TEXAS TECH UNIVERSITY

Multivariate Analysis

ISQS 6350

Project Update 1: Titanic Dataset with R

Group members:

Nicholas Small

Gatimbirizo Daniel Mucyo

Olivio Ogbebor

December 5th, 2020

Introduction (Author: Gatimbirizo Mucyo)

The Titanic movie is one of the most eventful and beautiful romantic true life story movies that has been released in the past few decades. Most people have watched this movie multiple times and dream to find their Jack or Rose one day. The Titanic was one of the largest ships built at the time of the incident (1912). It was 175 ft tall, had 9 decks (A-G) and cruised at about 39km/hr. When the massive ship drowned after bumping into an iceberg, there were only lifeboats for 1178 people out of the 2224 people who were on the Titanic. In the end, research has it that only 706 people survived the Titanic event.

Scope of Project

For this project, we chose to analyze the Titanic dataset so that we can answer some very important question such as:

1. What were the chances of surviving the Titanic event?
2. What was the age distribution on the Titanic?
3. What was the correlation between the people that survived and their financial situation (ticket class)?
4. What is the relationship between the fare price and the passengers that survived the event?

Data description

The dataset was found on Kaggle and consists of 12 columns and 891 rows.

The columns in the dataset are described below:

1. PasengerId: Passenger identification number
2. Survived: *Survival* (0 = No;1 = Yes)
3. Pclas*:* Ticket Class (1 = 1st, 2= 2nd; 3= 3rd)
4. Name*:* Passenger name
5. Sex*:* Gender (Male or Female)
6. Age*:* Passenger Age
7. SibSp*:* Number of sibling and/or spouses aboard
8. Parch*:* Number of parent(s) and/or children aboard

9. Ticket*:* Ticket number
10. Fare*:* Fare price (British Pound)
11. Cabin*:* Cabin
12. Embarked*:* Port of embarkation (C = Cherbourg: Q = Queenstown: S=

However, for our project we will be interested in column 2,3,5,6,7 and 10

Survived <int>	Pclass <int>	Sex <dbl>	Age <dbl>	Parch <int>	Fare <dbl>
0	3	1	22.00	0	7.2500
1	1	0	38.00	0	71.2833
1	3	0	26.00	0	7.9250
1	1	0	35.00	0	53.1000
0	3	1	35.00	0	8.0500
0	3	1	*NA*	0	8.4583
0	1	1	54.00	0	51.8625
0	3	1	2.00	1	21.0750
1	3	0	27.00	2	11.1333
1	2	0	14.00	0	30.0708

1-10 of 891 rows Previous 1 2 3 4 5 6 ... 90 Next

Fig: Titanic Dataset

Data Cleaning and Visualizations (Author: Nick Small)

Regarding data cleaning, Column 5, which is the column specifying the gender of the passenger contains binary values(male/female) that are converted into numbers and used for multivariate analysis. The age column was cleaned for its missing values using a linear regression model in Python shown below.

```
def inpute_age(cols):
    Age = cols[0]
    Pclass = cols[1]
    if pd.isnull(Age):
        if Pclass == 1:
            return 37
        elif Pclass == 2:
            return 29
        else:
            return 24
    else:
        return Age

train['Age'] = train[['Age','Pclass']].apply(inpute_age,axis=1)
```

Figure 1: Linear Regression Model: Data Cleaning

Our final dataset consists of 6 variables: Survived, Pclass, Sex, Age, Parch and Fare. After plotting a scatter plot of our data, we realized that we have multiple outliers that needed to be removed.

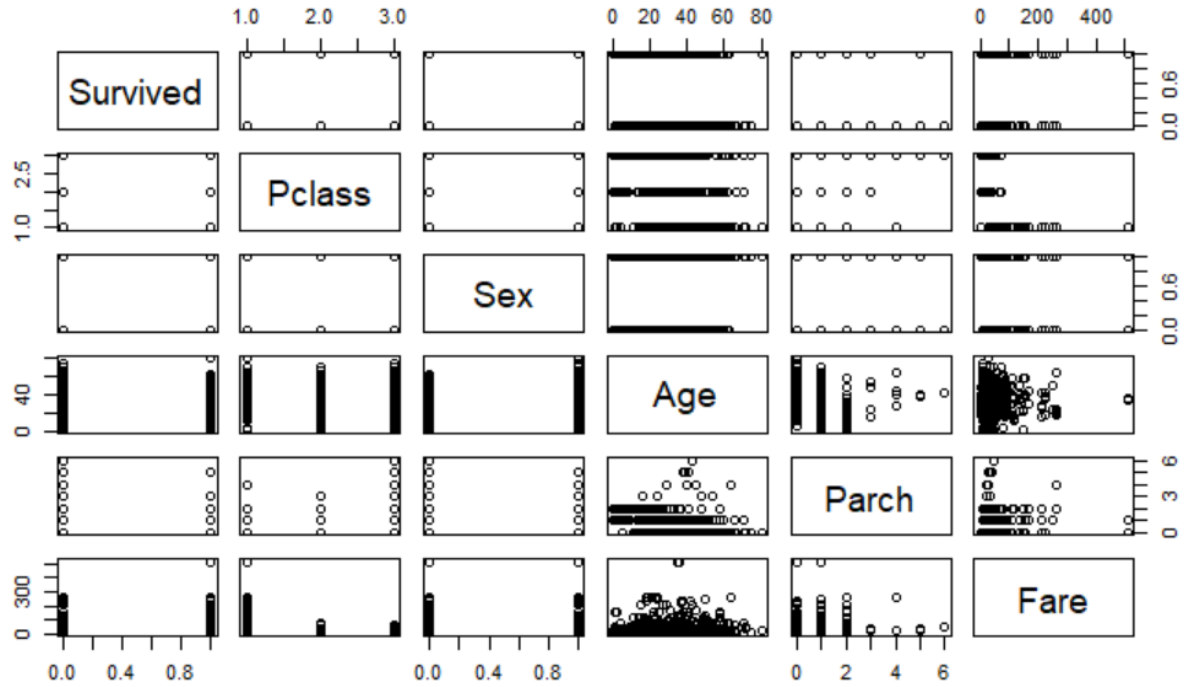

Fig: Scatter Plot Before Removing Outliers

Since we cannot identify each outlier by name, we will sort our data based on Mahalanobis distance, then remove those with the largest Mahalanobis distances. We will keep variables with the mahalanobis distance of less than 0.95. So, the cut point for our outliers is 0.95 quantile. After removing the outliers, the number of rows dropped from 891 to 846 rows.
Below is the new scatter plot after removing outliers

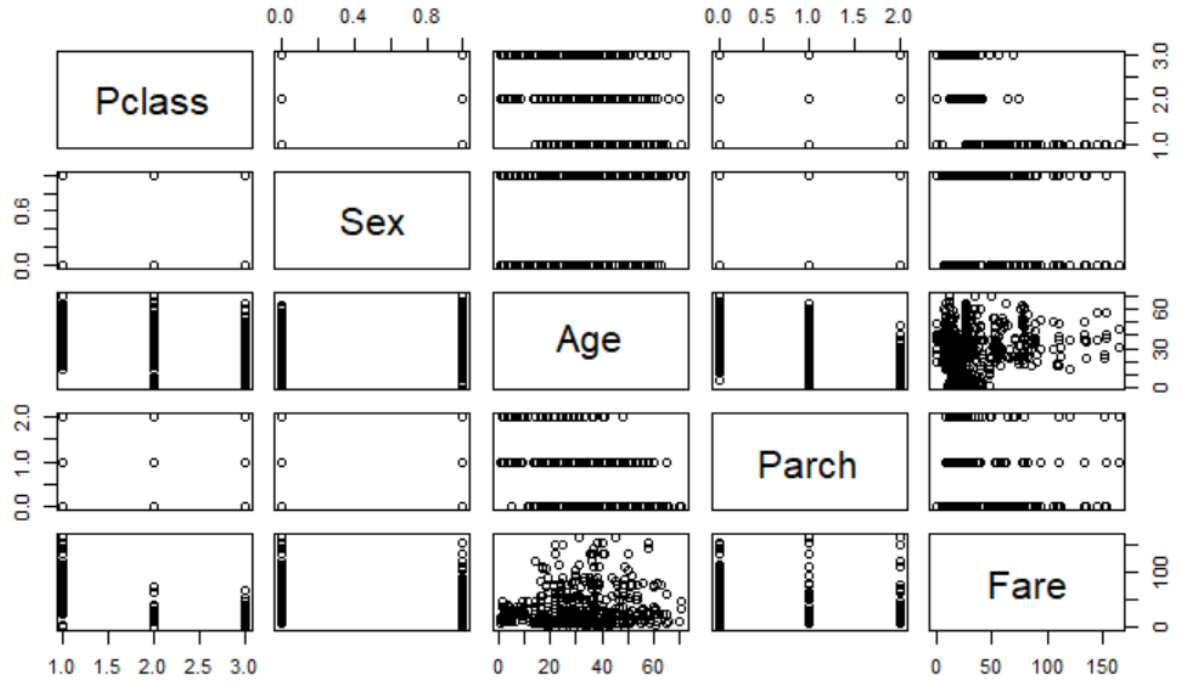

Fig: Scatter Plot After Removing Outliers

Dimension Reduction Using Principal Component Analysis (Authors: Nick Small, Gatimbirizo Mucyo, Olivio Ogbebor)

For the principal component analysis, we are using the eigen vectors of the correlation matrix to help find the direction of our dataset. We can now use those eigen vectors to make a linear construct and derive principal components from those eigen vectors that represent our dataset and explain the variability of the titanic dataset. These principal components are uncorrelated because the eigen vectors are orthogonal to each other unlike the variables in the original dataset which are correlated. The first principal component has the highest eigen values and therefore the highest variance compared to the second and third principal components and therefore is the strongest component.

	Survived	Pclass	Sex	Age	Parch	Fare
Survived	1.00000000	-0.32714358	-0.55912267	-0.03495482	0.11390113	0.3115311
Pclass	-0.32714358	1.00000000	0.13185666	-0.46901364	0.02879414	-0.6560765
Sex	-0.55912267	0.13185666	1.00000000	0.09755997	-0.24010424	-0.2395389
Age	-0.03495482	-0.46901364	0.09755997	1.00000000	-0.27730529	0.2203545

```
Parch      0.11390113  0.02879414 -0.24010424 -0.27730529  1.00000000  0.2398309
Fare       0.31153108 -0.65607652 -0.23953888  0.22035448  0.23983092  1.0000000
```

Fig: Correlation Matrix of Titanic Dataset

From the above correlation matrix, we can see that Pclass and Fare have a high inverse correlation (-0.656). This makes sense because as Pclas (Ticket class number (1 = 1st, 2= 2nd; 3= 3rd)) increases, we expect the fare price to decrease. Also, we can see a relatively high inverse correlation between Age and Pclass (-0.469) as we would expect the older passengers on the ship to be more financially stable and purchase higher class tickets (1st class) for more comfort than the younger passengers on the ship.

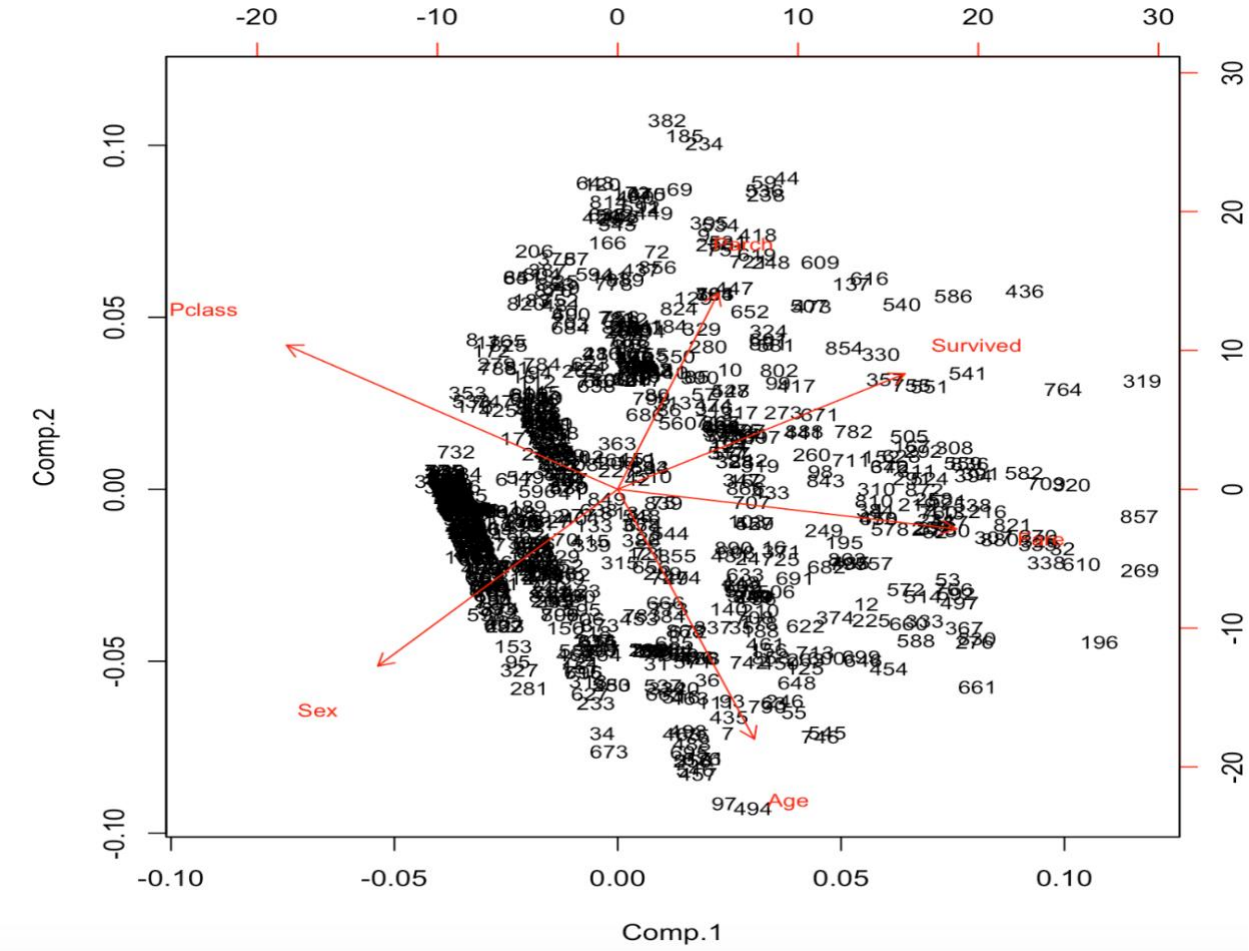

Fig: PCA Biplot

Analyzing the Biplot above, we can see that Parch and Survived are highly correlated and Survived and Fare is also highly correlated. This is due to the angles being relatively closer to

one another compared to the other variables. Angles forming at the 90 degrees are likely not correlated such as Survived and Age in regard to Fare. All other variables that diverge from a large angle or up to 180 degrees are negatively correlated such as Fare and Pclass or Pclass and Age which supports our correlation matrix analysis earlier.

```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6
Standard deviation     1.4949879 1.2683860 0.9754536 0.73923328 0.63235682 0.50828863
Proportion of Variance 0.3724982 0.2681339 0.1585849 0.09107764 0.06664586 0.04305956
Cumulative Proportion  0.3724982 0.6406320 0.7992169 0.89029459 0.95694044 1.00000000

Loadings:
         Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6
Survived  0.458  0.282  0.466  0.224  0.620  0.241
Pclass   -0.528  0.352  0.112 -0.220         0.730
Sex      -0.382 -0.431 -0.386  0.440  0.545  0.167
Age       0.218 -0.610        -0.685  0.206  0.249
Parch     0.160  0.479 -0.680 -0.386  0.352
Fare      0.540        -0.390  0.298 -0.386  0.556
```

Principal Component Analysis of Titanic Dataset

From the above principal component analysis results, we can see from the Cumulative Proportion that we can explain approximately 80% of the reality or variability of the titanic dataset with 3 principal components. Therefore, we have chosen the first 3 principal components for our data analysis.

```
            Comp.1      Comp.2      Comp.3
Survived  0.4579722  0.28212535  0.46621596
Pclass   -0.5281461  0.35180622  0.11228353
Sex      -0.3824803 -0.43137137 -0.38606906
Age       0.2178035 -0.60965154  0.08139921
Parch     0.1602993  0.47903659 -0.68001192
Fare      0.5402756 -0.09698117 -0.38979960
```

Titanic Dataset PCA Loading Coefficients

```
Comp.1      Comp.2       Comp.3
1 -1.6023464 -0.08375009 -0.007792743
2  2.9242238 -0.42773939  0.723765956
3  0.2324151  1.21698511  1.786945414
4  2.5287922 -0.22301896  0.953148615
5 -1.3659629 -0.70571835  0.063956363
6 -1.5454264 -0.18312690 -0.011583276
```

First 6 Passengers PC Scores

From the above loading coefficients, we can describe each of our principal components below:

PC1: From PC1, we can infer that this component represents passengers that survived and paid a higher fare price. We can also infer that this component represents passengers that purchased a higher ticket class (i.e. 1st class) since the value sign is negative (-0.528), which will make sense as to why the fare price will be higher for these passengers. Therefore, a passenger with a higher

PC score for component 1 most likely survived the event, paid a higher fair price and purchased a higher-class ticket.

PC2: We can infer that PC2 represents the sex and age of the passengers. Passengers with a higher PC score on component 2 will most likely be women (0) since the PC value for sex is negative (-0.431) and are younger in age since the PC value for age is also negative (-0.609).

PC3: This component gives us a good representation of 'Parch,' the number of parents and children onboard the Titanic ship. Since the PC value is negative (-0.680), we can infer that the higher the PC score on component 3 for a passenger, the lower the family count pertaining to that particular passenger onboard the Titanic ship.

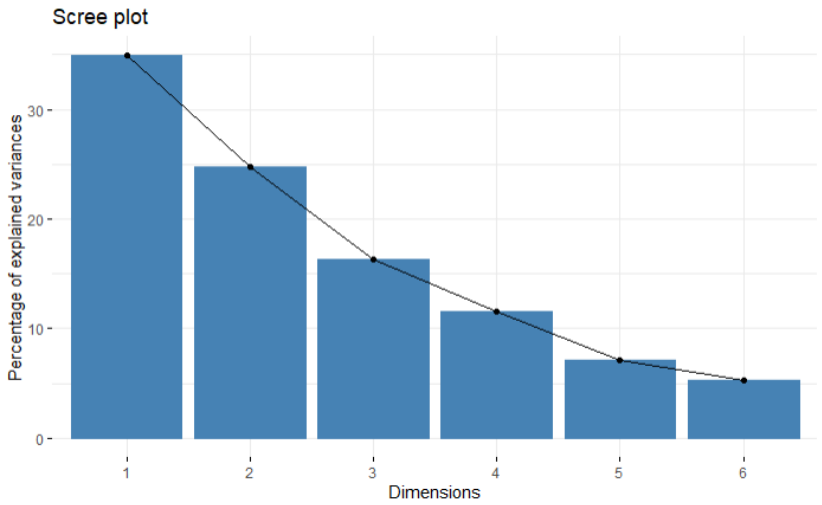

Fig: Scree Plot

Cluster Analysis (Authors: Nick Small, Gatimbirizo Mucyo, Olivio Ogbebor)

Hierarchical Clustering

For this project, clustering will be used in grouping similar items or variables together and in order to makes sure that the various groups are significantly different but with similar observations in each group. The distance matrix of the Titanic dataset will be utilized in creating the clustering for the Hierarchical cluster analysis. The complete Linkage cluster analysis method will be utilized, taking into account the maximum distance between two groups in determining the cluster analysis of the Titanic dataset.

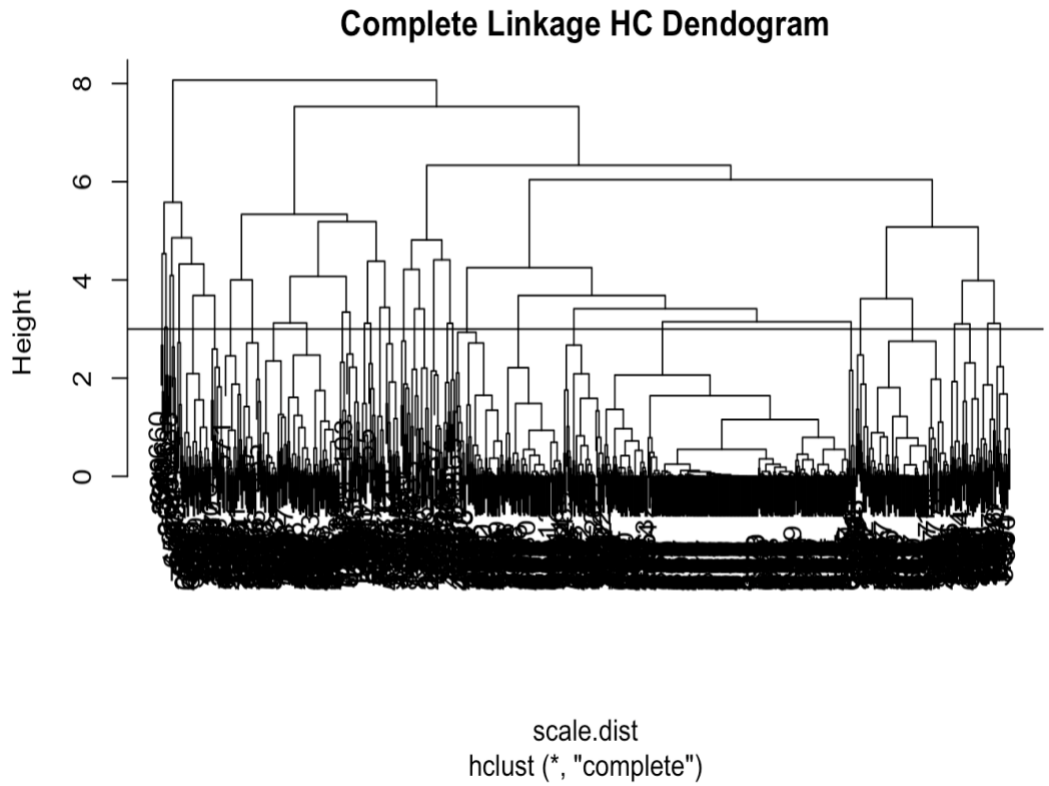

Fig: Complete Linkage HC Dendogram

From the dendrogram above, we can infer that there are four main cluster groups. Every other cluster group seems to fall under these 4 major cluster groups. It is impossible to read which individual in the Titanic dataset falls under a specific group using the dendogram because the dataset is too large. A scree plot can now be utilized to test our inferred number of clusters and find the actual number of clusters in our dataset. The complete linkage method was used because it gave us the best cluster analysis and a better spread of our dataset between all cluster groups. Below is the scree plot for our number of clusters analysis:

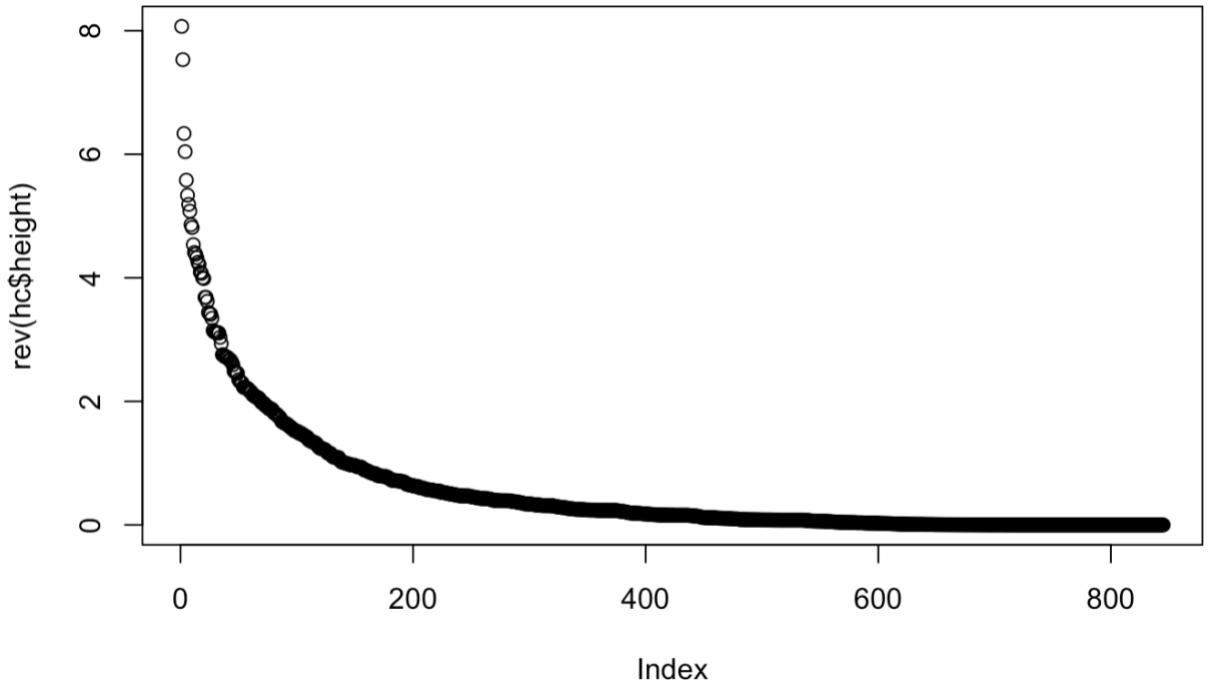

Fig: Scree Plot for Cluster Analysis (Elbow Test)

From the scree plot above, utilizing the elbow test, we can see that the biggest distance between clusters is between cluster 2 and 3. Therefore we can conclude that our observations in the Titanic dataset fall under 3 main cluster groups. As we can see, our inferred number of clusters from the dendogram is slightly different from the actual number of clusters we derived from the elbow test. This is why it is necessary to utilize the elbow test in determining the right number of clusters we have in any given dataset.

```
 1   2   3
607  64 175
```

Cluster distribution

From the above cluster distribution, we can see that 607 passengers fall under group 1, 64 passengers fall under group 2 and 175 passengers fall under group 3.

Kmeans Clustering

For this project, we also utilized Kmeans clustering, which is one of the most common cluster analysis methods, to analyze the clusters in the titanic dataset and see how it compares to the hierarchal cluster analysis method. The kmeans clustering utilizes the Within Group Sum of Squares (WGSS) method which actually measures the distance from the center of each cluster

group in the Titanic dataset. The WGSS algorithm is looking for the clusters that minimize the squared Euclidean distances of the points from centroids of the groups in the Titanic dataset.

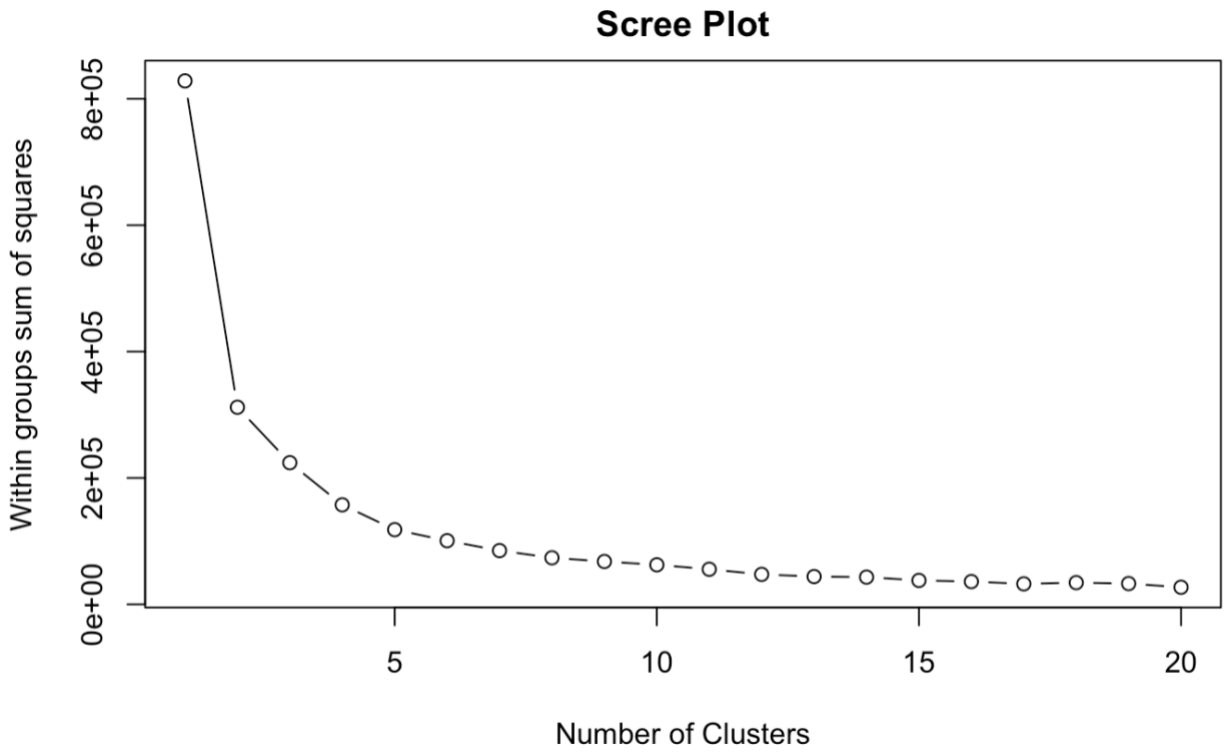

Fig: Scree Plot of First 20 Cluster Groups in The Titanic Dataset

From the above scree plot, we can see that the slope of the plot remained quite constant after the 4th cluster group. Therefore, we can conclude that there are 4 major cluster groups in the Titanic dataset that gives a good representation of the similar groups in the dataset.

```
1   2   3   4
523 117  29 177
```
Cluster distribution

From the above cluster distribution, we can see that 523 passengers fall under group 1, 117 passengers fall under group 2, 29 passengers fall under group 3 and 177 passengers fall under group 4.

Defining the Meaning of The Cluster Groups

We have utilized the results from the principal component analysis in determining the meaning of the cluster groups in our clustering analysis. We did this by making a plot of the first two pca scores of the passengers in the Titanic dataset and we color coded the plot by the cluster groups. The generated plot can be found below:

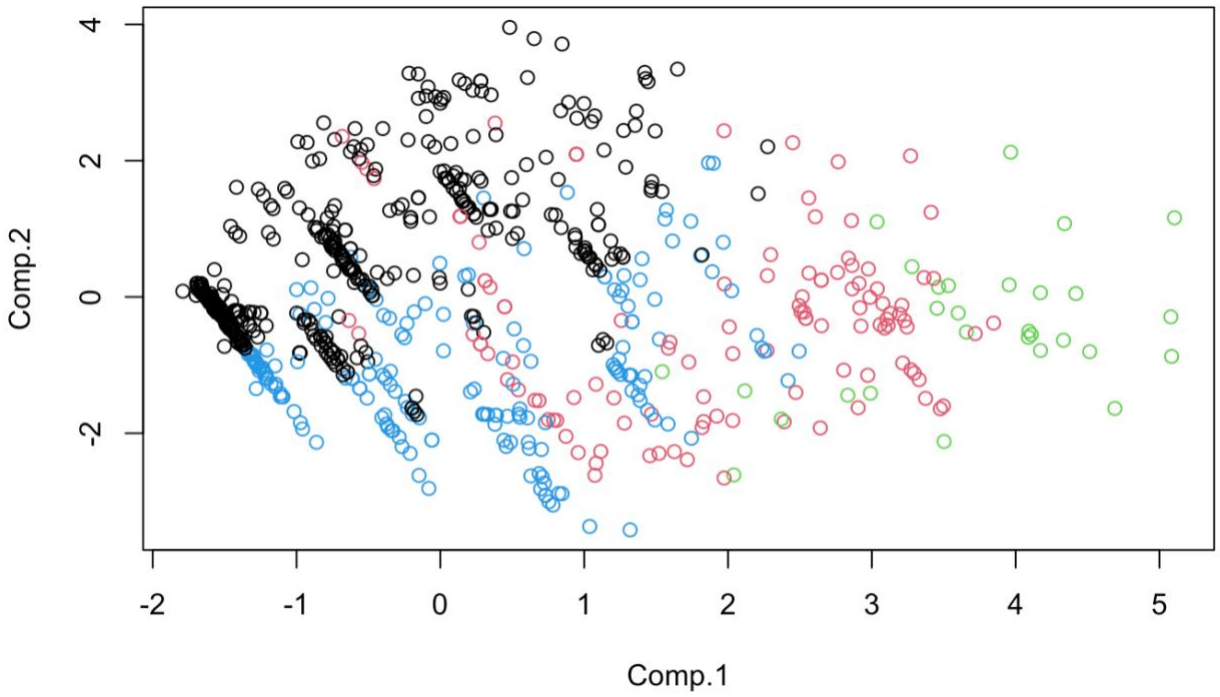

Fig: Plot of Pca scores color coded by cluster groups

From the above diagram, the following conclusions can be made:

Green group: This cluster group has the highest component 1 scores meaning that passengers in this group most likely survived the Titanic event, paid a higher fair price and purchased a higher-class ticket. We also notice that the number of passengers in this group are fewer compared to the other 3 groups which makes sense given that only few people survived the Titanic event.

Red Group: This cluster group has a weaker component 1 score as compared to the green group. Therefore, passengers in this group had a lower chance of surviving the Titanic event, paid a lower fair price and purchased a lower-class ticket that those passengers in the green group.

Blue Group: Majority of the passengers in this cluster group have an even weaker component 1 score as compared to the green and red group. Meaning that the majority of the passengers in this group had lower chance of surviving the titanic event, paid a lower fair price and purchased a lower-class ticket than those passengers in the green and red group.

Black group: The black group can be characterized majorly with passengers that did not survive the titanic event, paid the lowest fair prices and purchased the lowest class tickets as the passengers in this group have the weakest component 1 scores.

Note: We can notice some overlapping withing groups in the cluster plot which may be as a result of some uncertainty in the cluster analysis.

	Survived	Pclass	Sex	Age	Parch	Fare
1	0.3173996	2.755258	0.6730402	22.28761	0.2695985	12.71523
2	0.6153846	1.401709	0.5128205	34.19658	0.4786325	67.07415
3	0.7931034	1.000000	0.3103448	36.03448	0.6206897	129.37657
4	0.3107345	1.943503	0.7740113	43.64124	0.1864407	20.77971

Cluster Centroids

From the cluster centroid values, we can conclude that:

Group 1: we can see that cluster group 1 has the lowest mean centroid value from all the variables meaning that passengers in this group belong to the black group as they had the weakest component 1 scores. Also, from the cluster distribution above, we can see that there were 523 passengers in this group which makes absolute sense given that the majority of the passengers onboard the Titanic ship did not survive the event.

Group 2: Cluster group 2 has the second highest mean centroid value. This means that passengers in this group belong to the red group as they had the second highest component 1 scores. From the cluster distribution, we can see that there were 117 passengers in this group which is quite lower than the number of passengers in group 1. This makes sense given that fewer passengers on the Titanic ship had a good chance of surviving the event.

Group 3: This cluster group has the highest mean centroid value. This means that passengers in this group belong to the green group as they had the highest component 1 scores. We can also see from the cluster distribution that only 29 passengers were in this group which makes absolute sense given that these passengers have the highest chance of surviving. And this supports the reality as fewer number of passengers survived the titanic event.

Group 4: Group 4 has the second lowest mean centroid value. We can then conclude that passengers in this group belong to the blue group as they had the second lowest component 1 scores. From the cluster distribution, 177 passengers were in this group which is more than the number of passengers in group 2 and 3. This makes sense as the passengers in this group had a lower chance of surviving the event and supports the reality that majority of the people on the Titanic ship did not survive the event.

Model Based Clustering

Model based clustering was also utilized in this project for cluster analysis. In model base clustering we assume that if the data in the Titanic dataset comes from the same probability distribution, we can infer that that data belongs to the same cluster group. Therefore, we found the probability distribution parameters for each group and we found the best group based on assigning the best probability distribution parameter in the Titanic Dataset.

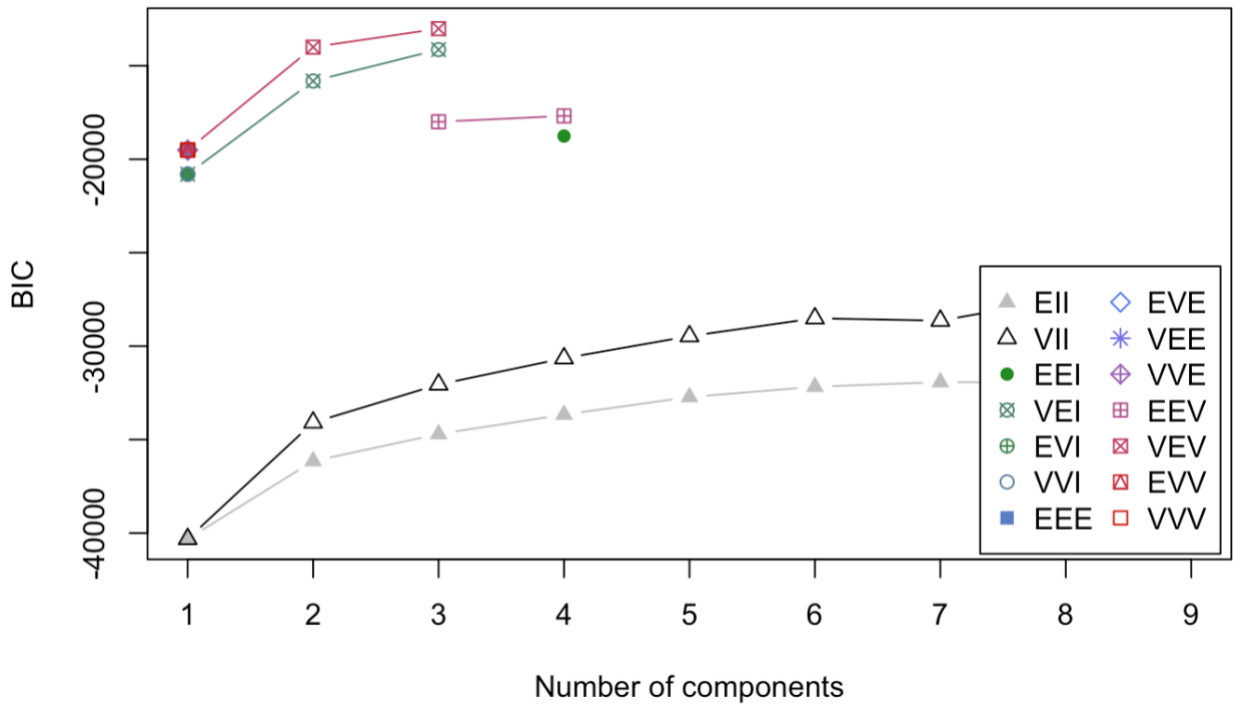

Fig: Beyesian Indicator (BIC) plot

The BIC is based on the tradeoff of the complexity of the model and the likelihood function result of the Titanic dataset. From the BIC plot we see that we have the maximum BIC value at component 3. Therefore, we can conclude that there are 3 cluster groups present in the Titanic dataset using the model-based clustering method.

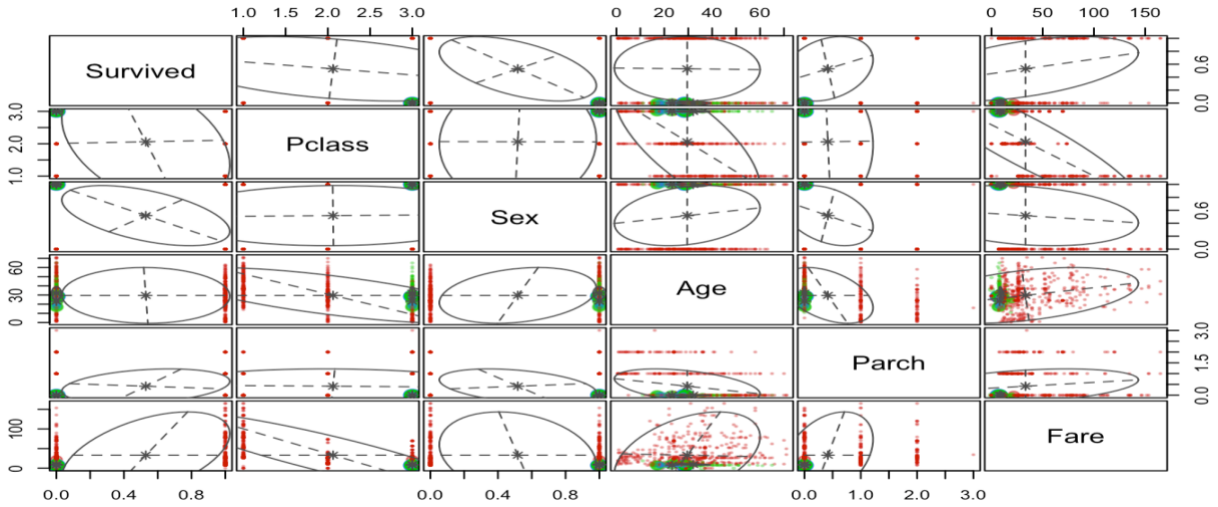

Fig: Uncertainty Plot

From the uncertainty plot, we can see that there seems to be a major uncertain point in our dataset represented by the bold green dot and occurs on various perspectives on the uncertainty plot.

314	"314"	"1"	"0.0394303441029911"
393	"393"	"1"	"0.0432991666537086"
478	"478"	"3"	"0.0491382566507076"
835	"835"	"1"	"0.0640527024478831"
81	"81"	"1"	"0.0757314245272015"
144	"144"	"1"	"0.0799295965594061"
722	"722"	"3"	"0.129737539529463"
91	"91"	"3"	"0.143818630366933"
49	"49"	"2"	"0.147199819446995"
351	"351"	"3"	"0.190555212898904"
434	"434"	"3"	"0.207068824315919"
423	"423"	"3"	"0.255483159897342"
232	"232"	"3"	"0.302443873700594"
58	"58"	"1"	"0.391911130064888"

Probability of Uncertainty Values

From the above probability of uncertainty values, the values in yellow represent the passengers with the highest uncertainty probabilities but we can clearly see that passenger 58 has the highest uncertainty of being in group 1 which explains the major uncertainty point in our uncertainty plot.

Confirmatory Factor Analysis (Authors: Nick Small, Gatimbirizo Mucyo, Olivio Ogbebor)

In confirmatory factor analysis, we know the factors representing the Titanic dataset and we build our model based off this knowledge. Nevertheless, unfortunately CFA was not very suitable for the Titanic Dataset as we could not derive meaningful factors (latent variables) from the EFA that we were supposed to use to create our CFA model. But below is our attempt in running the CFA regardless and the error code derived:

Loadings:

	Factor1	Factor2
Survived	0.254	0.672
Pclass	-0.991	-0.112
Sex	-0.782	
Age	0.498	-0.216
Parch	0.327	
Fare	0.629	0.288

EFA Factor Loadings

The first step in achieving our goal was to create a model that we called “titanic_model” which can be seen below:

```
People        -> Survived, lambda1, NA

People        -> Age, lambda2, NA

Category      -> Pclass, lambda3, NA

Category      -> Sex, lambda4, NA

Category      -> Parch, lambda5, NA

Category      -> Fare, lambda6, NA

People        <-> Category, rho, NA

Survived      <-> Survived, theta1, NA

Age           <-> Age, theta2, NA

Pclass        <-> Pclass, theta3, NA

Sex           <-> Sex, theta4, NA

Parch         <-> Parch, theta5, NA

Fare          <-> Fare, theta6, NA

People        <-> People, NA, 1

Category      <-> Category, NA, 1
```

After creating the model, we run the code to perform a confirmatory factor analysis base on two factors that we called: “People” and “Category”.

However, we ran into an error and we were not able to perform the analysis.

The error got was :

```
NA//Inf replaced by maximum positive valueOptimization may not have converged; nlm
return code = 4. Consult ?nlm.
Negative parameter variances.
Model may be underidentified.
maximum iterations exceeded
```

After long hours of trying to solve the error we did not succeed.

The code that we wrote to perform the analysis can be find in the appendix part of the report.

We wanted also to report a path diagram that shows coefficient estimates, report the SRMR, GFI, and AGFI. At the end we could have found the 95% confidence interval for the correlation between people and category returns.
The codes that could have executed our confirmatory factor analysis can be found in the appendix.
We would like to get a feedback on how to correct error that we got in the CFA.

Conclusion (Olivio Ogbebor)

1. What were the chances of surviving the Titanic event?
2. What was the age distribution on the Titanic?
3. What was the correlation between the people that survived and their financial situation (ticket class)?
4. What is the relationship between the fare price and the passengers that survived the event?

In conclusion, we will answer the questions posed in the introduction of this report which have again been stated above. From our data analysis of the titanic dataset, we can confidently say that the chances of surviving the Titanic event was quite low as fewer passengers fell under cluster groups 2 and 3 which were the groups with a higher survival expectancy. This also supports the real-life situation as only 706 out of 2224 passengers survived the event. We can also conclude from our correlation matrix that there were more younger passengers on the Titanic ship than seniors given that older citizens are expected to purchase a higher ticket class than younger citizens as they are more financially stable. And we can see an inverse relationship between Pclass and Age in the correlation matrix. Furthermore, from our data analysis the passengers that purchases a higher-class ticket fell under cluster group 2 and 3 and also had the fewer number of passengers, supporting the notion that there were fewer older passengers on the Titanic. The people that survived the Titanic event had a better financial status as they fell under cluster groups 2 and 3, meaning that they were able to purchase a higher-class ticket than the passengers in cluster groups 1 and 4. This also ties into the relationship with the fare price and passengers that survived the Titanic event as the people who survived the event purchased a higher-class ticket, and therefore paid a higher fare price. For post analysis recommendations, we recommend performing more analysis using Confirmatory Factor Analysis (CFA) in determining the latent variables that represent the variables in the Titanic dataset.

Appendix

Below are the R codes utilized for this project's data analysis:

```
#Reading the Titanic Dataset
train <- read.csv("/Users/Ewaen/Desktop/train.csv", stringsAsFactors=TRUE)
train
#selecting needed columns and plotting variables
mydata <-train[,c(0,1,2,3,4,5,6)]
mydata
mydata$score <- NULL
plot(mydata)
mydata
#removing outliers
#Checking the number of row before removing outliers
nrow(mydata)
# Find the mahalanobis distance
mdist <- mahalanobis(mydata, colMeans(mydata), cov = cov(mydata))
# find the 0.95 quantile of mahalanobis distances
cutPoint <- quantile(mdist, .95)
# Filter the data with the mahalanobis distance of less than 0.95 quantile
newdata <- mydata[(mdist<cutPoint),]
newdata
#Checking the number of rows after removing outliers
nrow(newdata)
#Plotting new scatter plot after removing outliers
plot(newdata)
cor<-cor(newdata)
cor
#performing principal component analysis and plotting biplot
pc<-princomp(newdata, cor=T)
pc
biplot(pc, col=c("black", "red"), cex =0.8)
#getting the pc loadings
```

```
summary(pc, loading = T)
round(100*sum((pc$sdev^2)[1:3])/sum(pc$sdev^2),1)
#scree plot
library(factoextra)
fviz_eig(pc)
#lets get the pc scores and loadings for the dataset
score<-pc$score[,1:3]
head(score)
pc$loading[,1:3]
#Performing hierarchical cluster analysis
scale.dist = dist(scale(newdata))
scale.dist
hc <- hclust(scale.dist, "complete")
plot(hc, main = "Complete Linkage HC Dendogram")
abline(h=3)
#getting scree plot of height values
names(hc)
hc$height
plot(rev(hc$height))
#Getting cluster distribution between groups
scale.dist = dist(scale(newdata))
hc <- hclust(scale.dist, "complete")
ct<-cutree(hc,3)
head(ct)
titanic.clust <- data.frame(ct)
table(titanic.clust)
#Kmeans
#scree plot to determine the number of clusters in the dataset.
plot.wgss = function(newdata, maxc) {
  wss = numeric(maxc)
  for (i in 1:maxc)
    wss[i] = kmeans(newdata,centers=i, nstart = 10)$tot.withinss
  plot(1:maxc, wss, type="b", xlab="Number of Clusters",
  ylab="Within groups sum of squares", main="Scree Plot")
```

```
}

plot.wgss(newdata, 20) #Elbow test; testing first 20 clusters

#repeating the iteration 10 times using the 'nstart' method to get a more reliable
analysis.

km <- kmeans(newdata, center = 4, nstart = 10)

table(km$cluster)

#using principal component analysis to determine the meaning of clusters by making a
plot of the pc scores

pc

pc$loadings[,1:3]

plot(pc$scores[, c(1:2)], col = km$cluster)

#getting cluster centroids

km$centers

#Model-based clustering

install.packages('mclust')

#library("mclust")

mc <- Mclust(newdata)

mc.clust <- mc$classification

table(mc.clust)

cat("\nCountries in cluster 1 \n")

names(mc.clust[mc.clust == 1])

cat("\nCountries in cluster 2 \n")

names(mc.clust[mc.clust == 2])

plot(mc, "BIC")

#uncertainty plot

plot(mc,  what = "uncertainty")

#Uncertainty probability values

options(max.print=1000000)

clust.data=cbind(rownames(newdata), mc$classification, mc$uncertainty)

clust.data[order(mc$uncertainty),]

efa <- factanal(newdata, 2)
print(efa$loadings, cut = 0.5)

# for CFA, we first need a model
library(sem)
titanic_model1 <- specifyModel(file="titanic_model1.txt")
titanic_sem <- sem(titanic_model1, cor(newdata), nrow(newdata))
options(fit.indices = c("GFI", "AGFI", "SRMR"))
```

```
summary(titanic_sem)

library(semPlot)
semPaths(titanic_stem, rotation = 2, 'std', 'est')

options(fit.indices = c("GFI", "AGFI", "SRMR")) # Some fit indices
criteria = summary(titanic_sem)
criteria$SRMR
criteria$GFI
criteria$AGFI
```