Class 9: Candy Mini-Project

Linh Tran (PID:A16435846)

For this mini-project, we will explore FiveThirtyEight's Halloween Candy dataset (538) to determine the top and worst ranked candies based on different factors such as price and the types.

Importing Data from FiveThirtyEight GitHub repo

First, we have to import candy data by downloading `candy-data.csv` and using `read.csv()`.

```
candy_file <- "candy-data.txt"
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 candy types

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

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

```
[1] 38
```

38 fruity candy types

What is your favorite candy?

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

We can look at `winpercent` to find the percentage of people who prefer a particular candy over another randomly chosen candy.

My favorite candy is Kit Kat and it's winpercent value is 47.83

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

```
[1] 49.52411
```

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

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

A useful function from "skimr" that can help giving a quick overview of a dataset is `skim()` first, we have to download `skimr` with `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 variables that look different from the majority are "s**ugarpercent", "pricepercent",** and **"winpercent".**

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

I believe a zero shows that the corresponding candy type based on the row is not chocolate whereas a one would show that it is chocolate.

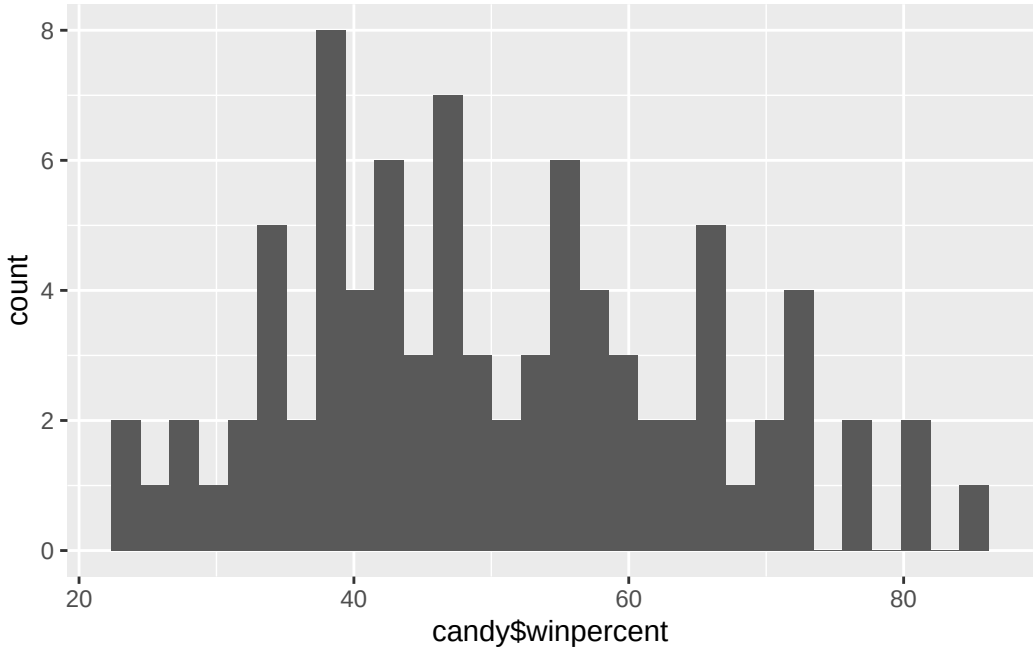
Q8. Plot a histogram of winpercent values

We can plot a histogram using either `hist()` or `ggplot()`

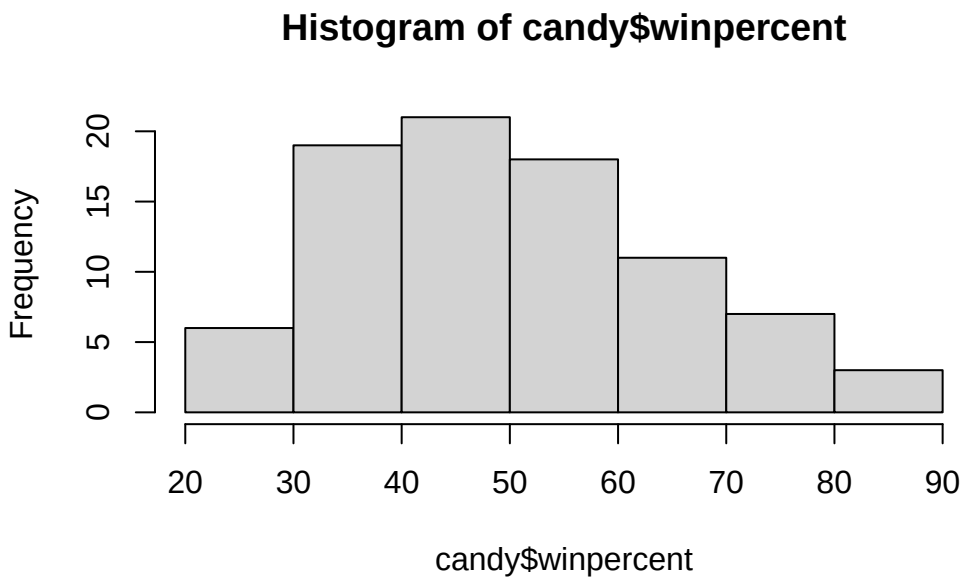
```
library("ggplot2")

ggplot(candy)+
  aes( x=candy$winpercent)+
  geom_histogram()
```

```
Warning: Use of `candy$winpercent` is discouraged.
i Use `winpercent` instead.
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




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



Q9. Is the distribution of winpercent values symmetrical?

The distribution of winpercent values is assymetrical, it is right skewed.

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

The center of the distribution is below 50%

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

Using `as.logical()`, we can turn 1's and 0's into logical and assign them to their corresponding category. Furthermore, we can take the average with `mean()`

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

From the results, we see that chocolate has a higher winpercent value on average than fruit candy; therefore, chocolate ranks higher

Q12. Is this difference statistically significant?

The t-test can be performed with `t.test()`

```
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
```

The t-test shows that there is a significant difference in the menas of the two candies, that is chocolate is likely ranked higher.

Overall Candy Rankings

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

we can use either bse R with `head()` or `dplyr` to arrange the candy types.

```
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
```

```
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
```

```
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
```

Both methods show that the least liked 5 are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters

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

```
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
```

```
candy %>% arrange(winpercent) %>% tail(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
```

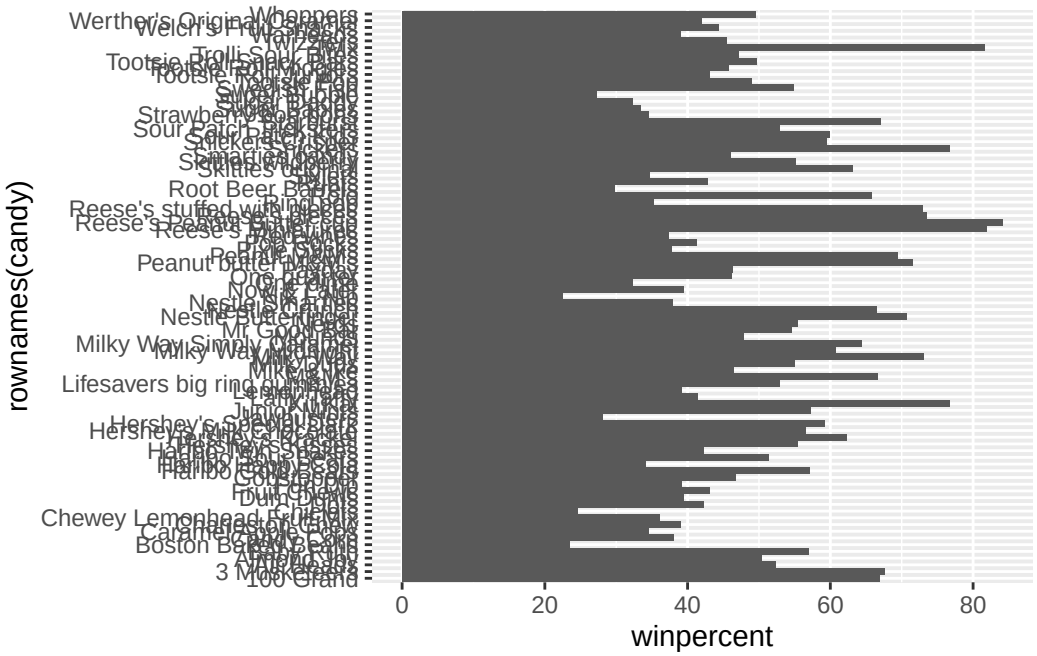
```
Snickers                            0.651    76.67378
Kit Kat                             0.511    76.76860
Twix                                0.906    81.64291
Reese's Miniatures                  0.279    81.86626
Reese's Peanut Butter cup           0.651    84.18029
```

Both methods showed that the top 5 are Snickers, Kit Kat, Twix, Reese's Miniatures, and Reese's Peanut Butter cup.

I like dplyr more, despite having to first retrieve it with `library()`, because the code is more straight forward.

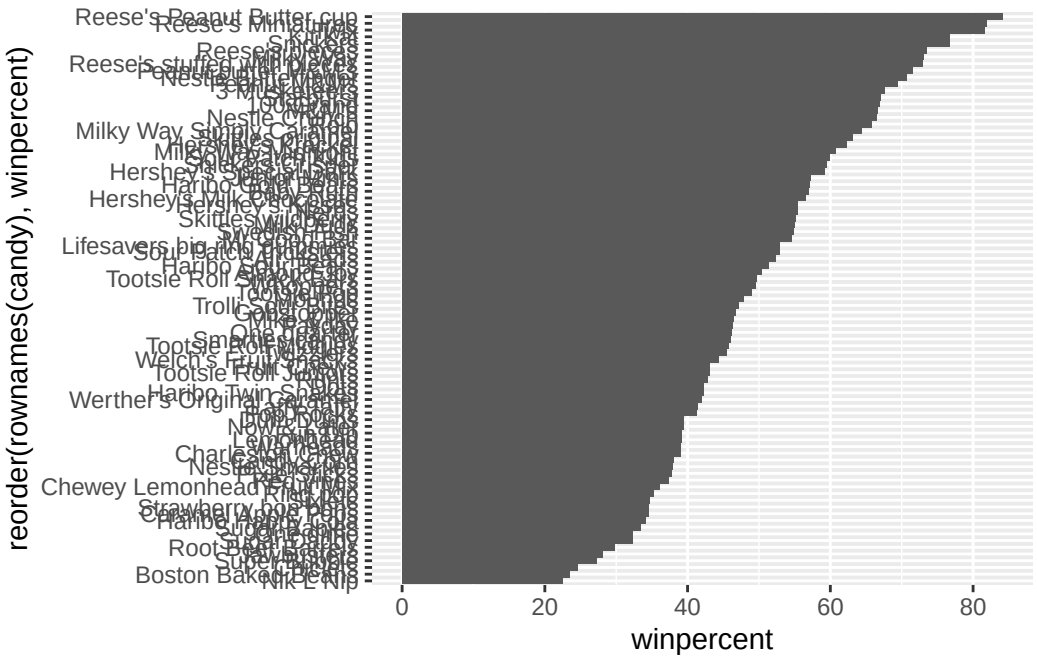
Q15. Make a first barplot of candy ranking based on winpercent values.

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



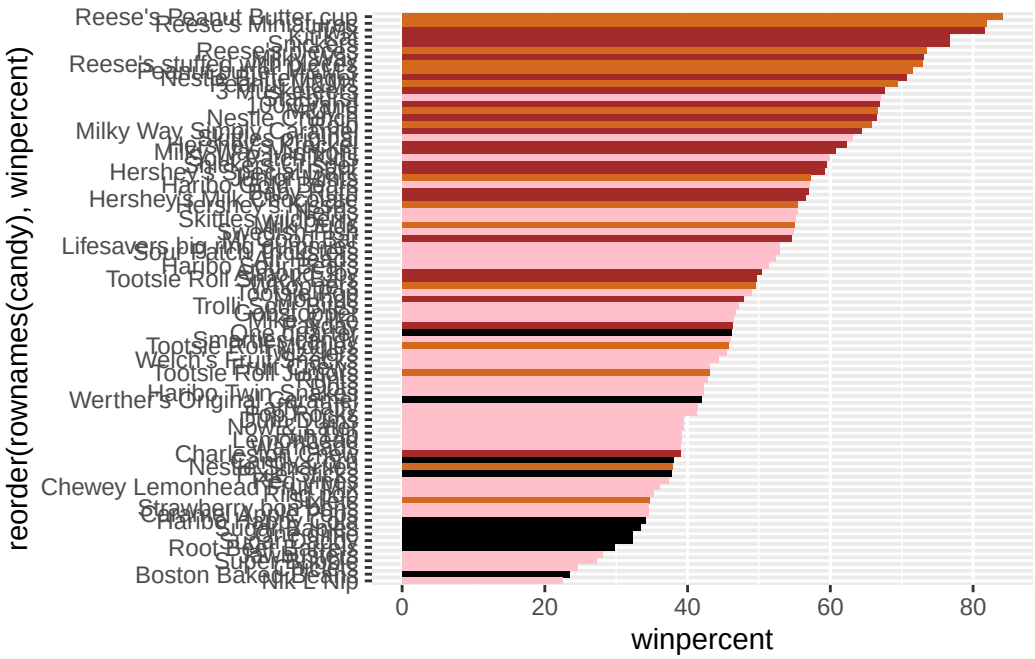
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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



We can also add 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?

The worst ranked chocolate candy is Sixlets.

Q18. What is the best ranked fruity candy?

The best ranked fruity candy is Starburst.

Taking a look at "*pricepercent*"

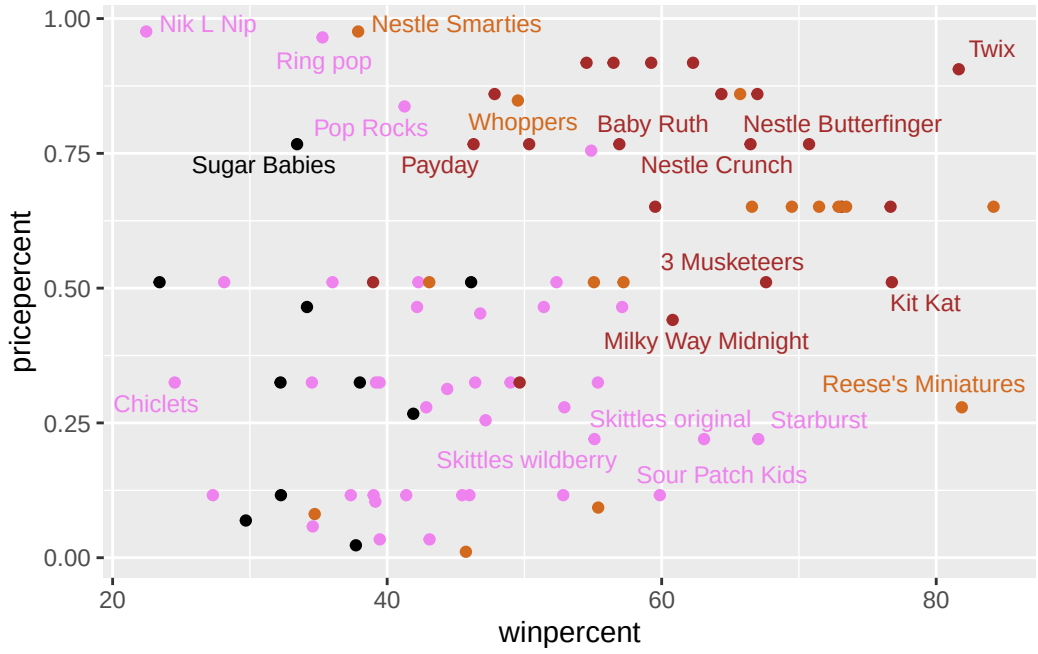
Q19. Which candy type is the highest ranked in terms of winpercent for the least money - i.e. offers the most bang for your buck?

To determine the highest ranked candy in terms of **winpercent** for the least money, we would have to plot "**winpercent**" vs"**pricepercent**" using `geom_label()` and `geom_text_repel()` from **ggrepel**

```
library(ggrepel)
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)] = "violet"
```

```
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
```



Based on the plot, I believe the highest ranked for the least money is Reeses Miniatures, which has a winpercent value of more than 80 and pricepercent value of slightly above 0.25.

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

We can use `order()` to arrange the candies based on **pricepercent**

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

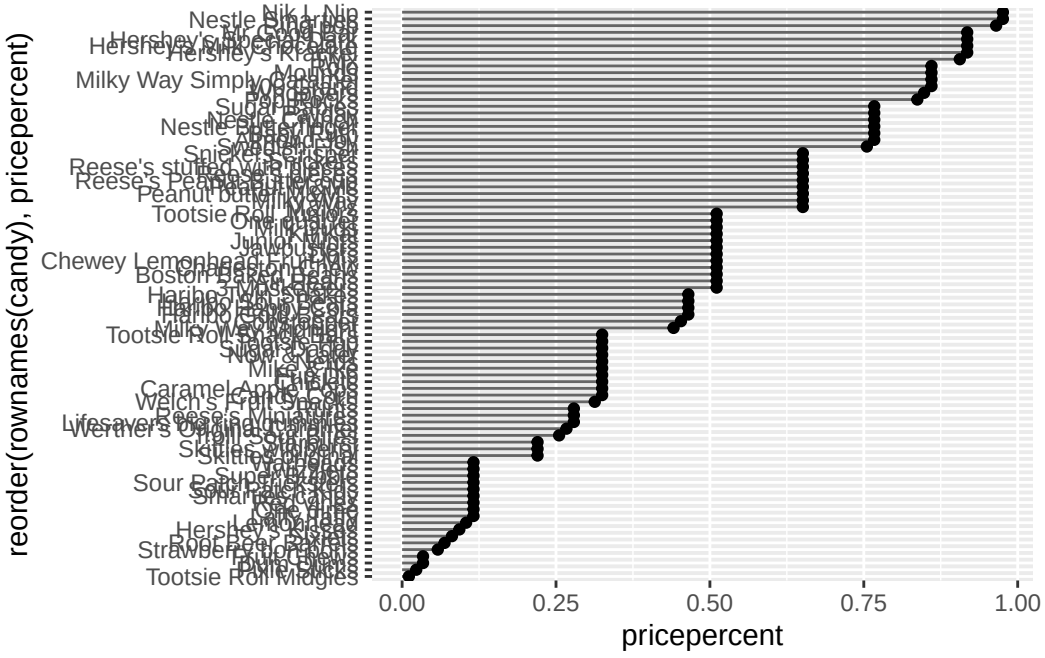
```
                pricepercent winpercent
Nik L Nip              0.976   22.44534
```

```
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
```

The most expensive candies are in the order of Nik L Nip, Nestle Smarties, Ring pop, Hersheys Krackel, Hersheys Milk Chocolate. Based on the result from **Q13.**, Nik L Nip is also the least popular

> 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()
```



Exploring the correlation structure

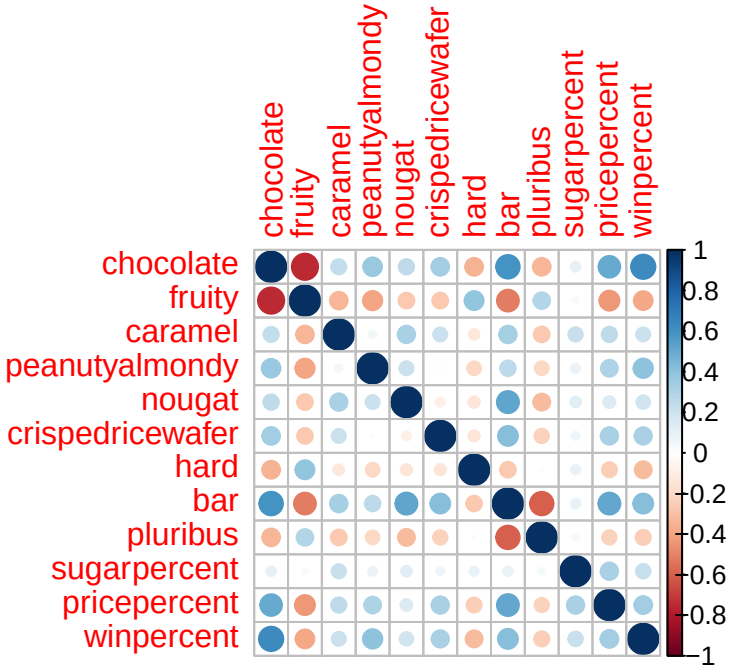
We can view how varianles interact by looking at correlation plots, which can be made with `corrplot()`

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

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



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

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

The two variables that are most positively correlated are bar candy and chocolate

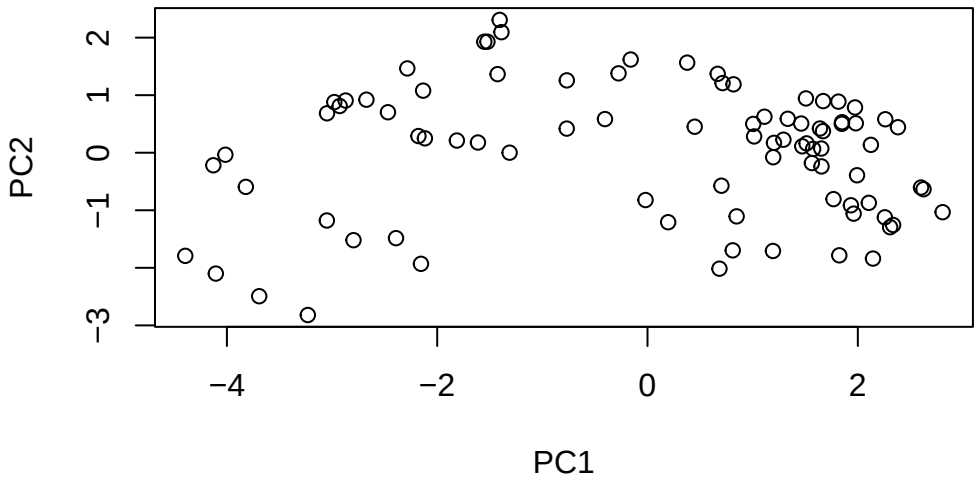
Principal Component Analysis

Lastly, we can apply PCA with `prcom()` and plot PC1 vs PC2

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

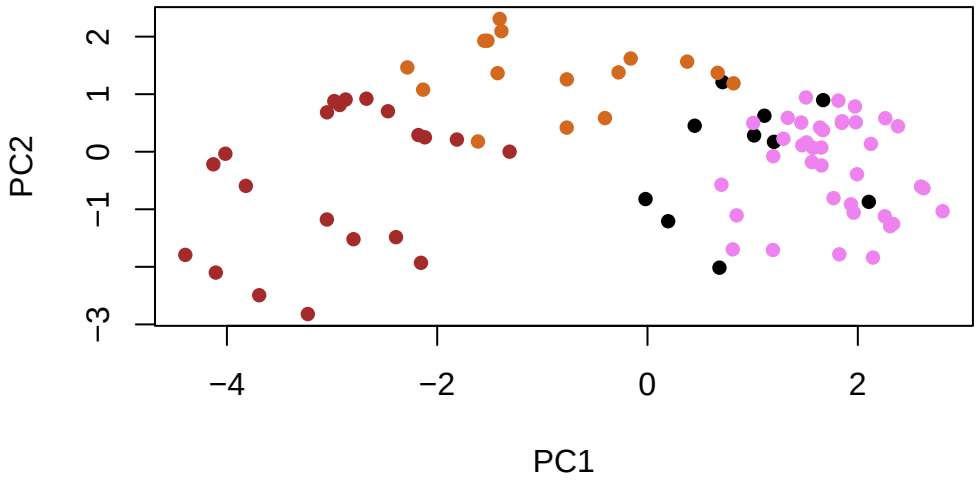
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 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
```

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



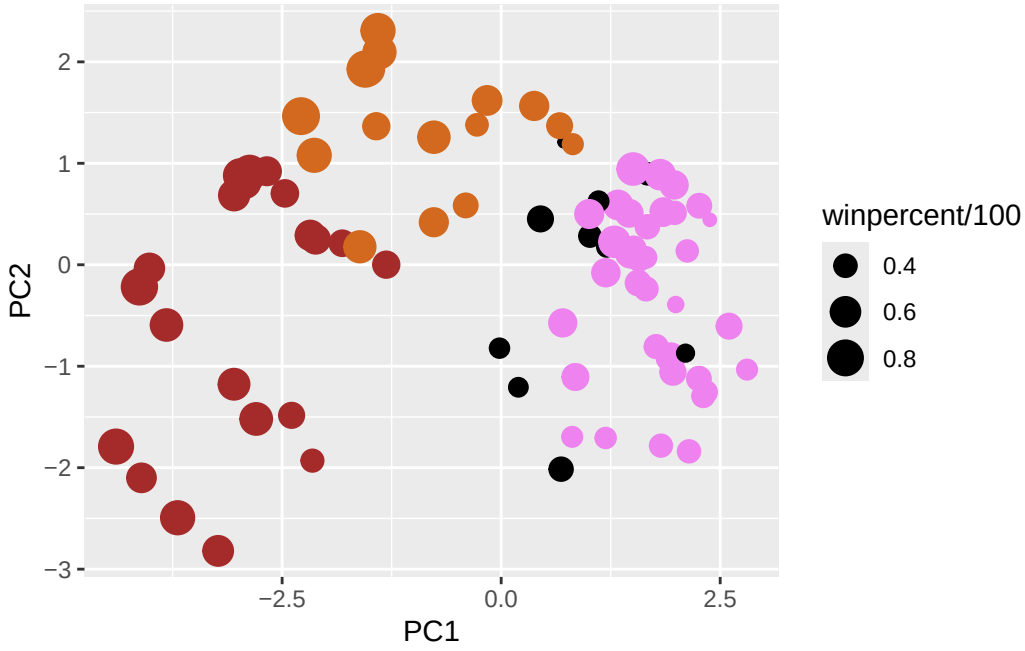
We can cahnge the plot character and add color

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



Using the PCA results and Candy data, we can make a new data-frame with `cbind()`

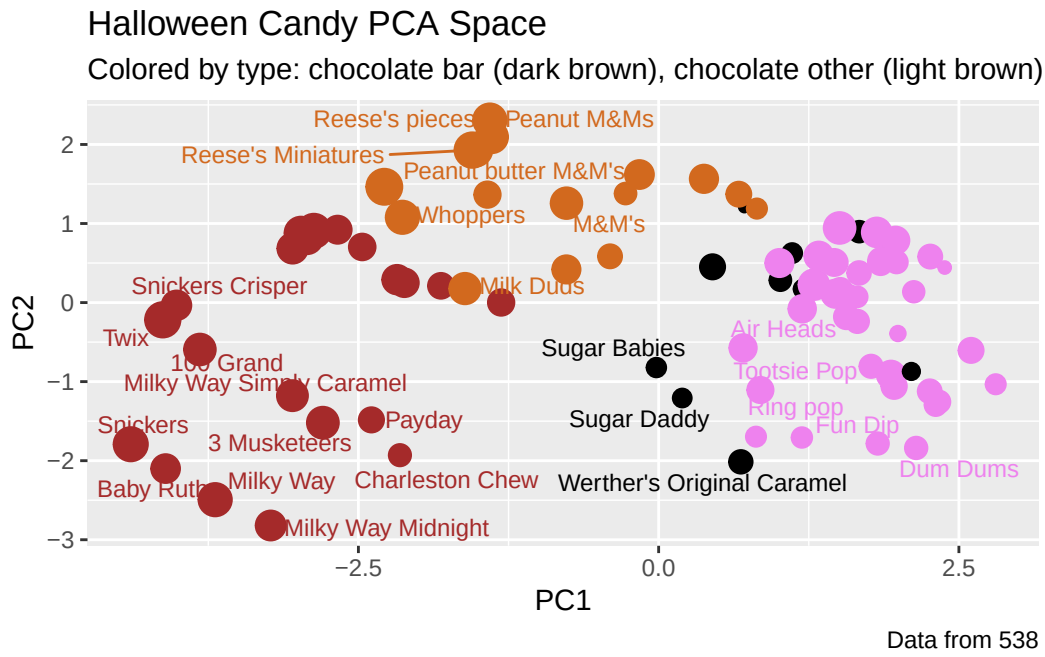
```
my_data <- cbind(candy, pca$x[,1:3])

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
```



We then add title and substitle, as well as further label the plot with **ggrepel**. An interactive plot can also be created with

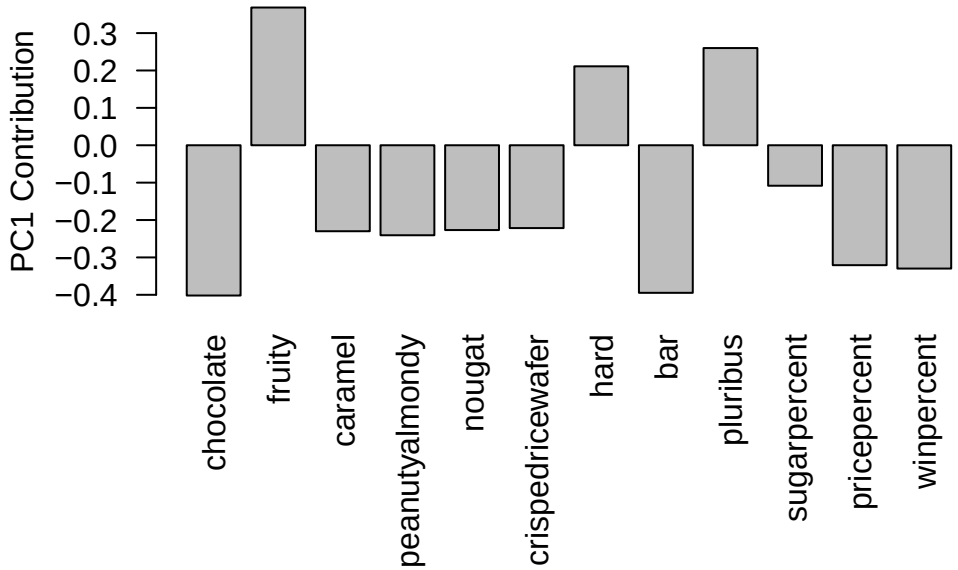
```
p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown),
       caption="Data from 538")
```

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



(I was unable to install ggplotly)

```
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?

THe variables picked up strongly by PC1 in the positive direction are fruit, hard, and pluribus. They make sense because it they seem to correlate and in real life I also associate them with one another even without looking at graphs.