````
---
title: "mini_project"
author: "Andrea Sama (A59010582)"
date: "10/27/2021"
output:
  html_document:
    df_print: paged
---
```{r}
#read.csv("WisconsinCancer.csv")
```
# Save your input data file into your Project directory
```{r}
fna.data <- "WisconsinCancer.csv"
```

# Complete the following code to input the data and store as wisc.df
```{r}
wisc.df <- read.csv(fna.data, row.names=1)

```
```{r}
# We can use -1 here to remove the first column
wisc.data <- wisc.df[,-1]
```

```{r}
#Setting diagnosis as a factor
diagnosis <- as.factor(wisc.df$diagnosis)
diagnosis
```

#Question 1: How many observations are there in this dataset?

```{r}
nrow(wisc.data)
```

#Question 2: How many observations have a malignant diagnosis?

```{r}
table(diagnosis)
```

#Question 3: How many variables/features in the data are suffixed with _mean?

```{r}
grepl("_mean", colnames(wisc.data))
```
````

There are 10 that are suffixed with mean

```{r}
# Check column means and standard deviations
colMeans(wisc.data)

apply(wisc.data,2,sd)
```

```{r}
# Perform PCA on wisc.data by completing the following code
wisc.pr <- prcomp(scale(wisc.data))

```
```{r}
summary(wisc.pr)
```

#Question 4: From your results, what proportion of the original variance is captured by the first principal components (PC1)?

44.27%

#Question 5: How many principal components (PCs) are required to describe at least 70% of the original variance in the data?

Three of the PCs are required to describe at least 70% of the original variance, as the PC3 cumulative proportion is reported as 72%.

#Q6. How many principal components (PCs) are required to describe at least 90% of the original variance in the data?

Seven PCs are required to describe at least 90% of the original variance, PC7's cumulative proportion is 91%.

#Plot this:

```{r}
biplot(wisc.pr)
```
#Q7. What stands out to you about this plot? Is it easy or difficult to understand? Why?

This plot is not easy to understand, it is very crowded and I don't know what anything means.

#Making a better plot:

```{r}
```

```
# Scatter plot observations by components 1 and 2
plot(wisc.pr$x[,1:2], col = diagnosis ,
     xlab = "PC1", ylab = "PC2")
```

#Question 8: Generate a similar plot for principal components 1 and 3. What do you notice about these plots?

```{r}
wisc.pr
plot(wisc.pr$x[,c(1,3) ], col = diagnosis,
     xlab = "PC1", ylab = "PC3")
```

The second plot is shifted down and there is more overlap between the diagnosis.
#making ggplot:

```{r}
# Create a data.frame for ggplot
df <- as.data.frame(wisc.pr$x)
df$diagnosis <- diagnosis

# Load the ggplot2 package
library(ggplot2)

# Make a scatter plot colored by diagnosis
ggplot(df) +
  aes(PC1, PC2, col=diagnosis) +
  geom_point()
```

#Calculating variance
```{r}
# Calculate variance of each component
pr.var <- wisc.pr$sdev^2
head(pr.var)
```

```{r}
# Variance explained by each principal component: pve
pve <- pr.var / sum(pr.var)

# Plot variance explained for each principal component
plot(pve, xlab = "Principal Component",
     ylab = "Proportion of Variance Explained",
     ylim = c(0, 1), type = "o")
```

```{r}
# Alternative scree plot of the same data, note data driven y-axis
barplot(pve, ylab = "Precent of Variance Explained",
     names.arg=paste0("PC",1:length(pve)), las=2, axes = FALSE)
axis(2, at=pve, labels=round(pve,2)*100 )
```

```{r}

```

#Q9. For the first principal component, what is the component of the loading vector (i.e. wisc.pr$rotation[,1]) for the feature concave.points_mean?

```{r}
wisc.pr$rotation[,1]

```

The component of concave.points_mean is -0.26085376.

#Q10. What is the minimum number of principal components required to explain 80% of the variance of the data?

```{r}
summary(wisc.pr)
```
You need 5 PCs to explain 80% of the variance in the data.

##Hierarchical clustering

```{r}
# Scale the wisc.data data using the "scale()" function
data.scaled <- scale(wisc.data)
```

```{r}
#calculating elucidian data
data.dist <- dist(data.scaled)
#data.dist
```
#Q11. Using the plot() and abline() functions, what is the height at which the clustering model has 4 clusters?
```{r}
wisc.hclust <- hclust(data.dist, method= "complete")
wisc.hclust
plot(wisc.hclust)
```

```
abline(h=19, col="red", lty=2)
```

The height at which there is 4 clusters is around 19.

```{r}
wisc.hclust.clusters <- cutree(wisc.hclust, 4)
wisc.hclust.clusters
```

```{r}
table(wisc.hclust.clusters, diagnosis)
```

#Q12. Can you find a better cluster vs diagnoses match by cutting into a different number of clusters between 2 and 10?

```{r}
wisc.hclust.clusters <- cutree(wisc.hclust, 4)
table(wisc.hclust.clusters, diagnosis)

```
We like 4 because that is where the diagnosis information really splits off from each other.

#Q13. Which method gives your favorite results for the same data.dist dataset? Explain your reasoning.

```{r}
wisc.hclust1 <- hclust(data.dist, method= "ward.D2")
wisc.hclust1
plot(wisc.hclust1)
abline(h=80, col='red')
```
We like ward.D2 because it splits it into two separate clusters initially.

Attempt at individually coding the below:
```{r}
data.scaled.90 <- scale(wisc.data[,1:7])
data.dist.90 <- dist(data.scaled.90)
wisc.pr.hclust1 <- hclust(data.dist.90, method= "ward.D2")
wisc.pr.hclust1
plot(wisc.pr.hclust1)
```

CLuster my PCA results
I will use 7 PCs and `hclust` and `dist()` as an input

```{r}
wisc.pr.hclust <-hclust(dist(wisc.pr$x[,1:7]), method="ward.D2")
wisc.pr.hclust
plot(wisc.pr.hclust)
abline(h=80, col='red')
```

Lets fnd our cluster memebership vecotr by cutting this tree into k=2 groups

```{r}
grps <-cutree(wisc.pr.hclust, k=2)
table(grps)
```
We can do a cross table by giving the `table()` two inputs.
We get True positives:188, False positives: 28, True negatives: 329, and false negatives: 24
```{r}
table(grps, diagnosis)
```
Accuracy, essentially how many did we get correct?
```{r}
(188+329)/nrow(wisc.data)
```
Sensitivity/Specificity:
sensitivity refers to tests ability to correctly detect ill patients who do have the condition. In our example here the sensitivity is the total number of samples in the cluster identified as predominantly malignant (cancerous) divided by the total number of known malignant samples. In other words: TP/(TP+FN).
```{r}
188/(188+24)
```

Specificity relates to a test's ability to correctly reject healthy patients without a condition. In our example specificity is the proportion of benign (not cancerous) samples in the cluster identified as predominantly benign that are known to be benign. In other words: TN/(TN+FN).

```{r}
329/(329+24)
```

We would want to increase the sensitivity as much as possible because it would be better to say a healthy patient is sick than a sick patient is healthy. That would be limited if we increased the sensitivity.

#7: Predictions

```{r}
url <- "new_samples.csv"
url <- "https://tinyurl.com/new-samples-CSV"
new <- read.csv(url)
npc <- predict(wisc.pr, newdata=new)
npc
```

plotting:

```{r}
plot(wisc.pr$x[,1:2], col=diagnosis)
points(npc[,1], npc[,2], col="blue", pch=16, cex=3)
text(npc[,1], npc[,2], c(1,2), col="white")
```

Part 5 continued:

```{r}
plot(wisc.pr$x[,1:2], col=grps)
```

```{r}
plot(wisc.pr$x[,1:2], col=diagnosis)
```

#Halloween Candy Project:

```{r}
candy=read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/
candy-power-ranking/candy-data.csv", row.names=1)

candy_file<-"https://raw.githubusercontent.com/fivethirtyeight/data/master/
candy-power-ranking/candy-data.csv"

head(candy)
head(candy)

```
Q1. How many different candy types are in this dataset?

```{r}
nrow(candy)
```

Q2. How many fruity candy types are in the dataset?

```{r}
rownames(candy)
rownames(candy)<-gsub("Õ","'", rownames(candy))
```

```{r}
sum(candy$fruity)
```

```{r}
candy["Twix",]$winpercent
```

Q3. What is your favorite candy in the dataset and what is it's winpercent value?
Twix actually is my favorite candy, so its winpercent is 81.64291

Q4. What is the winpercent value for "Kit Kat"?

```{r}
candy["Kit Kat",]$winpercent
```

Q5. What is the winpercent value for "Tootsie Roll Snack Bars"?

```{r}
candy["Tootsie Roll Snack Bars",]$winpercent
```

```{r}
#install.packages("skimr")
library("skimr")
skim(candy)
```

#Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?
The last three rows are percentage, and row 12 is multiplied by 100.

#Q7. What do you think a zero and one represent for the candy$chocolate column?
yes or no.

#Q8. Plot a histogram of winpercent values

```{r}
hist(candy$winpercent)
```

#Q9. Is the distribution of winpercent values symmetrical?
No. It is shifted to 40 ish percent with few at 90.
#Q10. Is the center of the distribution above or below 50%?
Below 50.
#Q11. On average is chocolate candy higher or lower ranked than fruit candy?
```{r}
chocolate <- candy[as.logical(candy$chocolate),]$winpercent
mean(chocolate)

fruity <- candy[as.logical(candy$fruity),]$winpercent
mean(fruity)
```
chocolate candy is higher ranked than fruity.
#Q12. Is this difference statistically significant?
```{r}
t.test(chocolate, fruity)
```
It is significant.

#Q13. What are the five least liked candy types in this set?

```{r}
#head(candy[order(candy$winpercent),], n=5)
```

#Q14. What are the top 5 all time favorite candy types out of this set?

```{r}
#head(candy[order(candy$winpercent, decreasing=TRUE),], n=5)
```
#Q15. Make a first barplot of candy ranking based on winpercent values.
HINT: Use the aes(winpercent, rownames(candy)) for your first ggplot like so:

```{r}
#library(ggplot2)

#ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col()

```
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```{r}
ggplot(candy) +
aes(winpercent, reorder(rownames(candy),winpercent)) +
```

```
  geom_col()
```
Making colors

```{r}
my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "brown"
my_cols[as.logical(candy$fruity)] = "pink"
```

```{r}
ggplot(candy) +
aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)
```
- Q17. What is the worst ranked chocolate candy?
The worst is Sixlets
- Q18. What is the best ranked fruity candy?
The best fruity candy is Starbursts

```{r}
#install.packages("ggrepel")
library(ggrepel)

# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 20)
```
#Q19. Which candy type is the highest ranked in terms of winpercent for the least money - i.e. offers the most bang for your buck?
Reeses Miniatures

#Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?
Nik L Nip, Smarties, Ring Pop, Hersheys Krackel, Hersheys Milk Chocolate. Nik L Nip is the least popular.

```{r}
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```
```{r}
#install.packages("corrplot")
#library(corrplot)
```
```{r}
```

```
#cij <- cor(candy)
#corrplot(cij)
```

#Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?
Chocolate and fruity

#Q23. Similarly, what two variables are most positively correlated?
Chocolate and winpercent

Section 6

```{r}
#pca<-prcomp(candy, scale=TRUE )
#summary(pca)
```

```{r}
#plot(pca$x[,1:2])
```

```{r}
#plot(pca$x[,1:2], col=my_cols, pch=16)
```
```{r}
# Make a new data-frame with our PCA results and candy data
#my_data <- cbind(candy, pca$x[,1:3])
```

```{r}
#p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

#p
```

```{r}
#library(ggrepel)

#p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
 # theme(legend.position = "none") +
  #labs(title="Halloween Candy PCA Space",
       #subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light  brown), fruity (red), other (black)",
```

```
       #caption="Data from 538")
```
```{r}
#library(plotly)
#ggplotly(p)
```

```{r}
#par(mar=c(8,4,2,2))
#barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```

Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

Fruity, hard, pluribus. Most fruity candies are hard and pluribus so it makes sense that those would be clustered similarly.

```{r}
rownames(candy)
rownames(candy)<-gsub("Õ","'", rownames(candy))
```