Halloween Mini Project

Max Strul

10-28-2022

Table of contents

Importing Candy data

```r
candy_url <- url("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-powe
candy <- read.csv(candy_url, header = TRUE, 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 candies are in this dataset?

85

Q2: how many fruity candy types are in this data set?

```
sum(candy[,2])
```

```
[1] 38
```

What is my favorite candy?

```
candy["Root Beer Barrels", ]$winpercent
```

```
[1] 29.70369
```

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

```
[1] 76.7686
```

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

```
[1] 49.6535
```

using library `skimr`

```
library("skimr")
skim(candy)
```

Table 1: Data summary

Name	candy
Number of rows	85

Table 1: Data summary

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 data set?

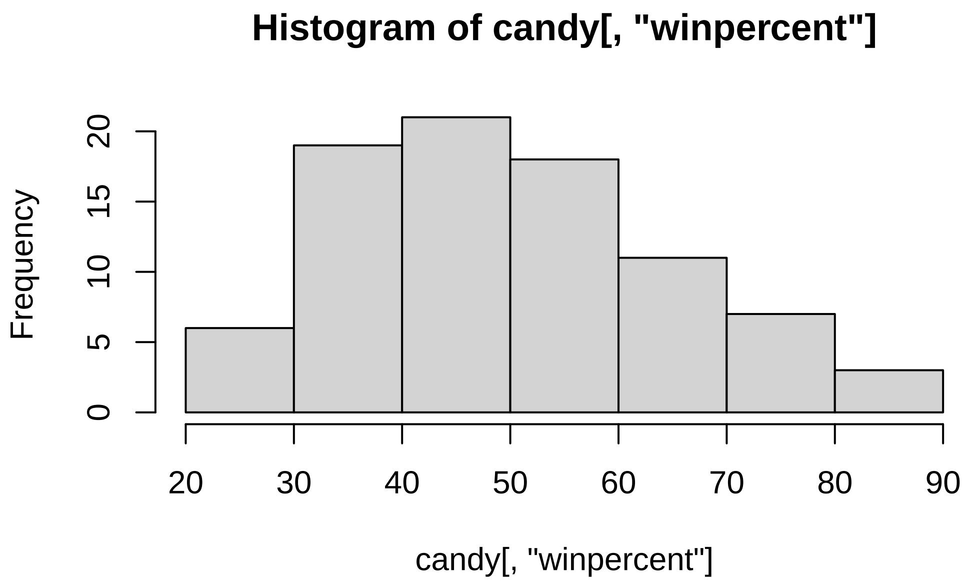
the winpercent is a different scale because the mean is 50 vs ~0-1

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

The zero vs one represents the identity of a candy being or containing chocolate

Q8: Plot a histogram of winpercent values

```
hist(candy[,"winpercent"])
```



Q9: is the distribution of winpercent values symetrical?

No

Q10 is the center of the distributino above or below 50%?

Below 50%

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

```
chocolate.inds <- as.logical(candy$chocolate)
chocolate.win <- candy[chocolate.inds,]$winpercent
mean(chocolate.win)
```

```
[1] 60.92153
```

```
fruity.inds <- as.logical(candy$fruit)
fruity.win <- candy[fruity.inds,]$winpercent
mean(fruity.win)
```

```
[1] 44.11974
```

Q12: is the difference stastically significant?

```
t.test(chocolate.win,fruity.win)
```

```

    Welch Two Sample t-test

data:  chocolate.win and fruity.win
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 
```

This data is significantly different, that chocolate, on average is ranked higher than fruity

Q13: What are the five least likely candy types in this set?

```
ordered_by_winpercent <- candy[order(candy[,"winpercent"]),]
head(ordered_by_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
```

```
# <- candy %>% group_by(winpercent)
#by_winpercent %>% arrange(desc(winpercent))
```

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

```
ordered_by_winpercent <- candy[order(candy[,"winpercent"],decreasing = TRUE),]
head(ordered_by_winpercent, n=5)
```

```
                          chocolate fruity caramel peanutyalmondy nougat
ReeseÕs Peanut Butter cup         1      0       0              1      0
ReeseÕs 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
ReeseÕs Peanut Butter cup                0    0   0        0        0.720
ReeseÕs Miniatures                       0    0   0        0        0.034
Twix                                     1    0   1        0        0.546
Kit Kat                                  1    0   1        0        0.313
```

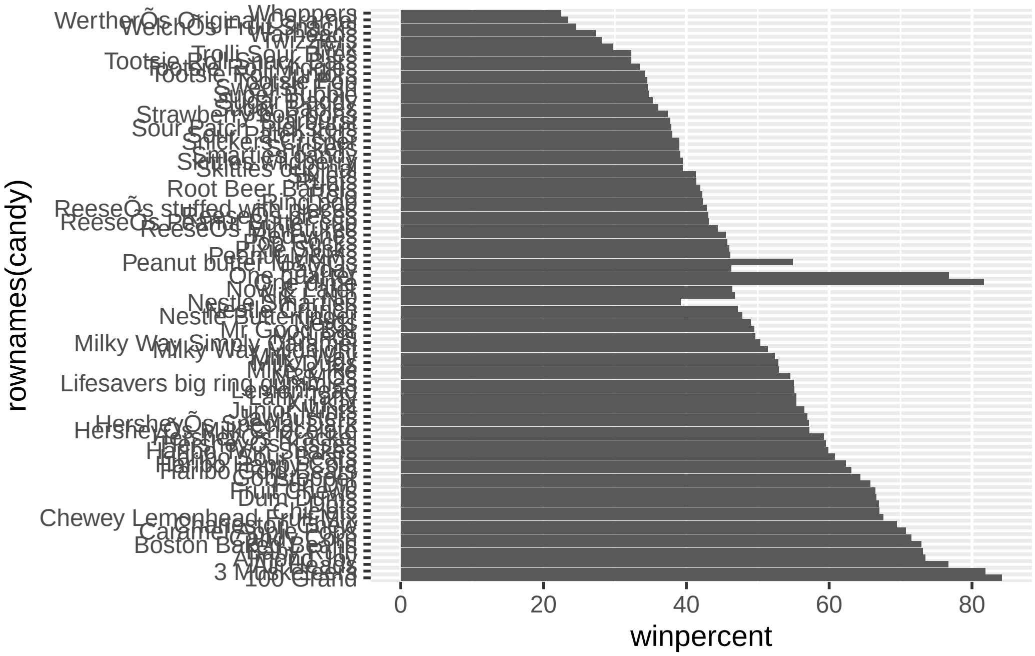
```
Snickers                                 0    0   1        0        0.546
                          pricepercent winpercent
ReeseÕs Peanut Butter cup        0.651   84.18029
ReeseÕs Miniatures               0.279   81.86626
Twix                             0.906   81.64291
Kit Kat                          0.511   76.76860
Snickers                         0.651   76.67378
```

```
library("dplyr")
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
```

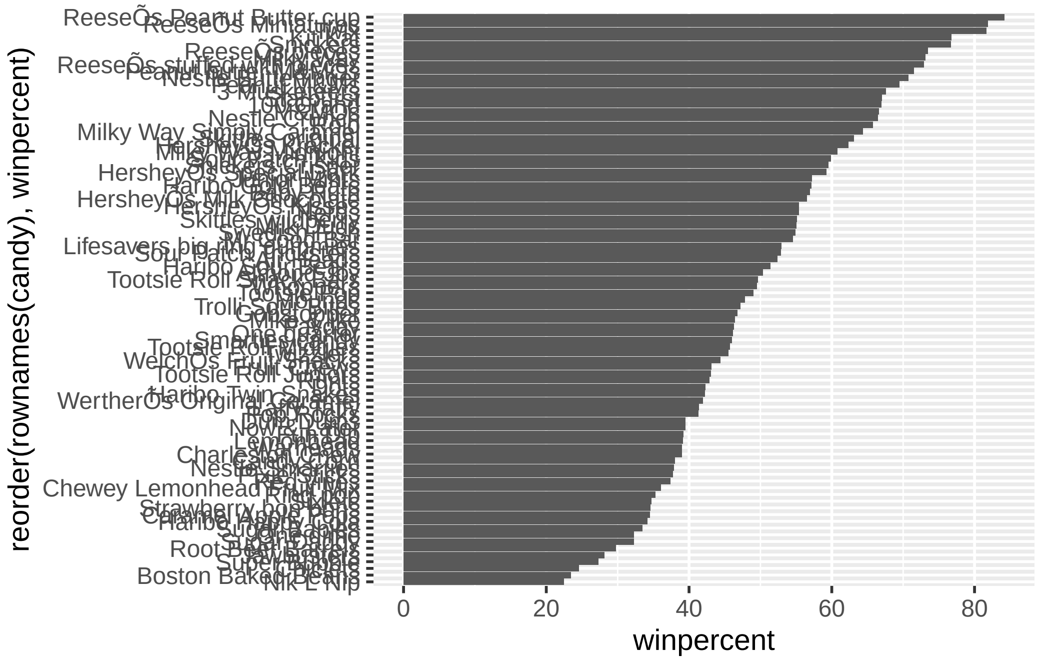
Q15/16: Make a barplot of candy rankings based on winpercent values

```
library("ggplot2")
ggplot(ordered_by_winpercent)+
  aes(winpercent,rownames(candy))+
  geom_col()
```



```
#You can do saving certain
#ggsave("mybarplot.png",height=1200,width=800,units=c("px"),dpi=300)
```

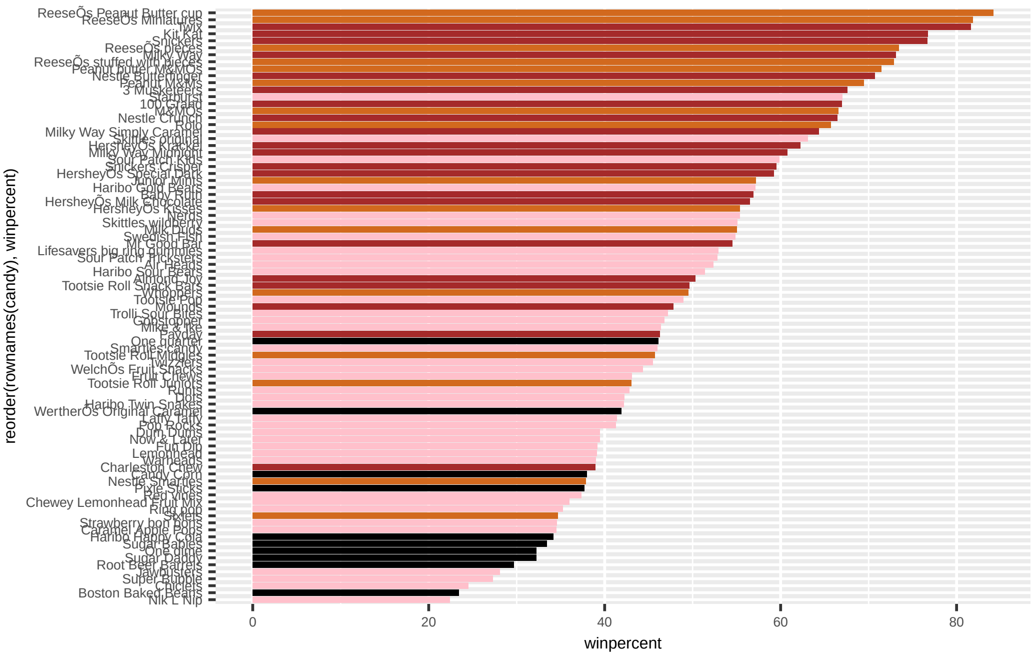
```
library("ggplot2")
ggplot(candy)+
  aes(winpercent,reorder(rownames(candy),winpercent))+
  geom_col()
```



Q17: Coloring the bar plot via type

```
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(text=element_text(size=6))
```



```
ggsave("mybarplot_withcol.png")
```

```
Saving 5.5 x 3.5 in image
```

Worst ranking chocolate is sixlets

Q18: What is the best ranked fruity candy

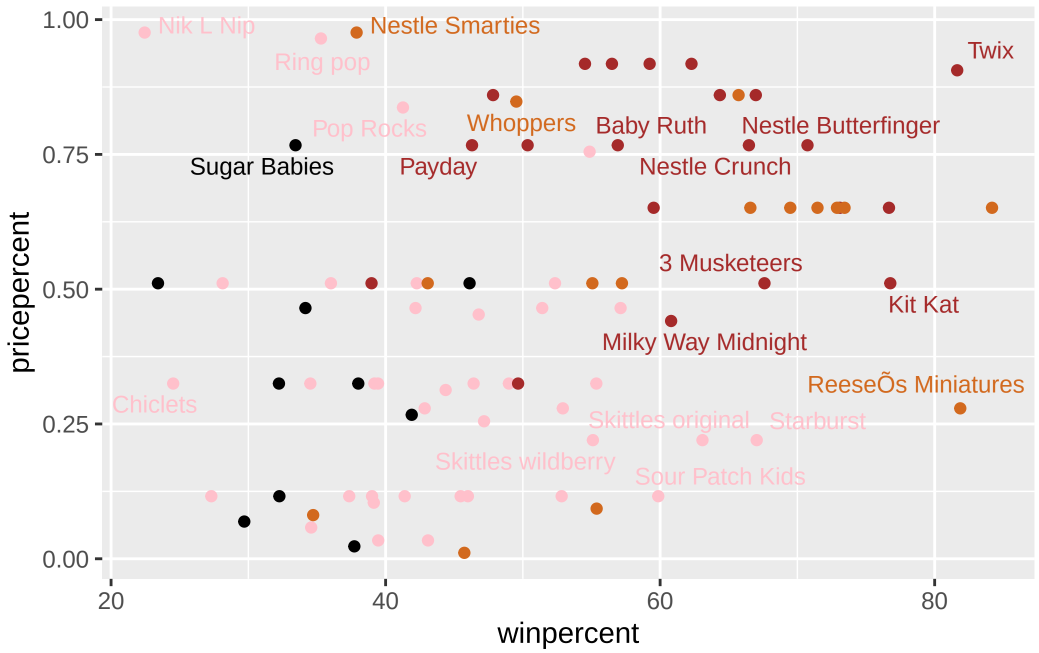
Star bursts

Looking at pricepercent

```
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 is the highest rank and the lowest price

Reese's miniatures

Q20: What are the top 5 most expensive candy types

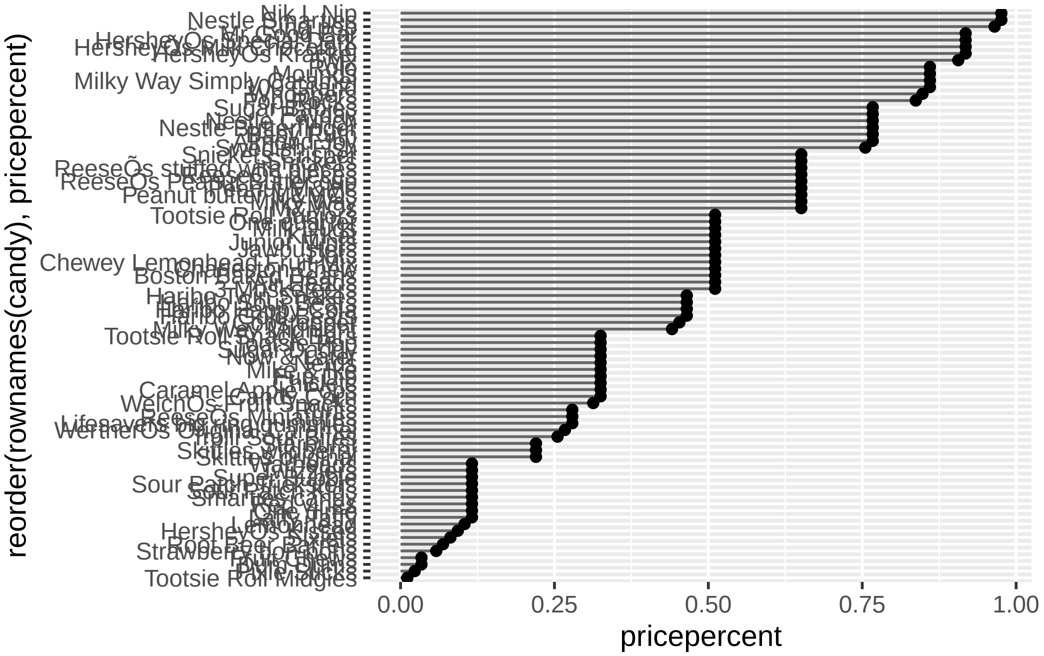
```r
ordered_by_pricepercent <- candy[order(candy[,"pricepercent"],decreasing = TRUE),]
head(ordered_by_pricepercent, n=5)
```

```
                         chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                        0      1       0              0      0
Nestle Smarties                  1      0       0              0      0
Ring pop                         0      1       0              0      0
HersheyÕs Krackel                1      0       0              0      0
HersheyÕs Milk Chocolate         1      0       0              0      0
```

```
                         crispedricewafer hard bar pluribus sugarpercent
Nik L Nip                               0    0   0        1        0.197
Nestle Smarties                         0    0   0        1        0.267
Ring pop                                0    1   0        0        0.732
HersheyÕs Krackel                       1    0   1        0        0.430
HersheyÕs Milk Chocolate                0    0   1        0        0.430
                         pricepercent winpercent
Nik L Nip                       0.976   22.44534
Nestle Smarties                 0.976   37.88719
Ring pop                        0.965   35.29076
HersheyÕs Krackel               0.918   62.28448
HersheyÕs Milk Chocolate        0.918   56.49050
```

Q21: Make a barplot again with `geom_col()` this time using `pricepercent`

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

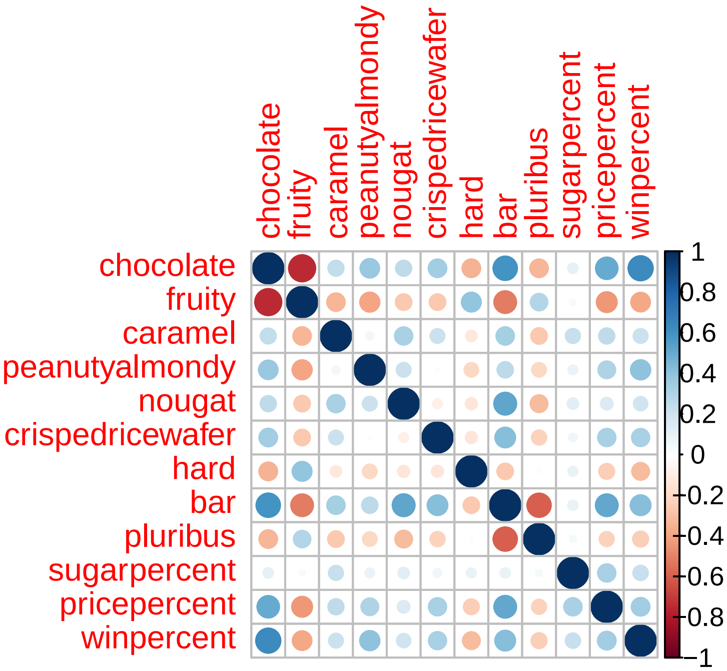


Exploring correlation structure

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



Q22: Examing this plot what two variables are anti-correlated?

Fruity with chocolate and pluribus with bar

which makes sense, very few candies are fruity and chocolate-y and very few candies are in a bar form with multiple bars

Q23:SWhat two variables are most positively correlated?

Chocolate with winpercent, and chocolate with Bar

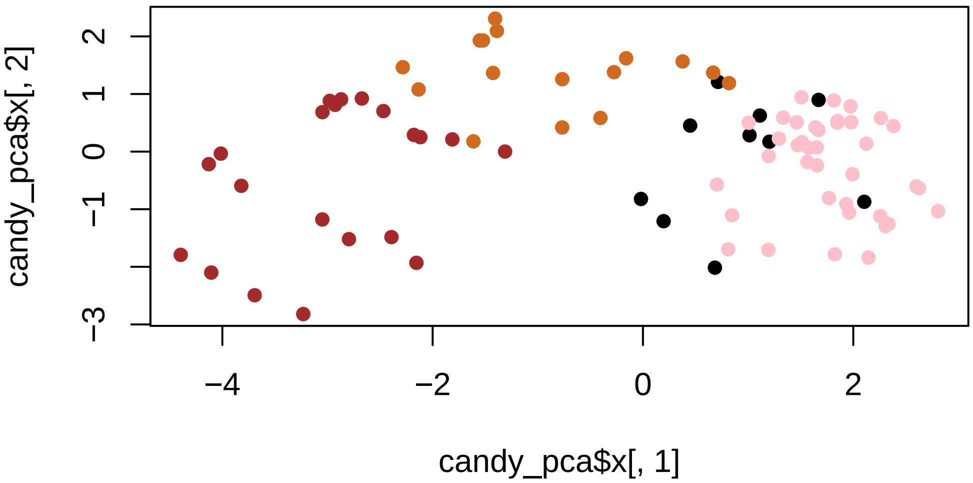
Principal component analysis

Note: we need `scale=TRUE` to be entered because we want to scale the winpercent which is abnormally scaled wtih the rest of the data.

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

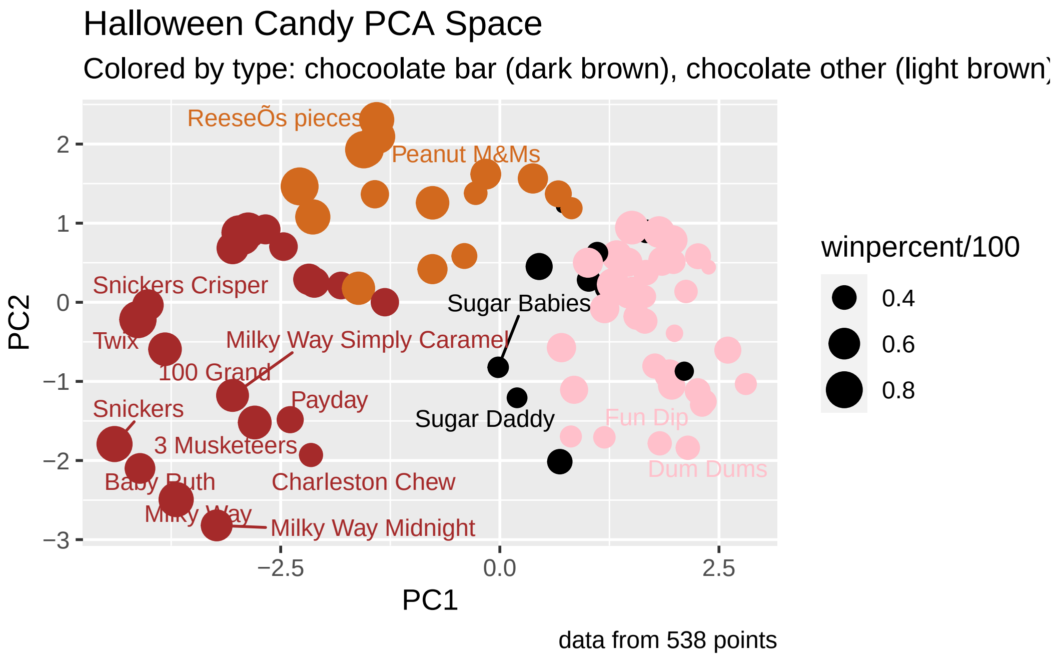
```
plot(candy_pca$x[,1],candy_pca$x[,2],col=my_cols,pch=16)
```



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

```
plot_1 <- ggplot(my_data)+
             aes(x=PC1, y=PC2,
                 size=winpercent/100,
                 text=rownames(my_data),
                 label=rownames(my_data))+
             geom_point(col=my_cols)+
              geom_text_repel(size=3.3,col=my_cols,max.overlaps = 7)+
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocoolate bar (dark brown), chocolate other (light brow
       caption="data from 538 points")
plot_1
```

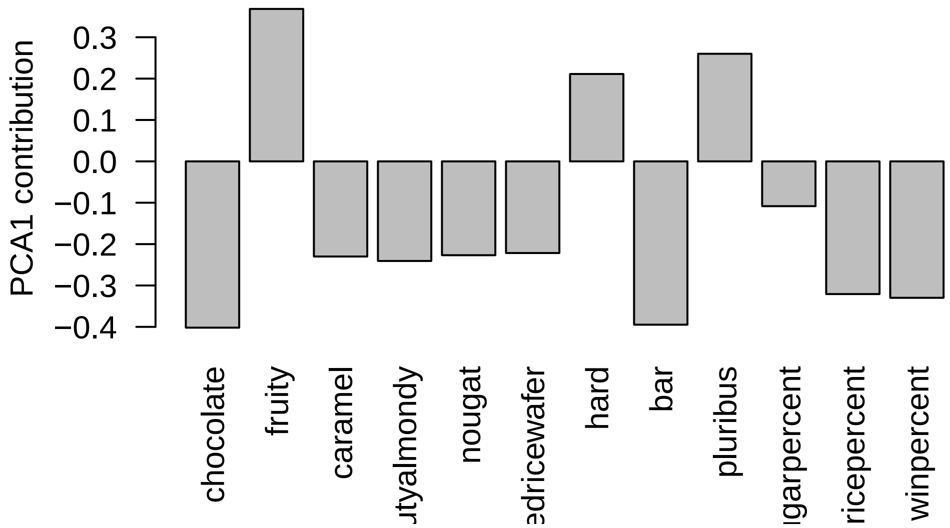
```
Warning: ggrepel: 68 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



```
#library(plotly)
#ggplotly(plot_1)
```

Q24: What original variables are picked up strongly by PC1 in the positive direction?

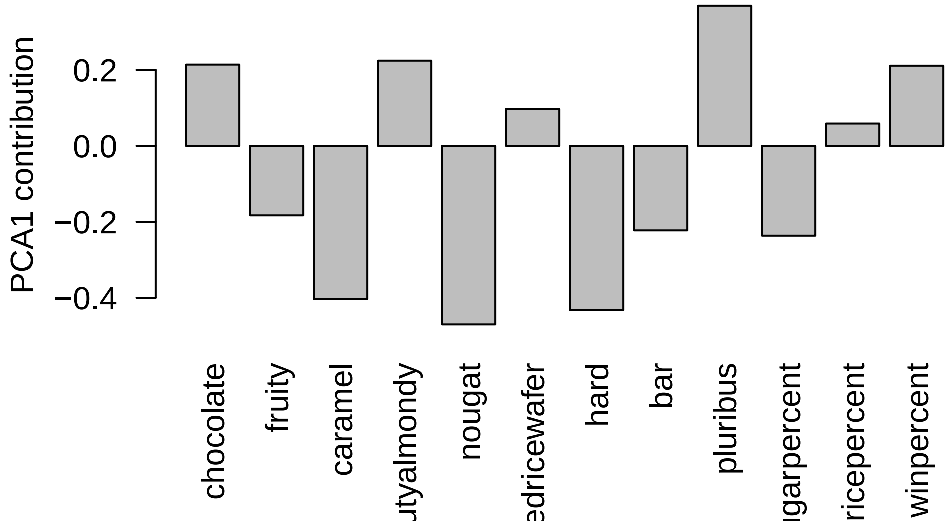
```
barplot(candy_pca$rotation[,1], las=2,ylab="PCA1 contribution")
```



This means that fruity, hard, and muptiple pieces of candy coming in a bag or a box are all tightly correlated with each other. Similary chocolate, caramel, peanuts & almods, nougat, crispy, bar, high sugar, high price, and high win percent are all correlated with eachother.

What about PC2?

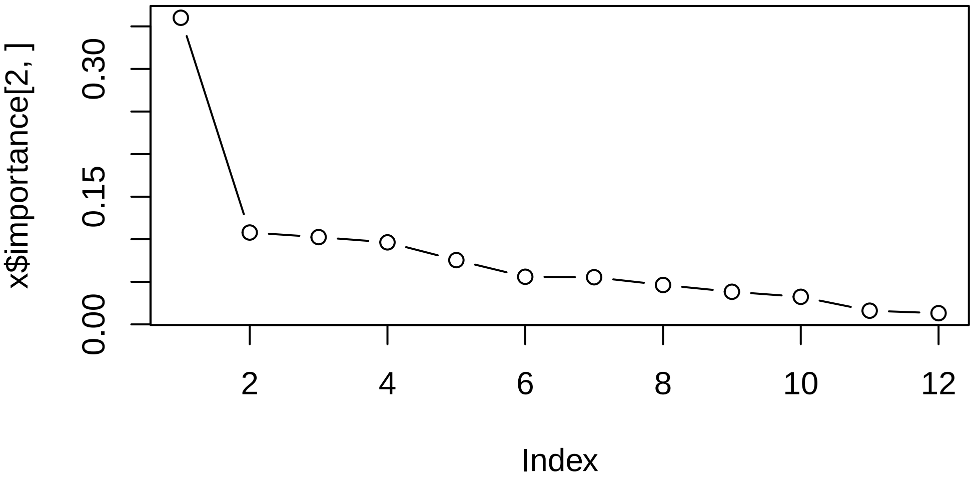
```
barplot(candy_pca$rotation[,2], las=2,ylab="PCA1 contribution")
```



PC2 shows us that caramel/nougat and hard are the opposite of nutty and multiple things in a single bag.

To find out what the intrinsic dimensionality of a data set is you can plot the `candy_pca` summary of variance to determine the inflection point

```
x <- summary(candy_pca)
plot(x$importance[2,],type="b")
```



This shows us that two dimensions is perfectly sufficient for data separation.