SmartTechTM

Smartphone Users Clustering Analysis

November 11, 2021

1. **Executive Summary:**

SmartTechTM’s management would like to learn whether there are any naturally occurring clusters among a smartphone user group. The purpose of this report is to help SmartTechTM to better understand smartphone users’ insightful inferences for future reference regarding future customers may be provided.

In order to do so, users’ data from a variety of Apple and Android users are provided to perform series of exploratory and clustering analyses. Ultimately, Three naturally occurring clusters are ultimately identified. However, further analyses of each identified cluster will need to be conducted if the management would like to learn more about each of its characteristics.

1. **Introduction:**

A handful of exploratory analyses such as identifying missing value and outliers, and explore correlations of each variable, as well as clustering analyses including partitional clustering, hierarchical clustering, and soft (fuzzy) clustering techniques, are performed among different parameters using R language, to classify, label, or group data points contained within the data set.

Lastly, implications and conclusions are drawn from the analyses so the SmartTechTM’s management team is able to orchestrate more robust and data driven business or product development decisions based on the results suggested.

1. **Data Analysis:**

First, let’s take a look at the data that is available. As we can see from the dataset, thirteen demographics and personality attributes are presented:

* **Smartphone**
* **Gender**
* **Age**
* **HEXACO personality variables1**
  + **Honesty-Humility (H)**
  + **Emotionality (E)**
  + **Extraversion (X)**
  + **Agreeableness (A)**
  + **Conscientiousness (C)**
  + **Openness to Experience (O)**
* **Avoidance Similarity2**
* **Phone as a status object**
* **Socioeconomic Status**
* **Time owned current phone**

To further examine the dataset’s characteristics, exploratory analysis such as descriptive analysis and data visualizations are performed in the below section.

* 1. **Exploratory Analysis**

By using the summarize() and describe() functions, several implications can already be made. Based on the outputs (Figure 1 and Table 1), there are 529 user entries. Moreover, two specific attribute, Age and Time Owned Current Phone, are worth examining further at this stage as they have skewness that are greater than 1, meaning the distributions of the attributes are highly skewed. Such indications suggest that while the majority of the data points lie within the left tail, there are values in the dataset that are much higher, potentially the outliers. On top of it, Time Owned Current Phone attribute also has a high kurtosis, suggesting Leptokurtic, meaning the distribution is longer and fatter, implying the attribute is heavy tailed and profusion of outliers. Outliers will be addressed in the later section.

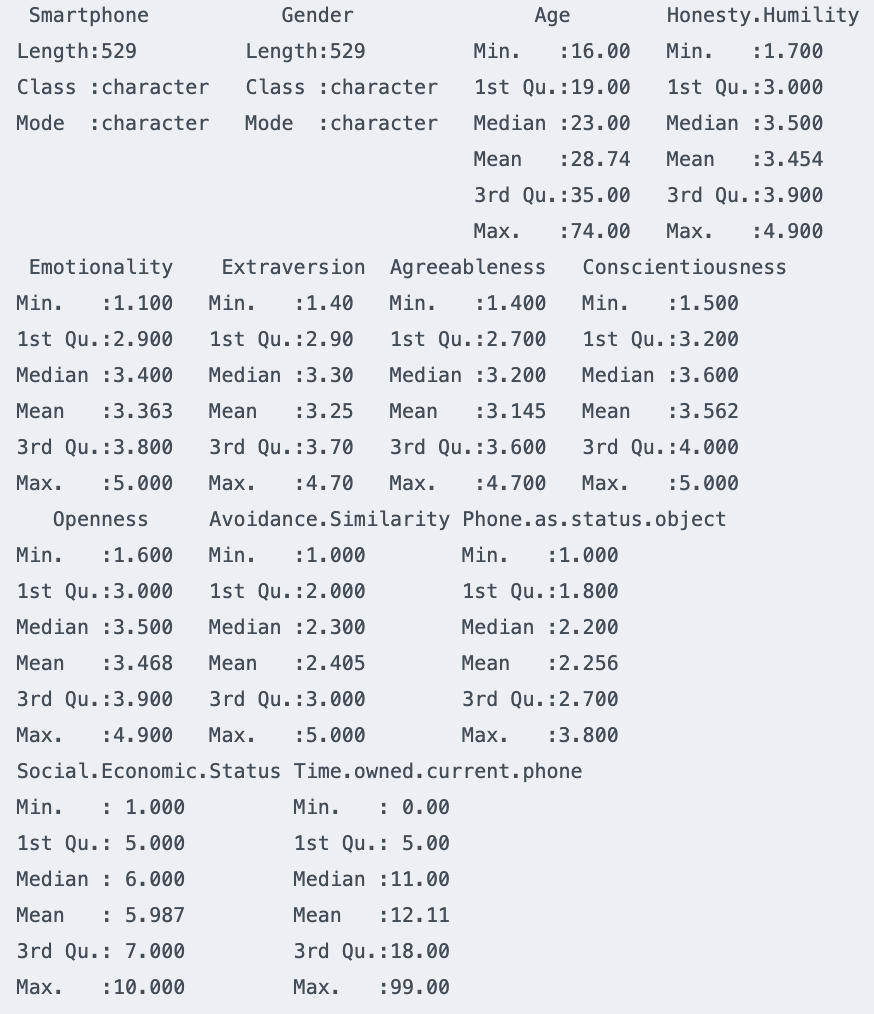


Figure 1. dataset summary with summarize()

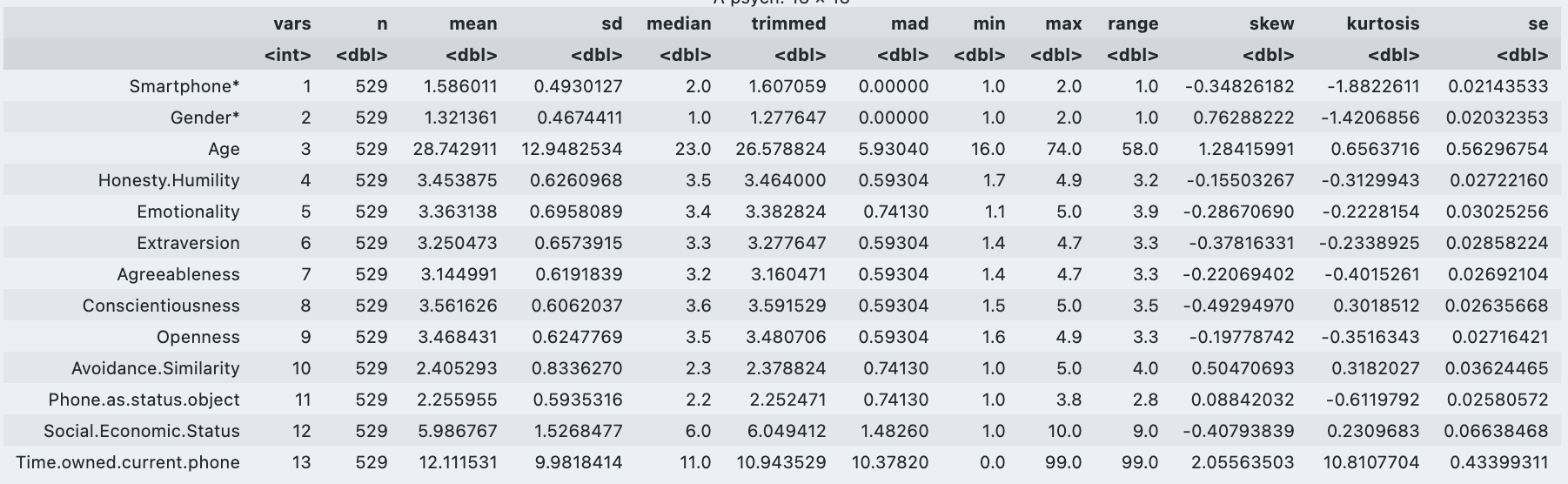


Table 1. dataset summary with describe()

Moreover, by using the missing\_plot() function, we can see from Figure 2 that there are no missing value presented in the dataset, so we can skip steps where we perform missing data tasks such as imputation to avoid information loss.

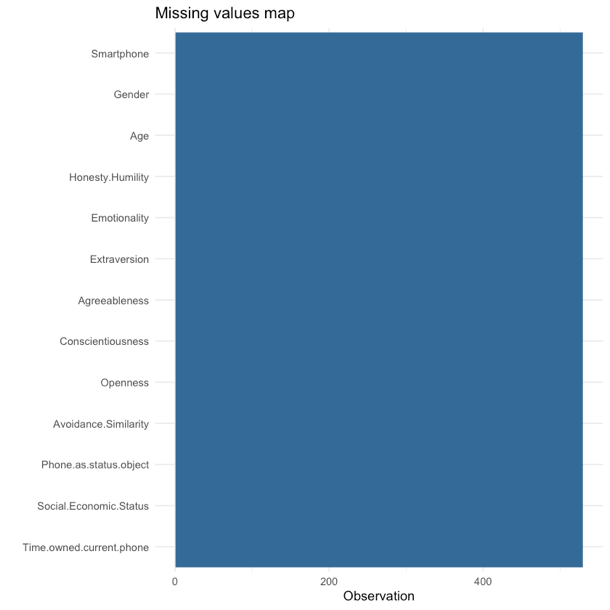
****

Figure 2. Missing value map with missing\_plot()

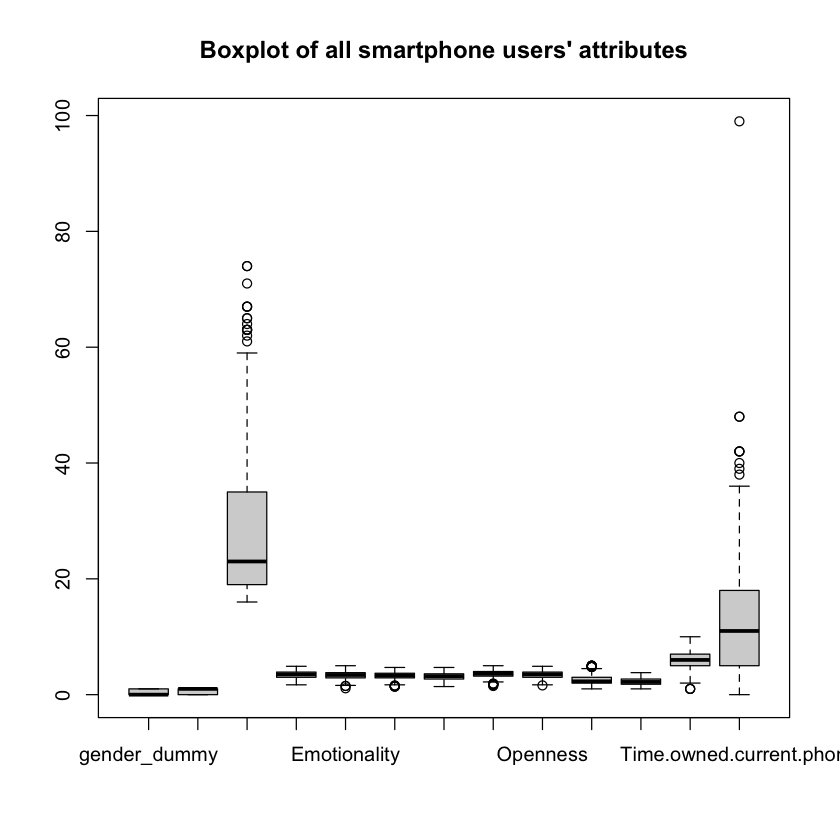
****

Figure 3. Boxplot of all smartphone users’ attributes with boxplot()

By plotting the boxplot across all attributes, we realize Age and Time.owned.current.phone are the two attributes that seem to have significant number of outliers. As it is often iterated that clustering methods like “K-Means algorithm does not give best results.” as “It is sensitive to outliers”, it is suggested that we remove the outliers in order to perform a more concise clustering analysis without having outliers jeopardizing the accuracy of the clustering analysis. After removing the outliers, we can see that, 477 user entries are still available in the dataset as shown in Figure 4.

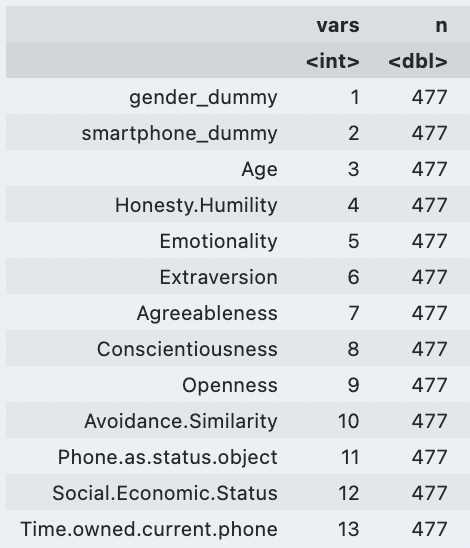


Figure 4. Number of entries after removing outliers

Lastly, before we get into clustering analysis, data visualizations and correlation analysis are performed. In Figure 5, we can see that a huge number of respondents are 30 years and younger, consisting of more female than male, whereas in Figure 6, More users who are 30 years and younger tend to use iPhone than Android, but suggested otherwise in the older age groups.

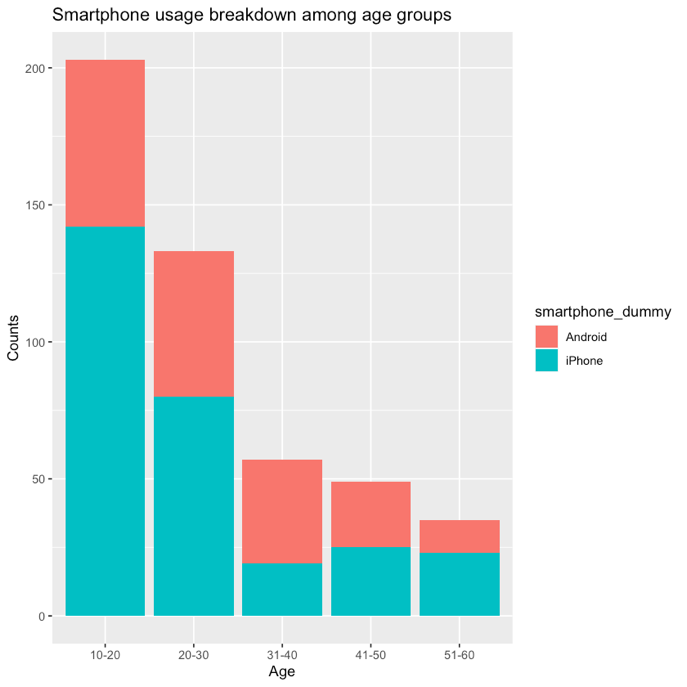
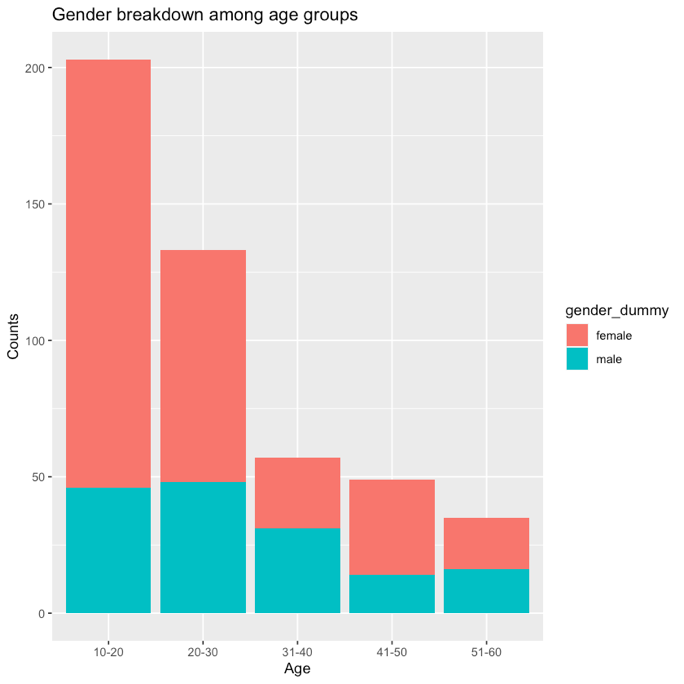
****

Figure 5 & 6. Gender/Smartphone breakdown among age groups

In Figure 7, we can also explore the correlation and the impact of every attribute on each other, in which 6 attribute pairs, positively or negatively, seem to have stronger correlations compare to other attribute pairs. Scatter plots of each attribute pairs are produced to help us examine the trend as we can see from Figure 8.

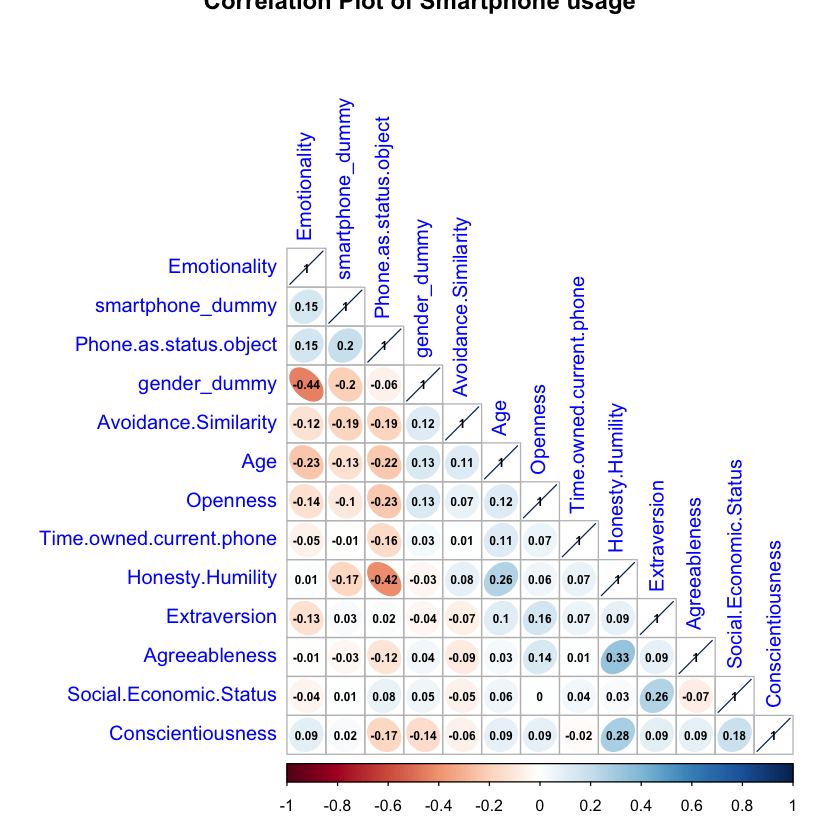
****

Figure 7. Correlation plot of smartphone usage

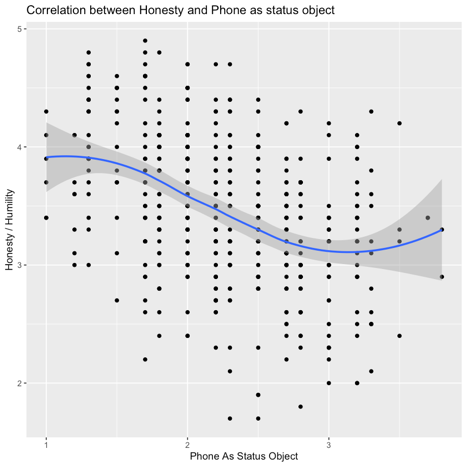
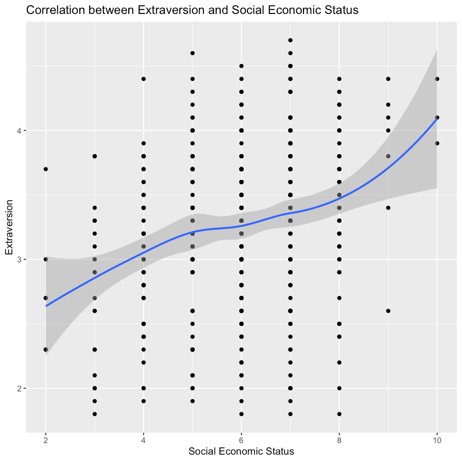
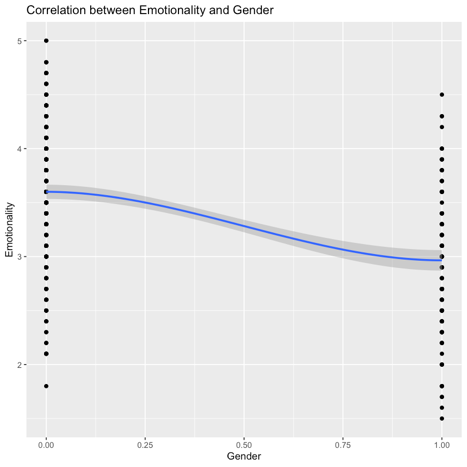
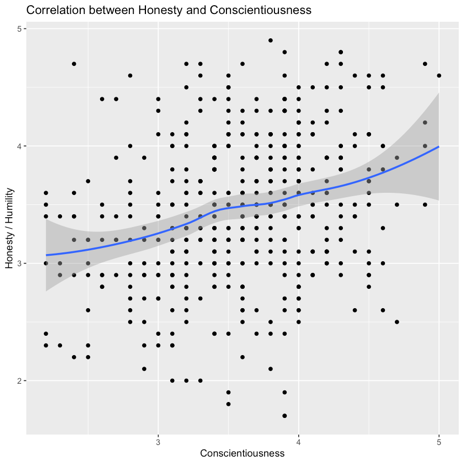
****

Figure 8. Attribute pairs that have strong correlations

* 1. **Clustering Analysis**

To make sure we can spot any naturally occurring clusters of people in this dataset, it is important that we remove pre-existing labels such as “Smartphone” or “Gender”, so a subset of the original dataset without the two mentioned attributes is created. Moreover, in order for us to compare different attributes on equal footing and applying clustering methods based on measures of how far the data points are, scaling is strongly suggested before performing clustering analysis. In the following sections, scaled dataset is used. Lastly, three clustering techniques, partitional clustering, hierarchical clustering, and soft clustering are chosen to determine the best clustering methods.

1. **Partitional Clustering**

Three partitional clustering techniques, K-Means clustering, K-Medoids clustering (PAM), and Clustering large applications (CLARA) are used.

When performing K-Means clustering, first we may use fviz\_nbclust() function to calculate the optimal number of clusters to have. However, by using “within sum of square” method, the 'bend' in the plot where the the bend indicates that additional clusters provide diminishing returns and value is not shown, hence, if we use the Silhouette method, where it essentially tries to measure the space between clusters, the optimal number of clusters seem to be 2 so far (Please refer to Figure 9), however, we will manually test k value of 2, 3, and 4 to compare the results.

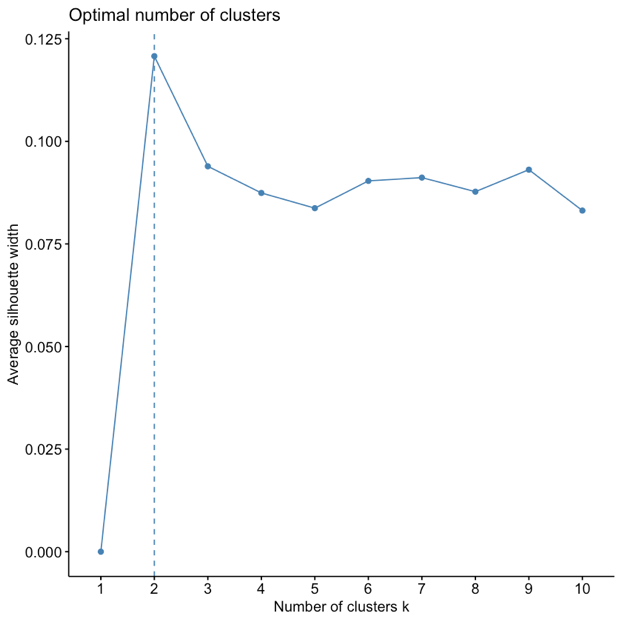
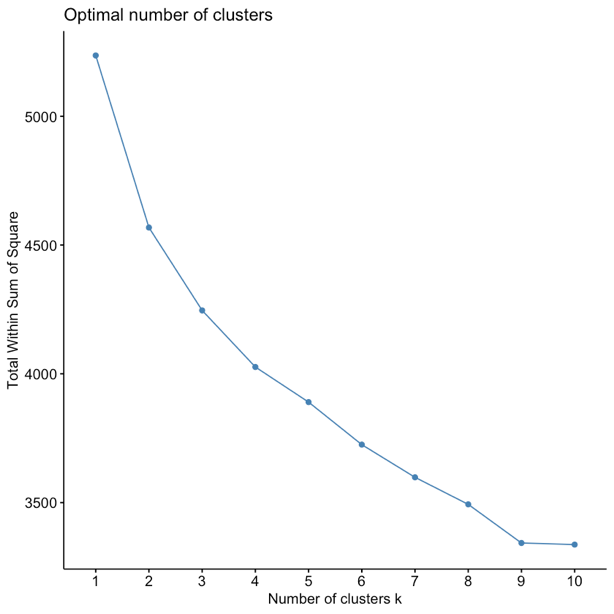


Figure 9. Finding optimal number of clusters using WSS and Silhouette methods

For k-mean method, data point is assigned with a cluster centroid that is calculated by finding mean value of all data point in the cluster that has minimum distance sum of square. Thus, as we can see, table 2, 3, and 4 show the centroids that are calculated for each parameter groups.

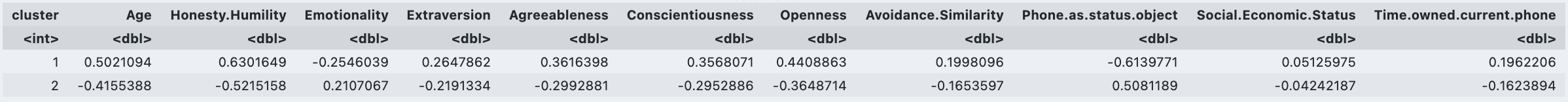


Table 2. Cluster centroids of k-mean clustering with k=2

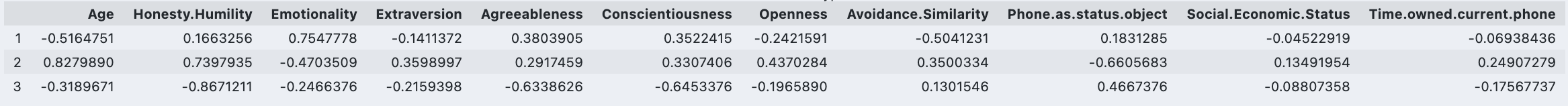


Table 3. Cluster centroids of k-mean clustering with k=3



Table 4. Cluster centroids of k-mean clustering with k=4

We can examine the clustering parameters efficacies by visualizing the clusters of all three parameter groups by referring to Figure 10. It seems like with k equals to 4, we are seeing overlaps of clusters, suggesting bad clustering. Thus, k equals to 3 seems to be the optimal number of clusters.

K-Medoids Clustering (PAM) and Clustering Large Application (Clara) methods on the other hand, as suggested by Figure 11, don’t seem to be a good clustering method for this dataset as even with number of cluster at 2, we are still seeing overlaps of clusters, suggesting the produced clusters by two methods are not the greatest.

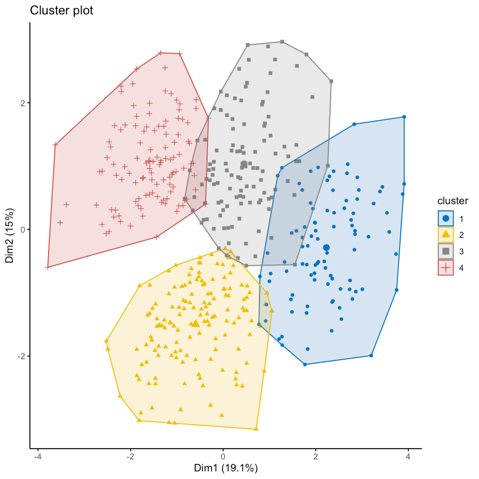
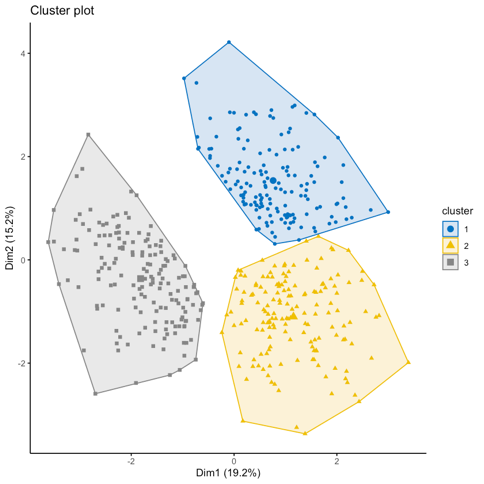
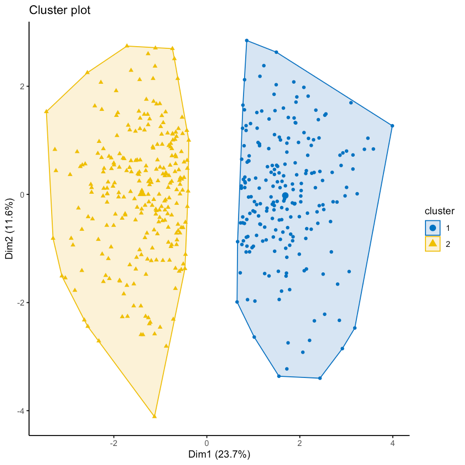


Figure 10. Clusters of k-mean clustering techniques with k=2, k=3, and k=4

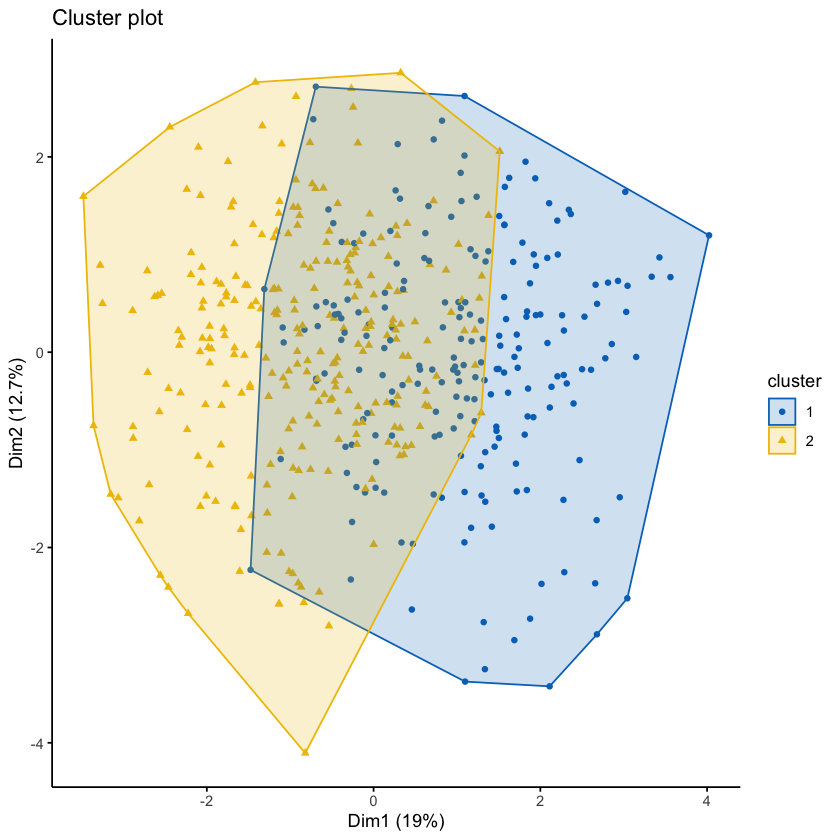
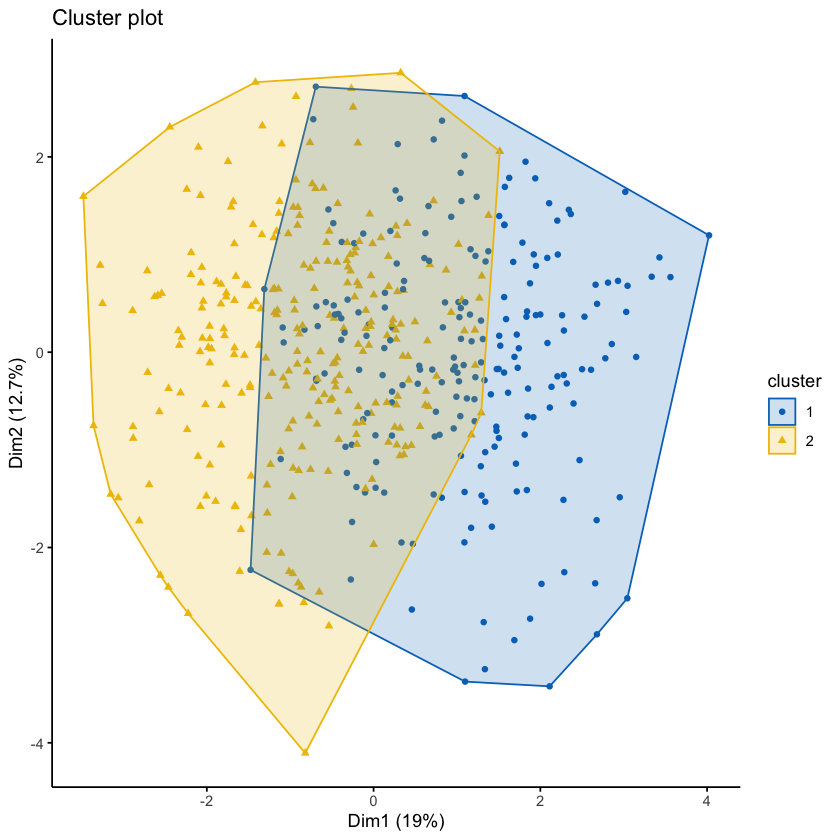


Figure 11. PAM and CLARA methods clustering technique with k = 2

On top of it, if we were to validate the results, we can use external clustering validation method to compare with the original smartphone user dataset. Figure 12 suggests that if we compare with the gender label, 125 of the female users has been classified as cluster 1, 96 as cluster 2, and 102 as cluster 3, whereas for male users, 28 has been classified as cluster 1, 64 as cluster 2, and 63 as cluster 3. On the other hand, if we compare with the smartphone label, 47 of the android users has been classified as cluster 1, 82 as cluster 2, and 60 as cluster 3, whereas for iPhone users, 107 users have been classified as cluster 1, 77 as cluster 2, and 105 as cluster 3. This indicates that Gender and Smartphone choices definitely cannot be used as

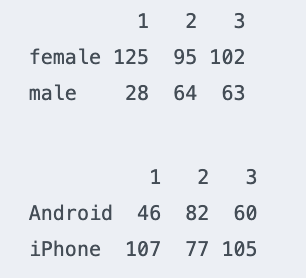


Figure 12. Cluster Validation Against original labels

1. **Hierarchical Clustering**

Agglomerative hierarchical clustering methods with different methods of measuring the similarity like Euclidean and Manhattan distances methods are also evaluated. by calculating the correlation between the cophenetic distances and the original distances. However, both methods suggest very low correlation coefficient (lower than 0.35), indicating hierarchical clustering methods might not be a good method for this dataset as If the value is high (near 1) the clustering result is an excellent representation of the original distances, if it is <<1 then it is not. In Figure 13, we can see hugely overlapped clusters when conducting hierarchical clustering.

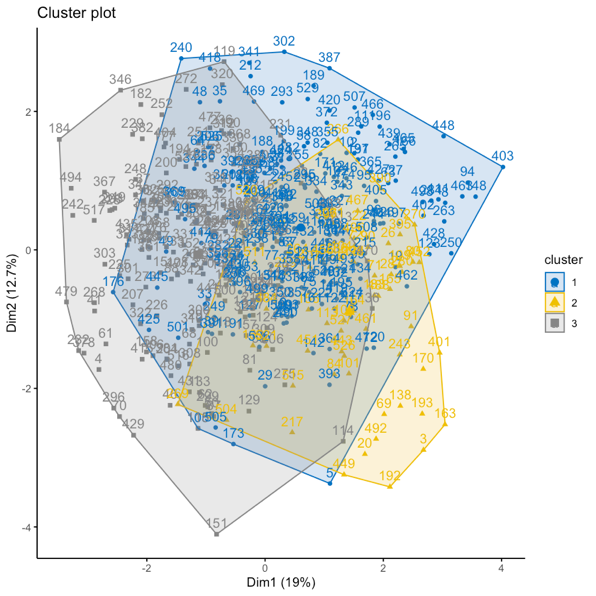
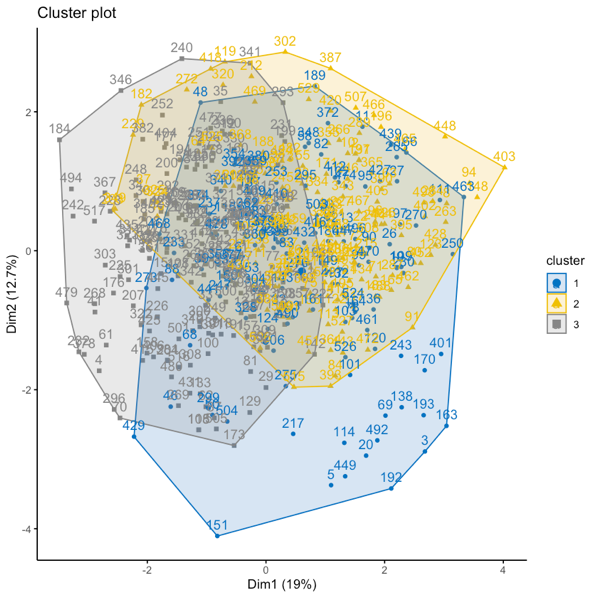


Figure 13. Hierarchical Clustering with Euclidean and Manhattan distance methods

1. **Soft Clustering**

Soft cluster, or fuzzy clustering’s centroid, is calculated as the mean of all points, weighted by their degree of belonging to the cluster. As suggested earlier by Silhouette method, the optimal number of clusters should be 2. With this information, in Figure 14, the algorithm suggests that each data point has a 50% chance of belonging to each cluster.

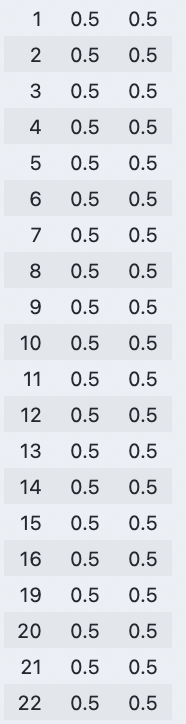


Figure 14. Membership (%) of each data point belong to each cluster

Moreover, by looking at Dunn’s partition coefficient, or the “fuzziness” indicator of the clustering, it suggests that two clusters are mediocrely classified as the Dunn’s partition coefficient is close 0.5, where A low value of Dunn's coefficient indicates a very fuzzy clustering, whereas a value close to 1 indicates a near-crisp clustering3. Figure 15 shows the overlapped clusters identified by soft clustering technique.

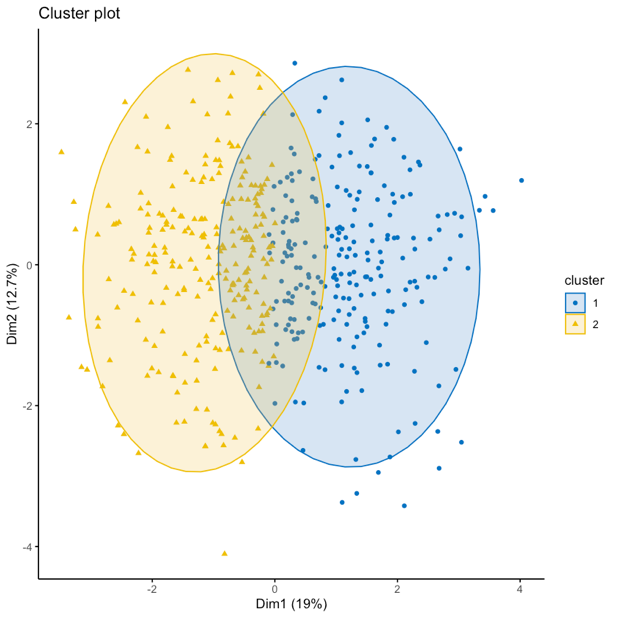


Figure 15. Clusters identified by Soft Clustering Technique

1. **Summary and Implication:**

After performing clustering techniques, a handful of insights may be drawn. Firstly, the more commonly, or more intuitive labels such as Gender or Smartphone may not be the best classification methods when trying to cluster a user base, or a target audience group. Such point was verified when performing clustering validation against the original labels.

Moreover, there seems to have naturally occurred clusters when using K-means partition clustering method with k value of 3 (optimal number of cluster). This configuration seems to be optimal for classifying the smartphone user dataset crisply. However, to learn more about each cluster’s characteristics, more explorations and analyses are required.

1. **References:**
2. Psychology Wiki. *HEXACO model of personality structure.* Wikia.org. Available from: <https://psychology.wikia.org/wiki/HEXACO_model_of_personality_structure> [Accessed on Oct 31st, 2021]
3. A. A. Ruvio, M.M. Brencic. *Consumers’ need for uniqueness: Short-form scale development and cross-cultural validation.* Research Gate. Available from: <https://www.researchgate.net/publication/235300028_Consumers'_need_for_uniqueness_Short-form_scale_development_and_cross-cultural_validation> [Accessed on Oct 31st, 2021]
4. ETH Zurich. *Fuzzy Analysis Object*. Available from: <https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/fanny.object.html> [Accessed on Oct 31st, 2021]