Class 10: Halloween Mini Project

Jessica Diaz-Vigil

Importing Candy Data

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

  There are 85 different types of candy in this dataset.

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

```
table(candy$fruity)
```

```
 0  1
47 38
```

There are 38 types of candy that are fruity in the dataset. We know 1 means `True` and 0 means `False`.

What is your favorite candy?

We can look at favorite candies by looking at larger values under the `winpercent` column.

```
candy["Twix", ]$winpercent
```

```
[1] 81.64291
```

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

  ```
  candy["Twix", ]$winpercent
  ```

  ```
  [1] 81.64291
  ```

  ```
  candy["Laffy Taffy", ]$winpercent
  ```

  ```
  [1] 41.38956
  ```

  My favorite type of candy is Twix which has a `winpercent` value of 81.62, but since it was already looked at, my second is Laffy Taffy which has a `winpercent` of 41.39.

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

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

  ```
  [1] 76.7686
  ```

  The `winpercent` value for Kit Kats is 76.77.

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

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

```
[1] 49.6535
```

The `winpercent` value for Kit Kats is 76.77.

Skim( ) Function

```
#install.packages("skimr")
library("skimr")
```

```
skim(candy)
```

Table 1: Data summary

Name	candy
Number of rows	85
Number of columns	12
_____	
Column type frequency:	
numeric	12
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.00	1.00	
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00	
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
nougat	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00	
crispedricewafer	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00	
hard	0	1	0.18	0.38	0.00	0.00	0.00	0.00	1.00	
bar	0	1	0.25	0.43	0.00	0.00	0.00	0.00	1.00	
pluribus	0	1	0.52	0.50	0.00	0.00	1.00	1.00	1.00	
sugarpercent	0	1	0.48	0.28	0.01	0.22	0.47	0.73	0.99	
pricepercent	0	1	0.47	0.29	0.01	0.26	0.47	0.65	0.98	
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18	

- **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 columns `n_missing` and `complete_rate` are different from the other columns since they are only ones and zeros. This is to ensure that the data was completed adequately and in our case, they all were since there is a one in every `complete_rate` row.

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

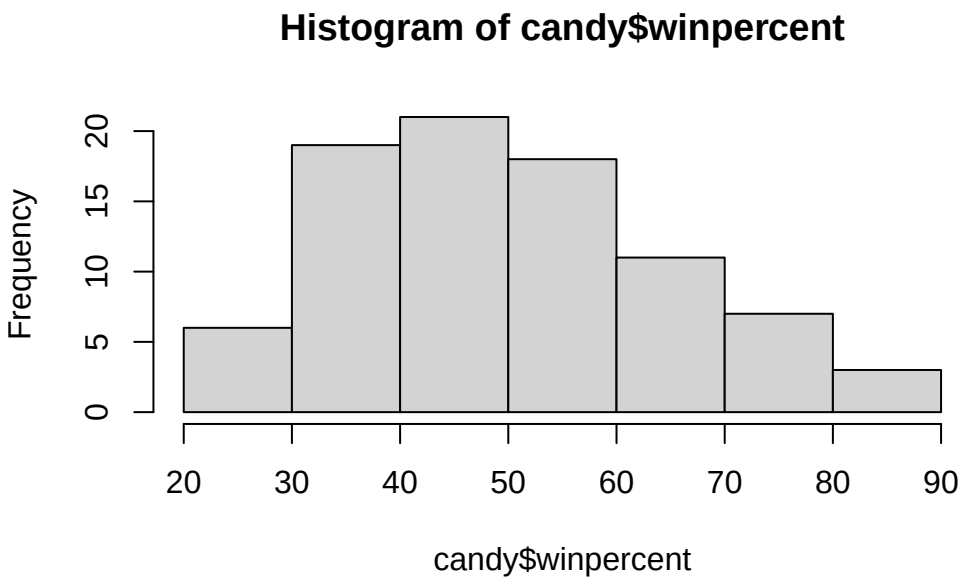
```
candy$chocolate
```

```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

  The zero in this case means that the candy does not have chocolate in it, whereas the one in this case means that candy does have chocolate in it.

- **Q8**. Plot a histogram of `winpercent` values

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



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

No, they are not. In this histogram we can see that there are a lot more candies with a `winpercent` of 30-60% and there is a decrease the higher the `winpercent` values are.

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

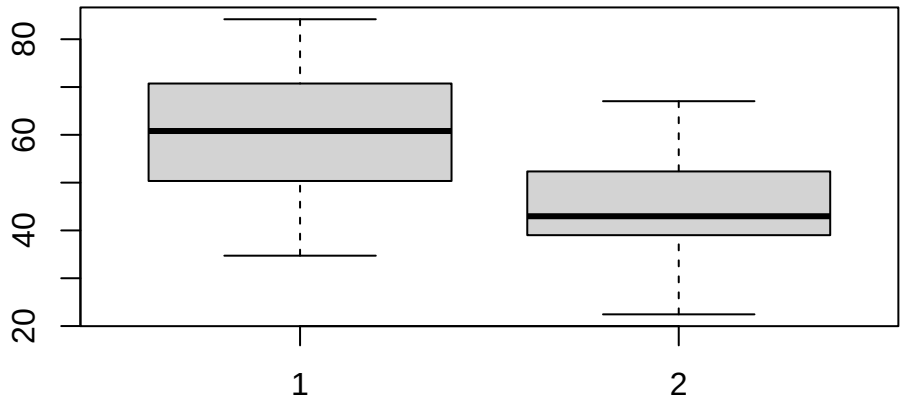
  Just by looking, we can see that the center of the distribution is below 50%.

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

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

```
    Welch Two Sample t-test

data:  chocolate and fruity
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
```

```
boxplot(chocolate, fruity)
```



By looking at the means, the mean of the chocolate candies are much higher that the means of the fruity candies (60.92% compared to 44.11%).

- **Q12**. Is this difference statistically significant?

  The difference is statistically significant since the p-value for the Welch Two Sample t-test is 2.871e-08 which is lower than 0.05.

Overall Candy Rankings

```
head(candy[order(candy$winpercent),], n=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
```

```
tail(candy[order(candy$winpercent),], n=5)
```

```
                          chocolate fruity caramel peanutyalmondy nougat
Snickers                          1      0       1              1      1
Kit Kat                           1      0       0              0      0
Twix                              1      0       1              0      0
Reese's Miniatures                1      0       0              1      0
Reese's Peanut Butter cup         1      0       0              1      0
                          crispedricewafer hard bar pluribus sugarpercent
Snickers                                 0    0   1        0        0.546
Kit Kat                                  1    0   1        0        0.313
Twix                                     1    0   1        0        0.546
Reese's Miniatures                       0    0   0        0        0.034
Reese's Peanut Butter cup                0    0   0        0        0.720
                          pricepercent winpercent
Snickers                         0.651   76.67378
Kit Kat                          0.511   76.76860
Twix                             0.906   81.64291
Reese's Miniatures               0.279   81.86626
Reese's Peanut Butter cup        0.651   84.18029
```

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

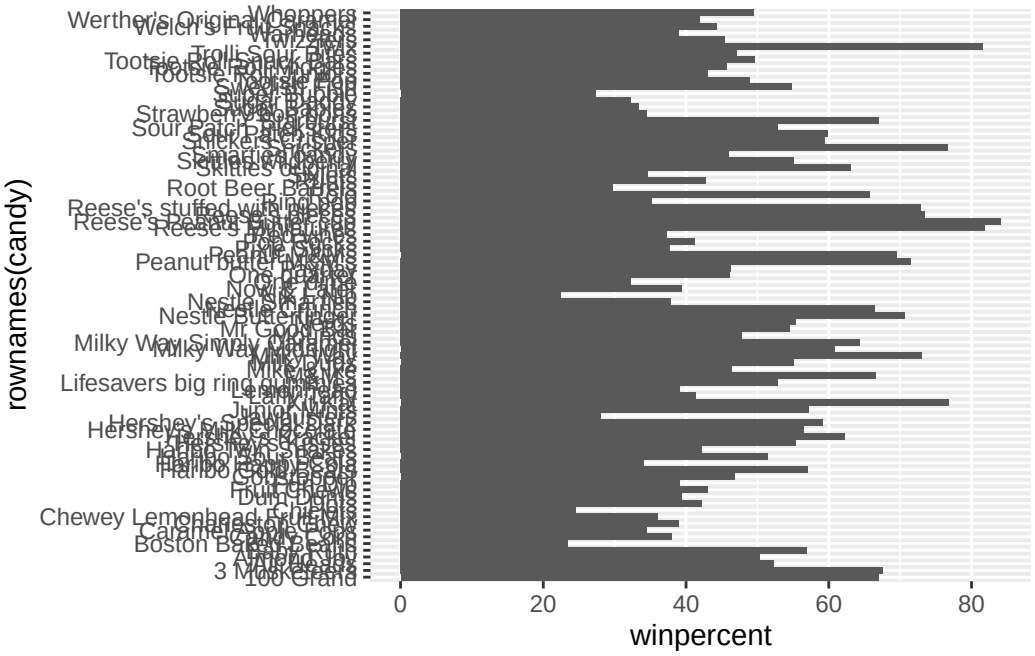
  The least favorite candies are Nik N Lip, Boston Baked, Chiclets, Super Bubble, and Jawbusters.

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

  The top favorite candy types are Snickers, Kit Kat, Twix, Reese's Miniature Cups, Reese's Peanut Butter Cups.

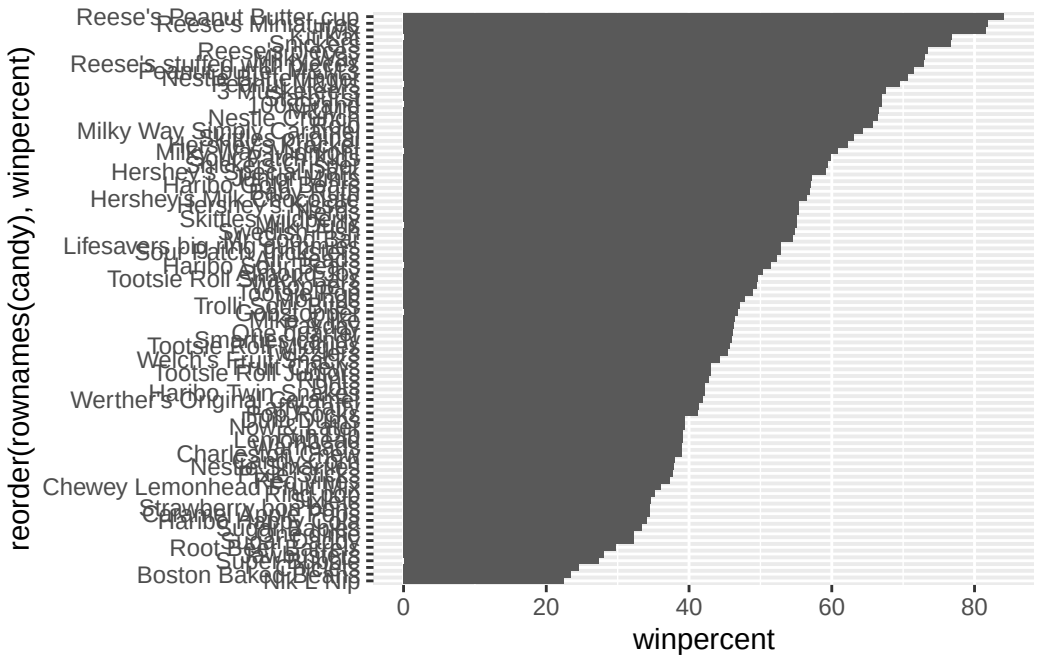
- **Q15**. Make a first barplot of candy ranking based on `winpercent` values.

```
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`?

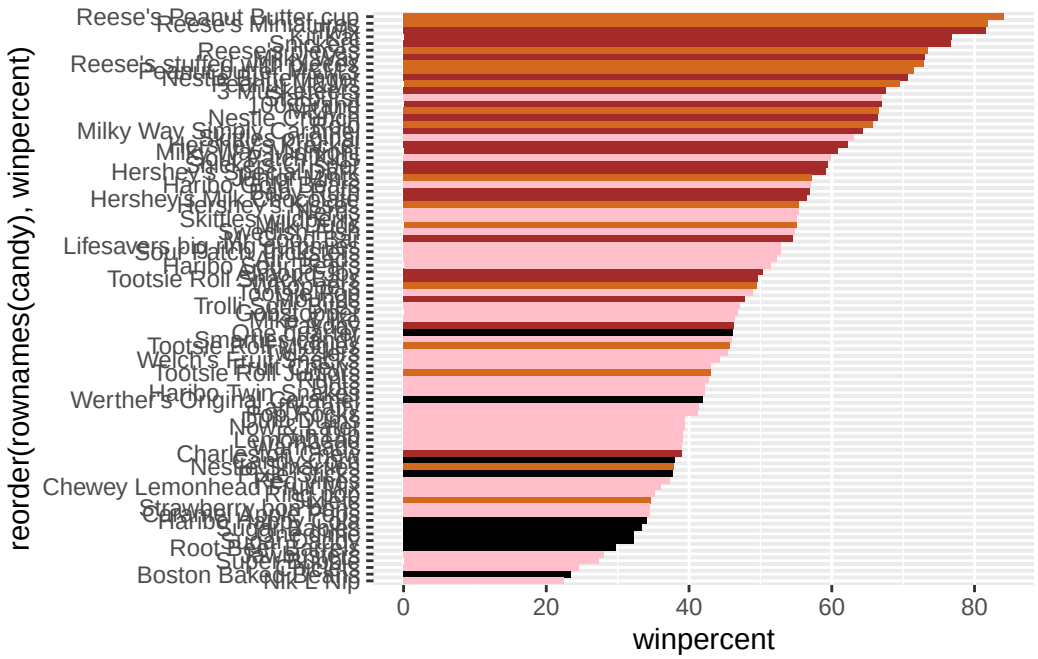
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col()
```



Adding Colors

```
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(fill=my_cols)
```



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

  The worst ranked chocolate candy was Sixlets

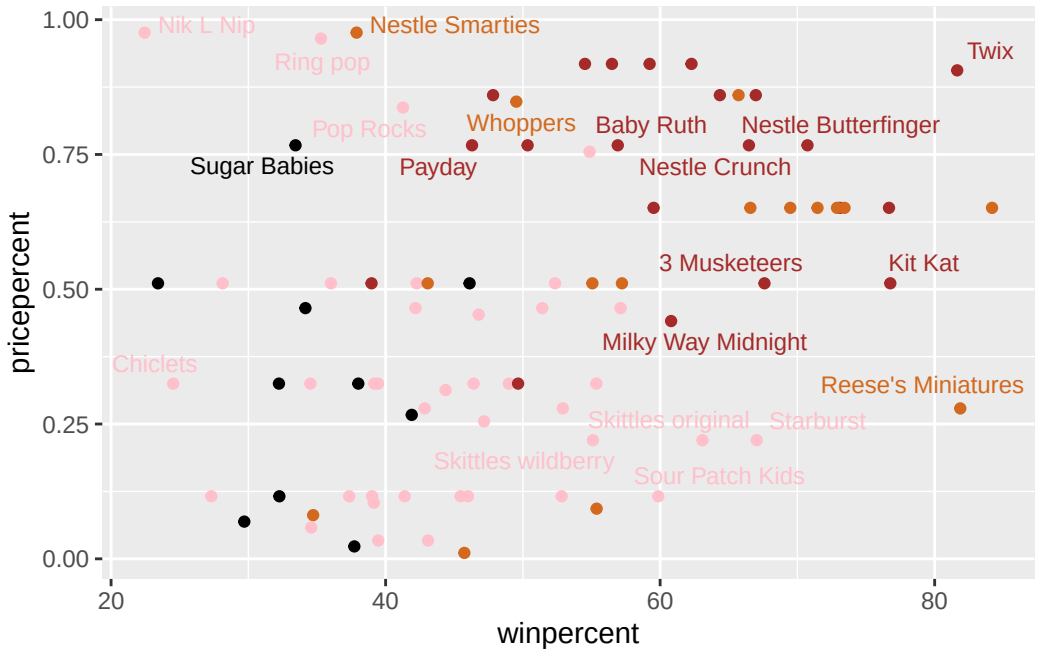
- **Q18**. What is the best ranked fruity candy?

  The worst ranked fruity candy was Nik L Nip

Taking a look at `Pricepercent`

```
#install.packages("ggrepel")
library(ggrepel)

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)
```

```
Warning: ggrepel: 65 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



- **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?

```
ord <- order(candy$pricepercent, decreasing = TRUE)
tail( candy[ord,c(11,12)], n=5 )
```

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

Considering that good candies have a `winpercent` higher than 80% and candies that are cheap have a `pricepercent` lower than .3, the candy with the most bang for your buck is Reece's Miniatures. Overall, fruity candies offer more bang for your buck though (higher win values, lower price values.

- **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(11,12)], n=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
```

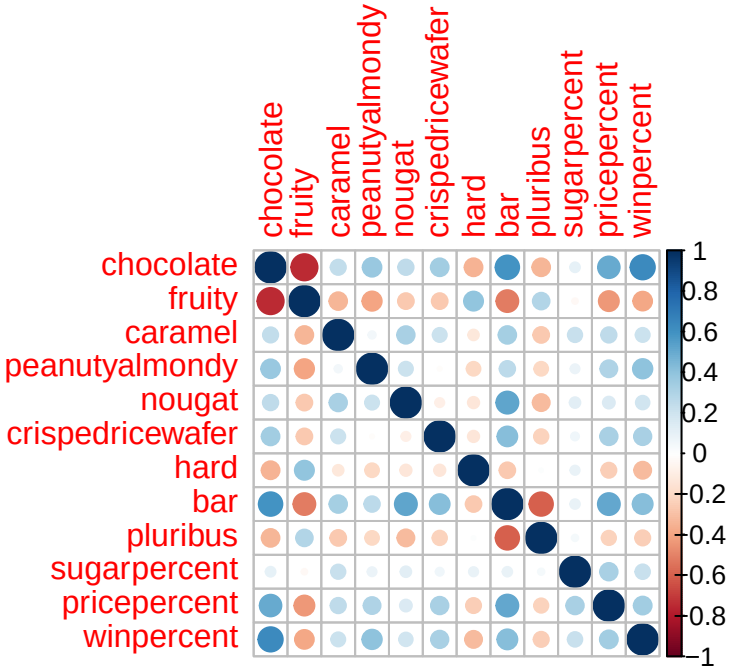
Considering that bad candies have a `winpercent` lower than 40% and candies that are expensive have a `pricepercent` higher than .75, the candies that are least popular and most expensive are: Nik L Nip, Ring pop, Nestle Smarties, Sugar Babies, and Pop Rocks.

Exploring the Correlation Structure

```
#install.packages("corrplot")
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



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

  The two variables that are anti-correlated are fruity and chocolate.

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

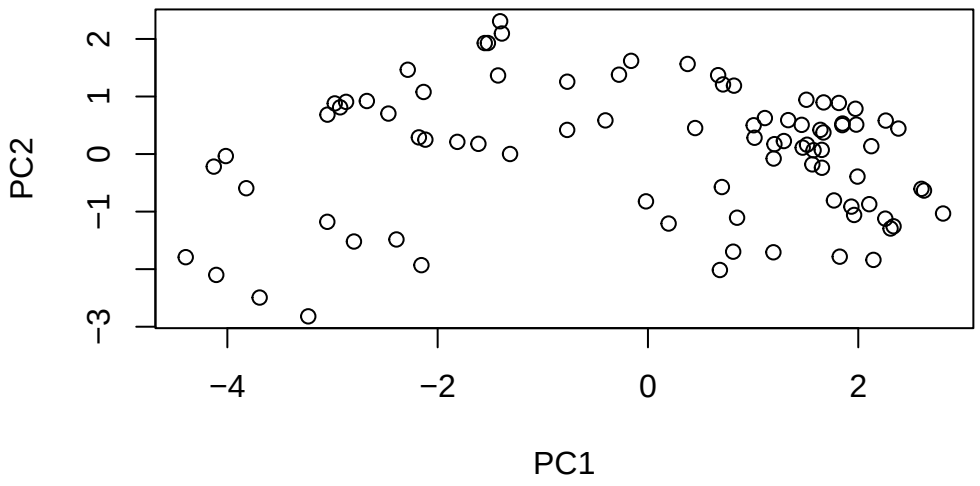
  The two variables that are correlated are chocolate and bar.

Principal Component Analysis

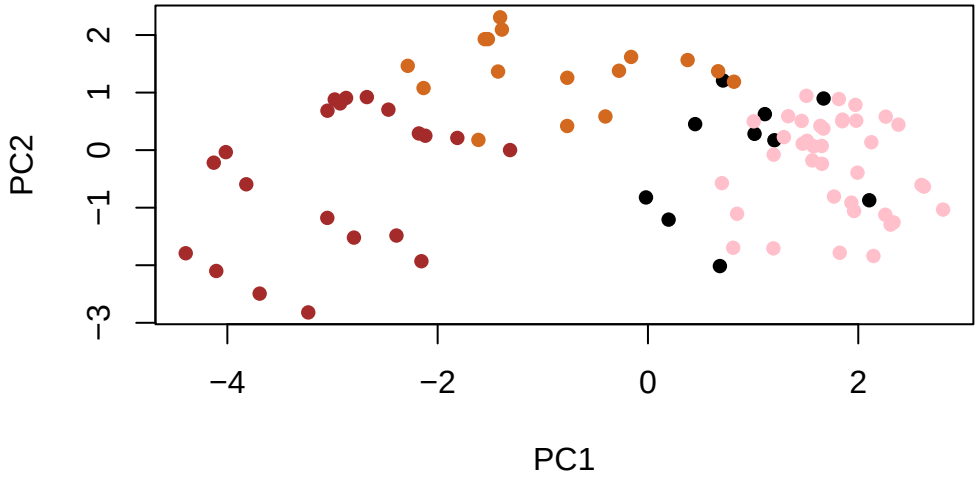
```
pca <- prcomp(candy, scale=T)
summary(pca)
```

```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 0.7424 0.79830 0.85369
                           PC8     PC9    PC10    PC11    PC12
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

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

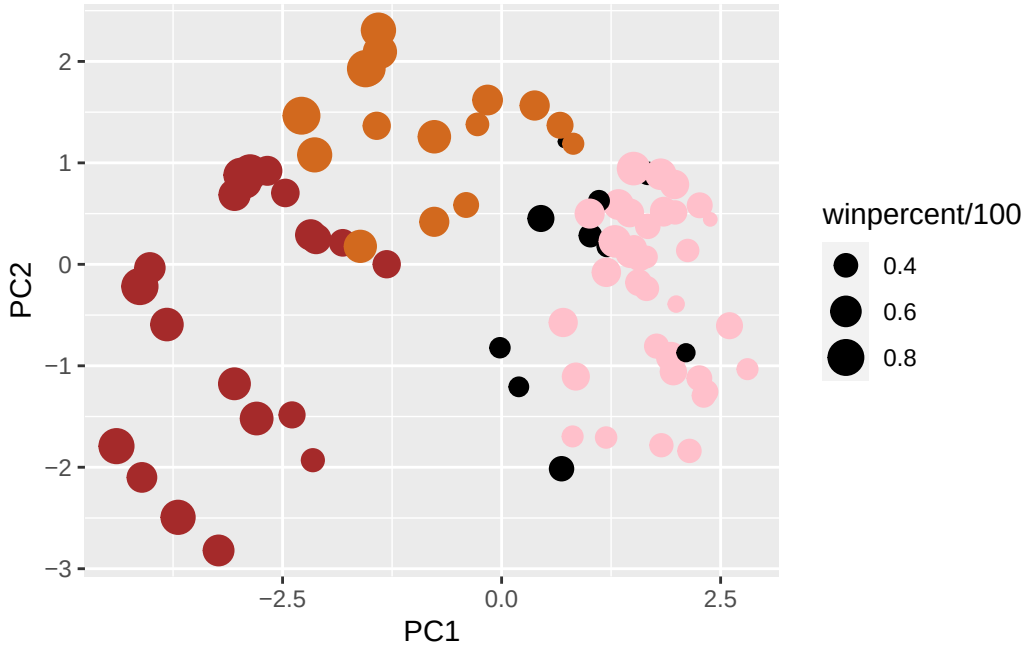



```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



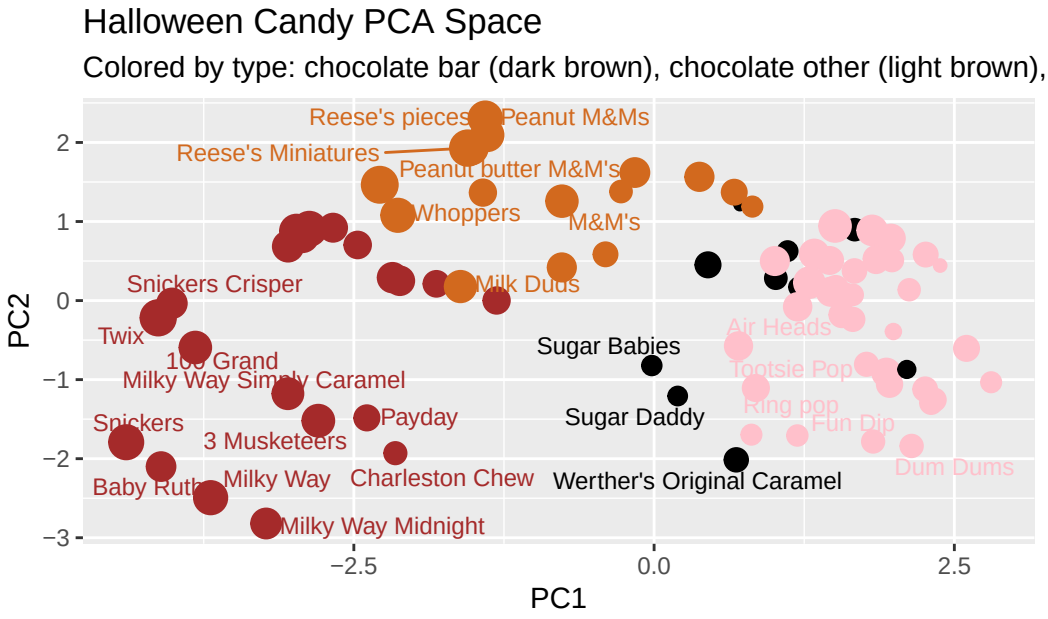
```
my_data <- cbind(candy, pca$x[,1:3])
```

```
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
```




```
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: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

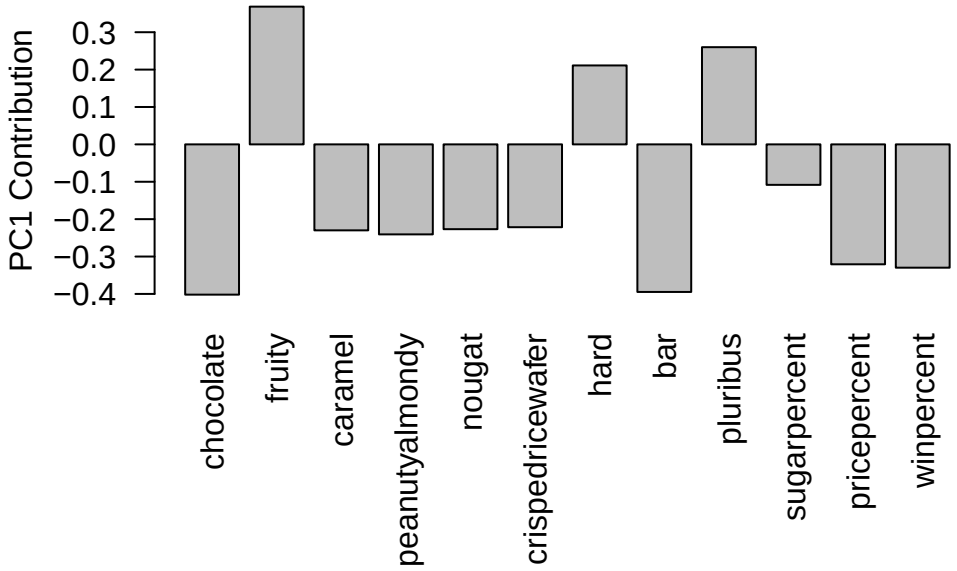



```
#install.packages("plotly")
library(plotly)
```

```
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)
```

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



- Q**24**. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

  The variables that are picked up strongly by PC1 in the positive direction are fruity, hard, and pluribus meaning that these values are more correlated together than the other variables. Fruity candies tend to be hard and come in packages of many.