Analysis A

Our goal is to find common hidden characteristics that may help explain or predict changes in stock prices. Thirteen variables, not including the date. I will assume that the factors are 8 because it is less than total of 13. By seeing the boxplots plotted by given data. The data needs to be standardized (fig1). And the variables are correlated highly (fig2).

Factor Analysis of both PC & ML Methods:

Employed both methods PC & MLE to comprehensively address the study goal of identifying underlying patterns in stock price behavior. Using the PC method, I observed that after considering the first 4 factors, there were no substantial effects on variables, as determined by a threshold of 0.3. Similarly, employing the MLE method, we found that after the first 3 factors, there were no significant impacts on variables, using the same threshold.

Despite the PC method typically indicating larger loadings, I opted for a 3-factor model because the consistency between the two methods and the number of significant groups identified in the correlation matrix. I finally concluded that a 3-factor model adequately captures the underlying patterns in the data.

After employing Varimax Rotation(fig3) of pc and mle for clarity, our analysis identified three key factors in stock price behavior. The first factor encompasses market leaders with diversified portfolios, indicating stability and growth potential. The second factor reflects companies influenced by consumer spending trends and online retail performance, particularly in e-commerce and consumer goods sectors. The third factor represents international giants focusing on technology, consumer goods, and finance, showcasing their dominance across multiple industries.

By Comparing the final models obtained from both PC and MLE methods results are shown below that there is a significant similarity, particularly for Factors 1 and 2. Acknowledged some differences in the specific companies captured by Factor 3 but overall found both methods identified similar underlying patterns in Factor analysis methods.

Principal Component Analysis (PCA):

The goal of the PCA is to determine the total number of principal components necessary to explain most of the variance in the data. Standardization of data is essential due to different ranges observed in the variables, as shown by box plots (Fig1). By Plotting scree plot (Fig 4) and examined result to identify that only two principal components are enough to cover approximately 82% of the data's variability.

In our Principal Component Analysis (PCA), I identified the key variables driving PC1 are Walmart, Apple, Toyota, Dow Jones, Google, and Amazon, with Walmart emerging as the most influential. These variables exhibit strong positive relationships with PC1. And also, Walmart and Apple carry the highest weights in defining PC1. while variables like HSBC and Honda show negative weights, suggesting an inverse relationship. PC1 captures common trends among these variables.

Additionally, our analysis of variable loadings reveals that certain variables, such as Walmart, Apple, and Toyota, exhibit the highest loadings in PC1, while others like HSBC and Honda demonstrate opposing trends.

From the PCA graph (Fig5), I observe that Walmart, Coca Cola, Pepsi, S&P 500, Google, Toyota, Apple, Chase, and Dow Jones form a tight cluster near the origin. However, Honda, Amazon, eBay, and HSBC stand out from this main cluster. This suggests that these companies that set them apart to share similar characteristics on the two principal components analyzed.

Comparison between the two parts PCA & FA:

By comparing the both PCA & FA methods. I found that both methods agree with each other in the results. Because they have similar patterns and groups. And also, both coveys the same interpretations. among the variables, companies clustering near the origin in the PCA graph, corresponding to the market leaders identified in the factor analysis. The outliers in the PCA graph align with the e-commerce and international giants factors.

Analysis B1:

The study goal is to identify if there is difference in Age, sleep duration and Heart Rate between the different genders Male, Female. The chi-sq QQ Plots & Royston test imply that all variables appear to be close enough to normality with some outliers.

For univariate analysis of the Age, sleep duration and Heart Rate plotted the boxplots. And for multivariate analysis also have plotted the 3d scatterplot and graph are in fig 8 to 12.

To address the study goal, we chose to analyze the hoteling t2 test. The hoteling's T2 test gives a p-value 1.624e-07 which is less than 0.05(α) stating that there is significant difference in atleast one of the variables between the groups at a significance of 0.05.

By checking the confidence intervals of 95% of Bonferroni. I found that there is difference in all the variables because there is no zero in between the resulted interval values.

It shows that three variable sleep duration, age, heart rate varies in gender male to female.

Analysis B2:

The study goal is to identify if there is difference in Age, sleep duration and Heart Rate between the different Occupations Accountant, Doctor & Lawyer. The chi-sq QQ Plots & Royston test imply that all variables appear to be close enough to normality with some outliers.

For univariate analysis of the Age, sleep duration and Heart Rate plotted the boxplots. And for multivariate analysis also have plotted the 3d scatterplot and graph.

To address the study goal, we chose to analyze the Manova test. The Manova test gives a p-value 5.143e-14 which is less than 0.05(α) stating that there is significant difference in atleast one of the variables between the groups at a significance of 0.05.

By checking the confidence intervals of 95% of Bonferroni. I found that there is difference in all the variables because there is no zero in between the resulted interval values.

It shows that three variable sleep duration, age, heart rate varies in different occupations (Accountant, Doctor, Lawyer).

variable	p-val
Age	9.94E-09
Sleep duration	3.55E-10
Heart rate	1.09E-05

Table: Anova test for occupation for other variables

Appendix:

Analysis A:

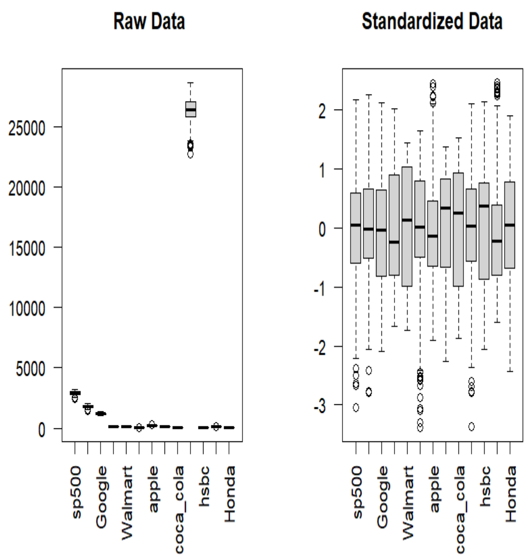

Fig1: Raw & Standardized data boxplots

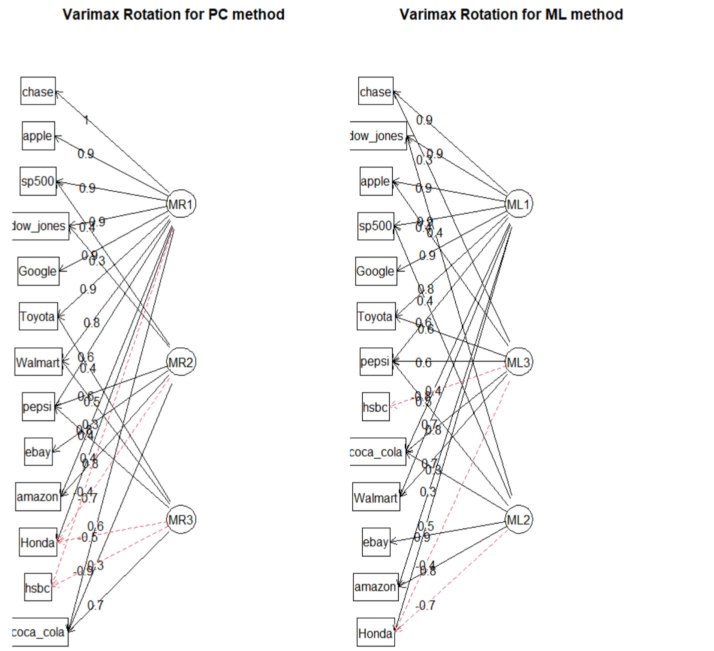

Fig3: Varimax Rotation

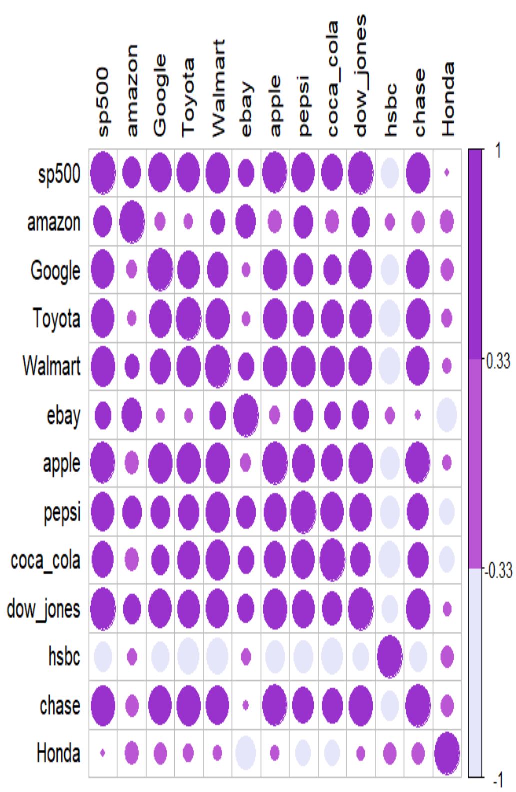

Fig2: Correlation plot

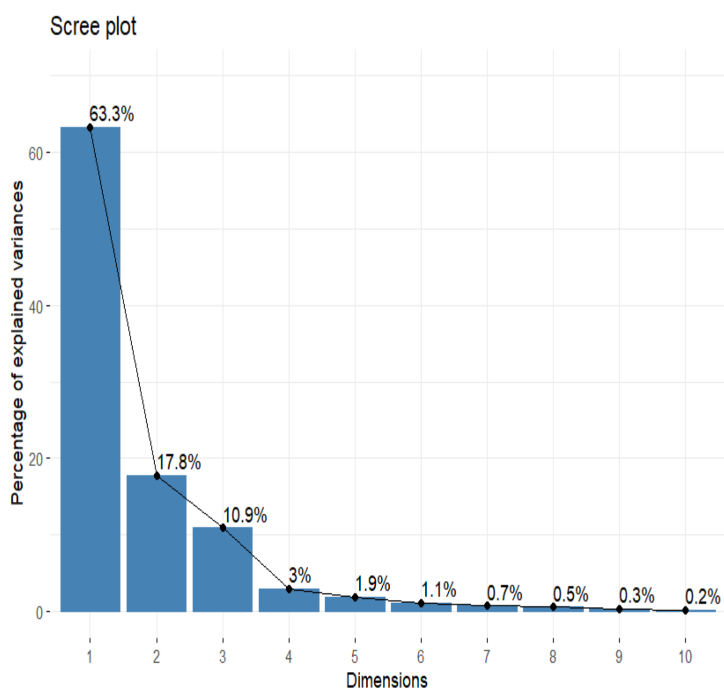

Fig4: screeplot

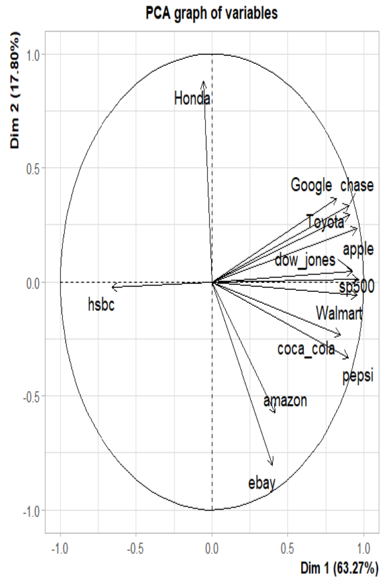

Fig5: PCA Graph

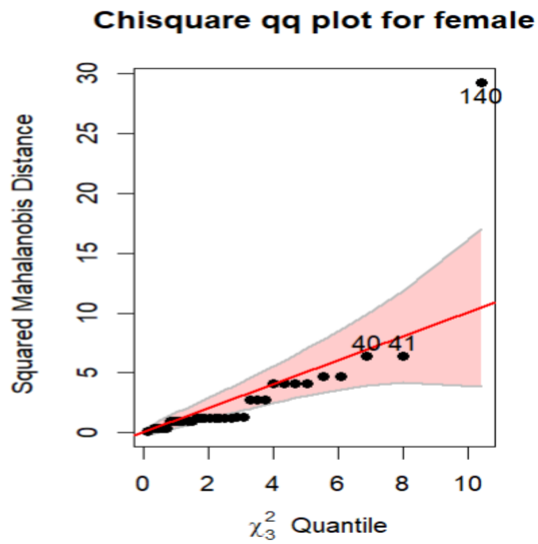

Fig7: chi-sq QQ-plot for female

Analysis b1:

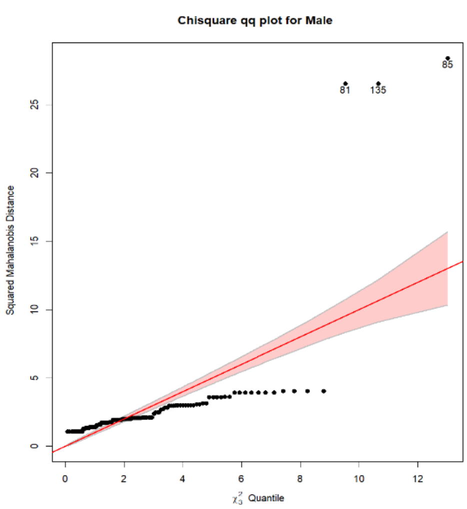

Fig 6: chisq qqplot for male

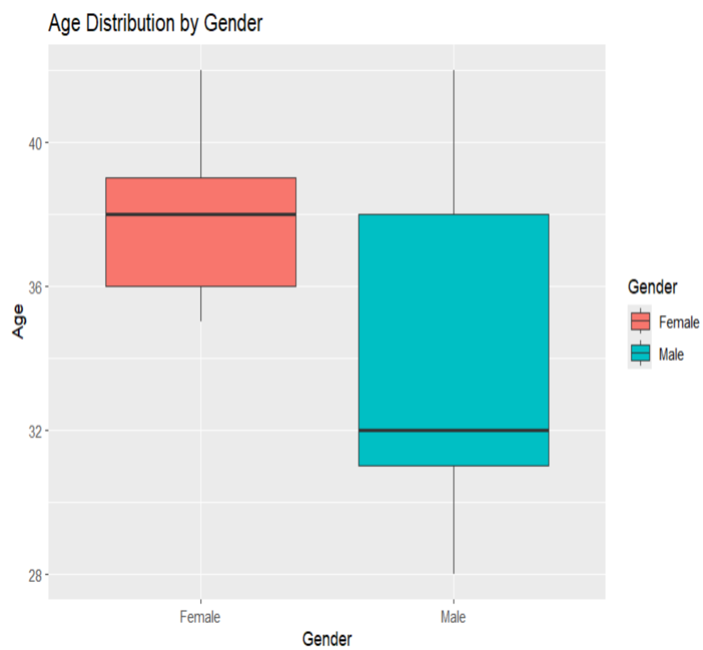

Fig 8: boxplot of age distribution by gender

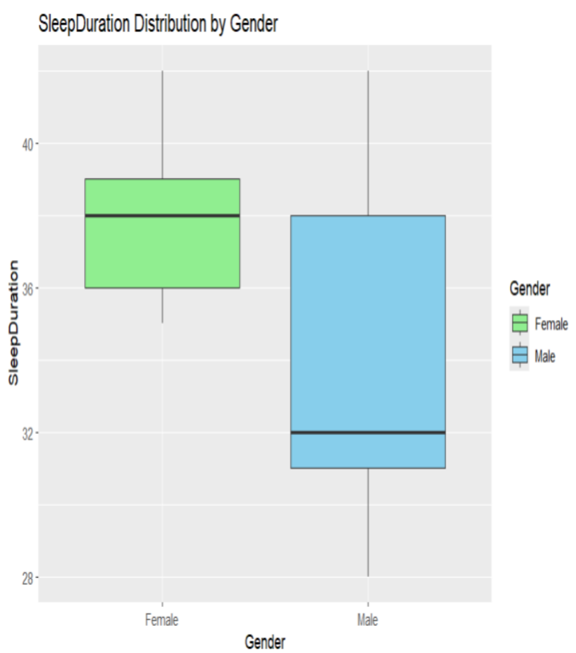

Fig 9: boxplot of sleep duration by gender

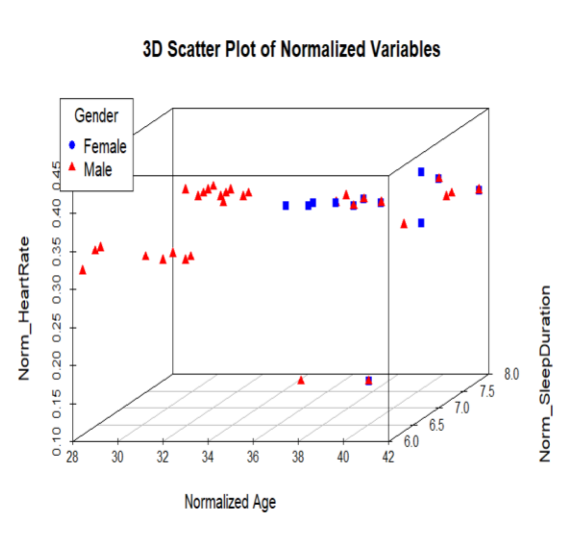

Fig 11: 3d sctterplot of numerical variables with gender

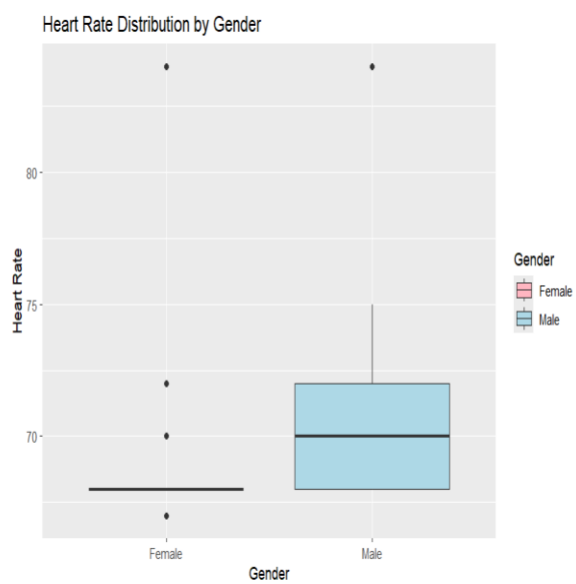

Fig 10: boxplot of heartrate distributed by gender

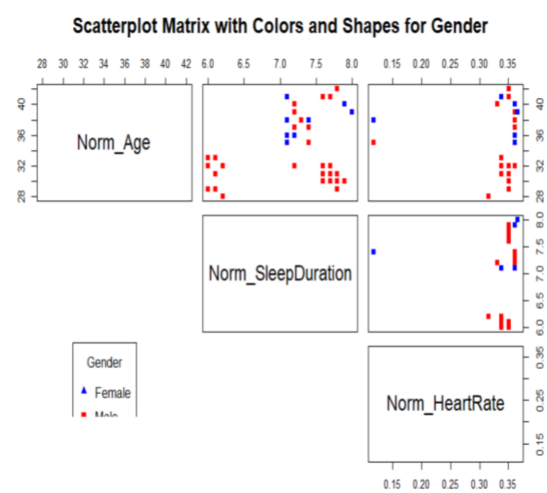

Fig 12: scatter plot matrix with gender

Analysis b2:

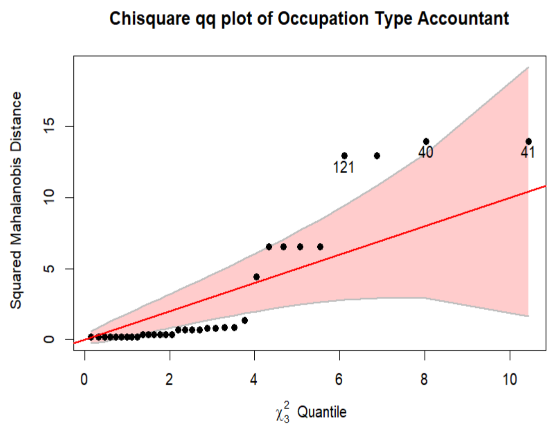

Fig 13: chi-sq of occupation type accountant

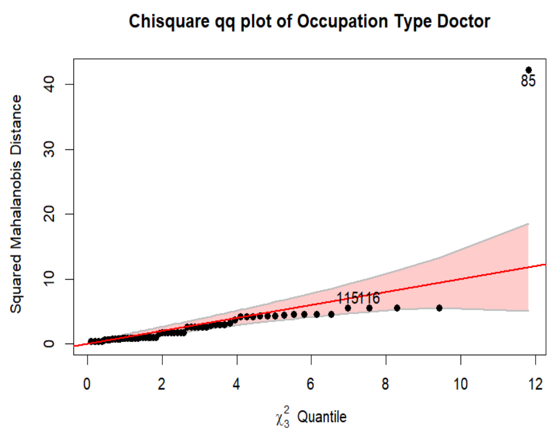

Fig 14: chisq of occupation type doctor

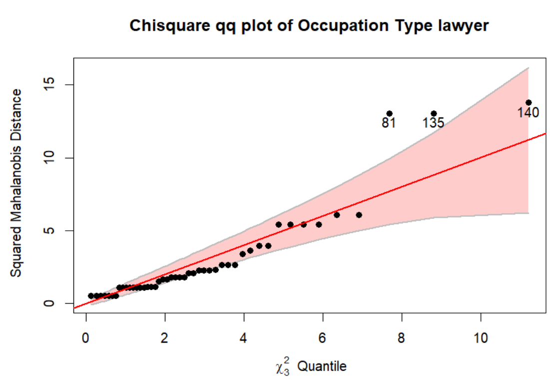

Fig 15: chi-sq of occupation type lawyer

Fig 16: boxplot of age distribution by occupation

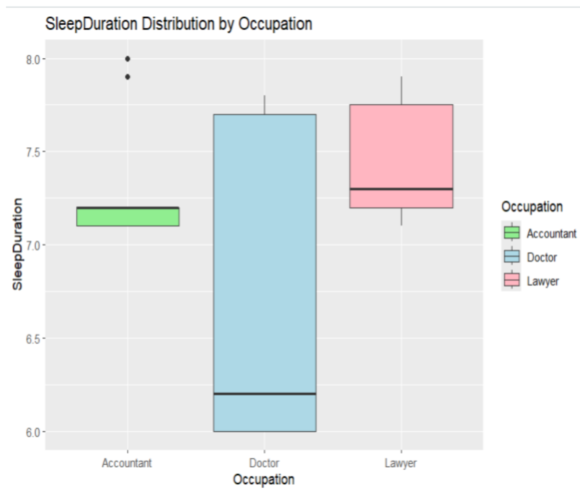

Fig 17: boxplot of sleep duration distribution by occupation

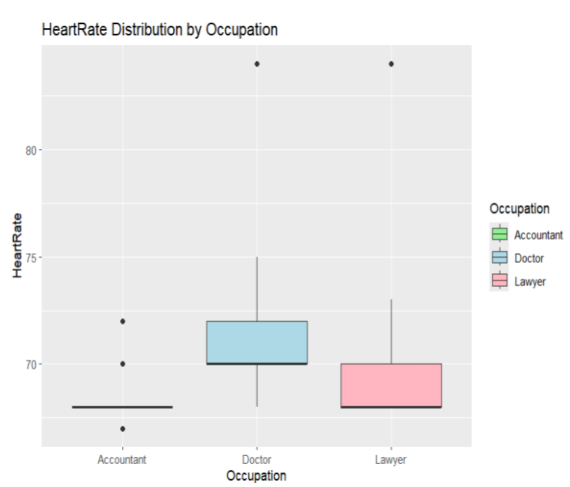

Fig 18: boxplot of heart rate distribution by occupation

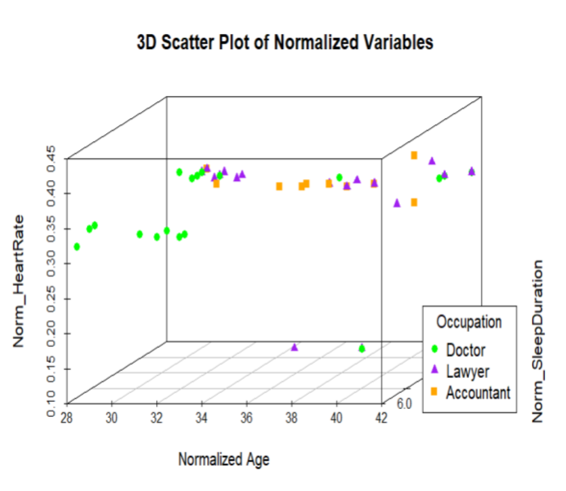

Fig 19: 3d scatterplot of occupation

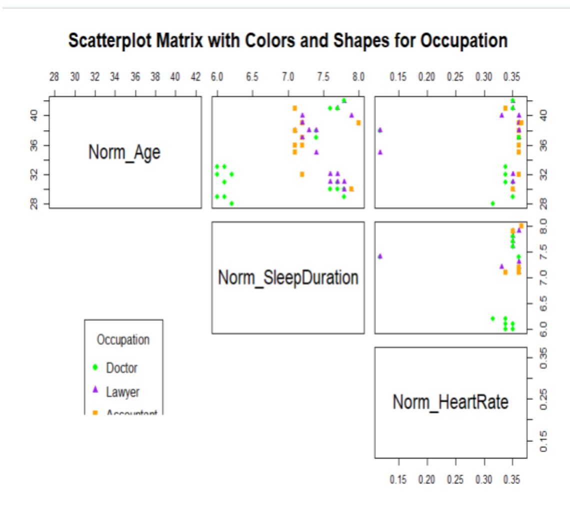

Fig 20: 3d scatterplot matrix of occupation

Index	Factor 1	Factor 2	Factor 3
1	sp500	amazon	amazon
2	amazon	pepsi	coca cola
3	Google	ebay	hsbc
4	Toyota	coca cola	
5	Walmart	Honda	
6	apple		
7	pepsi		
8	coca cola		
9	dow jones		
10	hsbc		
11	chase		

Table1: Companies under by each factor by ML method

Index	Factor 1	Factor 2	Factor 3
1	sp500	sp500	Toyota
2	Google	amazon	Walmart
3	Toyota	ebay	pepsi
4	Walmart	pepsi	coca cola
5	apple	coca cola	hsbc
6	pepsi	dow jones	Honda
7	coca cola	Honda	
8	dow jones		
9	hsbc		
10	chase		
11	Honda		

Table2: Companies under by each factor by pc method

Company	Dim.1	Dim.2
sp500	0.34	0.01
amazon	0.15	-0.38
Google	0.29	0.24
Toyota	0.32	0.19
Walmart	0.33	-0.04
eBay	0.14	-0.53
Apple	0.33	0.15
Pepsi	0.31	-0.22
Coca-Cola	0.30	-0.15
Dow Jones	0.32	0.03
HSBC	-0.23	-0.02
Chase	0.32	0.22
Honda	-0.02	0.58

Table 3: PCA Weights

MLE	P-Val
MLE1	NA
MLE2	0
MLE3	7.5 × 10−224
MLE4	1.6 × 10−157
MLE5	4.1 × 10−71
MLE6	1.0 × 10−45
MLE7	5.1 × 10−28
MLE8	3.9 × 10−23

Table 4: p-values for 8 factors

Pc method loadings for factors 9

sp500	-0.962	-0.011	-0.241	0.064	-0.012	-0.054	-0.017	0.073	-0.036
Amazon	-0.417	0.575	-0.625	-0.199	-0.098	0.226	-0.061	0.026	-0.003
Google	-0.822	-0.366	-0.130	0.305	-0.210	0.113	0.104	-0.098	0.032
Toyota	-0.907	-0.293	0.173	-0.089	0.156	0.128	0.005	0.010	0.082
Walmart	-0.956	0.058	0.160	-0.122	0.116	-0.042	-0.074	-0.057	0.105
eBay	-0.400	0.807	-0.160	0.315	0.239	-0.012	0.048	-0.038	0.003
Apple	-0.955	-0.233	-0.051	0.038	-0.114	-0.055	-0.009	0.025	-0.018
Pepsi	-0.899	0.332	0.021	-0.120	-0.105	-0.132	-0.096	-0.154	-0.048
Coca-Cola	-0.848	0.231	0.345	-0.235	0.081	0.054	0.189	-0.014	-0.086
Dow Jones	-0.927	-0.046	-0.297	0.129	0.100	-0.076	-0.036	0.093	-0.027
HSBC	0.659	0.025	-0.704	-0.160	0.029	-0.134	0.136	-0.050	0.058
Chase	-0.904	-0.337	-0.132	-0.124	-0.055	-0.105	0.065	0.070	0.028
Honda	0.058	-0.880	-0.387	-0.028	0.226	0.063	-0.049	-0.086	-0.076

ML method Loadings for 9 factors

sp500	0.969	0.147	0.145	0.096	-0.009	-0.069	-0.011	-0.019	0.021
Amazon	0.413	0.834	0.168	-0.299	0.063	-0.090	-0.002	-0.005	-0.008
Google	0.824	-0.185	0.219	0.201	0.106	-0.233	0.087	0.149	-0.060
Toyota	0.914	-0.31	-0.007	-0.085	0.159	0.143	0.043	0.030	-0.001
Walmart	0.952	-0.01	-0.176	-0.122	-0.079	0.152	-0.104	0.010	-0.022
eBay	0.383	0.75	-0.477	0.209	0.000	0.069	0.054	0.018	-0.021
Apple	0.959	-0.19	0.151	0.053	-0.051	-0.120	0.008	0.100	0.009
Pepsi	0.879	0.23	-0.189	-0.157	-0.254	-0.034	-0.058	0.034	-0.013
CocaCola	0.824	-0.32	-0.364	-0.267	-0.135	0.140	0.145	-0.026	0.180
DowJones	0.942	0.13	0.174	0.202	0.057	-0.005	-0.065	-0.051	0.032
HSBC	-0.640	0445	0.591	0.063	-0.048	0.174	0.004	0.038	0.014
Chase	0.916	-0.86	0.316	-0.004	-0.085	0.023	0.090	-0.057	-0.048
Honda	-0.013	-0424	0.741	0.206	0.332	0.162	-0.062	0.028	-0.035

B1 Analysis:

multivariateNormality

Test	H	p value
Royston	124.4435	1.279754e-15

Univariate Normality

variable	statistic	p-val
Healthsleep_data.Age	0.8265	0.0294
Healthsleep_data.SleepDuration	4.6335	<0.001
Healthsleep_data.HeartRate	7.4626	<0.001

CI (bonferroni):

LCBS	UCBS
-5.857	-2.24
-0.647	-0.012
0.374	3.143

B2 Analysis:

Multivariate Normality

Test	H	p value
Royston	77.58203	1.011546e-16

Univariate normality

variable	statistic	P-Val
Healthsleep_data.Age	1.6605	2.00E-04
Healthsleep_data.SleepDuration	6.5557	<0.001
Healthsleep_data.HeartRate	6.6117	<0.001

CI (Bonferroni):

Group comparisons	Age (LL,UL)	sleep duration (LL,UL)	Heart rate (LL,UL)
Accountant vs Doctor	(1.9665577, 6.3755733)	(0.2515194, 0.9533682)	(-4.6022878, -1.2677024)
Accountant vs Lawyer	(-2.0069251, 2.64006498)	(-0.5194463,-0.22028443)	(-3.4297534,-0.08481464)
Doctor vs Lawyer	(1.9665577,6.3755733)	(0.2515194, 0.9533682)	(-4.6022878, -1.2677024)

R codes:

```
# Load required libraries

library (openxlsx)

library(readxl)

library ( psych )

library ( tableone )

library ( heplots )

library (MVN )

library ( biotools )

library (car )

library ( GPArotation )

library ( corrr )

library ( ggplot2 )

library (FactoMineR)

library (factoextra)

# Read data from Excel file

stocks_data = read_excel("StockData.xlsx")

stocks_data = stocks_data [2:14] #exclude the first column as it 's probably an index

head ( stocks_data , 5) # display the first 5 rows

# Create a table of numerical summaries

numerical_summaries = CreateTableOne ( data = stocks_data , includeNA = FALSE )

print ( numerical_summaries )

# Standardize the data and create boxplots of raw and standardized data

Standardized_Data = as.data.frame ( scale (stocks_data ))

par ( mfrow = c(1 ,2))

boxplot ( stocks_data , las = 2, main = "Raw Data ")

R = cor( Standardized_Data ) # correlation matrix of standardized data

R2 = cor( stocks_data ) #correlation matrix of raw data

boxplot ( Standardized_Data , las = 2, main= " Standardized Data ")

# Create a correlation plot

par ( mfrow = c(1 ,1))

library ("corrplot")

corrplot (R, is.corr =TRUE , col = c("lavender","mediumorchid","darkorchid"), tl.col = " black ")

# Perform factor analysis and calculate p-values for different numbers of factors

pvals=c()

for (k in 2:8) {

  pvals[k]=factanal(covmat = R, n.obs = 252, factors = k)$PVAL

}

pvals

# Set the number of factors

factors = 3

# Perform principal component analysis

Eig.Res=eigen(R)

loadings=Eig.Res$vectors[, 1:factors]

eigenvalues = Eig.Res$values[1:factors]

L=round(loadings %*% diag(sqrt(eigenvalues)), 3)

PSI=diag(diag(R - L %*% t(L)))

EPS = R - L %*% t(L) - PSI
```

communalities=round(rowSums(L^2), 3)

diag(PSI)

Perform factor analysis using maximum likelihood estimation (MLE)

FA.Res=fa(stocks_data, nfactors = factors, rotate = "none", fm = "ml")

L1=round(FA.Res$loadings[, 1:factors], 3)

PSI1=diag(FA.Res$uniquenesses)

EPS1=R - L1 %*% t(L1) - PSI1

H1=FA.Res$communality

Perform factor analysis with different rotation methods (PC method)

FA.Res.None=fa(stocks_data, nfactors = factors, rotate = "none", fm = "pc")

FA.Res.Varimax=fa(stocks_data, nfactors = factors, rotate = "varimax", fm = "pc")

FA.Res.Quartimax=fa(stocks_data, nfactors = factors, rotate = "quartimax", fm = "pc")

Print the results

print(FA.Res.None)

print(FA.Res.Varimax)

print(FA.Res.Quartimax)

Create factor diagrams for different rotation methods

par(mfrow = c(1, 1))

fa.diagram(FA.Res.None, cut = 0.3, simple = FALSE, main = "No Rotation")

fa.diagram(FA.Res.Varimax, cut = 0.3, simple = FALSE, main = "Varimax Rotation")

fa.diagram(FA.Res.Quartimax, cut = 0.3, simple = FALSE, main = "Quartimax Rotation")

Print selected companies for each factor

for (col in names(selected_companies_list)) {

  cat("Companies for column", col, ":", toString(selected_companies_list[[col]]), "\n")

}

Perform PCA using FactoMineR

PCA.Res=PCA(Standardized_Data, graph = TRUE, ncp = 2, scale.unit = TRUE)

Print PCA results

PCA.Res

str(PCA.Res)

Summary of PCA analysis

summary(PCA.Res)

Print coordinates of variables

print(PCA.Resvarcoord)

Create a scree plot

fviz_eig(PCA.Res, addlabels = TRUE, ylim = c(0, 70))

Calculate factor loadings using PCA.Res

factor_loadings=round(sweep(PCA.Resvarcoord, 2, sqrt(PCA.Res$eig[1:ncol(PCA.Res$var$coord), 1]), FUN = "/"), 2)

factor_loadings

Perform PCA using prcomp function

PCA.Res1=prcomp(Standardized_Data, scale. = TRUE)

round(PCA.Res1$rotation, 3)

 # Analysis B1

Load required libraries for Analysis B1

source ("projectfile.R")

library (heplots)

```
library (MVN)

library ( DescTools )

#read excel

Healthsleep_data =
read_excel("Healthsleepdata.xlsx")

head ( Healthsleep_data , 5) # display the first 5 rows

# Subset the data gender

H1=subset(data.frame(Healthsleep_data$Age,
Healthsleep_data$SleepDuration,
Healthsleep_data$HeartRate),
Healthsleep_data$Gender == "Male")

n1=nrow(H1)

H1Bar=c(mean(H1[,1]), mean(H1[,2]), mean(H1[,3]))

SCOV1=cov(H1)

# Subset the data for individuals with heart disease



H2=subset(data.frame(Healthsleep_data$Age,

Healthsleep_data$SleepDuration,
Healthsleep_data$HeartRate),
Healthsleep_data$Gender == "Female")

n2=nrow(H2)

H2Bar=c(mean(H2[,1]), mean(H2[,2]), mean(H2[,3]))

SCOV2=cov(H2)

# Set up plotting layout

par(mfrow = c(1, 2))

# Plot qq plots and perform multivariate normality
tests for population without heart disease

cqplot(H1, id.n = 3, main = "Chisquare qq plot for
Male ")

mvn(H1, mvnTest = "royston")

mvn(H1, mvnTest = "hz")

# Plot qq plots and perform multivariate normality
tests for population with heart disease

cqplot(H2, id.n = 3, main = "Chisquare qq plot for
female")

mvn(H2, mvnTest = "royston")

mvn(H2, mvnTest = "hz")

# Reset plotting layout

par(mfrow = c(1, 1))

# Perform Hotellings T2 Test

HotellingsT2Test (H1 , H2 , test = "chi")

library(mvtnorm)

# Perform simultaneous confidence intervals

MVN2Sample.HT.CIs.NumSum.f(n1, H1Bar, SCOV1,
n2, H2Bar, SCOV2 , conf.level=0.95 , alpha = .05 ,
mu0 = rep (0, length( H1Bar )), ContrastMAT = NULL ,
SigDig =3, var.eq = TRUE )
```

#b2 Analysis

```
# Subset the data for different ocuupation types and
visualize multivariate normality

new_data=data.frame(Healthsleep_data$Occupatio
n, Healthsleep_data$Age,
Healthsleep_data$SleepDuration,
Healthsleep_data$HeartRate)

# Occupation type Accountant

O1=subset(data.frame(Healthsleep_data$Age,
Healthsleep_data$SleepDuration,
Healthsleep_data$HeartRate),
Healthsleep_data$Occupation == "Accountant")

cqplot(O1, id.n = 3, main = "Chisquare qq plot of
Occupation Type Accountant")

# Occupation type Doctor

O2= subset(data.frame(Healthsleep_data$Age,
Healthsleep_data$SleepDuration,
```

```
Healthsleep_data$HeartRate),
Healthsleep_data$Occupation == "Doctor")

cqplot(O2, id.n = 3, main = "Chisquare qq plot of
Occupation Type Doctor")

# Occupation type lawyer

O3=subset(data.frame(Healthsleep_data$Age,
Healthsleep_data$SleepDuration,
Healthsleep_data$HeartRate),
Healthsleep_data$Occupation == "Lawyer")

cqplot(O3, id.n = 3, main = "Chisquare qq plot of
Occupation Type lawyer")

# Plot qq plots and perform multivariate normality
tests for population without heart disease

cqplot(O1, id.n = 3, main = "chi")

mvn(O1, mvnTest = "royston")

mvn(O1, mvnTest = "hz")

#ANOVA test for Age

anova_age=aov(Healthsleep_data$Age ~
Healthsleep_data$Occupation)

summary(anova_age)

# ANOVA test for sleepduration

anova_SleepDuration=aov(Healthsleep_data$SleepD
uration ~ Healthsleep_data$Occupation)

summary(anova_SleepDuration)

# ANOVA test for hr

anova_HeartRate=aov(Healthsleep_data$HeartRate
~ Healthsleep_data$Occupation)

summary(anova_HeartRate)

# Manova with car package

# Computing Means and Var/Cov Mats per groups

VMeans=statList(new_data[, -1], new_data[, 1], FUN
= colMeans)

VMat=statList(new_data[, -1], new_data[, 1], FUN =
var)

Ns=table(new_data[, 1])  # Sample sizes

p=3  # Number of variables

g=3  # Number of groups

n=nrow(new_data)

# Manually computing the W matrix

W=(Ns[1] - 1) * VMat[[1]] + (Ns[2] - 1) * VMat[[2]] +
(Ns[3] - 1) * VMat[[3]]

# Load necessary packages

library(heplots)

library(MVN)

# Split data by group

new_data_gr=split(new_data, new_data[, 1])

# Checking Multivariate normality for each group

par(mfrow = c(2, 2))

for(i in 1:g){

  X = new_data_gr[[i]][, -1]

  group_name = names(new_data_gr)[i]  # Get the
group name

  cqplot(X, id.n=3, main = paste("Chi-square QQ plot
for", group_name))

  print(mvn(X, mvnTest = "royston"))

  print(mvn(X, mvnTest = "hz"))

}

# MANOVA with built in function

Y = as.matrix(new_data[, -1])

Gr = as.factor(new_data[, 1])

# First fit a linear regression
```

```
LM.res = lm(Y ~ Gr)

# Call Manova, This command gives all Tests

# Wilk's lambda, Pillai, Hotelling-Lawley and Roy

SUM = summary(Manova(LM.res), "Wilks")

SUM

# The W matrix using the build it functions

W = SUM$SSPE

# Check which groups differ from each other

library(biotools)

mvpaircomp(LM.res, factor1 = "Gr", test = "Wilks",
adjust = "bonferroni")



# Check each variable for differences across groups

summary.aov(LM.res)

# Calculate simultaneous confidence intervals

k=p * g * (g - 1) / 2

ta=qt(0.05 / (2 * k), df = n - g, lower.tail = FALSE)

# Simultaneous CIs

i = 1 # Group 1

j = 2 # Group 2

v = 1 # Variable

LL = VMeans [[i]] - VMeans [[j]] - ta *sqrt ((1 /Ns[i] +
1/Ns[j]) * diag (W) / (n - g))

UL = VMeans [[i]] - VMeans [[j]] + ta *sqrt ((1 /Ns[i] +
1/Ns[j]) * diag (W) / (n - g))

cbind (LL , UL)

i = 1 # Group 1

j = 3 # Group 3

v = 1 # Variable

LL = VMeans [[i]] - VMeans [[j]] - ta *sqrt ((1 /Ns[i] +
1/Ns[j]) * diag (W) / (n - g))

UL = VMeans [[i]] - VMeans [[j]] + ta *sqrt ((1 /Ns[i] +
1/Ns[j]) * diag (W) / (n - g))

cbind (LL , UL)

i = 1 # Group 2

j = 2 # Group 3

v = 1 # Variable

LL = VMeans [[i]] - VMeans [[j]] - ta *

  sqrt ((1/Ns[i] + 1/Ns[j]) * diag (W) / (

    n - g))

UL = VMeans [[i]] - VMeans [[j]] + ta *

  sqrt ((1/Ns[i] + 1/Ns[j]) * diag (W) / (

    n - g))

cbind (LL , UL)
```

Sign the statement below and attach it with your project. You project will not be graded without it.

This project is entirely my work. I have not discussed this project with anybody in or out of class. I did not use any unauthorized sources such as help from tutoring services (in person or online), nor have I used **generative AI/Online assignment assisting websites** (such as Chat GPT, DALL-E, Chegg, course Hero, etc.). I understand and have complied with the Rowan University's academic integrity policies outlined in (`https://confluence.rowan.edu/display/POLICY/Academic+Integrity+Policy`).

I understand that failure to adhere to this policy will be considered as violation of academic integrity and the appropriate university procedures will be invoked (`https://sites.rowan.edu/academic-affairs/academic_integrity`).

NAME: Akhila Vitta

DATE: 04/13/2024

SIGNATURE: V.Akhila