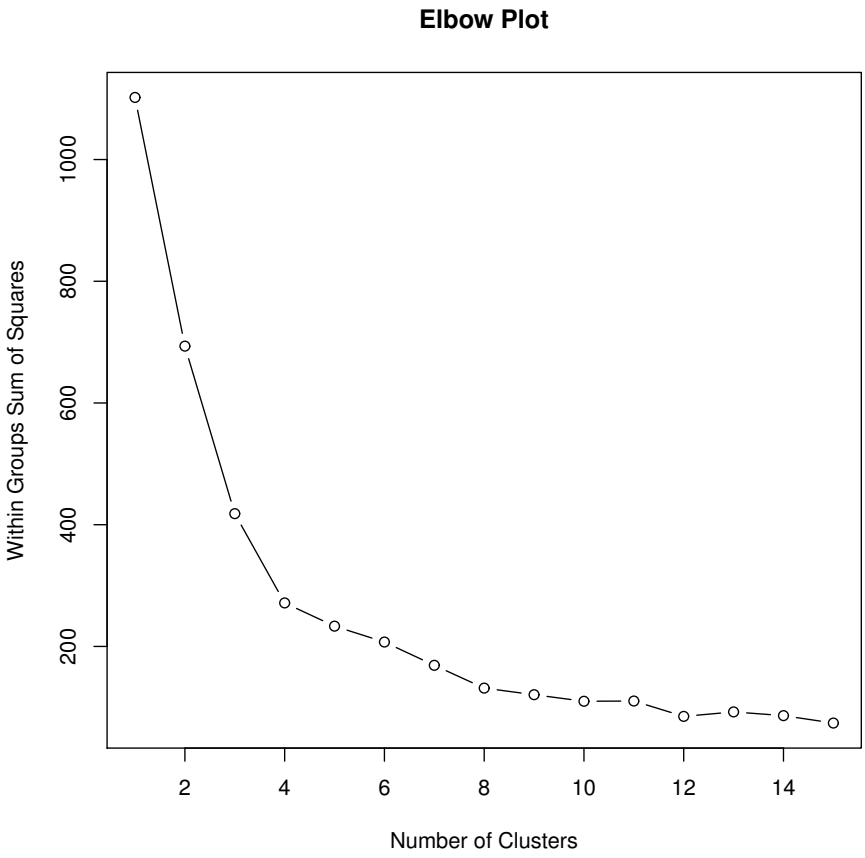


Figure 2: Clus Plot for Male/Female Clustering

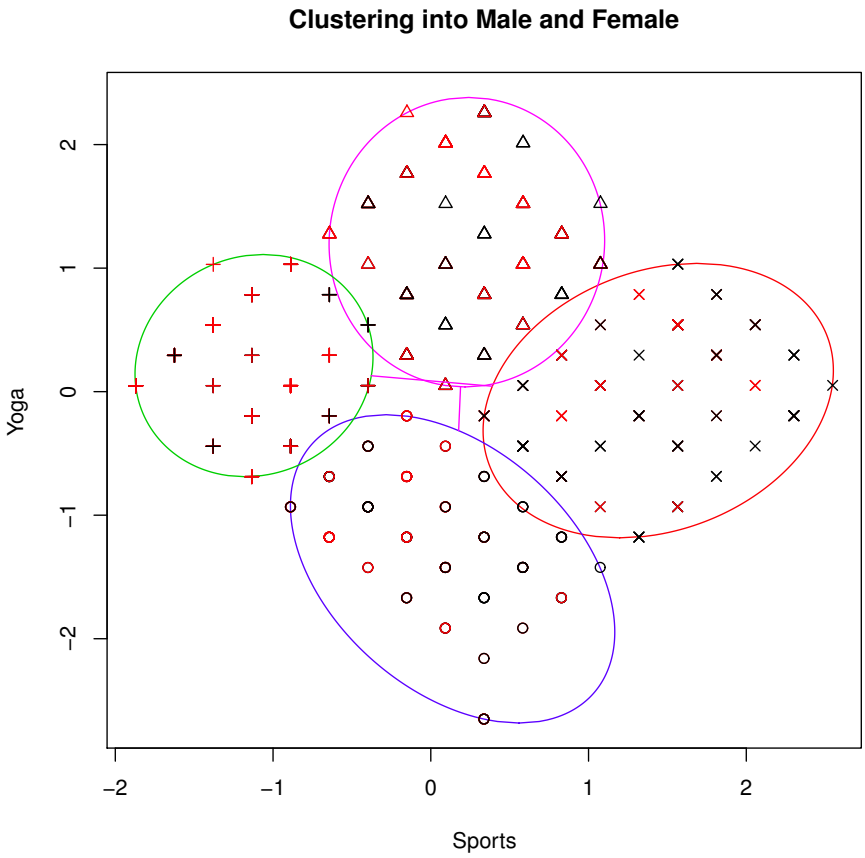
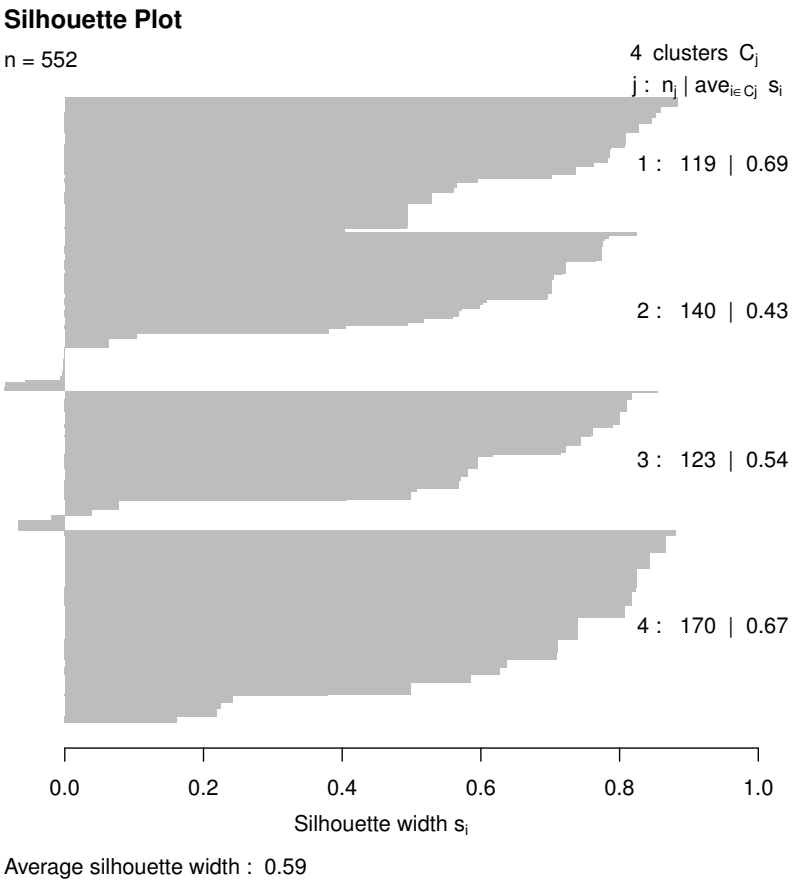


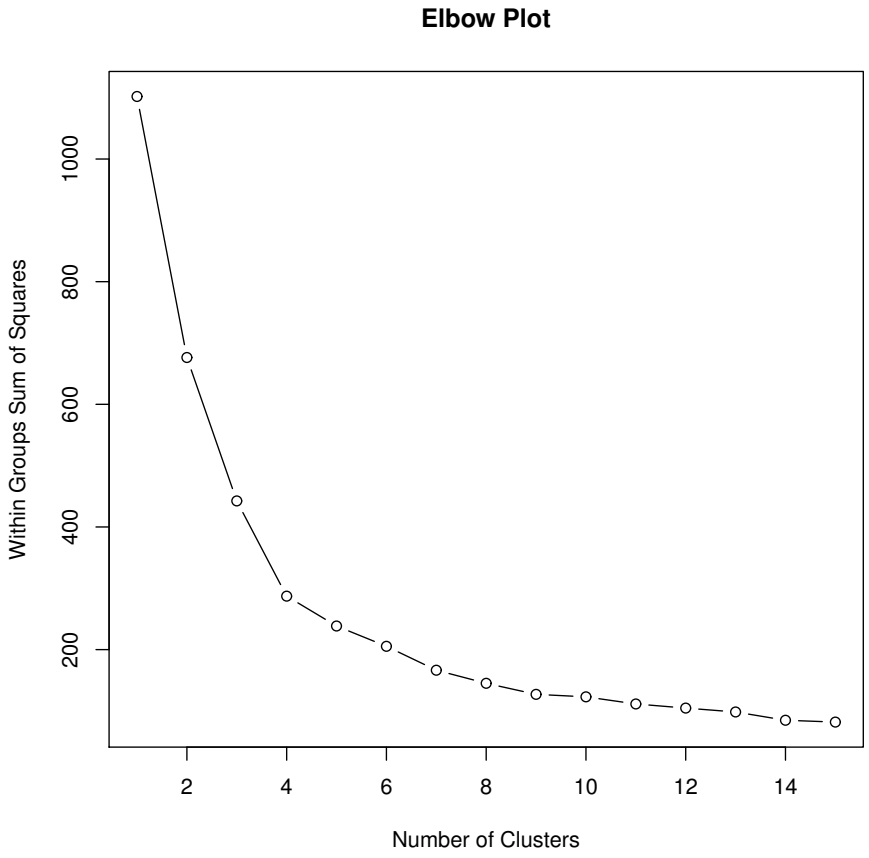
Figure 3: Silhouette Plot for Male/Female Clustering



3 Clustering into Attractive and Non attractive Participants

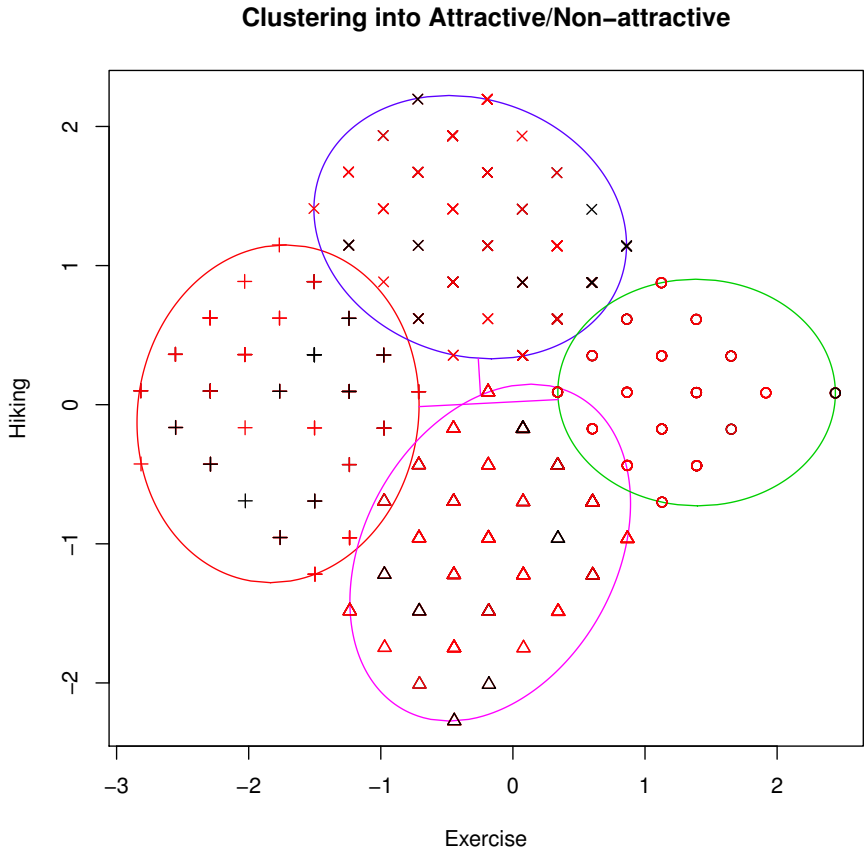
Our second best clustering result corresponded to clustering the participants as rated attractive or non attractive by their date partners. We again used our descriptive analysis results to find the best data attributes. After experimenting with different attributes, we chose two ordinal data attributes namely (1) hiking and (2) exercise. These two attributes produced the best clustering mainly because of their high correlation between a participant's interest in them and their attractiveness rating by their date partner. The external validation of our clustering was 67% when we compared the cluster labels to the real labels in our data set.

Figure 4: Elbow Plot for Attractive/Non-attractive Clustering



Our elbow plot 4 was used to find the optimum k. The x-axis represents the number of clusters k and the y-axis represents the within groups sum of squares.

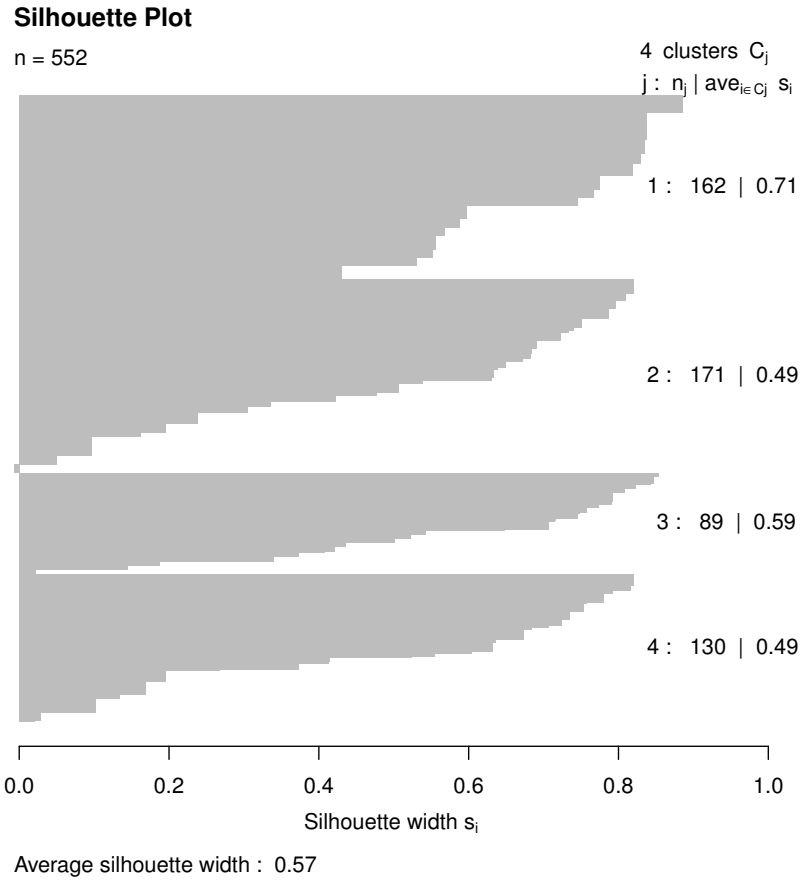
Figure 5: Clus Plot for Attractive/Non-attractive Clustering



From our elbow plot 4 with the x-axis representing the number of clusters k, we chose k to maximize the average silhouette width which corresponded to $k = 4$. We then applied a k-means clustering on our 2 dimensional data with ordinal attributes (1) hiking and (2) exercise. We later clustered the obtained 4 clusters into 2 clusters namely attractive and non attractive rating. A rating between $0-7$ by the date partner was considered non attractive while a rating between $8-10$ was considered attractive. Our clusplot 5 displays the clustering obtained with the two axis representing hiking and exercise. The black and red colours were used to externally label the data according to the real attractiveness rating of the participant. Black refers to non attractive and red refers to attractive. Our

clustering when grouped in this manner, clusters *red* as cluster *non attractive* and clusters *blue*, *pink* and *green* as cluster attractive produced an external validation of 67%.

Figure 6: Silhouette Plot for Attractive/Non-attractive Clustering



We had chosen k from our elbow plot to maximize the average silhouette width. This corresponded to $k = 4$ with an average silhouette width equal to 0.57. The silhouette plot with our result is included 6