Class 8: Halloween

Benjie Miao (PID: A69026849)

1. Import the dataset

```
candy_file <- "candy-data.csv"
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?

Since each row represents a different type of candy, we can just count the row number:

```
nrow(candy)
```

```
[1] 85
```

There are 85 types of candy.

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

Since each fruity candy has a 1 in the `fruity` columns, we can sum that columns.

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

```
[1] 38
```

There are 38 fruity candy types.

2. What's your favourite candy?

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

```
[1] 81.64291
```

Q3. What is your favorite candy in the dataset and what is its winpercent value?

I don't know any kind of the candy but let's assume that my favorite candy is 100 Grand. Let's find the winpercent value of chocolate:

```
candy["100 Grand", ]$winpercent
```

```
[1] 66.97173
```

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

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

```
[1] 76.7686
```

It's 76.77.

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

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

```
[1] 49.6535
```

It's 49.65.

We can use `skim` to get a quick view of the dataset:

```
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	

From the `skim` funciton, we can see:

Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

It is the winpercent. It is around 50 where other columns are fractions.

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

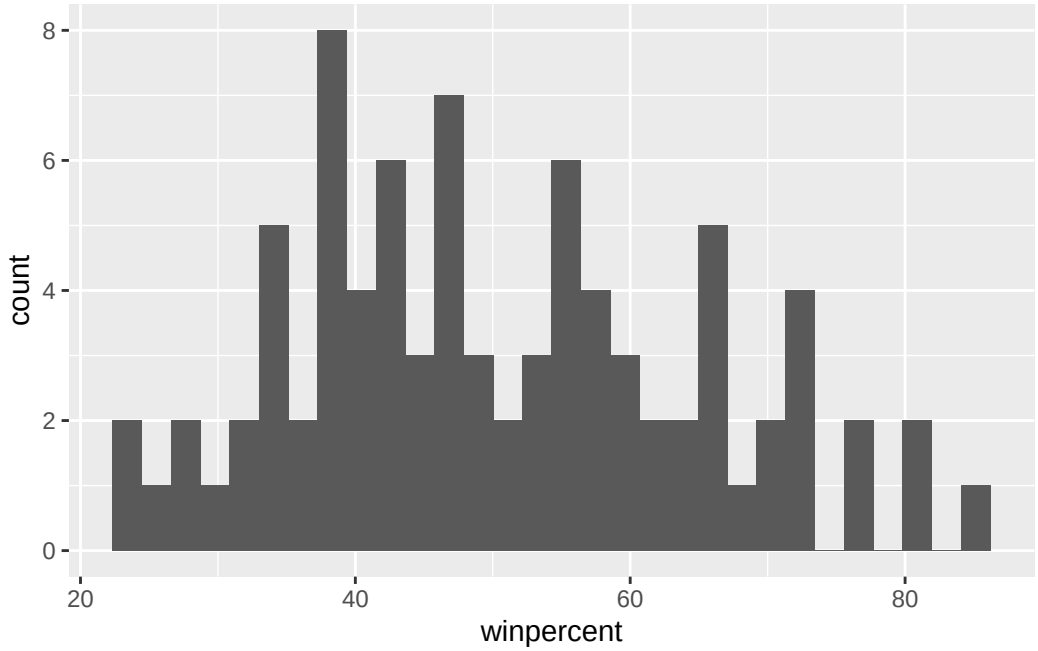
The column is exclusively zero or one, which means it is a binary variable which indicates whether or not this type of candy contains chocolate.

Then we can use `ggplot` to start exploratory analysis.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)

ggplot(candy, aes(winpercent)) +
  geom_histogram()
```

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



Q9. Is the distribution of winpercent values symmetrical?

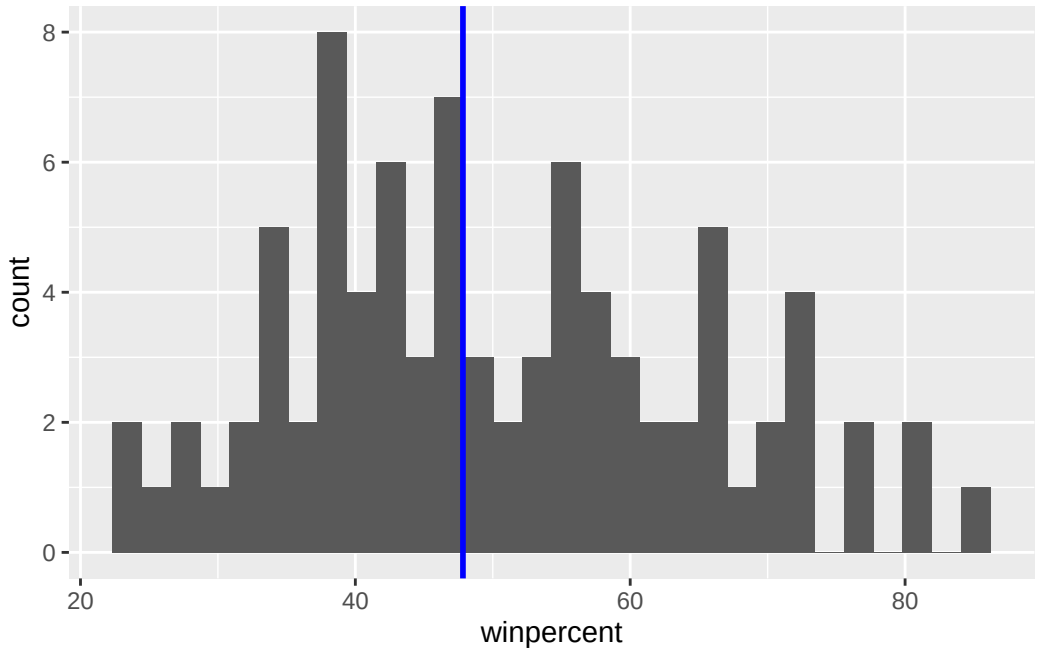
Apparently it is not symmetrical.

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

We can draw the median line to this distribution:

```
library(ggplot2)

ggplot(candy, aes(winpercent)) +
  geom_histogram() +
  geom_vline(aes(xintercept = median(winpercent)),col='blue',linewidth=1)
```

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



It is below 50.

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

To test this, we need first to extract the chocolate candy and fruity candy, and calculate their average rank. We can use `as.logical()` to transform the 0-1 binary variable to boolean variable.

```
chocolate_winpercent <- candy[as.logical(candy$chocolate), ]$winpercent
fruity_winpercent <- candy[as.logical(candy$fruity), ]$winpercent
sprintf("Cholocate average winpercent: %.2f", mean(chocolate_winpercent))
```

```
[1] "Cholocate average winpercent: 60.92"
```

```
sprintf("Fruit candy average winpercent: %.2f", mean(fruity_winpercent))
```

```
[1] "Fruit candy average winpercent: 44.12"
```

Therefore, chocolate candy has a higher average rank than fruit candy.

Q12. Is this difference statistically significant?

```
t.test(chocolate_winpercent, fruity_winpercent)
```

```

    Welch Two Sample t-test

data:  chocolate_winpercent and fruity_winpercent
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
```

Since `p = 2.9e-08`, we consider it is significant.

3. Overall Candy Rankings

Now we can use the `order()` function to sort the whole dataset.

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

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

```
[1] "Nik L Nip"           "Boston Baked Beans" "Chiclets"
[4] "Super Bubble"        "Jawbusters"
```

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?

```
rownames(head(candy[order(candy$winpercent, decreasing = TRUE), ], n = 5))
```

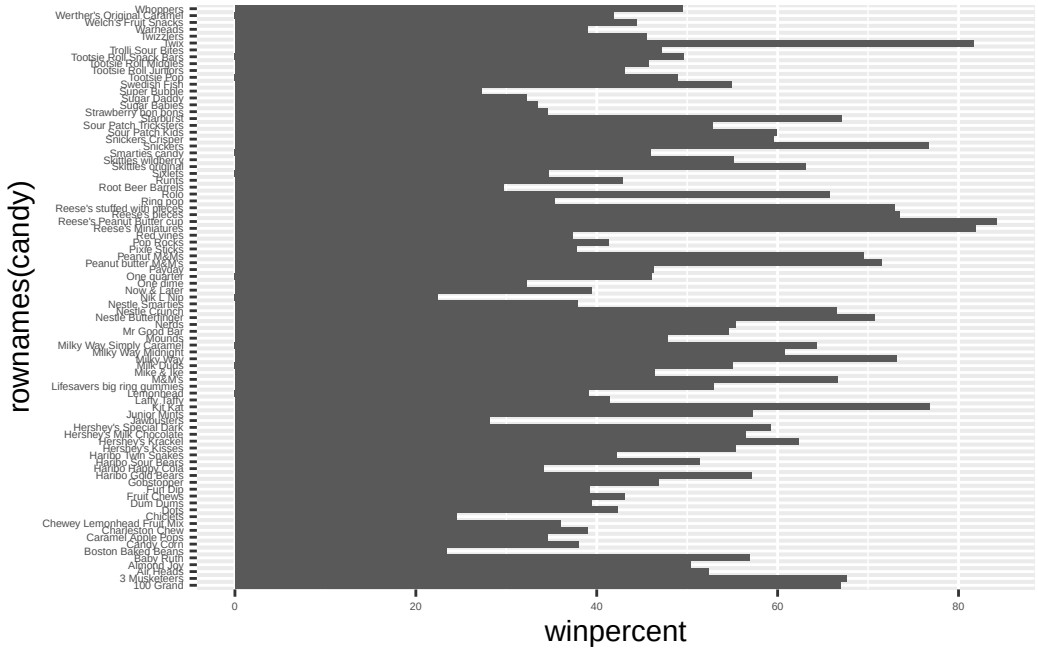
```
[1] "Reese's Peanut Butter cup" "Reese's Miniatures"
[3] "Twix"                      "Kit Kat"
[5] "Snickers"
```

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

Then we can visualize the rank to make it clear.

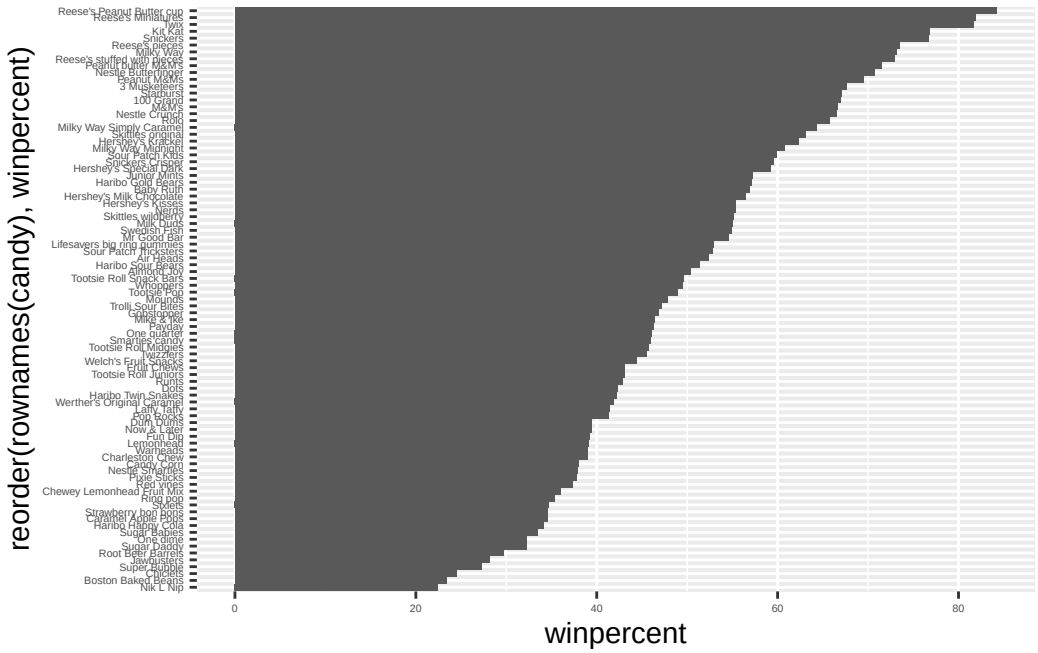
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_bar(stat = "identity") +
  theme(
    axis.text = element_text(size = 4)
  )
```



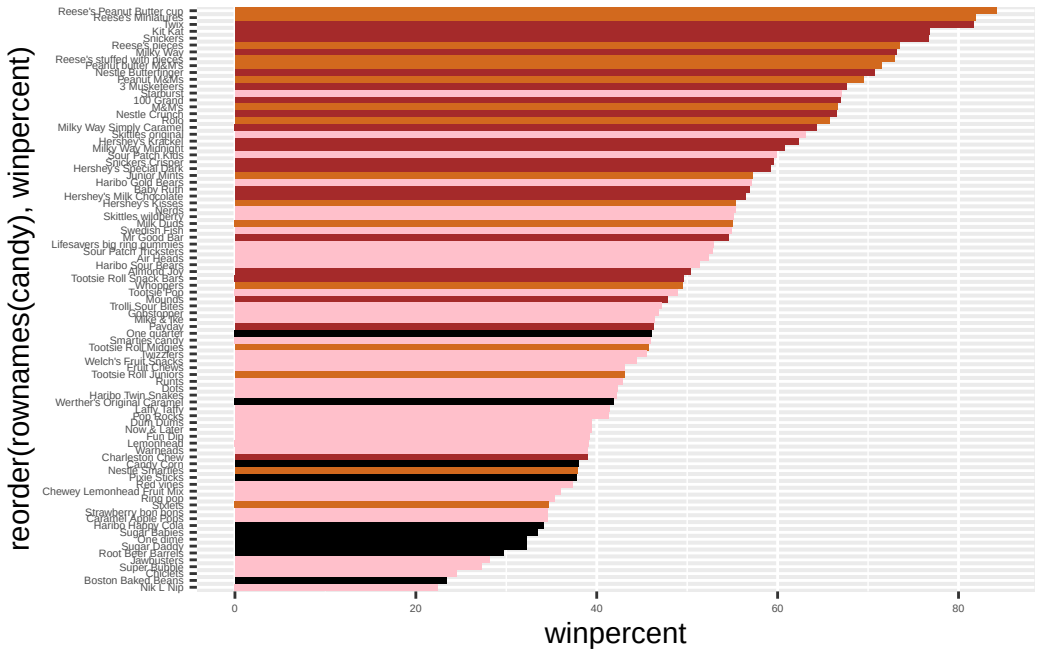
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_bar(stat = "identity") +
  theme(
    axis.text = element_text(size = 4)
  )
```



We add more colors to this chart:

```
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)  +
  theme(
    axis.text = element_text(size = 4)
  )
```



Now, for the first time, using this plot we can answer questions like:

Q17. What is the worst ranked chocolate candy?

Sixlets.

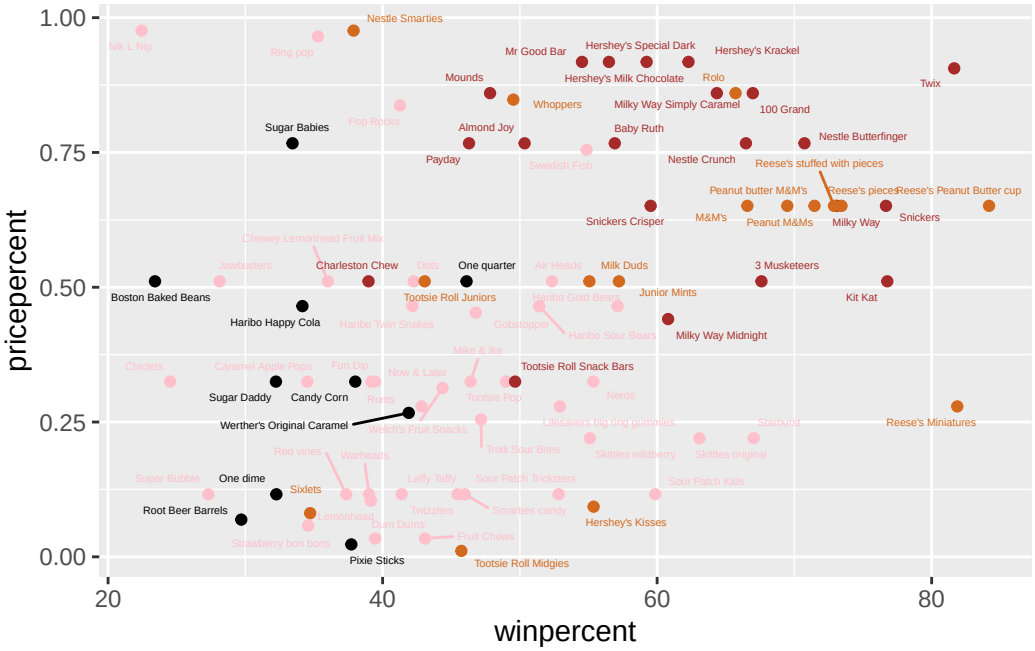
Q18. What is the best ranked fruity candy?

Starburst.

4. Taking a look at pricepercent

We would like to test more about value for money.

```
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=1.3, max.overlaps = 15)
```



(This plot is cool!)

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?

I think it will be "Reese's Miniatures". It has an 80 plus winpercent where it's price percent is just 0.25. It has the lowest price among the candy with higher than 70 percent money.

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

(My eyesight is bad so I used a piece of code for auxiliary analysis.)

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

```
[1] "Nik L Nip"                "Nestle Smarties"
[3] "Ring pop"                 "Hershey's Krackel"
[5] "Hershey's Milk Chocolate"
```

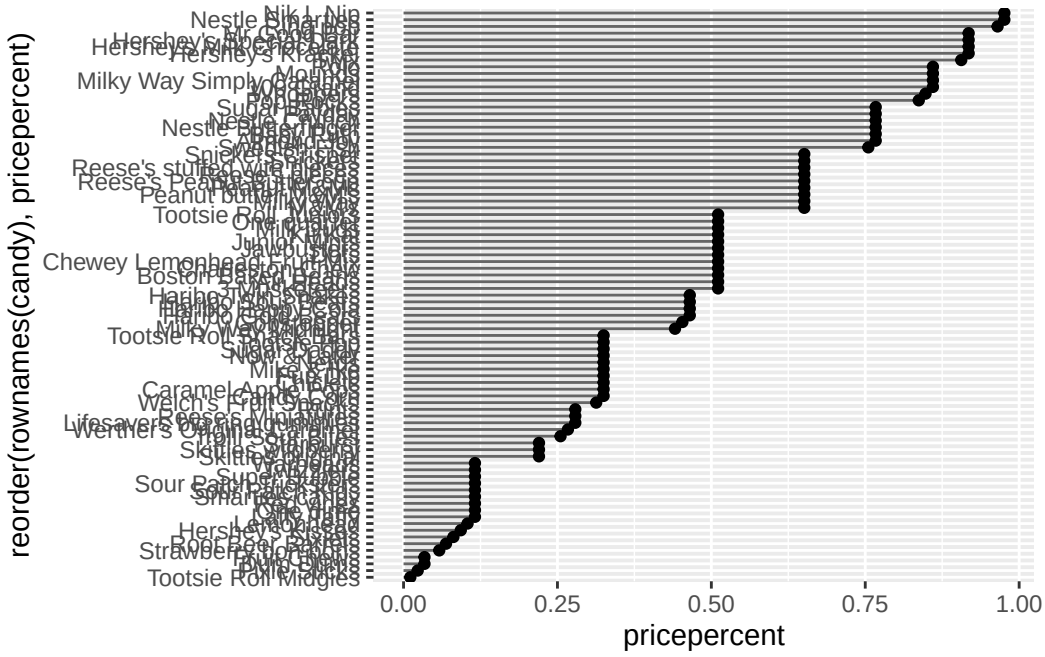
The top 5 most expensive candy types are: Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, Hershey's Milk Chocolate.

The Nik L Nip is the least popular candy.

(Optional) 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()
```



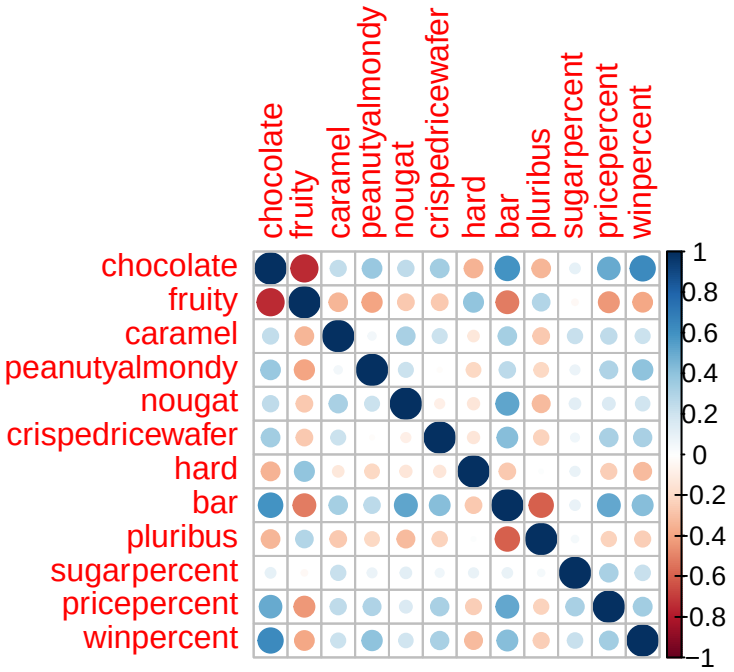
5. Exploring the correlation structure

Now we would like to see how the variables interacts with each other.

```
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)?

A few pairs of them are highly anti-correlated:

- Chocolate v.s. fruity (Chocolate fruit candy is uncommon)
- Pluribus v.s bar (For sure)
- Fruity v.s. bar (Makes sense)
- Fruity/hard/pluribus v.s. pricepercent and winpercent (Those kinds of candy are cheap and unpopular.)

There exist other pairs of them are also anti-correlated.

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

It is chocolate v.s. winpercent. Everyone loves chocolate!

6. Principal Component Analysis

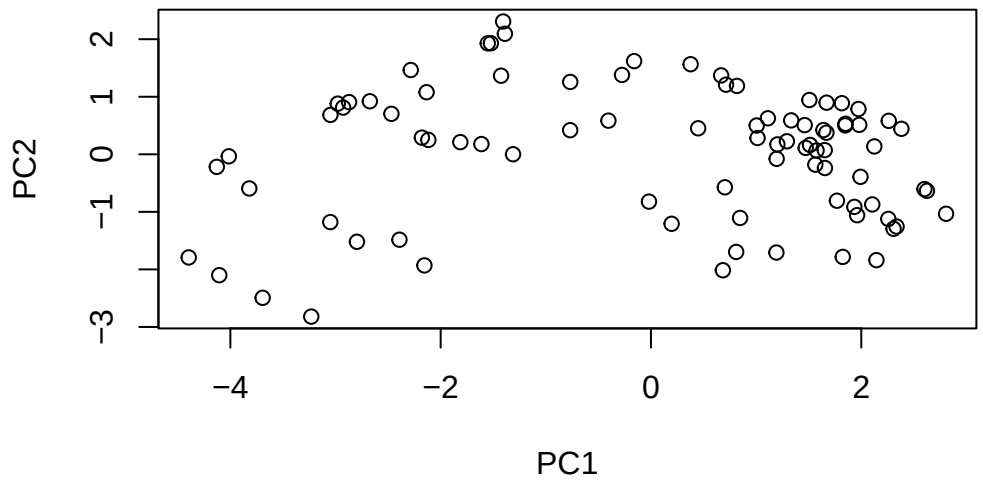
Now we would like to do PCA on this dataset to get more insights.

```
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
```

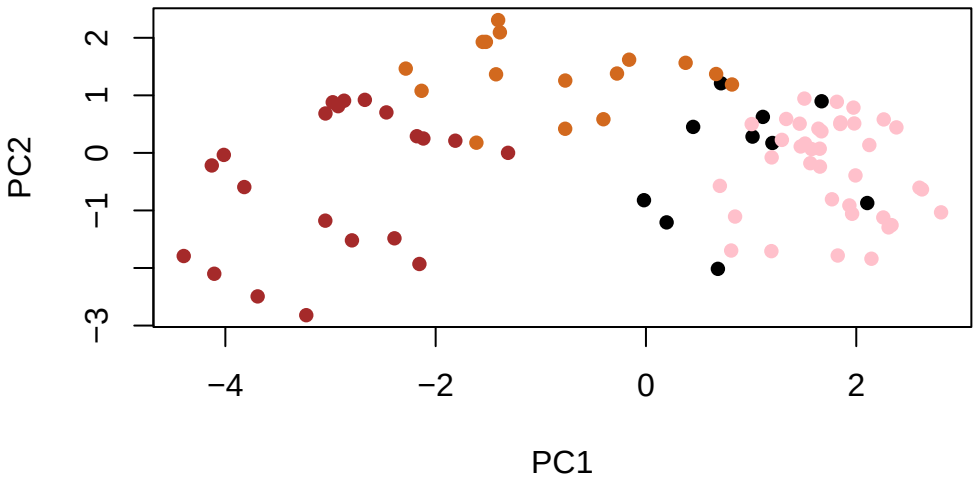
Then we plot the first two principle components (PCs).

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



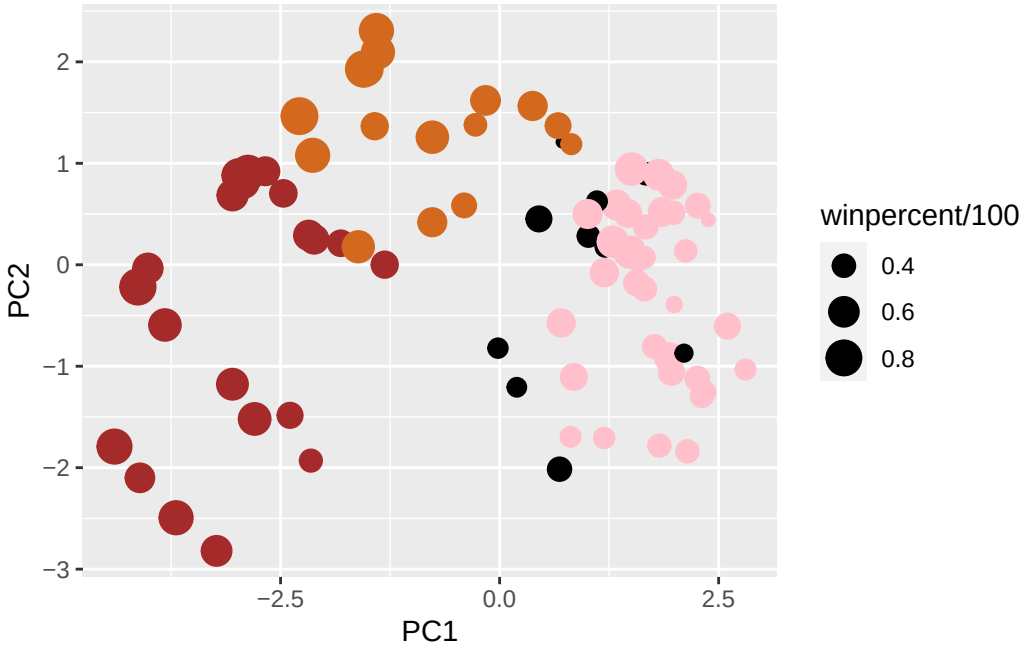
We can change the plotting character and add some color:

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



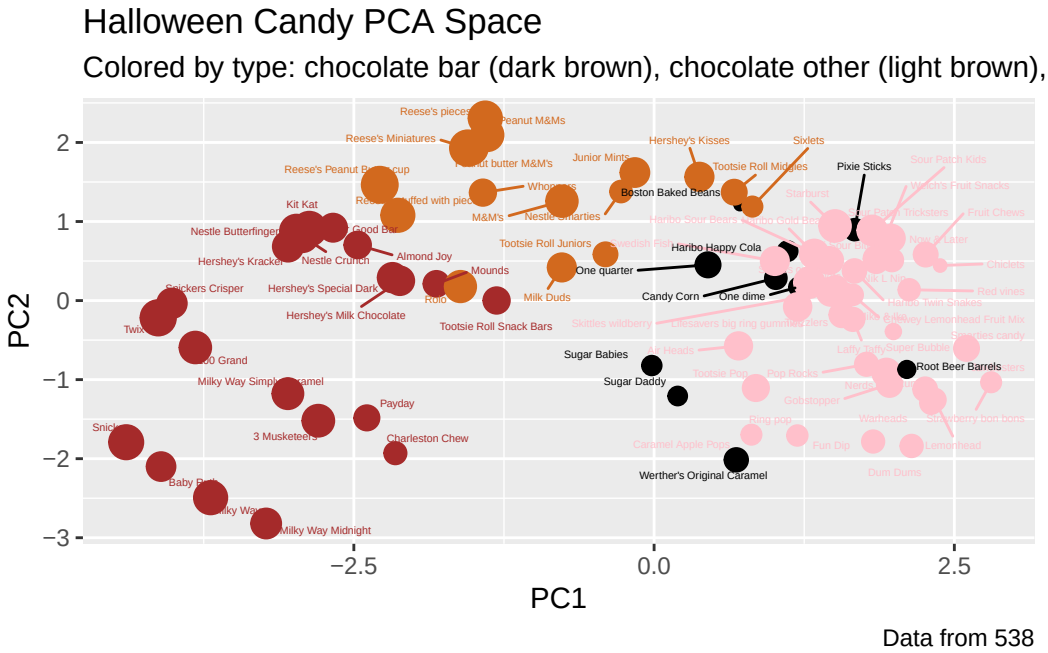
We now would like to use ggplot2 to make a nicer plot. To achieve this we would rather consruct the dataset into a data.frame.

```
# Make a new data-frame with our PCA results and candy data
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
```



Again, let's add some labels.

```
library(ggrepel)

p + geom_text_repel(size=1.3, col=my_cols, max.overlaps = 30)  +
  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")
```

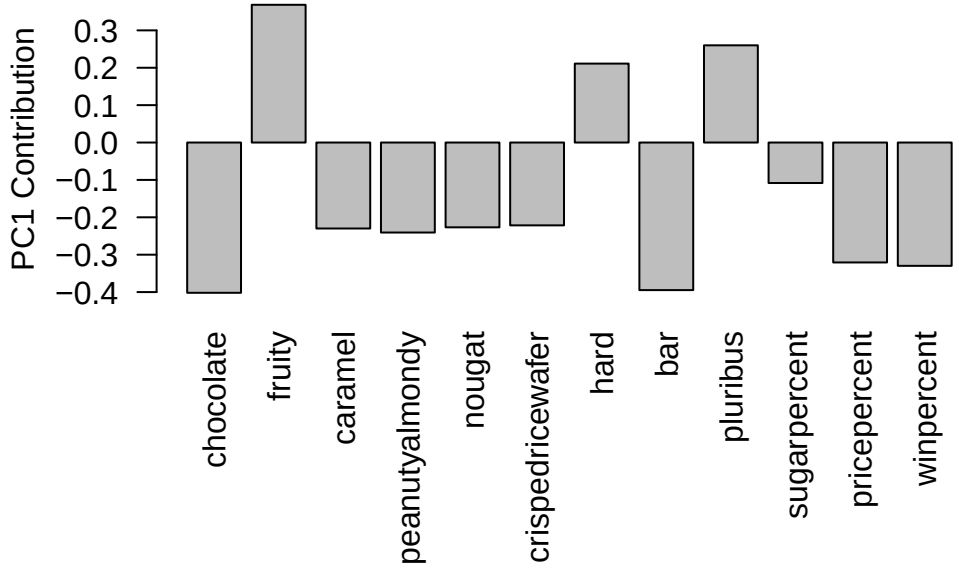


We can also use `plotly` to generate an interactive plot.

```
library(plotly)
ggplotly(p)
```

We can then take a look at the direction of the first PC (the most essential component in some sense:)

```
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. It makes sense since those three variables are positively correlated with each other, and more or less negatively correlated with other variables. It makes sense since along those three directions the variance will be large, so it should be the first PC.