University of Bath

MN50752 Data Mining & Machine Learning

SMARTPHONE USER ANALYSIS REPORT

iPhone VS Android

Word Count: 2155

By Ningkai Zheng

22/4/2022

Table of Contents

Abstract 1

Introduction 1

Descriptive Statistics 1

iPhone VS Android, Round 1: Gender 2

iPhone VS Android, Round 2: Age 3

iPhone VS Android, Round 3: HEXACO 4

iPhone VS Android, Round 4: Avoidance Similarity 5

iPhone VS Android, Round 5: Phone as Status Object 6

iPhone VS Android, Round 6: Socioeconomic Status 7

iPhone VS Android, Round 7: Time Owned Current Phone 8

Clustering 8

Implement Partitional Clustering 8

Characteristics of Clusters 9

Classification 10

Variables Sets 10

Modelling and Evaluation Methods 10

Logistic Regression (LR) 10

k-Nearest Neighbours (kNN) 11

Naïve Bayes (NB) 11

Random Forest (RF) 11

Support Vector Machine (SVM) 11

Comparison of Classification Models 11

Discussion 12

Limitations 12

Conclusion 12

Appendix 13

Bibliography 18

Abstract

This report examines the relationship between smartphone systems (iPhone and Android) and their users. Our data includes users' gender, age and personality, etc. We first visualise and analyse the data using descriptive statistics and then use k-means to find natural clusters in the data. Based on the first two steps, we can determine which variables are significant. Finally, we use logistic regression, k Nearest Neighbours, Naïve Bayes, Random Forest and Support Vector Machine to build classification models to predict which smartphone system the users use. There are seven significant differences between iPhone and Android users: gender, age, Honesty-Humility, Emotionality, Conscientiousness, Avoidance Similarity, and Phone as Status Object. The Naive Bayes method performs best in the classification model using these significant variables.

Introduction

Since the 21st century, smartphones have snowballed and quickly taken over most mobile phone markets (Ross, 2011). According to recent research, around 80% of UK adults own a smartphone (Ofcom, 2019), and the iPhone and Android systems account for 100% of the global smartphone market (IDC, 2021). There have been a number of studies on the link between smartphone use and user personality (Chen, et al., 2021) and (Clayton, et al., 2015), and Forbes (2014) has also investigated the differences between users of different smartphone systems, with results including higher average education and higher average salaries for iPhone users. Furthermore, a study (Shaw, et al., 2016) shows that iPhone users are more likely to be female and consider their phones as status objects, while Android users are more likely to be male and more concerned about what mobile devices are used by the majority of people. This paper will explore more visualised factors within the data from this study (Shaw, et al., 2016) and apply clustering and classification algorithms to this data.

Descriptive Statistics

There are 13 variables in this data, as shown in Table 4. There are 529 records in this data, and through preliminary checks, we confirm that the records are well recorded (there are no missing data and other issues). Next, we will analyse and visualise each variable to go through the battle between iPhone and Android. Fig. 1 shows an overview of the iPhone and Android users in our data. Apparently, iPhone wins by a 59% of users.

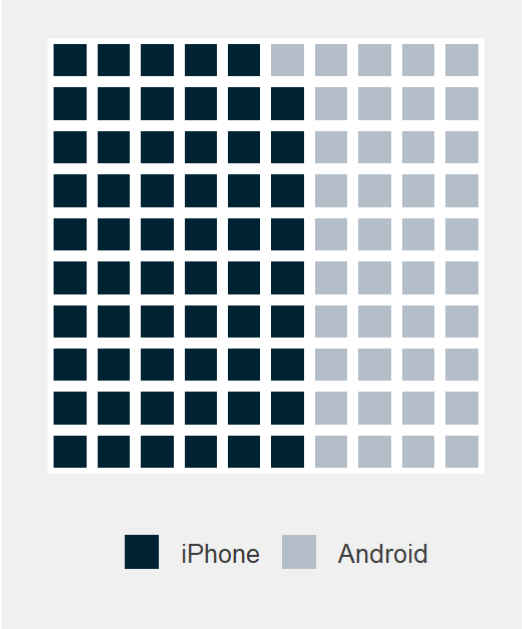

Fig. 1. Percentage of iPhone and Android users

iPhone VS Android, Round 1: Gender

Females account for 68% of our data, but they take up 75% of iPhone users and only account for 58% of Android users, as shown in Fig. 2 and Table 1, which suggests that females prefer iPhone while males prefer Android. The same results can be reached in the battle between iPhone and Android in each gender group (Table 2).

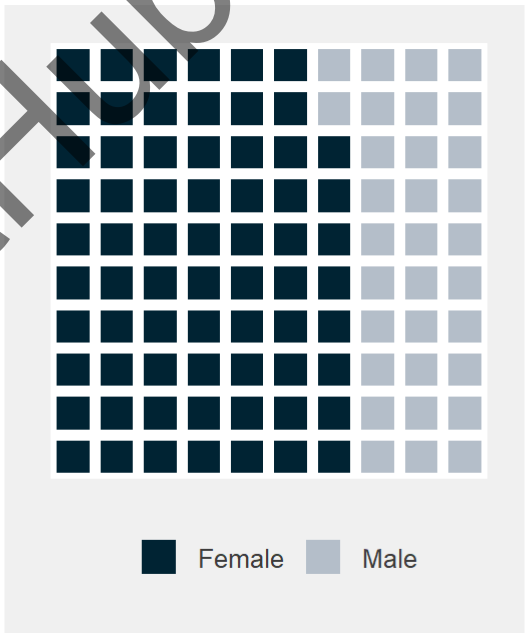

Fig. 2. Percentage of Female and Male

Table 1

Percentage of each gender group in different operating systems

System	Gender	Total	Percentage
iPhone	Female	233	75%
	Male	77	25%
Android	Female	126	58%
	Male	93	42%

Table 2

Percentage of each operating system in different gender group

System	Gender	Total	Percentage
Female	iPhone	233	65%
	Android	126	35%
Male	iPhone	77	45%
	Android	93	55%

iPhone VS Android, Round 2: Age

From Fig. 3, we can see that young people in their 20's dominate our data. Fig. 4 shows that iPhone overwhelmingly beat Android among young generations under 25, while these two operating systems almost play out a draw in other age groups.

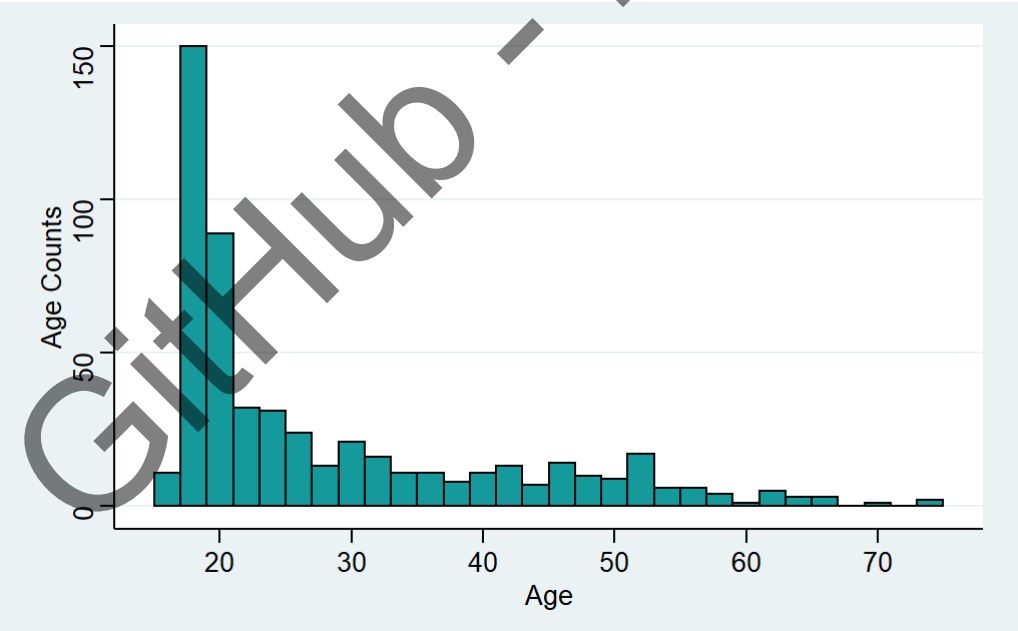

Fig. 3. Distribution of age

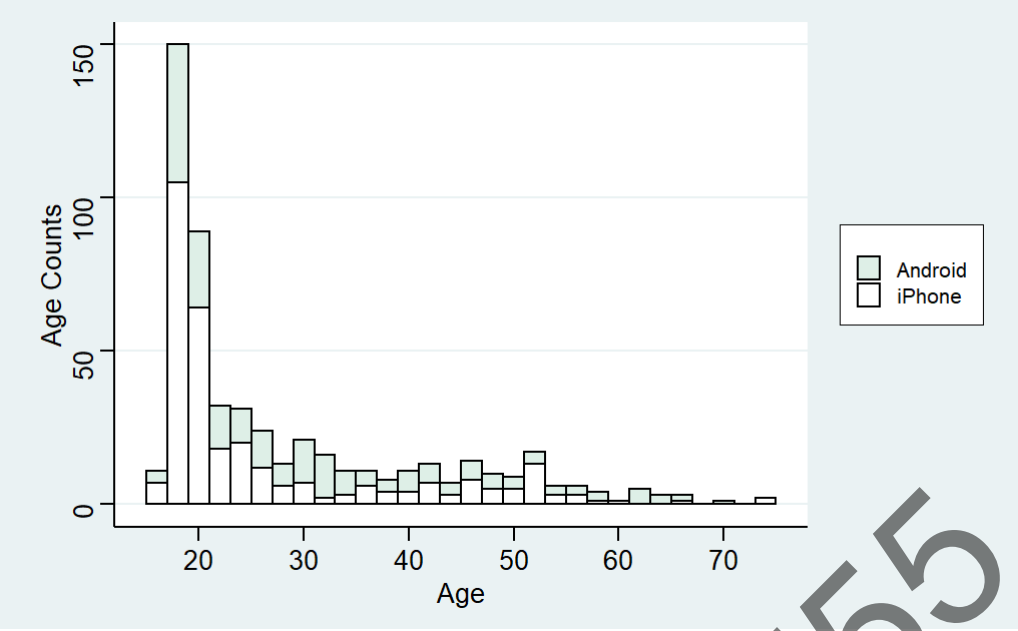

Fig. 4. Age distribution and operating system group

iPhone VS Android, Round 3: HEXACO

HEXACO is a kind of personality assessment that consists of 6 dimensions: Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C) and Openness (O) (Ashton and Lee, 2009).

First, we need to investigate the correlation between each variable in HEXACO in case of the influence of multicollinearity. Fig. 5 shows little intercorrelations between HEXACO factors, with all correlation coefficients under 0.35. Thus, we are free to use all these factors to analyse and fit models.

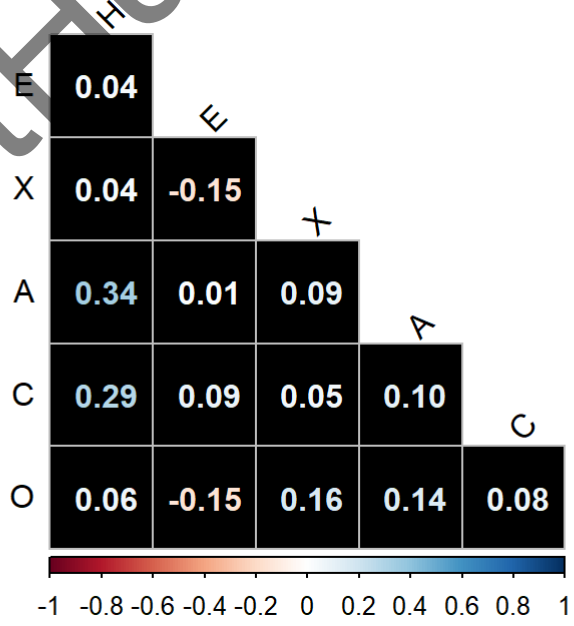

Fig. 5. The correlations between HEXACO factors

According to Fig. 6 and Fig. 7, iPhone users are more likely to have a high level of Emotionality, while Android users tend to have a high level of Honesty-Humility.

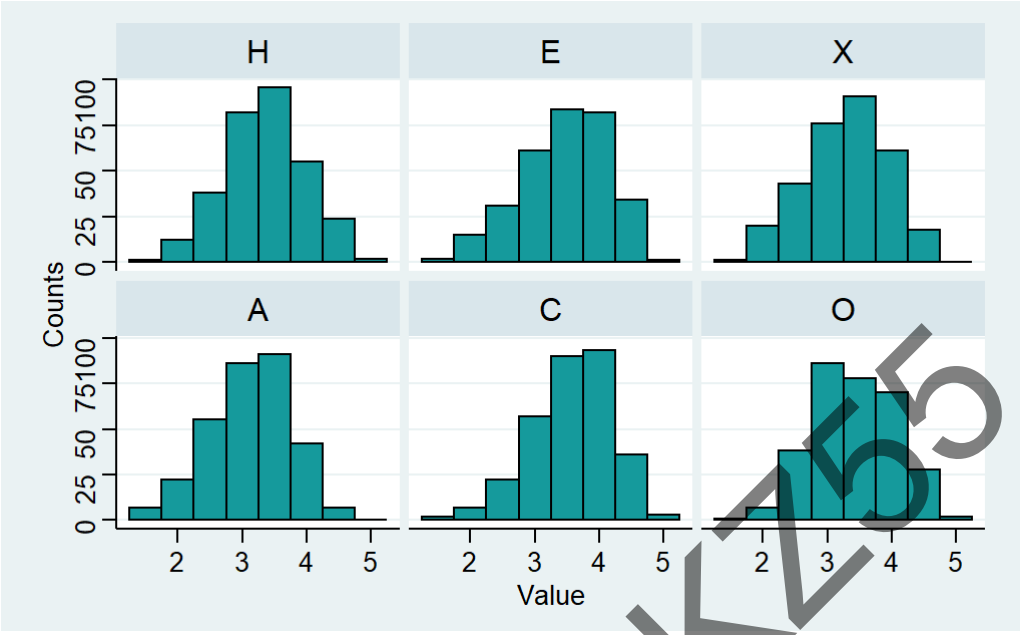

Fig. 6. HEXACO distributions of iPhone user

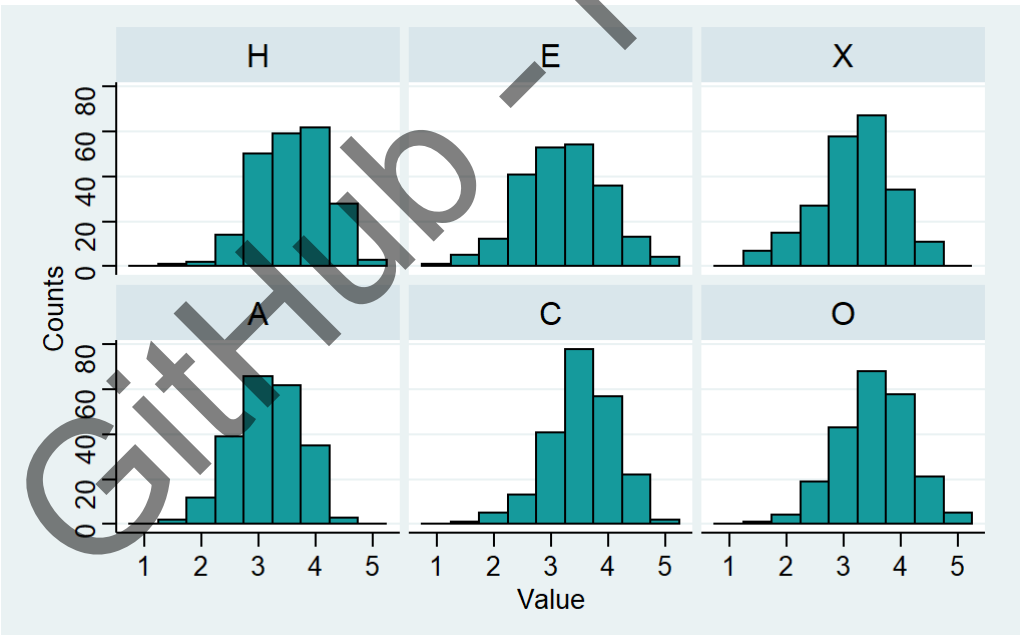

Fig. 7. HEXACO distributions of Android user

iPhone VS Android, Round 4: Avoidance Similarity

Compared to iPhone users, Android users are more likely to be concerned about what type of smartphone most people are using, as shown in Fig. 8.

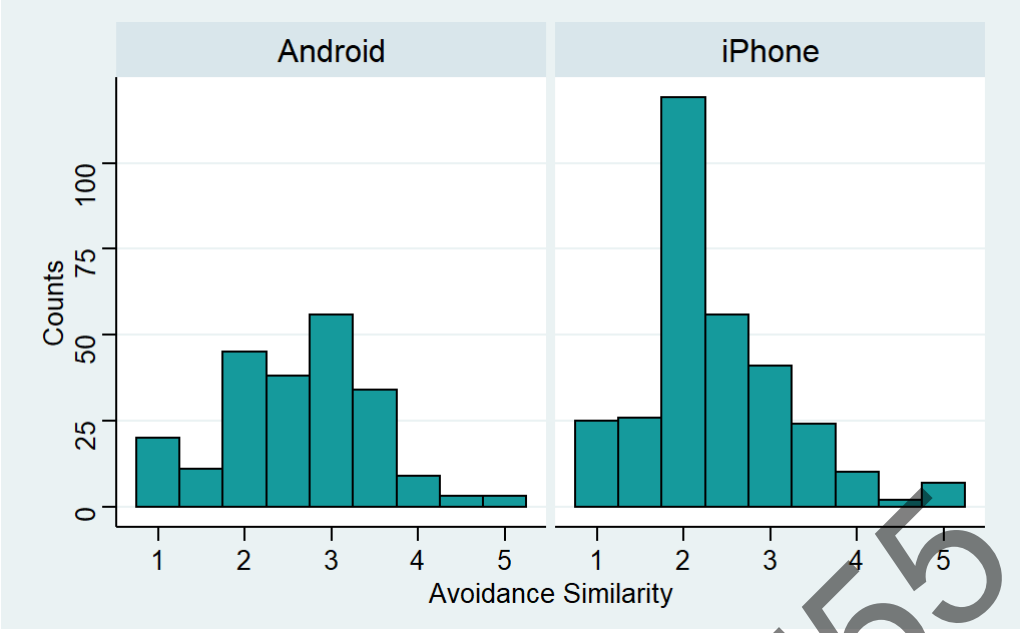

Fig. 8. Distribution of Avoidance Similarity in Android and iPhone users

iPhone VS Android, Round 5: Phone as Status Object

Compared to Android users, iPhone users are more likely to consider their phones as status objects, as shown in Fig. 9.

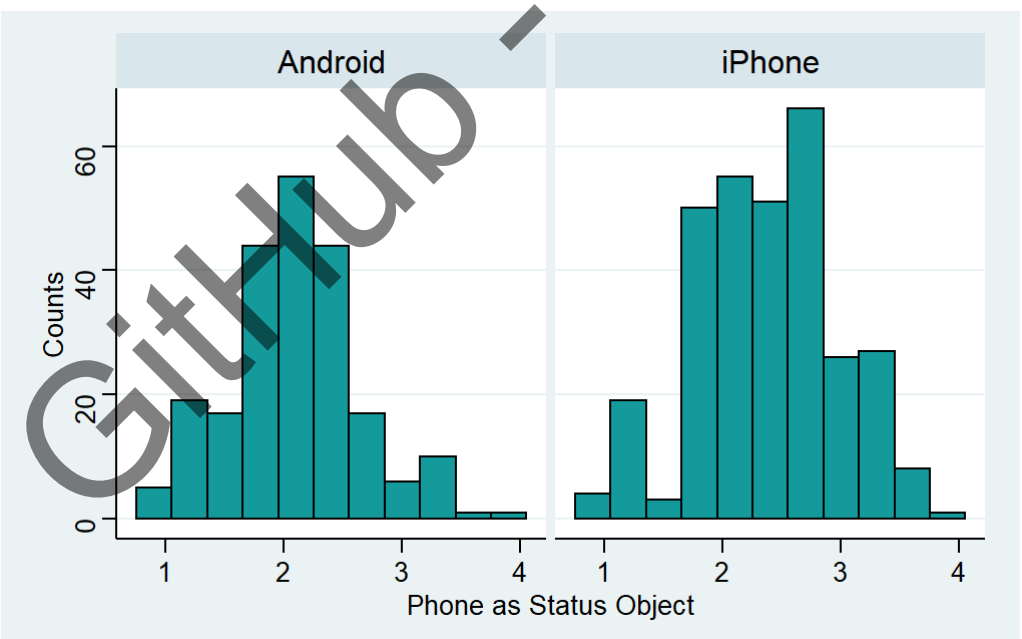

Fig. 9. Distribution of Phone as Status Object in Android and iPhone users

iPhone VS Android, Round 6: Socioeconomic Status

Before we step into the battle between iPhone and Android, we wonder if there is a relationship between age and socioeconomic status. Fig. 10 shows that the relationship between age and socioeconomic status is not clear.

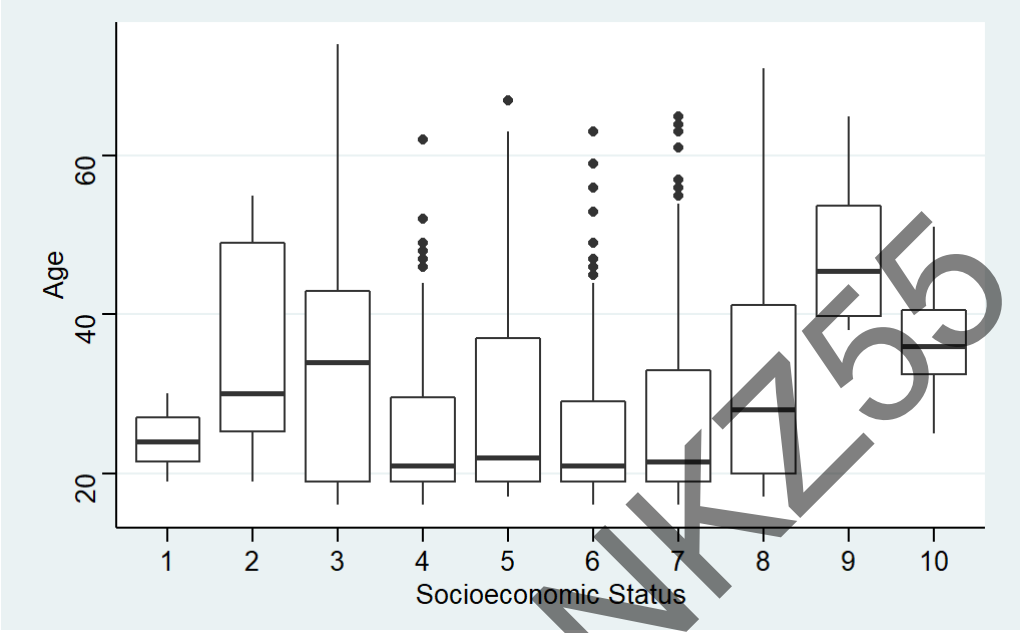

Fig. 10. Age distribution of different socioeconomic status

The distributions of Socioeconomic Status in Android and iPhone users are similar in Fig. 11, so we consider Socioeconomic Status as a not significant factor.

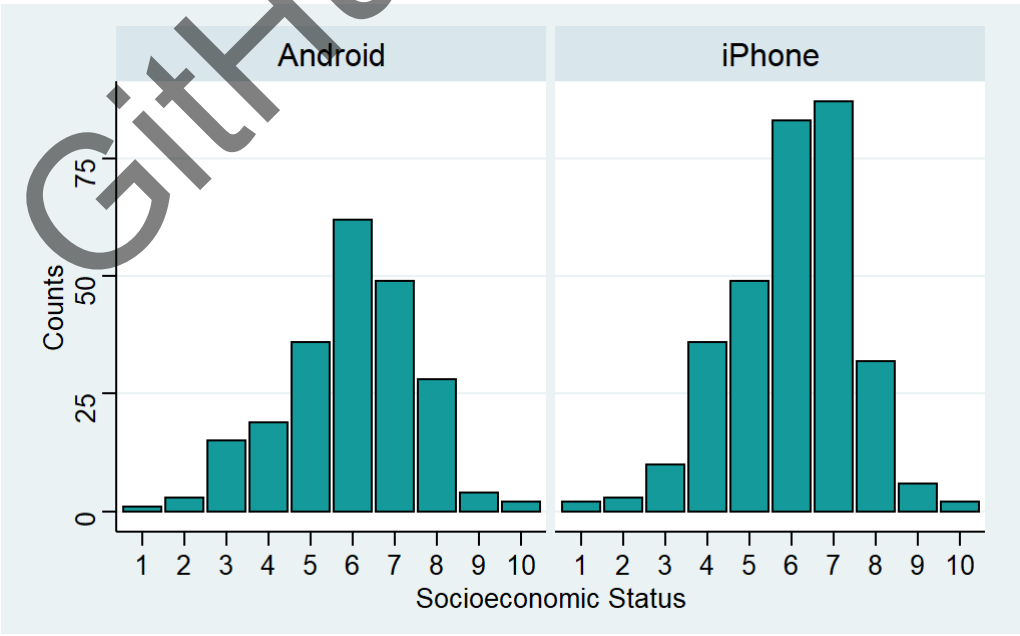

Fig. 11. Distribution of Socioeconomic Status in Android and iPhone users

iPhone VS Android, Round 7: Time Owned Current Phone

According to Fig. 12, the distributions are similar, except for the outlier. We exclude the outlier and consider this factor as a not significant factor. After excluding the outlier, we have 528 records in our data.

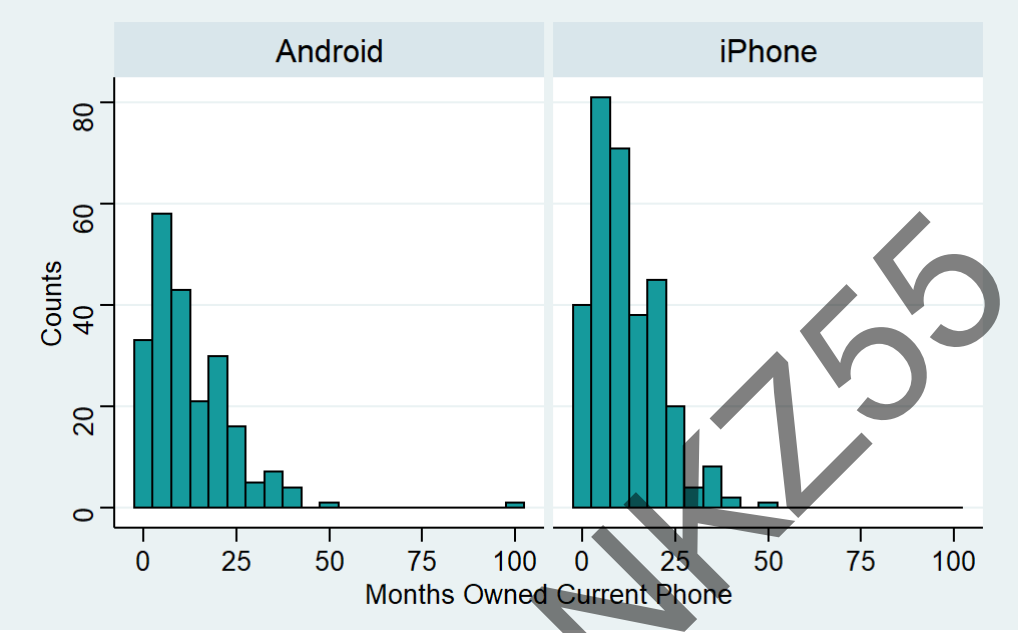

Fig. 12. Distribution of Months Owned Current Phone in Android and iPhone users

Clustering

Clustering is about separating records into several different groups by their characteristics. We have 528 records in the data, which is too much for hierarchical clustering, so we first use the partitional clustering method, and then we will discuss whether there is a need to use other clustering methods.

Implement Partitional Clustering

The partitional clustering method focuses on assigning individual data points to the k clusters by calculating the distance between the data point and the k centroids and selecting the closest centroid group. We use the k-means method here and choose to separate the data into 5 clusters (Fig. 15, Fig. 16).

According to Table 3, the clusters show significant differences by only considering two factors - the operating system and gender, which divide the data into female iPhone users, female Android users, male Android users, etc.

Table 3

The structure of each cluster while separating the data into 5 clusters using k-means

Cluster No.	Percentage of iPhone User	Percentage of Male
1	74.53%	18.87%
2	38.04%	2.17%
3	92.97%	6.25%
4	37.23%	34.04%
5	38.89%	99.07%

Characteristics of Clusters

Let us take all features into consideration (Fig. 13). Clusters 1 and 3 are mainly iPhone users (74.53% and 92.97%, respectively). However, Cluster 1 contains iPhone users who strongly consider their phones as status objects with a deficient level of Honesty-Humility and Conscientiousness. In contrast, Cluster 3 contains iPhone users with a high level of Emotionality and Conscientiousness. In Clusters 1 and 3, most are young generations and females.

Clusters 2, 4 and 5 are mainly Android users (around 62%). Cluster 2 contains Android users with a low level of socioeconomic status. Cluster 4 mainly contains older people. Cluster 5 mainly contains males with a low level of Emotionality.

As these clusters already have high intra-class similarity and low inter-class similarity, we do not need to apply the distribution-based clustering method since it requires hard-to-define statistics hypotheses and often suffers from overfitting. The data points are almost evenly distributed, so we do not need to apply the density-based clustering method.

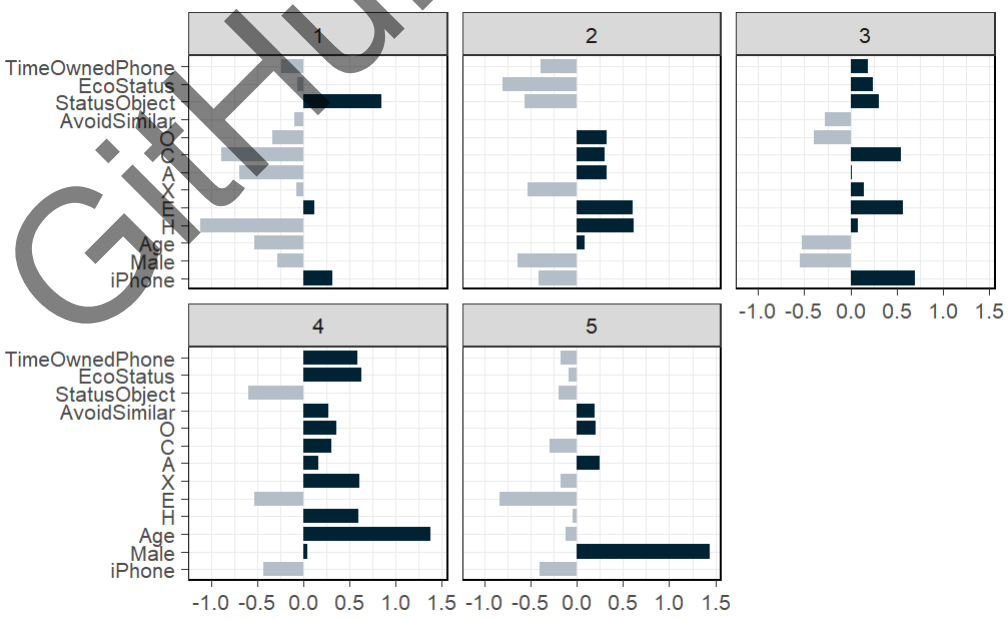

Fig. 13. Characteristics of clusters

Classification

To predict whether a user is an iPhone or Android user by their characteristics, we need to build a classification model.

Variables Sets

We will use all features in the data except the operating system factor for modelling and reduce the number of features based on the results of descriptive statistics and clustering analysis.

According to the descriptive analysis and clustering process, females tend to use iPhone while males prefer the Android. Younger generations prefer iPhone. The patterns of HECO in HEXACO factors' distributions are different between different systems. iPhone users are much more likely to neglect what smartphones most people use. iPhone users tend to consider their mobile phone as a status object. Therefore, we need to take Gender, Age, H, E, C, Avoidance Similarity and Phone as Status Object into consideration.

We also fit all the data to the logistic regression model, and according to the p-value, significant variables include Gender, Honesty-Humility, Avoidance Similarity and Phone as Status Object. Thus, we will test two more variable sets: Intersection (Gender, H, Avoidance Similarity and Phone as Status Object) and Union (Gender, Age, H, E, C, Avoidance Similarity and Phone as Status Object)[1].

Modelling and Evaluation Methods

We use the Logistic Regression method, k-Nearest Neighbours method, Naïve Bayes method, Random Forest method and Support Vector Machine method for classification. We train and test these models using the same randomly generated train and test data[2]. Except for the Logistic Regression model, other models are automatically tuned by the algorithms[3]. For each method, we select the model with the highest AUC[4] value.

Logistic Regression (LR)

Our best Logistic Regression model[5] is:

$$P(iPhone) = 0.47 - 0.89 * Gender - 0.01 * Age - 0.49 * H + 0.06 * E + 0.32 * C - 0.22 * AS + 0.63 * SO,$$

suggesting that iPhone users are likely to be female and have a low level of Honesty-Humility and a high level of Conscientiousness, they are less concerned about what mobile devices most people use, and they tend to consider their phones as status objects. Age and Emotionality show less influence on predicting the user type.

[1] In this case, the Intersection data set is exactly the data set suggested by the logistic regression, and the Union data set is exactly the data set suggested by the descriptive statistics.

[2] We take 75% of the data as train data, and the rest as test data.

[3] We use the cross-validation method to tune the models.

[4] Area Under the Curve, the trade-off between sensitivity and specificity, the higher the better.

[5] The logistic regression model calculates the probability of being an iPhone user. Gender: 1 is Male and 0 is Female. H: Honesty-Humility. AS: Avoidance Similarity. SO: Phone as Status Object.

k-Nearest Neighbours (kNN)

The k-Nearest Neighbours method does not produce a model for classification, which prevents us from understanding the classification process and is one of its disadvantages. The best number of neighbours chosen by the algorithm is ten (Table 23).

Naïve Bayes (NB)

As we already explored in the descriptive statistics, the variables are not at the risk of autocorrelation, which meets the independence assumption of the NB method. If this assumption holds, the NB method will outperform other methods.

Random Forest (RF)

The Random Forest method is a combination of many decision trees. In our best Random Forest model (Table 5), there are 100 decision trees are built and combined. Although the Decision Tree model is quite clear, the model of Random Forest is also hard to visualise and interpret.

Support Vector Machine (SVM)

Since the methods above do not perform well (AUC value all under or around 0.7), including the Random Forest method that usually performs well on large data sets, we assume that the data set is small, suggesting the SVM method may fit well. However, this method also has poor predictive power (Table 5).

Comparison of Classification Models

Among the three data sets we used to fit classification models, the Union set shows the strongest prediction power (Table 6), demonstrating that the data characteristics observed by our descriptive statistics are meaningful. Also, the Intersection set shows the poorest prediction power (Table 6), suggesting the inadequacy of focusing only on the significant variables in the logistic regression.

The best model is the NB model trained using the Union data set in terms of the AUC value (Table 5), suggesting that the variables in the Union data set are independent. Moreover, the Random Forest method is usually the best unless the training data is small; on the contrary, the NB model works well on small training data sets. Thus, we consider this data to be a small data set that suitable for applying the NB method.

However, as with all other models, the best ones have difficulty in accurately predicting Android users (Table 20). These models all tend to predict users as iPhone users, probably because Android users are more diverse and include users similar to iPhone users due to the large number of Android phone manufacturers.

Discussion

Our research finds that iPhone and Android users are significantly different in Gender, Age, H, E, C, Avoidance Similarity and Phone as Status Object. Specifically, compared to Android, iPhone users are younger, are more likely to be female, have a low level of Honesty-Humility, have a high level of Emotionality, have a low level of Conscientiousness, are less concerned about what smartphone most people use, and are more likely to consider their phones as status objects.

Limitations

- The data do not contain the information about phone manufacturers of Android phones, so we cannot tell if some Android phone brands have similar users to iPhone users.
- We only tested one k-value in the k-means approach based on the Elbow chart. However, the k-value corresponding to the Elbow point is not necessarily the best. Testing multiple k-values and comparing metrics such as the Silhouette Coefficient may yield better clustering results.
- We only experimented with three combinations of independent variables for the classification model, for which there may be better combinations.

Conclusion

Overall, users of different smartphone systems differ in gender, age and personality, which can divide smartphone users into natural clusters. We experimented with several classification models, and the results showed that the best model was formed using the independent variables Gender, Age, H, E, C, Avoidance Similarity and Phone as Status Object using the Naïve Bayes method. There are still some limitations to the data we collected and the method we used, such as lack of smartphone manufacturer data, experimenting with only one k-value in the k-means method and only trying three combinations of independent variables in the classification model.

Appendix

Table 4

Variables and their details and explanations in the original data

Variable	Details	Explanation
Smartphone (operating system)	iPhone	
	Android	
Gender	Female	
	Male	
Age		
Honesty-Humility		This is a personality assessment called HEXACO, with a higher score representing a higher level.
Emotionality		
Extraversion		
Agreeableness		
Conscientiousness		
Openness		
Avoidance Similarity		People's concern about avoiding the kind of smartphone the majority of people are using, with a higher score representing a higher level.
Phone as status object		People's concern about considering their smartphone as a status object, with a higher score representing a higher level.
Socioeconomic status		A higher score represents a higher level of socioeconomic status.
Time owned current phone (months)		

Table 5

Evaluations of classification models in test data (the asterisk following the method name means it is the best model among models built using this method)

Features	Method	Accuracy (%)	AUC (%)
All	LR	67.42	67.39
All	kNN*	63.64	64.26
All	NB	65.91	69.41
All	RF	65.91	64.94
All	SVM	66.67	64.47
Intersection	LR	62.88	66.05
Intersection	kNN	63.64	61.68
Intersection	NB	65.15	67.41
Intersection	RF	66.67	63.58
Intersection	SVM	55.30	62.50
Union	LR*	68.18	68.23
Union	kNN	61.36	62.92
Union	NB*	67.42	70.54
Union	RF*	67.42	67.79
Union	SVM*	65.15	66.02

Table 6

Average evaluation indexes for different variable sets

Features	Average Accuracy (%)	Average AUC (%)
All	65.91	66.09
Intersection	62.73	64.24
Union	65.91	67.10

Table 7

Average evaluation indexes for different models

Features	Average Accuracy (%)	Average AUC (%)
LR	62.88	62.95
kNN	66.16	67.22
NB	66.16	69.12
RF	66.67	65.44
SVM	62.37	64.33

Table 8

Confusion matrix: All - LR

Predicted \ Actual	Android	iPhone
Android	30	17
iPhone	26	59

Table 9

Confusion matrix: All - kNN

Predicted \ Actual	Android	iPhone
Android	27	19
iPhone	29	57

Table 10

Confusion matrix: All - NB

Predicted \ Actual	Android	iPhone
Android	24	13
iPhone	32	63

Table 11

Confusion matrix: All - RF

Predicted \ Actual	Android	iPhone
Android	24	13
iPhone	32	63

Table 12

Confusion matrix: All - SVM

Predicted \ Actual	Android	iPhone
Android	26	14
iPhone	30	62

Table 13

Confusion matrix: Intersection - LR

Predicted \ Actual	Android	iPhone
Android	23	16
iPhone	33	60

Table 14

Confusion matrix: Intersection - kNN

Predicted \ Actual	Android	iPhone
Android	28	20
iPhone	28	56

Table 15

Confusion matrix: Intersection - NB

Predicted \ Actual	Android	iPhone
Android	24	14
iPhone	32	62

Table 16

Confusion matrix: Intersection - RF

Predicted \ Actual	Android	iPhone
Android	26	14
iPhone	30	62

Table 17

Confusion matrix: Intersection - SVM

Predicted \ Actual	Android	iPhone
Android	19	22
iPhone	37	54

Table 18

Confusion matrix: Union - LR

Predicted \ Actual	Android	iPhone
Android	30	16
iPhone	26	60

Table 19

Confusion matrix: Union - kNN

Predicted \ Actual	Android	iPhone
Android	24	19
iPhone	32	57

Table 20

Confusion matrix: Union - NB

Predicted \ Actual	Android	iPhone
Android	25	12
iPhone	31	64

Table 21

Confusion matrix: Union - RF

Predicted \ Actual	Android	iPhone
Android	26	13
iPhone	30	63

Table 22

Confusion matrix: Union - SVM

Predicted \ Actual	Android	iPhone
Android	26	16
iPhone	30	60

Table 23

Best k-value for kNN

Features	Best k-value
All	10
Intersection	10
Union	10

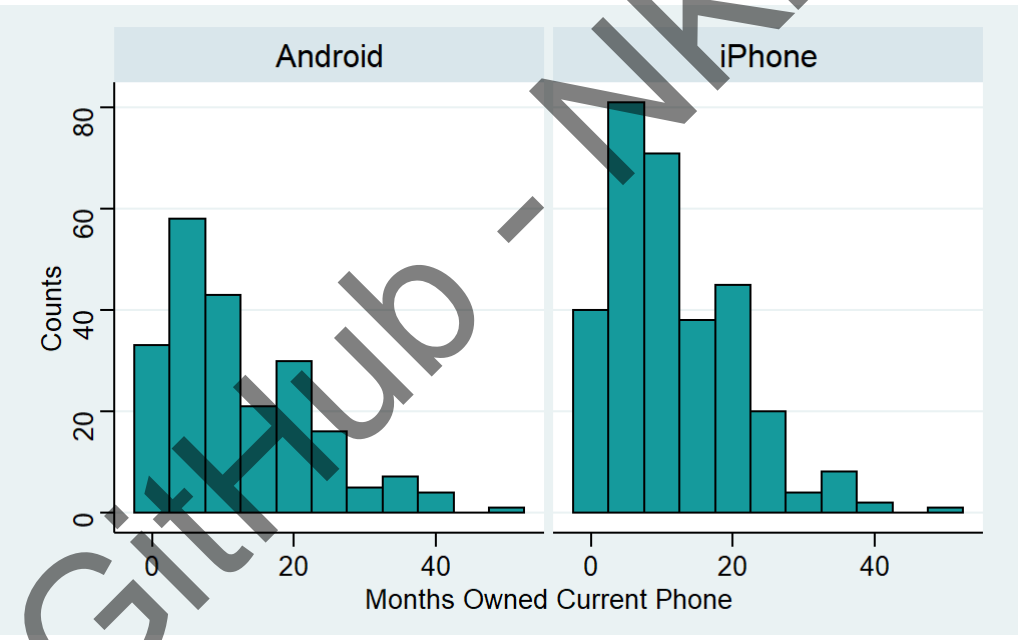

Fig. 14. Distribution of Months Owned Current Phone in Android and iPhone users (exclude the outlier)

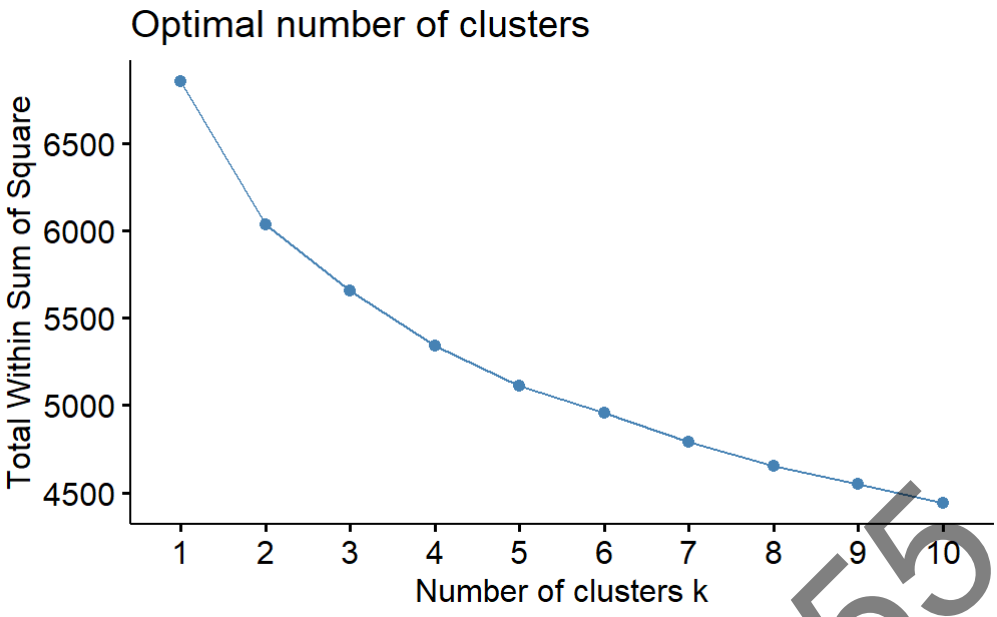

Fig. 15. Elbow chart for the k-value (slope decreases after 5 clusters)

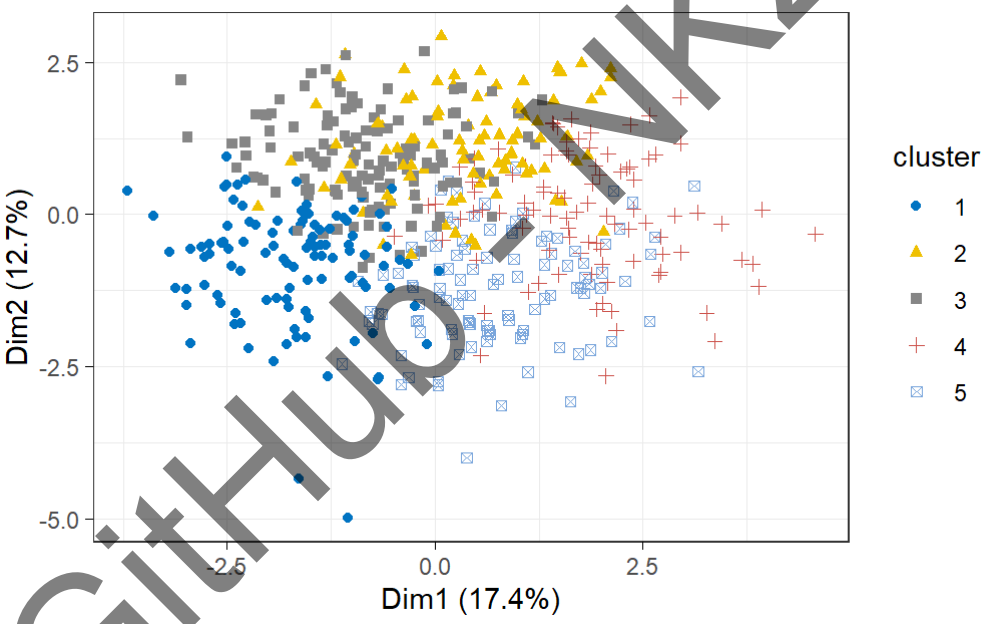

Fig. 16. Clusters created by the k-means method

Bibliography

Ross, P.E., 2011. Top 11 technologies of the decade. *IEEE Spectrum*, 48(1), pp.27–63.

Ofcom, 2019. *The Communications Market 2019* [Online]. Available from: https://www.ofcom.org.uk/research-and-data/multi-sector-research/cmr/cmr-2019 [Accessed 14 April 2022]

IDC, 2021. *Smartphone Market Share* [Online]. Available from: https://www.idc.com/promo/smartphone-market-share [Accessed 14 April 2022].

Chen, X., Wang, Y., Tao, D., Jiang, L. and Li, S., 2021. Antecedents of smartphone multitasking: roles of demographics, personalities and motivations. *Internet Research*, 31(4), pp.1405–1444.

Clayton, R.B., Leshner, G. and Almond, A., 2015. The Extended iSelf: The Impact of iPhone Separation on Cognition, Emotion, and Physiology. *Journal of computer-mediated communication*, 20(2), pp.119–135.

Forbes, 2014. *What Kind Of Person Prefers An iPhone?* [Online]. Available from: https://www.forbes.com/sites/toddhixon/2014/04/10/what-kind-of-person-prefers-an-iphone [Accessed 14 April 2022]

Shaw, H., Ellis, D.A., Kendrick, L.R., Ziegler, F. and Wiseman, R., 2016. Predicting Smartphone Operating System from Personality and Individual Differences. *Cyberpsychology, Behavior and Social Networking*, 19(12), pp.727–732.

Ashton, M.C. and Lee, K., 2009. The HEXACO-60: A Short Measure of the Major Dimensions of Personality. Journal of personality assessment, 91(4), pp.340–345.