Mobile Price Classification & Factor Evaluation

Süleyman Buğra Gülsoy
Department of Statistics
Middle East Technical University
Ankara, Turkey
bugra.gulsoy@metu.edu.tr

Abstract—**Mobile phone prices are fluctuating market. That everyone tries to take advantage of. In the research paper mobile prices and some attributes that might have an impact on are examined. The relationship between them is explained and important features are subtracted. The research aimed to explain those relationships to create awareness. And propose a method to successfully classify mobile prices. In the study multivariate methods are used to evaluate and test. And resulted the there exit three main factors "camera", "screen" and "pixel to explain the prices of mobile phones. Also, in the study it is found that the multinomials principal component logistic regression and LDA successfully classifies the mobile prices. It is aimed that with this study individuals can avoid getting exploited with over-underpriced transactions.**

Keywords—EDA, HotellingT, MANOVA, PCA, FA, LDA, CLUSTERING, MCD

I. INTRODUCTION

Mobile phone trading is a rapidly expanding and important industry. Especially for developing countries and the countries which accessing technology is expensive. Prices for second-hand mobile phones are diverse and are exposed to fluctuations. Many factors might influence these prices. Hence, determining the price of a second-hand mobile phone is a complex process which can be exploited. In the study we aim to reveal the factors that impact mobile phone prices and explain the relation between some features of mobiles and price. Therefore, individuals' awareness of these factors and features can be increased. Moreover, the study aims to provide a model to classify the mobile prices. With the motivation of individuals can utilize the model to classify their devices into their corresponding price segments to avoid any exploitation. The findings of the study can not only be used for individuals who want to classify their devices but also as a method to determine the optimal price range for new mobile products. In addition, the study can aid in revealing the under-overpriced products.

II. DATA PREPROCESS AND PREPARATION

A. *Data Description*

The data consists of information about mobile phone features and prices. Data is shared multiple times in Kaggle from different people without any provenance or resource. Hence, the anonymously sourced data is available at "https://www.kaggle.com/datasets/iabhishekofficial/mobile-price-classification". The data has in total 21 variables. Six of them are binary categories, one of them is a target variable with four categories and fourteen numerical features. The binary categories are namely as;
blue: Identifier of Bluetooth, dual_sim: Identifier of dual sim support, four_g: Identifier of 4G technology, three_g:

Identifier of 3G technology, touch_screen: Identifier of touch screen availability, wifi: Identifier of wifi availability
The fourteen numerical features are namely as;
battery_power: The amount of electrical energy stored in the battery, clock_speed: The number of instructions the processor can execute per second, fc: Megapixel of front camera, int_memory: Memory of device, m_dep: Depth of device, mobile_wt: Weight of device, n_cores: Processor cores, pc: Megapixel of back camera, px_height: Resolution height of pixel, px_width: Resolution width of pixel, ram: Random access memory, sc_h: Height of phone screen, sc_w: Width of phone screen, talk_time: Time of full charged battery to deplete with only talking
And the four-class target variable is "price_range" from zero to three {0,1,2,3} each class represents a price segment from cheapest to expensive.

B. *Exploratory Data Analysis*

We can observe from the Fig. 1. that the data does not suffer significantly from correlation problems. There exist only three combinations for variables that are meaningfully correlated with each other. Eveny they also do not exceed a problematic limit (0.8) of correlation. The correlation matrix of numeric features indicated such low correlated data can be used well into principal component analysis and factor analyses.
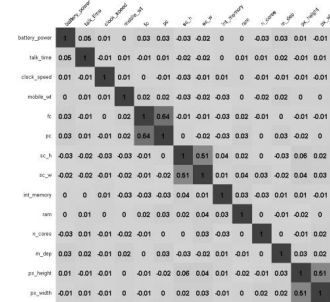

Fig. 1.   *Corrolation Plot*

The target variable of the data price range can be examined in the Fig 2. The price range itself is distributed evenly among its levels. Each level has the same proportion of data. On the other hand, the relationship between numerical features and target variable price range is quite various. Some features show no change in median values as can be observed from Fig. 2 for different levels of price range. But some numerical features affect price range. Those variables included battery power, pixel height, pixel weight and ram. The median values of these features are different amongst different price levels. Hence, those valuebsles might conveying a valuable information. Moreover the Fig.2. signal that there might be

outliers in the data but with slight margin. The grahs give brief information about the data but furhet analyses is ineeded to drow staitstical inferecnes and resuts.
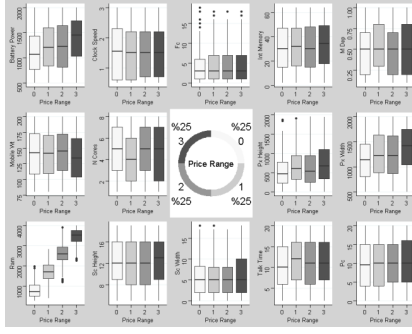

Fig. 2. *Numeric Feature by Price Range*

C. *Missing Value Analysis*

The data does not have any missing information.

D. *Outlier Analysis*

The outlier values create great risk for multivariate analysis. As, majority of multivariate methods utilize some kind of distance measurement in their algorithm. And outlier values significantly alter the distance measurements. To avoid the misleading effect of outliers in the analyses, they should be managed. Mahalanobis distance measurement is the most common method to detect multivariate outliers. Although, it conveys a paradox. The method is being used to detect outliers which the method itself is also affected by. The mahalanobis distance takes the covariance structure of data into account. Thus, the outlier values might disturb the mean and covariance of data which makes mahalanobis distance unreliable. Many resources propose using a robust mahalanobis distance. The robust mahalanobis distance utilizes "MCD" minimum covariance determinant, instead of using the covariance structure of data directly. The MCD grants estimates of covariance matrix and mean of data such that the influence of anomalies is minimized. Which makes MCD robust estimator for multivariate covariance and mean. In the project the proposed robust mahalanobis distance is used to detect outliers. The distance for every data point is calculated and compared with the threshold score corresponding to the 0.05 p-value (23.68). Forty-two observations that found exceeding the cutoff value of 23.68, are being removed. The removal of outliers is not recommended most of the time. However, the presence of such values poses a risk for all the applications of study. And the observations that are marked as outliers are only 2.6% of all observations. Thus, the removal of such observations does not lead to huge information loss.

E. *Standardization*

Data has features that do not share common scale. The difference between units of variables causes problems with machine learning algorithms and most of the multivariate methods. The problem is the variable with greater unit draws more emphasis on itself even if it does not convey valuable information. This problem can be solved by adjusting all the features into common, uniform units. The process is called scaling. The two most common scaling methods are z-score scaling (standardization) and min-max scaling

(normalization). The standardization method is preferable for the principal components analyses which will be conducted. Thus, standardization is chosen among the two methods and applied to the data. All the numerical features are scaled with their center and standard deviation such that,

$$(1) \quad x' = \frac{x - \mu_x}{\sigma_x}$$

Features of data shares common mean of zero and variance one after the scaling. Which makes integrations of PCA and model building applicable.

F. *Principal Component Analysis*

Principal component analyses transform the numerical features into orthogonal components which is linear combination of original features. The method ensures that multicollinearity is not present, and It reduces the dimensionality of data to make it easier to interpret. For applying the method, the categorical variables are subtracted from the data. Since the PCA can only be applied to numerical features. The remaining data which has only numerical variables has fourteen dimensions. So, PCA created in total fourteen components.
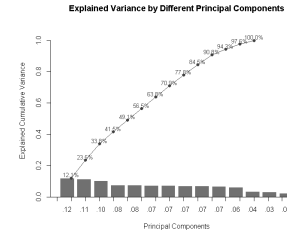

Fig. 3. *Explained Variances by Components*

The total and cumulative variability explained by each component can be observed in Fig.3 the bars represent the variability explained by each component and line represents the cumulative variability. We can observe from figure that 10 components are needed to explain enough variability. The ten components successfully explain the 84% of total variability. The components explain somehow similar amount of variability as can be observed from bars at Fig.3. The sufficient 10 components were extracted to further pursue analyzing. Each component is a linear combination of already existing 14 features, and they are all orthogonal to each other. In other words, they are linearly independent and correlation between components are zero (see Appendix 4). The dimensions were successfully reduced from fourteen to ten with only losing 16% of the information. The original variables contribute differently to the 10 components. This contribution can be observed from the biplots (see Appendix5). When the first components are examined, it can be observed that "pc", "fc", makes a similar contribution to the first three principal components. Also, "width", "px_height" and "sc_w", "sc_h" makes the exact similar contribution. These similarities are very natural. The three pairs pc-fc, px_width- px_heigh, sc_w-sc_h was found correlated. Thus, they contain similar information. The six variables have the highest contribution for the first three components. It can already be observed that those three pairs behave similarly and can be grouped into newly variables as "camera", "screen" and "pixel". This information will be rather more valuable when conducting the factor analysis.

The principal components can also be used for classification purposes (see Appendix6). Using the first two components is not optimal. Classification with all the components will be discussed later. As a result, the PCA successfully reduced the dimensionality from 14 to 10, solved the correlation problem between three pairs of variables. Explored the potentially important variables to explain structure of data and created linearly independent variables to be used later in linear model.

G. Factor Analysis and Factor Rotation

Factor analysis is also a method for dimensional reduction like PCA. Yet, it distinguishes from PCA in a way that FA does not create new variables, it combines the correlated variables into groups with the major goal being explaining the relationship between several variables. Thus, FA can be expressed as variable clustering method. Before applying FA, a statistical test needs to be conducted to determine whether the data is suitable for such analysis. To check it, Kaiser-Meyer-Olkin test (KMO) is conducted (see Appendix7). The overall MSA value from KMO test is found to be 0.5 which indicates FA is appropriate for the data. To determine the number of factors, parallel analysis scree plot is conducted (see Appendix8). The number of factors can be determined by looking at the elbow shape from the scree lot which indicated three factors are enough. Or the parallel analyses compare the eigen values of data with the randomly generated eigen values and decide whether a specific eigen value is from a meaningful factor. The analyses also show similar result such that three factor is enough. Following that maximum likelihood factor analyses are conducted to statistically prove whether three factor is enough. The p-value of test Is obtained as 0.75 which leads to do rejection of null hypothesis and statistically prove that 3 factor indeed adequate. Calculation of the factors can be done with orthogonal or oblique rotations. The difference between them and the selection of the better of two is very complex topic. After analyzing the results from two rotations solutions were almost equal. And the choice is made for acquiring a simpler structure. The oblique rotation gives a simpler solution when the loading for two rotations is being compared (see Appendix9). Also factor correlation matrix obtained from oblique rotation does not signal any critical value ±0.32 proposed by Tabachnick & Fidell in 2007. Thus, the oblique rotation is preferred for analysis. The obtained three factors successfully explains the 28% of the total variation in the data. The number does not seem much but when we consider that the number of total variables decreased from 14 to 3, and by using only 3 factor 28% of total variability can be explained than the analyses prove its significance. It can be observed from the loading matrix (see Appendix9) only 2 variables contribute to each factor. So, each factor is just a combination of 2 variables. Which was indicated in the Principal Component Analysis, the result from factor analysis is complementary and indicated that first factor is "fc" and "pc" second factor is "px_height" , "px_width" and third factor is "sc_h" and "sc_w". Hence a parallel conclusion can be made that those three pairs behave similarly and can be grouped into newly variables as "camera", "screen" and "pixel" (see Appendix10). Cronbach's alpha value is checked for each factor to measure internal consistency. The value for

all three factors were greater than .70 indicates variables are overall consistent within the factor. Also, when the correlation plot (see Appendix11) of factor is examined it can be observed that the factor is indeed uncorrelated. Thus, factor analysis successfully reduced the dimensionality of numerical features from fourteen to three. And the resulting uncorrelated factor might be further used for regression analysis.

III. Results and findings

A. Comparisons Of Several Multivariate Means

The main research interest is identifying the features and factor who has a significant impact on the mobile phone' price range and coming up with a model to explain it. After completing the necessity preprocessing. The research can be pursued. Firstly, for the study to have meaningful goal. There should be a difference among different price ranges. The price range has 4 levels. Thus, the multivariate mean of those 4 levels should be tested. To create such test MANOVA should be conducted. Since we are dealing with 4 levels, we compare multivariate means at each level. The statistical hypothesis is:

$$(2) \quad H_0: \mu_0 = \mu_1 = \mu_2 = \mu_3$$

With the alternative being at least one means is different than other. The motivation of the test is deciding whether there exists a meaningful price range. MANOVA has its two critical assumptions. One it assumes multivariate normality and two it assumes each group shares a common covariance matrix. First, the multivariate normality assumptions should be tested. To test multivariate normality Henze-Zirkler's multivariate test was conducted (see Appendix12). The result shows that none of the groups have multivariate normality. To aid the problem log transformation was applied to each feature. Then the HZ test was re-conducted on the log transformed data. The problem of non-linearity was still present even after the log transformed data. But the sample size for each level was large enough to relax the multivariate normality assumptions by Central Limit Theorem. Secondly, the common variance assumption was tested using the Box's M test for multivariate equality of var-cov structure. The p-value of the test was found to be greater than .05 and test resulted proving the equality of common var-cov structure. Thus, the assumptions are satisfied for multivariate analyses of variance. The hypothesis given (2) tested using MANOVA. The test gave a p-value less than .05 and it is concluded that at least one of the price ranges has different multivariate means. But the test is conducted with the means of all features. So, it is inconclusive that which variables mean are different among price ranges to have such claim a post-hoc analysis is conducted testing difference in means for price ranges for each numerical variable separately. The analysis showed that the mean of "battery_power", "int_memory", "px_height", "px_width"," ram" and "pc" is different among price ranges. Hence, these variables hold significant information about price ranges.

B. Inferences About Mean Vector

The lacking point about the MANOVA analyses is that we concluded a specific set of variables have different means for

different price ranges. Also, we concluded price ranges are different overall. But which price ranges create this distinction was not known. To further analyze it. We compared each multivariate mean for 4 levels with overall mean. In other words, we conducted a Hotelling T test for hypothesis of each four-treatment effect τ_i. The Hotelling t test had two assumptions multivariate normality and common covariance matrix. In the previous part it was showed that price range groups have common covariance matrix and due to the big sample size, we could relax the normality assumption with CLT. Each multivariate treatment mean is tested with a numerical mean vector.

$$(3) \quad H_0: \tau_i = 0$$
$$(4) \quad H_0: \mu_i = \mu$$

The Hotelling T test showed that each level of price range conveys a valuable information (p<.05) thus significant. Recall that post-hoc analysis from MANOVA showed mean of "battery_power", "int_memory", "px_height", "px_width"," ram" and "pc" are different among price ranges. But it was unclear which levels of price ranges has significant different mean for each variable. To test the hypothesis a T test is conducted. Null hypothesis being that for a given a single variable the mean of i'th level is equal to overall mean of that variable. This T test is conducted at each 4 levels of price range for every variable that is found significant in MANOVA post-hoc. In total 24 T tests are conducted. Resulted that, battery power of price range 0 and price range 4 is significantly different than overall battery power. Whereas price range 1 and 2 does not have significantly different battery power. Only the price range 4 is found having significantly different device memory. Similarly for pixel height and weight price ranges 1 and 4 have significantly different values than overall mean. Surprisingly, the ram values of each price range are significantly different than mean ram value. The different price ranges are examined for all the variables. For the second emphasize of the study the predictive model for price range was created.

C. Principal Components Regression

The target variable price range was multi class categorical variable thus a multinomial logistics regression was conducted using orthogonal principal components. Thus, for the multinomial regression target variable was price range whereas exploratory variables were the ten principal components obtained before. The model was used to predict the "test" data set. To avoid data leakage the data was splinted into a train test at the beginning of the analyses. All the preprocessing analyses and modelling was done to the train data. The same preprocessing is later applied to the test data using the values obtained from train. The model correctly classified 93% of the test observations.

TABLE I.   CONFUSION MATRIX OF PCA REGRESSION

Act./Pred.	0	1	2	3
0	95	2	0	0
1	5	94	6	0
2	0	4	90	4
3	0	0	4	96

D. Discrimination And Classification

Fisher discriminant analyses were conducted to classify and discriminate the data into 4 different price range groups. The LDA finds a linear function such that it discriminates the data into desired groups regarding the target variable. The LDA was applied to the train data (see Appendix13). Found that the prior probabilities of all groups are equal (0.25). Since the target had four levels to classify LDA produced three linear discrimination functions. Each function classifies whether a data point is from the level "i" or from the reference level. One of the category levels is assigned as "reference" the other three levels are compared with the reference level regarding their own linear discriminant functions. Each linear function can be obtained as coefficients of linear discriminants from appendix13. The performance of LDA was evaluated with prediction of test data. The model correctly classified 93% of the test observations. Similarly, to the PCA model.

TABLE II.   CONFUSION MATRIX OF LDA

Act./Pred.	0	1	2	3
0	96	2	0	0
1	4	93	8	0
2	0	5	92	6
3	0	0	0	94

E. Clustering

The clustering algorithm does not require a target variable instead it is a natural process of grouping the observation based on their similarity or distance. All the analysis that has being done was based on the fact that the target variable had 4 categories. To validate whether using such categories aligns the nature of the data clustering was conducted. The clustering algorithms were divided into two main groups hierarchical and non-hierarchical. The hierarchical cluster is more advantageous in terms of cluster interpretability. Thus, it is preferred. The agglomerative hierarchical clustering we applied to the data using Euclidean distance and complete linkage. The result Fig.4 shows that the data can be naturally classified into three groups.
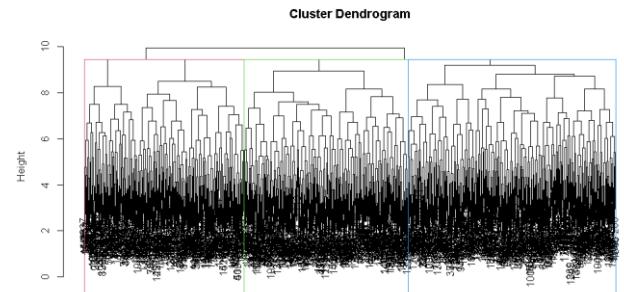

Fig. 4.   Hierarchical *Clustering Dendogram*

IV. CONCLUSION

In this research paper two main research interests were describing the relation between features of mobile phones,

mobile phone prices and proposing method to classify price range of mobile phones. The study employed PCA, FA method to describe the relationship. Found that three main factors "camera", "screen", and "pixel" have the highest impact on mobile prices. Thus, for one to evaluate or group the mobile phone these factors need to be considered. Also, statistical tests are conducted to determine validity of price ranges. Found that four prices ranges are all significant. The research paper utilized multivariate methods for all analyses. According to the analyses, in addition to variables mentioned above, the batter power, memory and ram also found important for determining the price range. When all things consider individuals should be aware of this parameter to avoid being exploited. Also, the research paper proposed two methods to classify the price range of mobile phones. Both methods (PCA multinominal logistic regression and Fisher's Linear Discriminant) is found extremely helpful at classifying the price range both methods predicted the test observation with 93% accuracy. Thus, it can be concluded that any of the methods can be used to classify the price range for a mobile phone. For further work, it could be helpful to repeat the study with different data sources to validate the conclusions.

Appendix 2

Appendix 3

V. REFERENCES

M. Durairaj and S. Sivagowry, "A Pragmatic Approach of Preprocessing the Data Set for Heart Disease Prediction"

S. Gopal Krishna Patro, Kishore Kumar sahu, "Normalization: A Preprocessing Stage"

Bhavisha Suthar, Hemant Patel and Ankur Goswami, "A Survey: Classification of Imputation Methods in Data Mining"

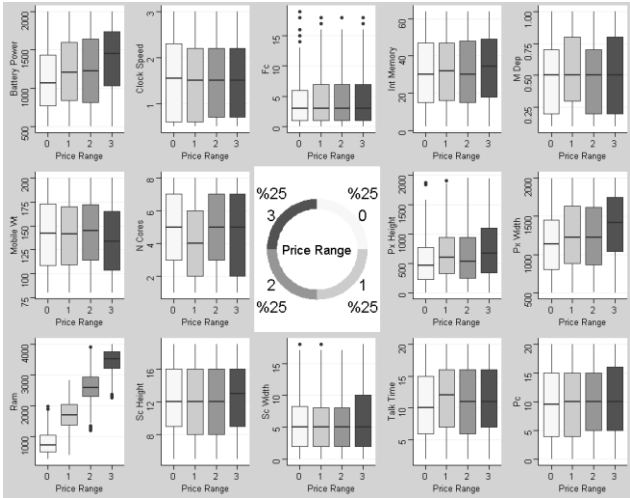Sarab AlMuhaideb, "An Individualized Preprocessing for Medical Data Classification"

M. Schumacher, R. Robner and W. Vach, "Neural networks and logistic regression,"
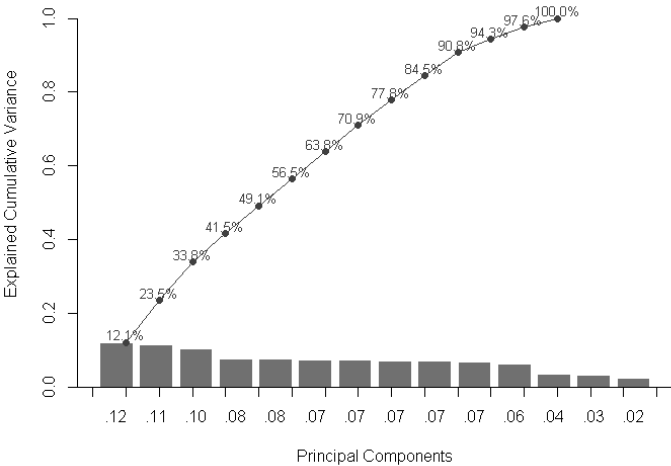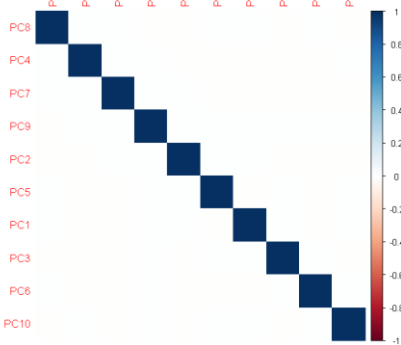
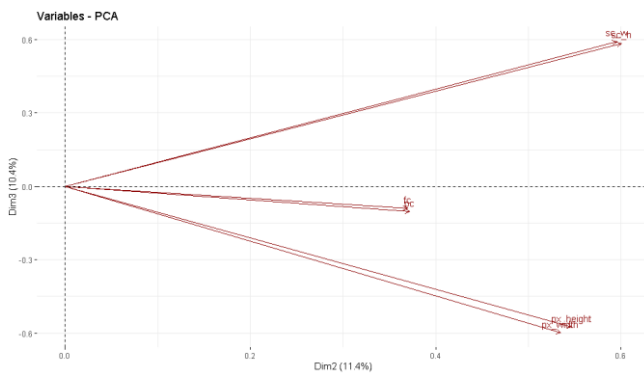## VI. APPENDICES

Appendix 1

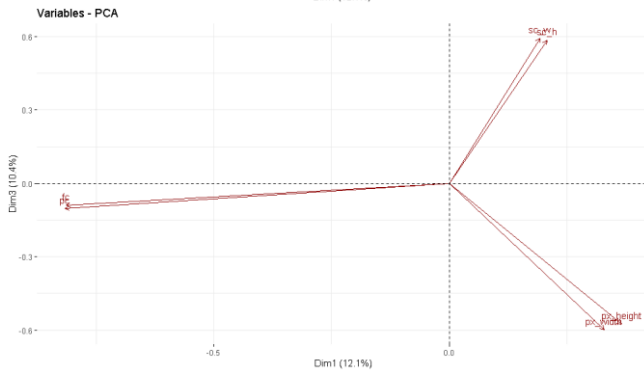**Explained Variance by Different Principal Components**

Appendix 4

Appendix 5

Variables - PCA

Variables - PCA
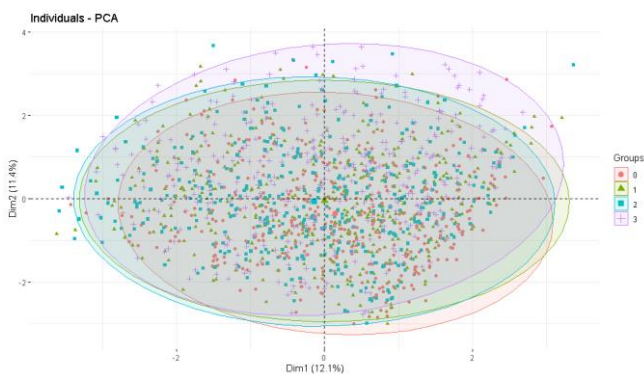

Variables - PCA

Appendix 6
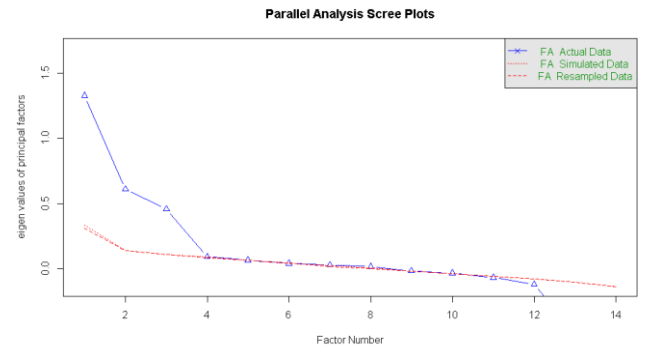

Individuals - PCA
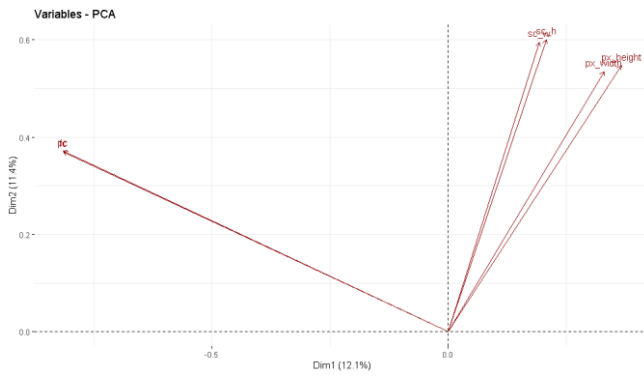
Appendix 7

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cm)
Overall MSA =  0.5
MSA for each item =
battery_power   clock_speed           fc   int_memory       m_dep
         0.56          0.49         0.50         0.46        0.37
    mobile_wt       n_cores           pc    px_height    px_width
         0.56          0.47         0.50         0.50        0.50
          ram          sc_h         sc_w    talk_time
         0.60          0.50         0.50         0.48
>
```

Appendix 8

Parallel Analysis Scree Plots

Appendix 9

```
Call:
factanal(x = non_cat_data, factors = 3, scores = "regression",     rotation = "varima
x")

Uniquenesses:
battery_power   clock_speed           fc   int_memory       m_dep
        0.996         1.000        0.572        0.997       0.997
    mobile_wt       n_cores           pc    px_height    px_width
        0.998         1.000        0.005        0.005       0.716
          ram          sc_h         sc_w    talk_time
        0.996         0.005        0.734        0.999

Loadings:
              Factor1 Factor2 Factor3
battery_power
clock_speed
fc             0.631  -0.127   0.117
int_memory
m_dep
mobile_wt
n_cores
pc             0.965  -0.167   0.190
px_height     -0.214           0.971
px_width      -0.107           0.522
ram
sc_h           0.170   0.983
sc_w                   0.511
talk_time

               Factor1 Factor2 Factor3
SS loadings     1.429   1.282   1.270
Proportion Var  0.102   0.092   0.091
Cumulative Var  0.102   0.194   0.284

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 44.74 on 52 degrees of freedom.
The p-value is 0.752
```

```
Call:
factanal(x = non_cat_data, factors = 3, scores = "regression",     rotation = "proma
x")

Uniquenesses:
battery_power   clock_speed           fc   int_memory       m_dep
        0.996         1.000        0.572        0.997       0.997
    mobile_wt       n_cores           pc    px_height    px_width
        0.998         1.000        0.005        0.005       0.716
          ram          sc_h         sc_w    talk_time
        0.996         0.005        0.734        0.999

Loadings:
              Factor1 Factor2 Factor3
battery_power
clock_speed
fc             0.654
int_memory
m_dep
mobile_wt
n_cores
pc             0.998
px_height              0.997
px_width               0.534
ram
sc_h                           0.997
sc_w                           0.516
talk_time

               Factor1 Factor2 Factor3
SS loadings     1.432   1.283   1.265
Proportion Var  0.102   0.092   0.090
Cumulative Var  0.102   0.194   0.284

Factor Correlations:
         Factor1 Factor2 Factor3
Factor1  1.0000  0.0227 -0.0557
Factor2  0.0227  1.0000 -0.0325
Factor3 -0.0557 -0.0325  1.0000

Test of the hypothesis that 3 factors are sufficient.
The chi square statistic is 44.74 on 52 degrees of freedom.
The p-value is 0.752
```
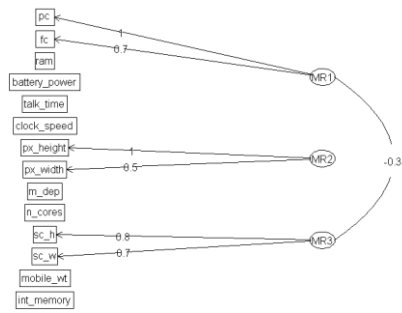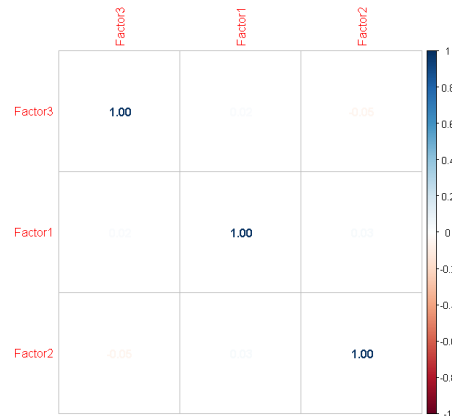
Appendix 10

Factor Analysis

Appendix 11

Appendix 12

```
$`0`
          Test          HZ p value MVN
1 Henze-Zirkler 1.020545        0  NO

$`1`
          Test          HZ      p value MVN
1 Henze-Zirkler 1.010632 4.062217e-08  NO

$`2`
          Test          HZ      p value MVN
1 Henze-Zirkler 1.017453 1.332268e-15  NO

$`3`
          Test          HZ      p value MVN
1 Henze-Zirkler 1.01574 1.864064e-13  NO
```

Appendix 13

```
Prior probabilities of groups:
        0         1         2         3
0.2509628 0.2490372 0.2503209 0.2496791

Group means:
   battery_power      blue  clock_speed  dual_sim          fc    four_g
0    -0.29110423 0.4961637  0.035800236 0.5140665 -0.09695252 0.5038363
1    -0.02995238 0.4871134 -0.085493000 0.5051546  0.02802966 0.5360825
2    -0.01846183 0.4794872  0.052063362 0.5128205  0.05086113 0.4769231
3     0.34098558 0.4987147 -0.002908276 0.5038560  0.01850151 0.5604113
    int_memory      m_dep   mobile_wt     n_cores          pc  px_height
0 -0.04273901 -0.02604404 -0.02120954 0.01942844 -0.09433185 -0.21810572
1  0.01272576  0.04161025  0.02812942 -0.10883597 0.01165720  0.03633254
2 -0.08443896 -0.03507952  0.06723324 0.05694283 0.01533180 -0.07590223
3  0.11492173  0.01984436 -0.07414460 0.03193865 0.06781841  0.25908530
     px_width        ram          sc_h        sc_w   talk_time   three_g
0 -0.19677284 -1.2380017 -0.007441824 -0.02186474 -0.071481969 0.7391304
1 -0.05174468 -0.4022911 -0.014994771 -0.03363945  0.065606638 0.7680412
2 -0.04646664  0.4291891 -0.078411324 -0.03327303  0.001770781 0.7717949
3  0.29598228  1.2153313  0.101049204  0.08888869  0.004636169 0.7814910
   touch_screen      wifi
0    0.5294118 0.4987212
1    0.5231959 0.5000000
2    0.4743590 0.5076923
3    0.4755784 0.5192802

Coefficients of linear discriminants:
                       LD1           LD2          LD3
battery_power  0.7106202826  0.363217906 -0.04891417
blue          -0.0486950445  0.103632017 -0.15555191
clock_speed   -0.0250931841 -0.222241425 -0.32413845
dual_sim      -0.0961081707  0.014843903  0.01910811
fc            -0.0017637373 -0.148748569  0.41524676
four_g        -0.0127326690  0.910293769 -0.29559563
int_memory     0.0551313772  0.354344983 -0.11843865
m_dep         -0.0124681288  0.116952678  0.13208016
mobile_wt     -0.1026441042 -0.200263060  0.28094726
n_cores        0.0201639377 -0.219618259 -0.41941722
pc            -0.0003277509  0.168143609 -0.11411368
px_height      0.4335559346  0.402947641  0.42685490
px_width       0.3652160127  0.160290736 -0.44203698
ram            3.3407746642 -0.205520776  0.03288326
sc_h           0.0315092433  0.232712721 -0.20658026
sc_w          -0.0127898779  0.003830543 -0.12781287
```