~~~
---
title: "Class 10"
author: 'Shivani Khosla (PID: A59010433)'
date: "10/29/2021"
output:
  html_document: default
  pdf_document: default
---
```{r}
candy_file <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/candy-data.csv"
candy = read.csv(candy_file, row.names=1)
head(candy)
rownames(candy) <- gsub("Õ", "", rownames(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}
sum(candy$fruity)

```

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

```

> Q3. What is your favorite candy in the dataset and what is it’s winpercent value?

```{r}
candy["Almond Joy", ]$winpercent

```
> 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}
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?

Columns 10-12 are not binary values since they are percentages

> Q7. What do you think a zero and one represent for the candy$chocolate column?
~~~

They are yes(1) or no(0) values.

> Q8. Plot a histogram of winpercent values

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

> Q9. Is the distribution of winpercent values symmetrical?

The distribution is skewed right

> Q10. Is the center of the distribution above or below 50%?

The center is 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)
```
```{r}
fruity <- candy[as.logical(candy$fruity), ]$winpercent
mean(fruity)
```

> Q12. Is this difference statistically significant?

```{r}
t.test(chocolate, fruity)
```

p = 2.871e-08 so the difference is statistically 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)
```

Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, Snickers

> Q15. Make a first barplot of candy ranking based on winpercent values.

```{r}
library(ggplot2)
rownames(candy) <- gsub("Õ", "'", rownames(candy))
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}
library(ggplot2)

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

```

for 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"
```

incorporate colors

```{r}
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +      geom_col(fill=my_cols)
```

> Q17. What is the worst ranked chocolate candy?

Sixlets

> Q18. What is the best ranked fruity candy?

Starburst

```{r}
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 = 5)
```

> 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?

Reese's miniatures

> Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

```{r}
head(candy[order(candy$pricepercent, decreasing = TRUE), ], n = 5)

```

Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, Hershey's Milk Chocolate

least popular: Nik L Nip

```{r}
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

```{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)

```
```{r}
#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. We discussed that this makes sense since fruity candies tend to be hard and come packaged with many other fruity candies.