Class 10: Halloween Mini-Project

Norman Lee A18086849

Background

In this mini-project, you will explore FiveThirtyEight's Halloween Candy dataset. FiveThirtyEight, sometimes rendered as just 538, is an American website that focuses mostly on opinion poll analysis, politics, economics, and sports blogging. They recently ran a rather large poll to determine which candy their readers like best. From their website: "While we don't know who exactly voted, we do know this: 8,371 different IP addresses voted on about 269,000 randomly generated candy matchups". task is to explore their candy dataset to find out answers to these types of questions - but most of all your job is to have fun, learn by doing hands on data analysis, and hopefully make this type of analysis less frightining for the future! Let's get started.

1. Importing candy data

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

candy <- read.csv(candy_file, row.names = 1)
head(candy)
```

```
             chocolate fruity caramel peanutyalmondy nougat crispedricewafer
100 Grand            1      0       1              0      0                 1
3 Musketeers         1      0       0              0      1                 0
One dime             0      0       0              0      0                 0
One quarter          0      0       0              0      0                 0
Air Heads            0      1       0              0      0                 0
Almond Joy           1      0       0              1      0                 0
             hard bar pluribus sugarpercent pricepercent winpercent
100 Grand       0   1        0        0.732        0.860   66.97173
3 Musketeers    0   1        0        0.604        0.511   67.60294
```

```
One dime         0   0         0         0.011         0.116   32.26109
One quarter      0   0         0         0.011         0.511   46.11650
Air Heads        0   0         0         0.906         0.511   52.34146
Almond Joy       0   1         0         0.465         0.767   50.34755
```

- **Q1**. How many different candy types are in this dataset

```
nrow(candy)
```

```
[1] 85
```

85

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

```
sum(candy$fruity)
```

```
[1] 38
```

38

2. What is your favorate candy?

One of the most interesting variables in the dataset is `winpercent`. For a given candy this value is the percentage of people who prefer this candy over another randomly chosen candy from the dataset (what 538 term a matchup). Higher values indicate a more popular candy.

We can find the `winpercent` value for Twix by using its name to access the corresponding row of the dataset. This is because the dataset has each candy name as `rownames` (recall that we set this when we imported the original CSV file). For example the code for Twix is:

candy[“Twix”,]$winpercent

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

```
candy["Nestle Crunch",]$winpercent
```

```
[1] 66.47068
```

66.47

- **Q4**. What is the `winpercent` value for “Kit Kat”?

```
candy["Kit Kat",]$winpercent
```

```
[1] 76.7686
```

76.77

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

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

```
[1] 49.6535
```

49.65

load skimr package

```
#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?

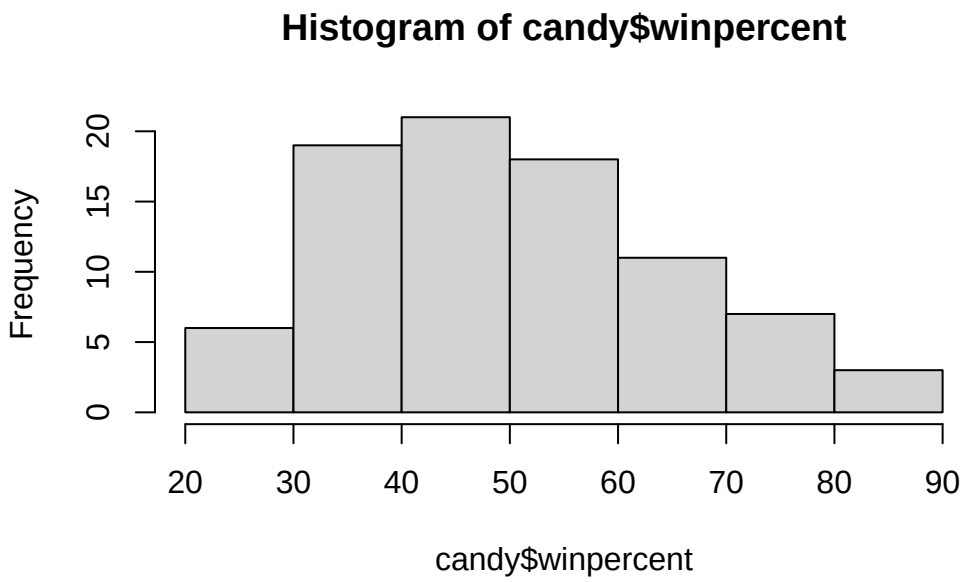
The winpercent variable is on a different scale (0–100) while most others are binary (0 or 1)

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

A value of 1 means the candy contains chocolate, and 0 means it does not contain chocolate.

Q8. Plot a histogram of `winpercent` values

```
hist(candy$winpercent,)
```



Q9. Is the distribution of `winpercent` values symmetrical?

no its slightly skewed to the right

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

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

slightly above

Q11. On average is chocolate candy higher or lower ranked than fruit candy?

```
mean(candy$winpercent[as.logical(candy$chocolate)])
```

```
[1] 60.92153
```

```
mean(candy$winpercent[as.logical(candy$fruity)])
```

```
[1] 44.11974
```

chocolate>candy

Q12. Is this difference statistically significant?

```
t.test(candy$winpercent[as.logical(candy$chocolate)],
       candy$winpercent[as.logical(candy$fruity)])
```

```
    Welch Two Sample t-test

data:  candy$winpercent[as.logical(candy$chocolate)] and candy$winpercent[as.logical(candy$fr
t = 6.2582, df = 68.882, p-value = 2.871e-08
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 11.44563 22.15795
sample estimates:
mean of x mean of y
 60.92153  44.11974
```

p<0.01, the difference is statistically significant

3. Overall Candy Rankings

Let's use the base R `order()` function together with `head()` to sort the whole dataset by `winpercent`. Or if you have been getting into the tidyverse and the **dplyr** package you can use the `arrange()` function together with `head()` to do the same thing and answer the following questions:

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

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
# Least liked
candy %>% arrange(winpercent) %>% head(5)
```

```
                  chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                 0      1       0              0      0
Boston Baked Beans        0      0       0              1      0
Chiclets                  0      1       0              0      0
Super Bubble              0      1       0              0      0
Jawbusters                0      1       0              0      0
                  crispedricewafer hard bar pluribus sugarpercent pricepercent
Nik L Nip                        0    0   0        1        0.197        0.976
Boston Baked Beans               0    0   0        1        0.313        0.511
Chiclets                         0    0   0        1        0.046        0.325
Super Bubble                     0    0   0        0        0.162        0.116
Jawbusters                       0    1   0        1        0.093        0.511
                  winpercent
Nik L Nip           22.44534
Boston Baked Beans  23.41782
Chiclets            24.52499
Super Bubble        27.30386
Jawbusters          28.12744
```

```r
# Most liked
candy %>% arrange(desc(winpercent)) %>% head(5)
```

```
                        chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           1      0       0              0      0
Snickers                          1      0       1              1      1
                        crispedricewafer hard bar pluribus sugarpercent
Reese's Peanut Butter cup               0    0   0        0        0.720
Reese's Miniatures                      0    0   0        0        0.034
Twix                                    1    0   1        0        0.546
Kit Kat                                 1    0   1        0        0.313
Snickers                                0    0   1        0        0.546
```

	pricepercent	winpercent
Reese's Peanut Butter cup	0.651	84.18029
Reese's Miniatures	0.279	81.86626
Twix	0.906	81.64291
Kit Kat	0.511	76.76860
Snickers	0.651	76.67378

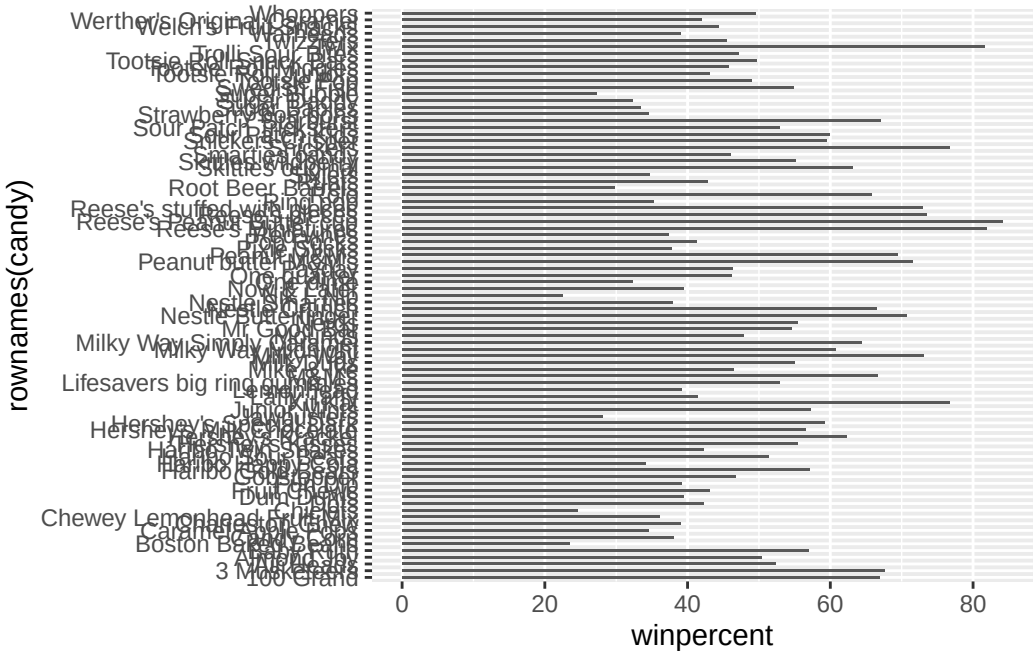
least liked: nik l nip, boston baked beans, chiclets, super bubble, jawbusters

- **q14.** What are the top 5 all time favorite candy types out of this set?

most liked: reese peanut butter cu and miniatures, twix, kitkat , snickers

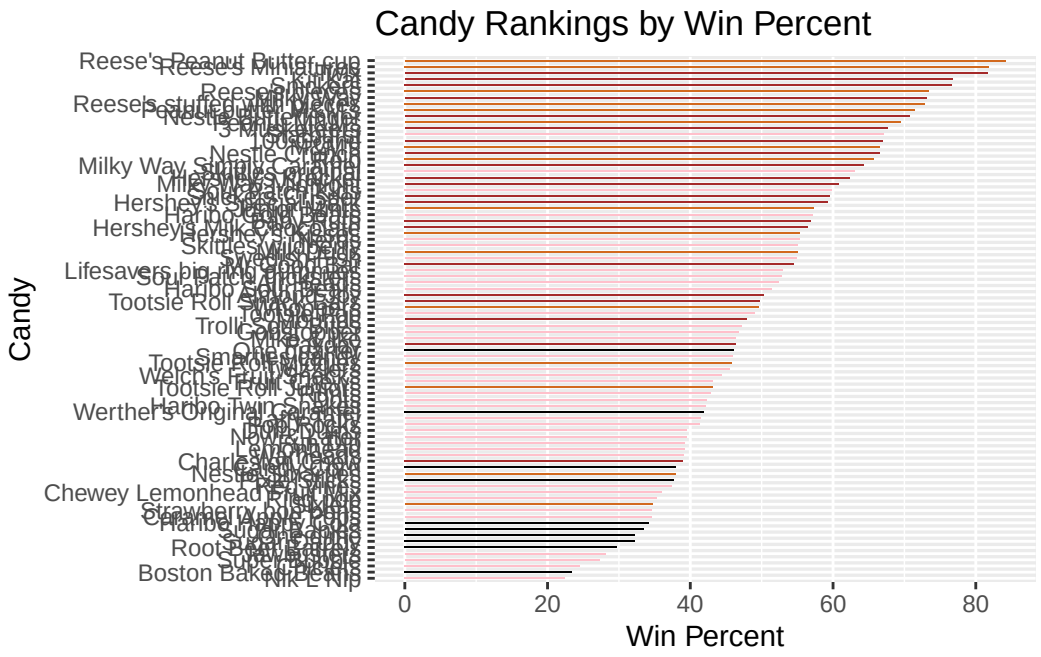
Q15. Make a first barplot of candy ranking based on `winpercent` values.

```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col(width = 0.3, position = position_dodge(0.7))
```



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

```
library(ggplot2)
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"
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col(width = 0.3, position = position_dodge(30),fill=my_cols) +
  labs(
    title = "Candy Rankings by Win Percent",
    x = "Win Percent",
    y = "Candy"
  )
```



- **Q17**. What is the worst ranked chocolate candy?
nip l nip

- **Q18**. What is the best ranked fruity candy?

reeses peanut butter cup

4. Taking a look at pricepercent

```
#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?

```
candy$value <- candy$winpercent / candy$pricepercent
candy[order(-candy$value), ][1:5, c("winpercent", "pricepercent")]
```

```
                     winpercent pricepercent
Tootsie Roll Midgies   45.73675        0.011
Pixie Sticks           37.72234        0.023
Fruit Chews            43.08892        0.034
Dum Dums               39.46056        0.034
Strawberry bon bons    34.57899        0.058
```

- Tootsie Roll Midgies

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

```
ord <- order(candy$pricepercent, decreasing = TRUE)
head(candy[ord, c("pricepercent", "winpercent")], 5)
```

```
                        pricepercent winpercent
Nik L Nip                      0.976   22.44534
Nestle Smarties                0.976   37.88719
Ring pop                       0.965   35.29076
Hershey's Krackel              0.918   62.28448
Hershey's Milk Chocolate       0.918   56.49050
```

nik l nip

5 Exploring the correlation structure

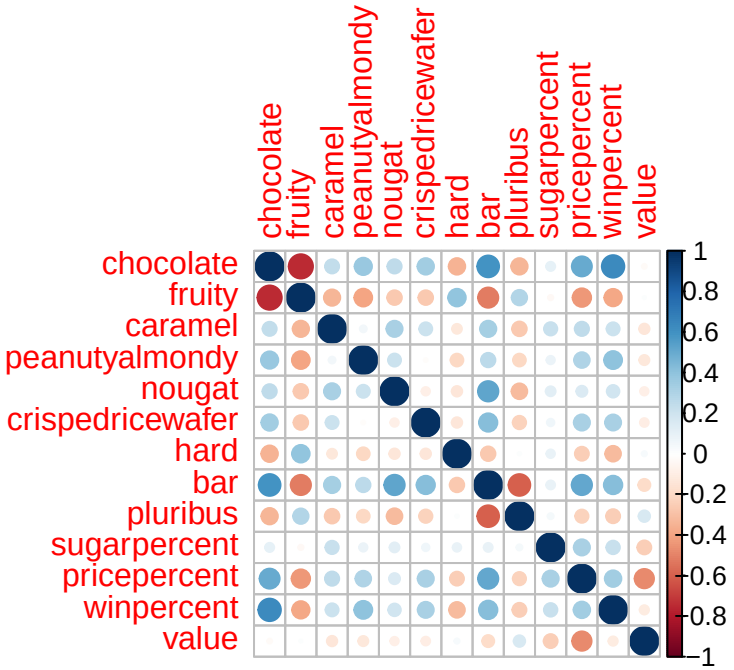
Now that we've explored the dataset a little, we'll see how the variables interact with one another. We'll use correlation and view the results with the **corrplot** package to plot a correlation matrix.

```
library(corrplot)
```

```
Warning: package 'corrplot' was built under R version 4.5.2
```

```
corrplot 0.95 loaded
```

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



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

  fruity and chocolate

- **Q23**. Similarly, what two variables are most positively correlated?

chocolate and bar

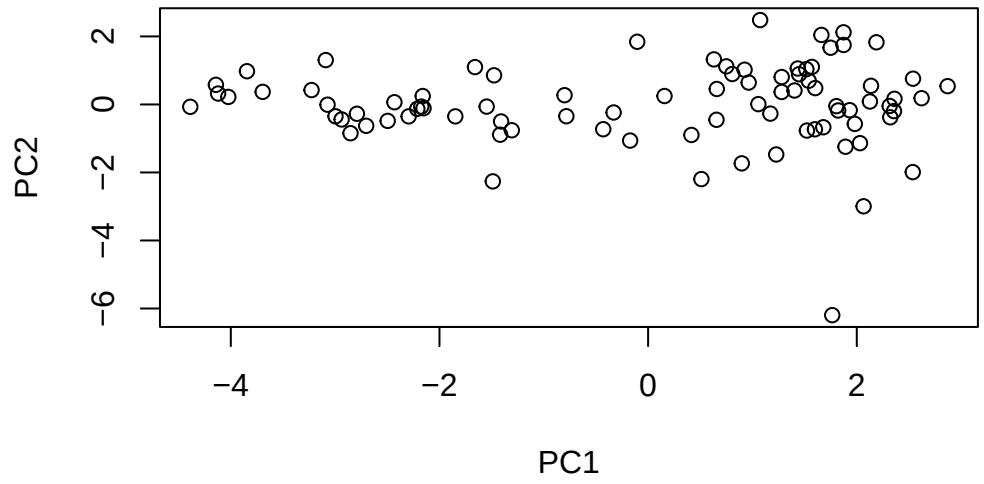
6. Principal Component Analysis

Let's apply PCA using the `prcom()` function to our candy dataset remembering to set the `scale=TRUE` argument.

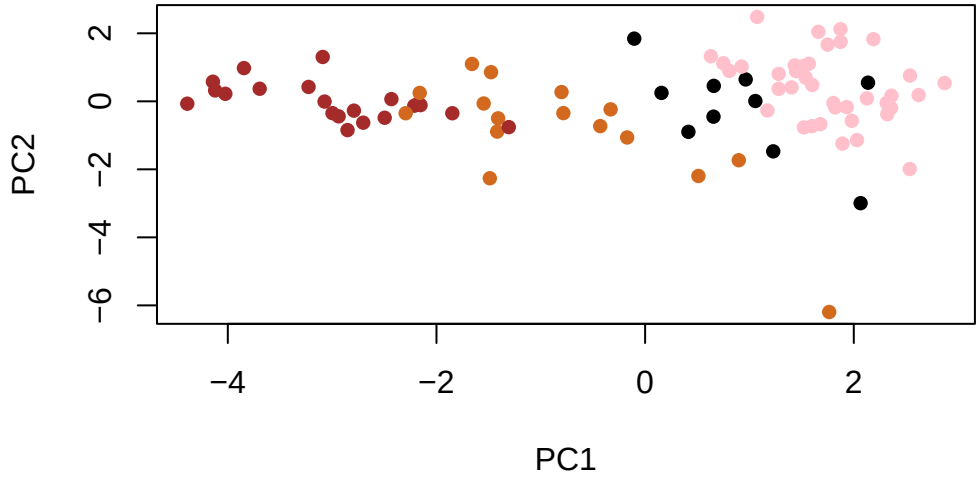
```
pca <- prcomp(candy, scale = TRUE)
summary(pca)
```

```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7
Standard deviation     2.0938 1.2127 1.13054 1.0787 0.98027 0.93656 0.81530
Proportion of Variance 0.3372 0.1131 0.09832 0.0895 0.07392 0.06747 0.05113
Cumulative Proportion  0.3372 0.4503 0.54866 0.6382 0.71208 0.77956 0.83069
                           PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     0.78462 0.68466 0.66328 0.57829 0.43128 0.39534
Proportion of Variance 0.04736 0.03606 0.03384 0.02572 0.01431 0.01202
Cumulative Proportion  0.87804 0.91410 0.94794 0.97367 0.98798 1.00000
```

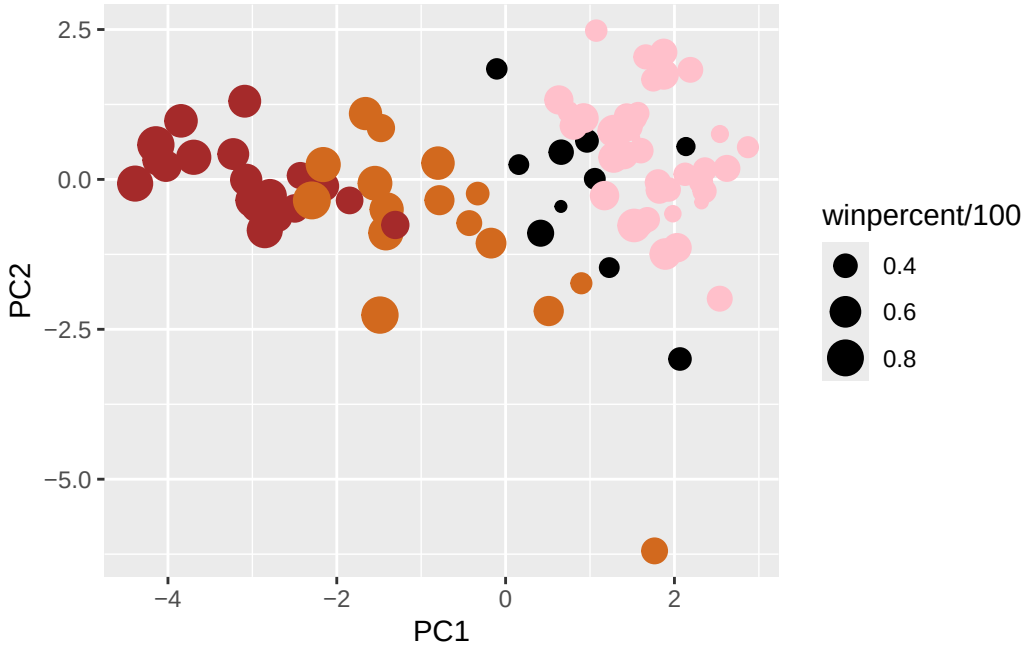
```
plot(pca$x[, 1:2])
```



```
plot(pca$x[,1:2], col=my_cols, pch=16)
```




```
my_data <- cbind(candy, pca$x[,1:3])
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
```

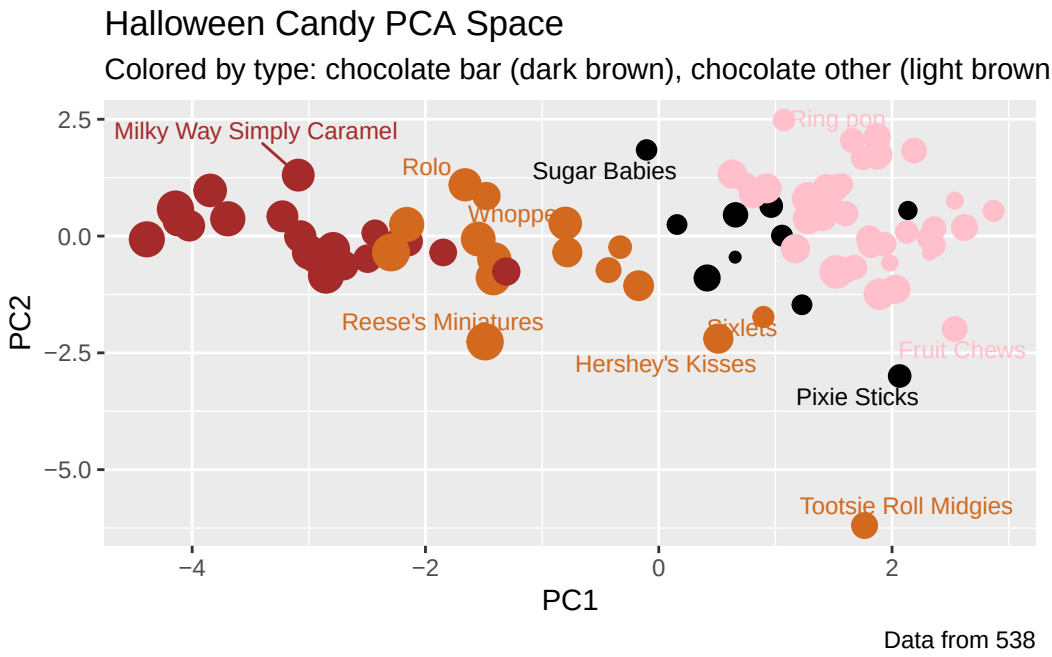



```
library(ggrepel)
```

```
Warning: package 'ggrepel' was built under R version 4.5.2
```

```
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),
       caption="Data from 538")
```

```
Warning: ggrepel: 74 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




```
library(plotly)
```

```
Warning: package 'plotly' was built under R version 4.5.2
```

```
Attaching package: 'plotly'
```

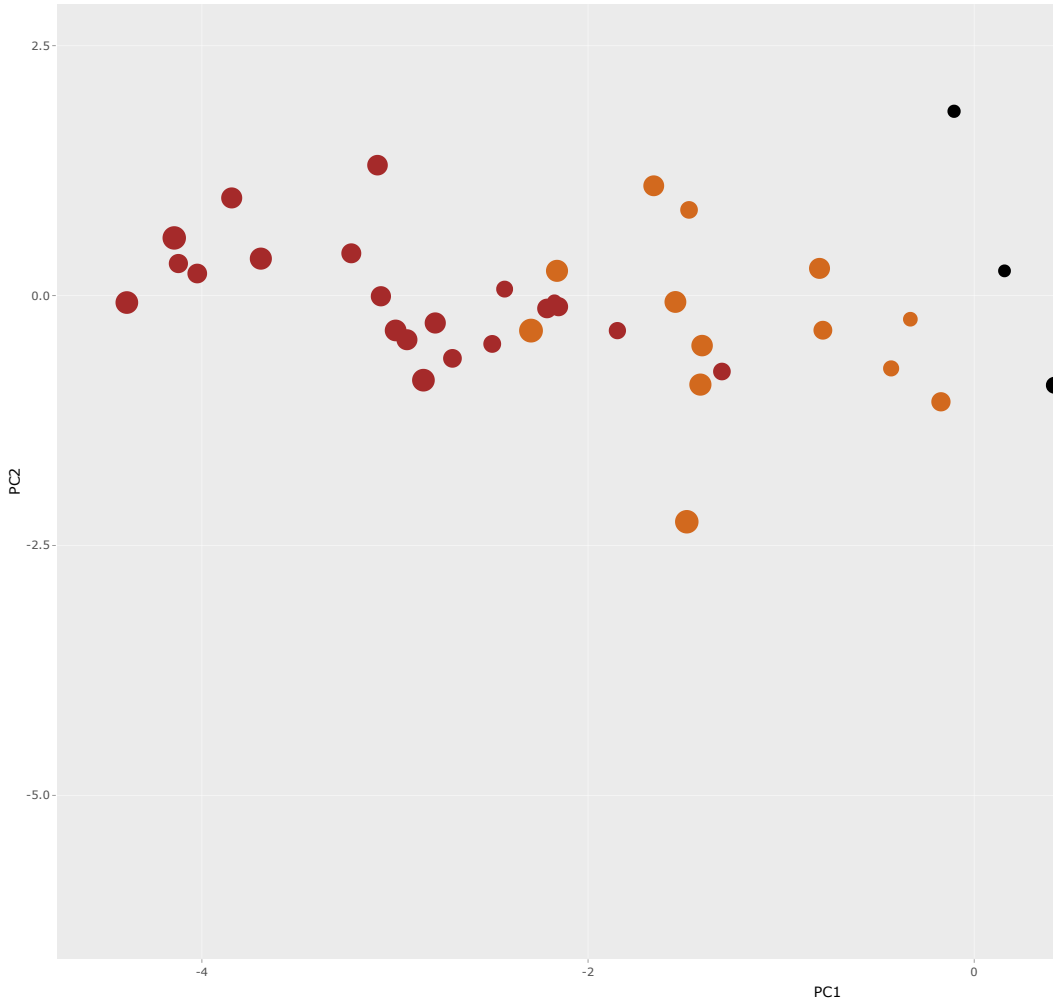
```
The following object is masked from 'package:ggplot2':

    last_plot
```

```
The following object is masked from 'package:stats':

    filter
```

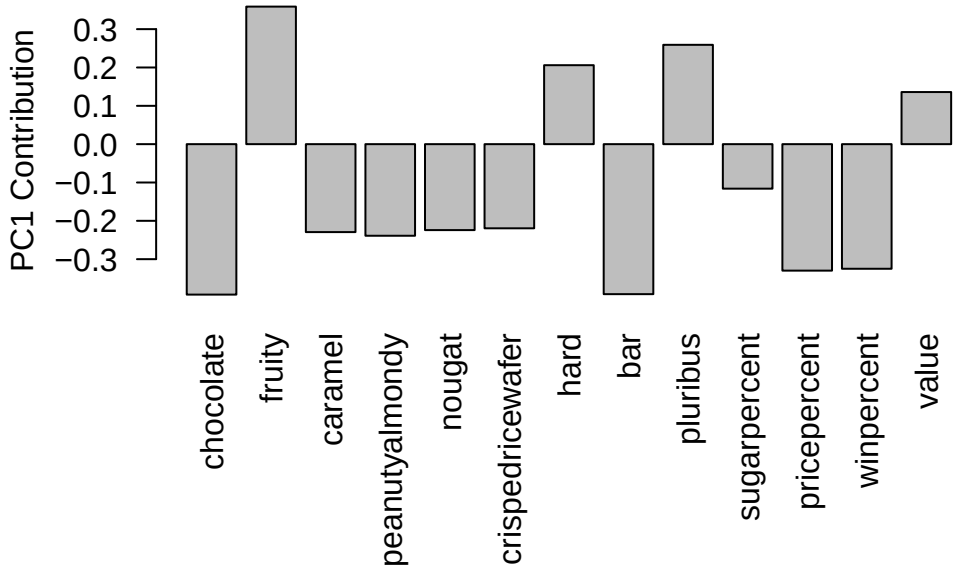
```
The following object is masked from 'package:graphics':

    layout
```

```
ggplotly(p)
```

```
file:///C:/Users/leeho/AppData/Local/Temp/RtmpoZ3lX4/file64bc755c280e/widget64bcbb462bf.html
```


2.5
0.0
-2.5
-5.0
PC2
-4
-2
0
PC1

```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?

  Positive side = fruity, hard candies, often sold in bags (pluribus) e.g., Skittles, Starburst, Jolly Ranchers. Pc1 shows a contrast between fruity bagged candies and chocolate bars