Class 10: Halloween Mini-Project

Nicole Alfonso, (PID: A16429176)

Exploratory Analysis of Halloween Candy

1. 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 data set?

There are 85 different brands, and 12 different candy types in the data set.

```
dim(candy)
```

```
[1] 85 12
```

Q2. How many fruit candy types are in the data set?

There are 38 fruit candies in the data set.

```
fruit_candy <- table(candy$fruity)
fruit_candy
```

```
 0  1
47 38
```

2. What is your favorite candy?

Q3. What is your favorite candy in the data set and what is it's winpercent value?

My favorite candy is the **Reese's Peanut Butter Cup**, and its win percent value is 84.18%

```
reeses_winpercent <- round(candy["Reese's Peanut Butter cup", ]$winpercent, 2)
reeses_winpercent
```

```
[1] 84.18
```

Q4. What is the winpercent value for KitKat?

76.77%

```
kitkat_winpercent <- round(candy["Kit Kat", ]$winpercent, 2)
kitkat_winpercent
```

```
[1] 76.77
```

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

49.65%

```
tootsieroll_winpercent <- round(candy["Tootsie Roll Snack Bars", ]$winpercent, 2)
tootsieroll_winpercent
```

```
[1] 49.65
```

```
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?

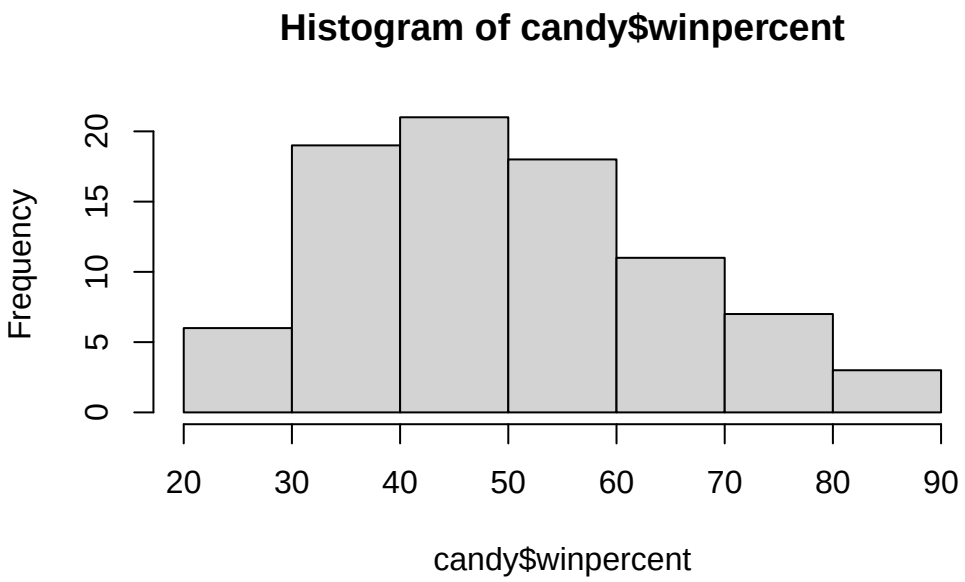
The winpercent variables looks to be on a different scale, as most of the other variable's values are between 0 and 1, whereas the winpercent values range anywhere from 0-100, as you look as the mean, standard deviation, p0, etc.

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

The zero in the n_missing() column signifies the sum of any missing values, and the 1 in the complete_rate() column signifies that all values are present – none are missing.

Q8. Plot a histogram of `winpercent` values

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



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

Simply looking at the histogram, you can observe that the winpercent values are not symmetrical, as the data is right-skewed.

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

The mean of the winpercent values is 50.32, so the center of the distribution is above 50.

```
round(mean(candy$winpercent), 2)
```

```
[1] 50.32
```

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

On average, chocolate candy is ranker higher than fruit candy, as it has an average winpercent of 60.92, with fruit candy hacing an average winpercent of 44.12.

```
chocolate_mean <- round(mean(candy$winpercent[as.logical(candy$chocolate)]), 2)
chocolate_mean
```

```
[1] 60.92
```

```
fruit_mean <- round(mean(candy$winpercent[as.logical(candy$fruity)]), 2)
fruit_mean
```

```
[1] 44.12
```

Q12. Is this difference statistically significant?

Because the p-value is less than 0.05, the difference between chocolate and fruity candy is statistically significant.

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

```
    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
```

3. Overall Candy Rankings

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

Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters

```
least_liked_candy <- head(candy[order(candy$winpercent), ], n = 5)
least_liked_candy
```

```
                  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
```

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

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

```
most_liked_candy <- head(candy[order(-candy$winpercent), ], n=5)
most_liked_candy
```

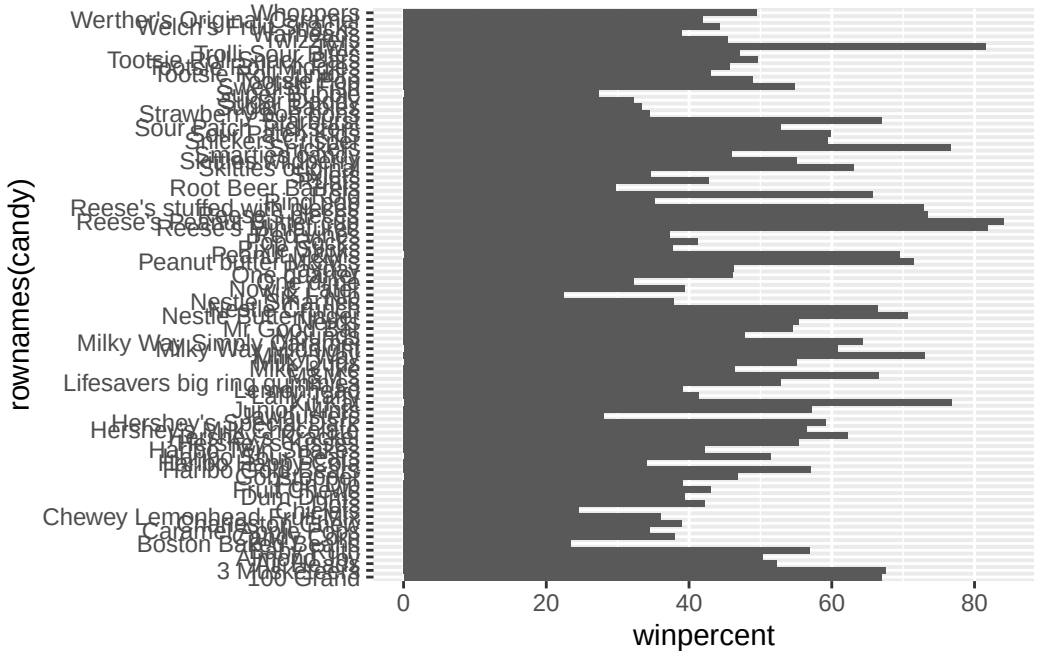
```
                        chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           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
```

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

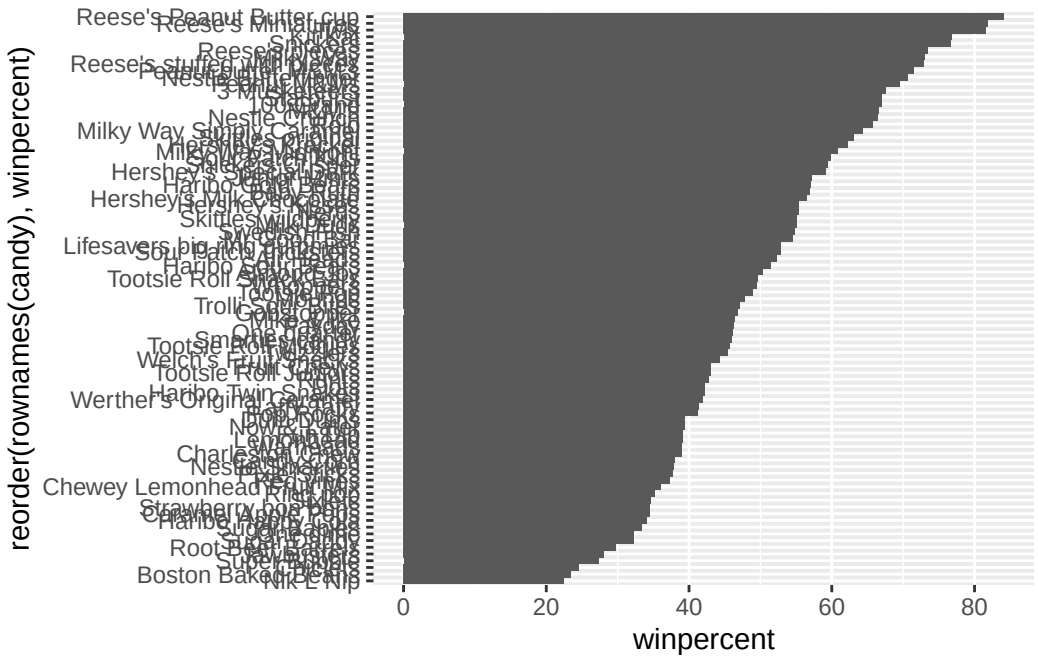
```
library(ggplot2)

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



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

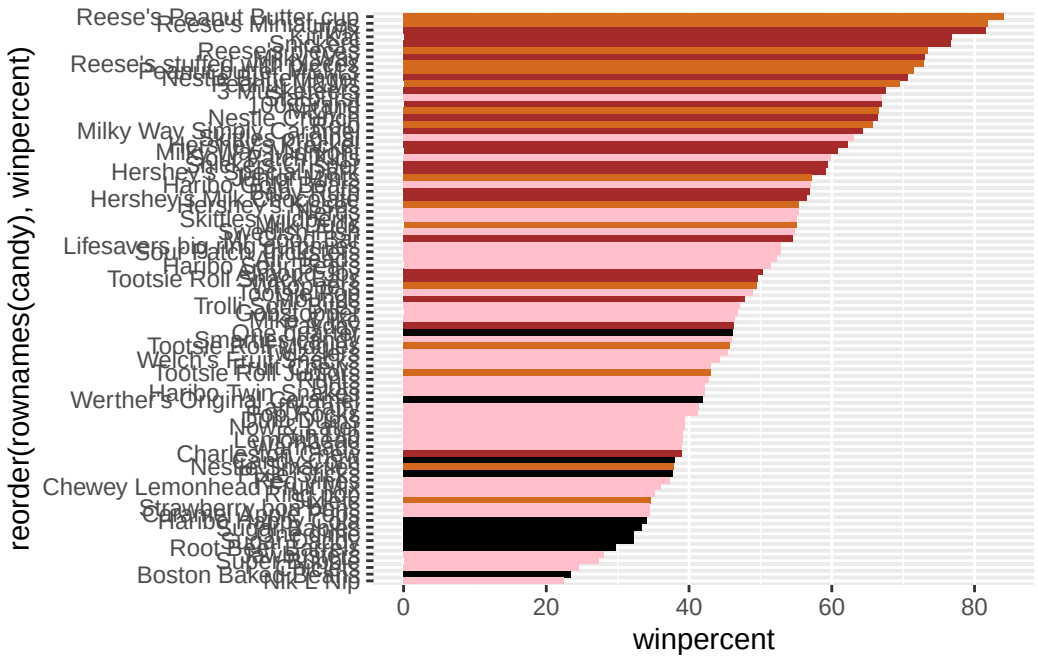
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col()
```



Time to add some useful color

```
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?

Sixlets

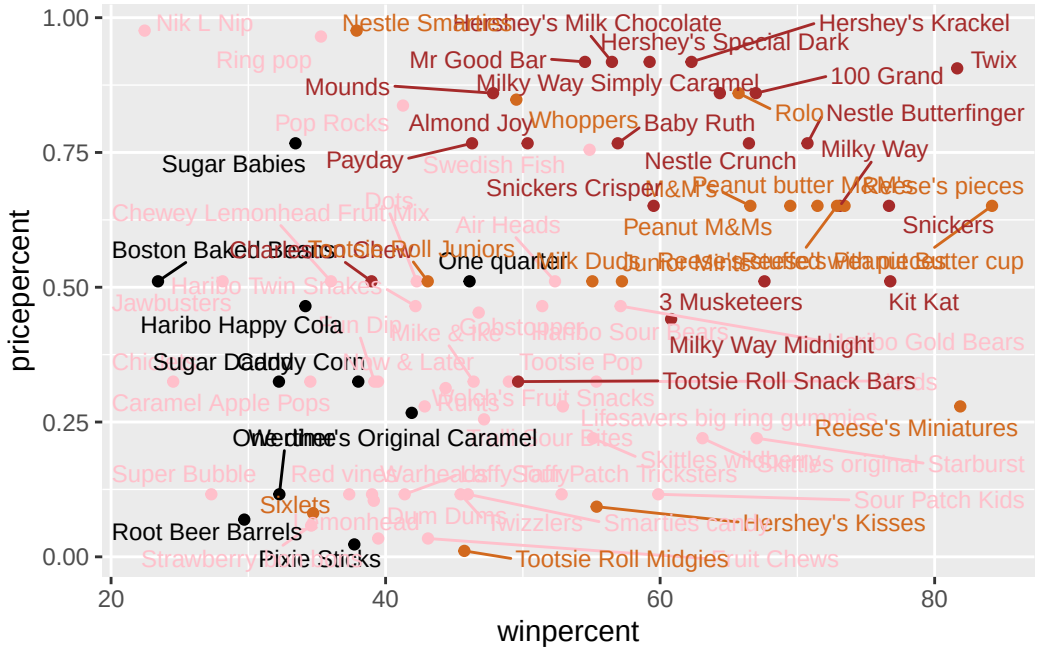
Q18. What is the best ranked fruity candy?

Starburst

4. Taking a look at pricepercent

```
library(ggrepel)

# Plotting 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 = 50)
```



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?

5 most expensive: Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate

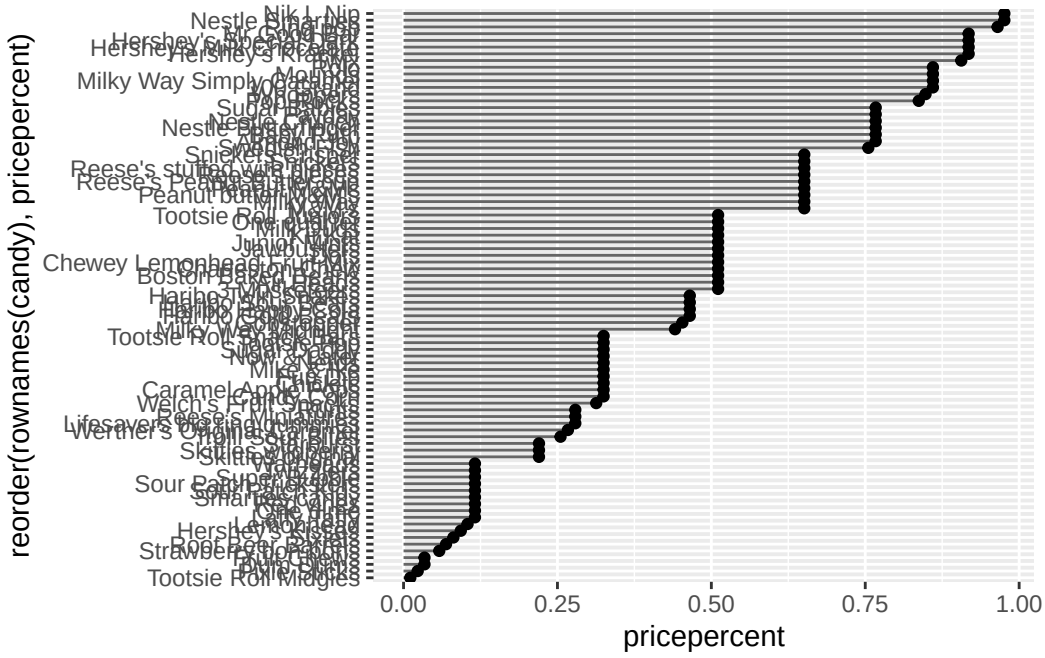
Least popular: Ring pop

```
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
```

Q21. Make a barplot again with `geom_col()` this time using `pricepercent` and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping `geom_col()` for `geom_point()` + `geom_segment()`.

```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```

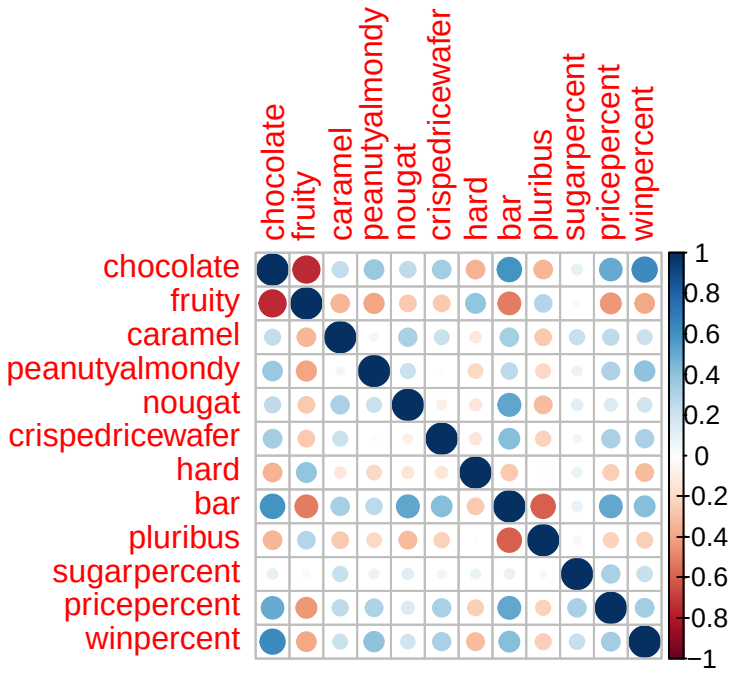


5. Exploring the Correlation Structure

```
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)?

chocolate X fruity (most anti-correlated), pluribus x bar, fruity X bar

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

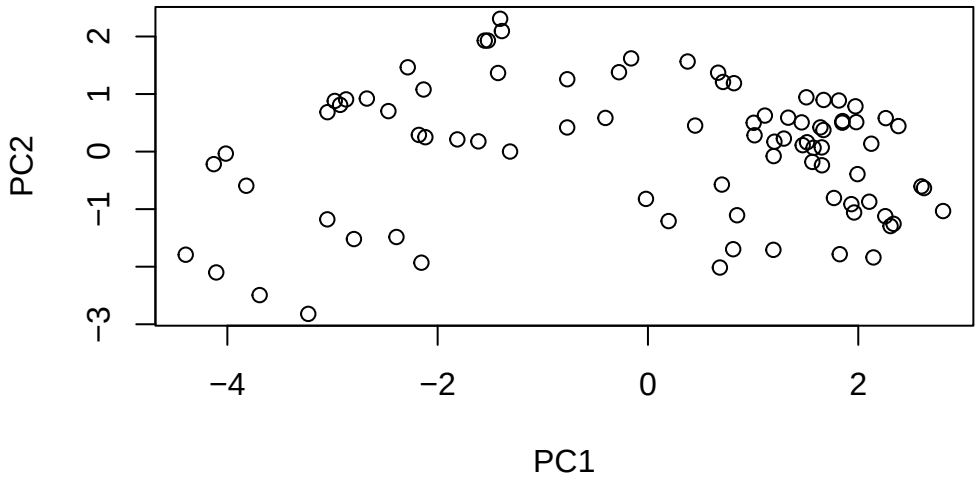
chocolate x bar

6. Principal Component Analysis

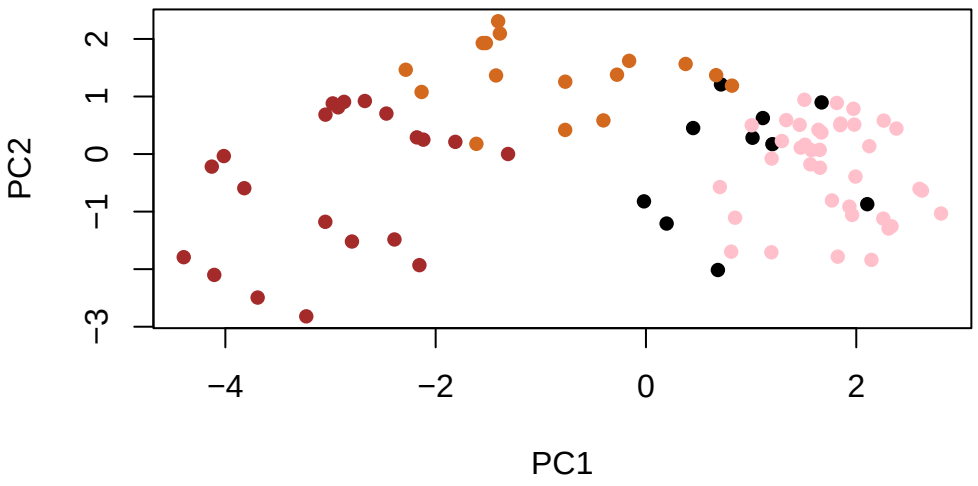
```
pca <- prcomp(candy, scale. = TRUE)
summary(pca)
```

```
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
```

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



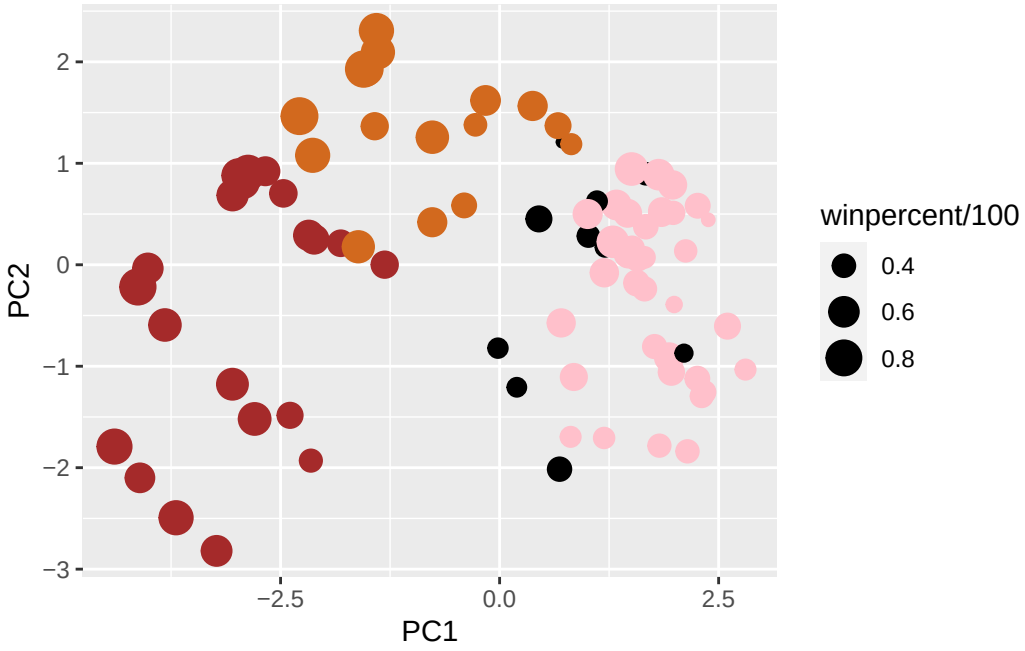

```r
plot(pca$x[,1:2], col=my_cols, pch=16)
```




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

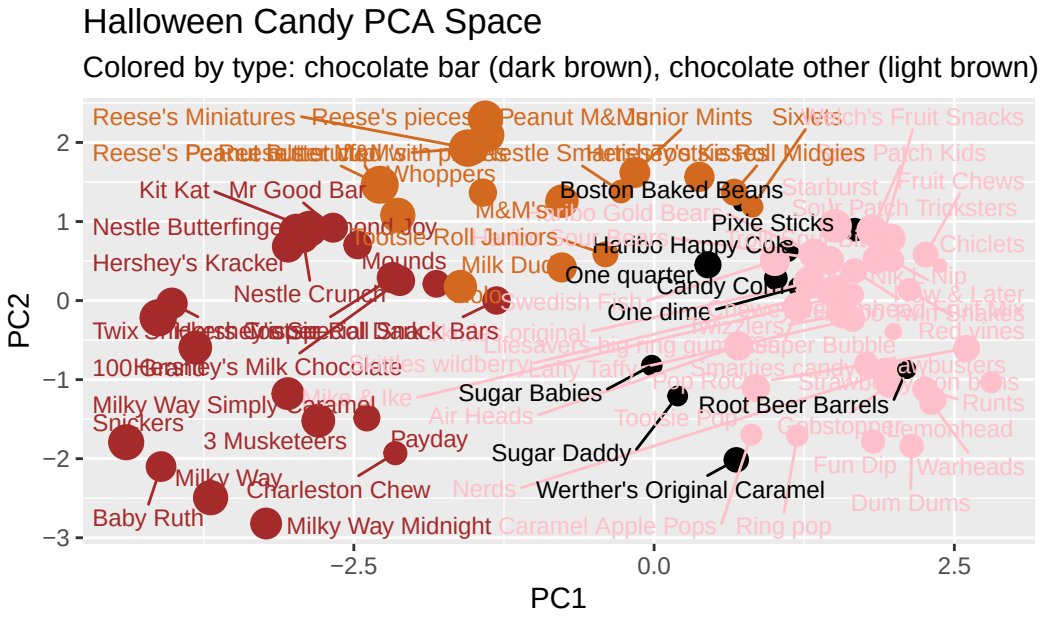
```
p <- ggplot(candy_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(candy_data),
            label=rownames(candy_data)) +
        geom_point(col=my_cols)

p
```



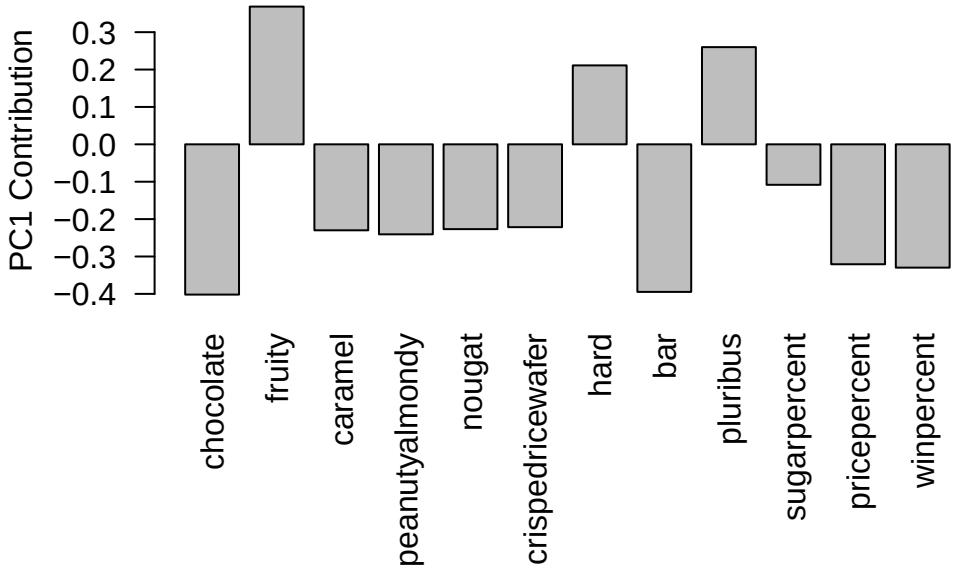

```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 50)  +
  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")
```



Data from 538

```
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, and pluribus. This makes sense as the most popular fruit candy such as Starbursts and Skittles, are hard candies which come packaged amongst larger quantities. On the contrast, chocolates with different fillings such as caramel, peanuts, and nougat, are more likely to be sold as individual bars. Thus, it makes sense that these variables are clustered together as they are most highly correlated.