```
# Load all the required packages
library(readr)
library(psych)
library(tidyverse)
library(Hmisc)
library(stats)
library(car)
library(dae)
library(corrplot)
library(dplyr)
library(moments)
library(ggplot2)
library(bios2mds)
library(factoextra)
library(PerformanceAnalytics)
library(gridExtra)

# Set Working Directory
setwd("C:/Users/subha/Desktop/GRAD695")


# Health Outcomes data analysis
health_out<-read.csv("health_outcomes.csv")
# Subset the data to only look at county data
health_out_cty <- subset(health_out, County!="")

#Summary statistics of health outcomes
knitr::kable(summary(health_out_cty[,c(4:7)]))

my_data <- health_out_cty[,c(4:7)]

# Correlation plot of health outcomes
chart.Correlation(my_data, histogram=TRUE, pch=19)


# Health Behaviors data analysis
health_beh<-read.csv("health_behaviours.csv")
health_beh_cty <- subset(health_beh, County!="")
summary(health_beh_cty[,c(4:8)])

#Summary statistics of health behaviours
knitr::kable(summary(health_beh_cty[,c(4:8)]))

my_data1 <- health_beh_cty[, c(4:8)]

# Correlation plot of health behaviours
chart.Correlation(my_data1, histogram=TRUE, pch=19)


# Clinical Care access data analysis
health_acc<-read.csv("clinical_care.csv")
health_acc_cty <- subset(health_acc, County!="")
summary(health_acc_cty[,c(4:6)])

# Summary statistics of clinical care access
knitr::kable(summary(health_acc_cty[,c(4:6)]))

my_data2 <- health_acc_cty[, c(4:6)]

# Correlation plot of clinical care access
chart.Correlation(my_data2, histogram=TRUE, pch=19)

# Socio-economic data analysis
health_soc<-read.csv("socio_economic_data.csv")
health_soc_cty <- subset(health_soc, County!="")

# Summary of socio-economic data
knitr::kable(summary(health_soc_cty[,c(4:10)]))

my_data3 <- health_soc_cty[, c(4:10)]

# Correlation plot of socio-economic data
chart.Correlation(my_data3, histogram=TRUE, pch=19)


#demographic data analysis
health_demo<-read.csv("demographic_data.csv")
health_demo_cty <- subset(health_demo, County!="")

# Summary of demographic data
knitr::kable(summary(health_demo_cty[,c(4:8)]))

my_data4 <- health_demo_cty[, c(4:8)]

# Correlaton plot of demographic data
chart.Correlation(my_data4, histogram=TRUE, pch=19)


health_ind<-read.csv("final_datasetV1.csv")
health_ind_cty <- subset(health_ind, County!="")

park_access<-read.csv("data_191957.csv")
#head(park_access)

hist(park_access$park_access, main = "Distribution of % people within half mile of a park")
knitr::kable(summary(park_access$park_access))
summary(park_access$park_access)

# Merge the park access and other data
merged_data<-merge(park_access, health_ind_cty, by.x = "countyFIPS", by.y = "FIPS", all.x = TRUE)

summary(merged_data)

clust_data <- merged_data[, c(6,9:12, 14:24)]
clust_data1 <- clust_data[complete.cases(clust_data),]
```

```
#Principal Component Analysis
x<-prcomp(clust_data1[,c(6:16)], retx=TRUE, center=TRUE, scale=TRUE)
summary(x)

# Scree Plot
fviz_screeplot(x)
biplot(x,scale=0, cex=1.3)

pcs<-data.frame(x$x[,1:5])

data1<-data.frame(clust_data1,pcs)


set.seed(123)

# function to compute total within-cluster sum of square

pa<-fviz_nbclust(data1[,c(17:21)], kmeans, method="wss")
pb<-fviz_nbclust(data1[,c(17:21)], kmeans, method="silhouette")
pc<-fviz_nbclust(data1[,c(17:21)], kmeans, nstart = 25,  method = "gap_stat", nboot = 50)

grid.arrange(pa,pb, pc, nrow=2)

# Plot the different clusters
k2 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 2, nstart = 25)
k3 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 3, nstart = 25)
k4 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 4, nstart = 25)
k5 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 5, nstart = 25)
k6 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 6, nstart = 25)
k7 <- kmeans(scale(clust_data1[, c(6:16)]), centers = 7, nstart = 25)

# plots to compare
p1 <- fviz_cluster(k2, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 2")
p2 <- fviz_cluster(k3, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 3")
p3 <- fviz_cluster(k4, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 4")
p4 <- fviz_cluster(k5, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 5")
p5 <- fviz_cluster(k6, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 6")
p6 <- fviz_cluster(k7, geom = "point",  data = scale(clust_data1[, c(6:16)])) + ggtitle("k = 7")

grid.arrange(p1, p2, p3, p4, p5, p6, nrow = 3)


final <- kmeans(scale(data1[, c(17:21)]), 3, nstart = 25)

#Print the centers, size
print(final$centers)
print(final$size)


final_dataset0<-data.frame(data1,final$cluster)
final_dataset<- final_dataset0[complete.cases(final_dataset0),]


#Summarize the median data for each cluster
final_dataset %>%
  group_by(final.cluster) %>%
  summarise(across(c(6:10), median))


final_dataset %>%
  group_by(final.cluster) %>%
  summarise(across(c(11:16), median))



# Box Plots
x1<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_college))+geom_boxplot()+ ggtitle("Distribution of % with college
education")+xlab("Cluster")+ylab("% with college education")

x2<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_unemp))+geom_boxplot()+ ggtitle("Distribution of %
unemployed")+xlab("Cluster")+ylab("% unemployed")

x3<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=median_inc))+geom_boxplot()+ ggtitle("Distribution of median household
income")+xlab("Cluster")+ylab("Median household income")

x4<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_gt_65))+geom_boxplot()+ ggtitle("Distribution of % greater than 65 years of
age")+xlab("Cluster")+ylab("% more than 65 years of age")

x5<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_black))+geom_boxplot()+ ggtitle("Distribution of %
Black")+xlab("Cluster")+ylab("% Black")

x6<-ggplot(data=final_dataset, aes(x=as.factor(final.cluster),y=perc_female))+geom_boxplot()+ ggtitle("Distribution of %
female")+xlab("Cluster")+ylab("% female")

grid.arrange(x1, x2, x3, x4, x5, x6, nrow = 3)


# Regression Plots for non socio-economic/demographic variables
y1<-
  ggplot(final_dataset, aes(x=park_access, y=perc_fair_poor_health, color=factor(final.cluster), shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "Park Access", color="Clusters", shape =
"Clusters")

y2<-
  ggplot(final_dataset, aes(x=perc_obese, y=perc_fair_poor_health, color=factor(final.cluster), shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "% Obese", color="Clusters", shape = "Clusters")
```

```
y3<-
  ggplot(final_dataset, aes(x=perc_food_insecure, y=perc_fair_poor_health, color=factor(final.cluster), shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "% Food Insecure", color="Clusters", shape =
"Clusters")

y4<-
  ggplot(final_dataset, aes(x=perc_uninsured, y=perc_fair_poor_health, color=factor(final.cluster), shape=factor(final.cluster))) +
  geom_smooth(method=lm, se=FALSE, fullrange=TRUE)+
  scale_shape_manual(values=c(3, 16, 17))+
  scale_color_manual(values=c('Green','#E69F00', '#56B4E9'))+
  theme(legend.position="top")+
  labs(title = "Impact of Variable on % with fair/poor health", y = "% with fair/poor health", x = "% Uninsured", color="Clusters", shape =
"Clusters")

grid.arrange(y1, y2, y3, y4, nrow = 2)



# Regression analysis by clusters
final_dataset$log_perc_fph<-log(final_dataset$perc_fair_poor_health)
cluster1<-subset(final_dataset, final.cluster == 1)
cluster2<-subset(final_dataset, final.cluster == 2)
cluster3<-subset(final_dataset, final.cluster == 3)


chart.Correlation(final_dataset[,c(1:5,14)], histogram=TRUE, pch=19)


# Cluster 1 analysis
c1_model1<-lm(log_perc_fph~park_access+perc_obese*perc_food_insecure*perc_uninsured, data=cluster1)
kable(tidy(c1_model1))

c1_model2<-lm(log_perc_fph~park_access+perc_food_insecure*perc_uninsured, data=cluster1)
kable(tidy(c1_model2))
anova(c1_model1, c1_model2)

par(mfrow=c(2,2))
plot(c1_model2)


# Cluster 2 analysis
c2_model1<-lm(log_perc_fph~park_access+perc_obese*perc_food_insecure*perc_uninsured, data=cluster2)
kable(tidy(c2_model1))

c2_model2<-lm(log_perc_fph~park_access+perc_food_insecure*perc_uninsured, data=cluster2)
kable(tidy(c2_model2))
anova(c2_model1, c2_model2)

par(mfrow=c(2,2))
plot(c2_model2)


# Cluster 3 analysis
c3_model1<-lm(log_perc_fph~park_access+perc_obese*perc_food_insecure*perc_uninsured, data=cluster3)
kable(tidy(c3_model1))

c3_model2<-lm(log_perc_fph~park_access+perc_food_insecure*perc_uninsured, data=cluster3)
kable(tidy(c3_model2))
anova(c3_model1, c3_model2)

par(mfrow=c(2,2))
plot(c3_model2)

c3_model3<-lm(log_perc_fph~park_access+perc_food_insecure+perc_uninsured, data=cluster3)
kable(tidy(c3_model3))
anova(c3_model1, c3_model3)
anova(c3_model2, c3_model3)

par(mfrow=c(2,2))
plot(c3_model3)
```