Class 8: Halloween Mini-Project

Jason Hsiao (PID: A15871650)

Importing the Candy Dataset

```
candy <- read.csv("candy-data.csv", 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
```

What's in the dataset?

```
nrow(candy) #rows in the dataset (candy types)
```

```
[1] 85
```

> Q1: How many different candy types are in this dataset?

There are 85 different candy types in the dataset

> Q2: How many fruity candy types are in the dataset?

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

```
[1] 38
```

There are 38 fruity candy types in the dataset.

> Q3: What is your favorite candy in the dataset and what is its `winpercent` value?

```r
candy["Peanut M&Ms", ]$winpercent
```

```
[1] 69.48379
```

Win % of Peanut M&Ms is 69.5%

> Q4: What is the winpercent value for "Kit Kat"?

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

```
[1] 76.7686
```

Win % of Kit Kat is 76.8%.

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

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

```
[1] 49.6535
```

Win % of Tootsie Roll Snack Bars is 49.7%.

Using the `skim()` function to get an overview of a given dataset…

```r
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?

winpercent values are in whole percentages, while others are in proportions.

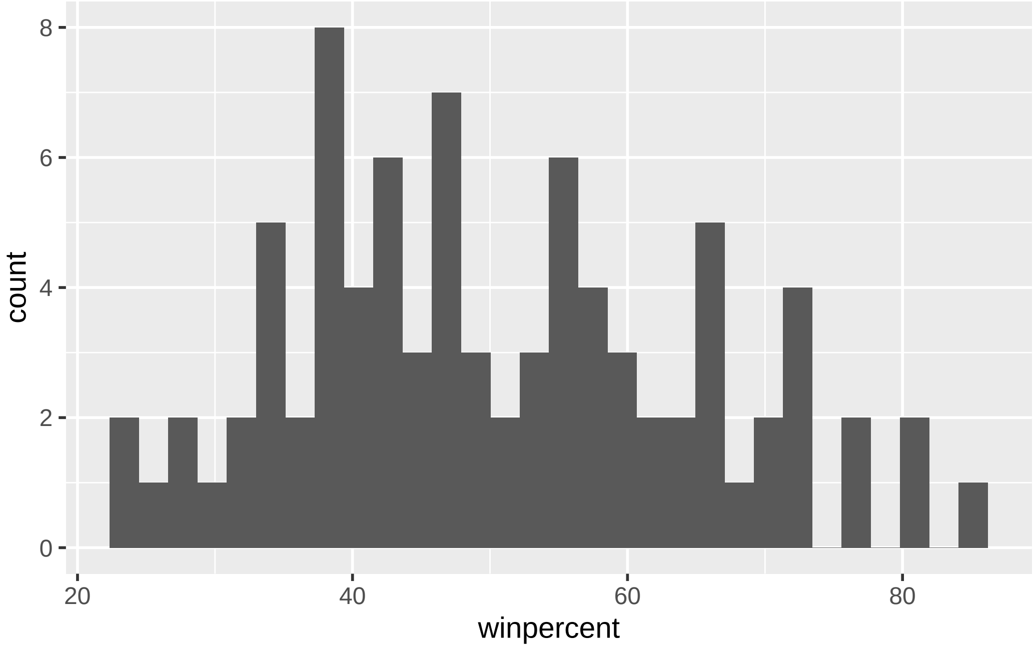
> Q7: What do you think a zero and one represent for the candy$chocolate column?

Whether or not the given candy from a given row is (one) and isn't (zero) under the chocolate category.

> Q8: Plot a histogram of winpercent values

```r
library(ggplot2)
ggplot(candy, aes(x = winpercent)) +
  geom_histogram()
```

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



Q9: Is the distribution of winpercent values symmetrical?

No, it appears to have a right-skew.

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

Below 50%.

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

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

Chocolate candy is ranked higher than fruit candy!

> Q12: Is this difference statistically significant?

```
#assign 'choc' object as vector of win percentages of chocolate candies
choc <- candy$winpercent[as.logical(candy$chocolate)]
#assign 'fruit' object as vector of win percentages of fruity candies
fruity <- candy$winpercent[as.logical(candy$fruity)]

t.test(choc, fruity)
```

```

    Welch Two Sample t-test

data:  choc 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 
```

The difference is statistically significant!

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

```
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
```

Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters are the 5 least liked candies!

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

```
candy %>%
  arrange(-winpercent) %>%
  head(5)
```

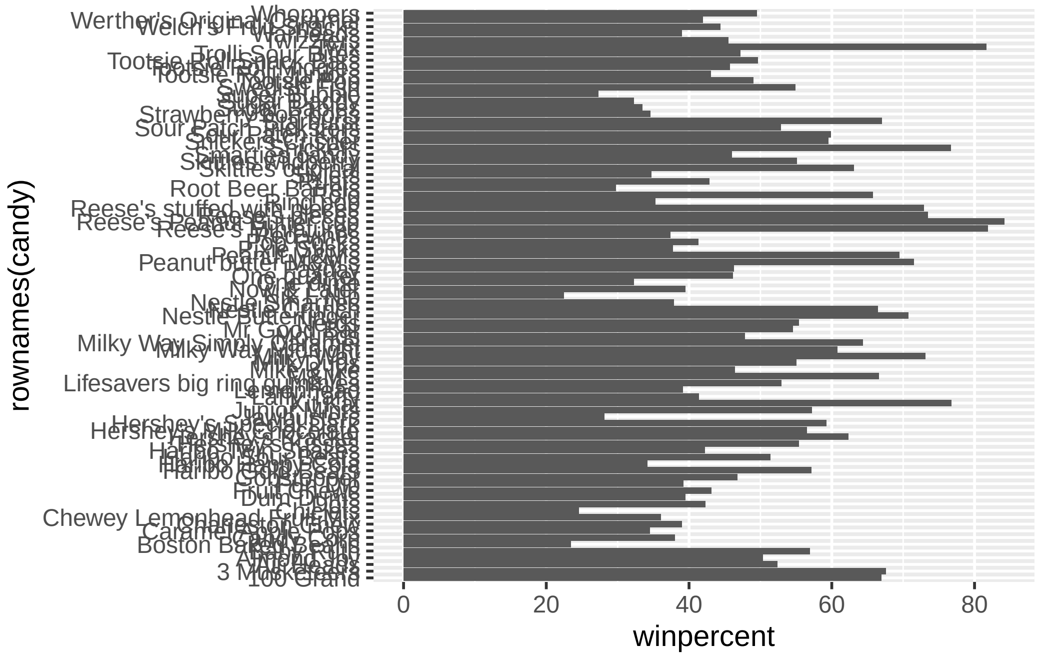
```
                          chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat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

Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, Snickers are the top 5 favorites!

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

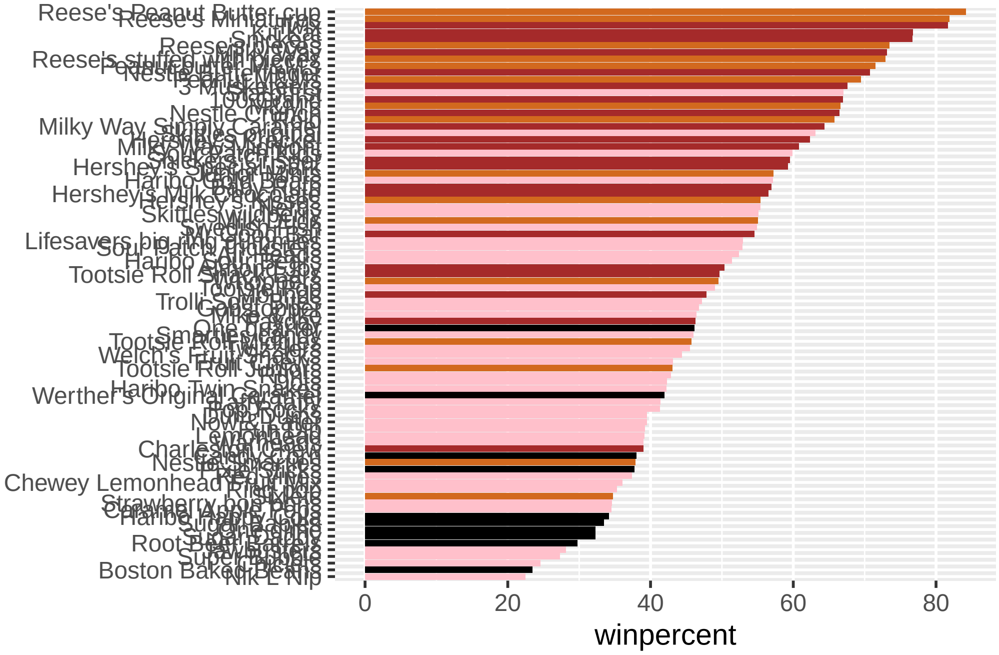
```
ggplot(candy, aes(winpercent, rownames(candy))) +
  geom_col()
```



Reordering by winpercent:

```
#setting color parameters
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?

Nik L Nip

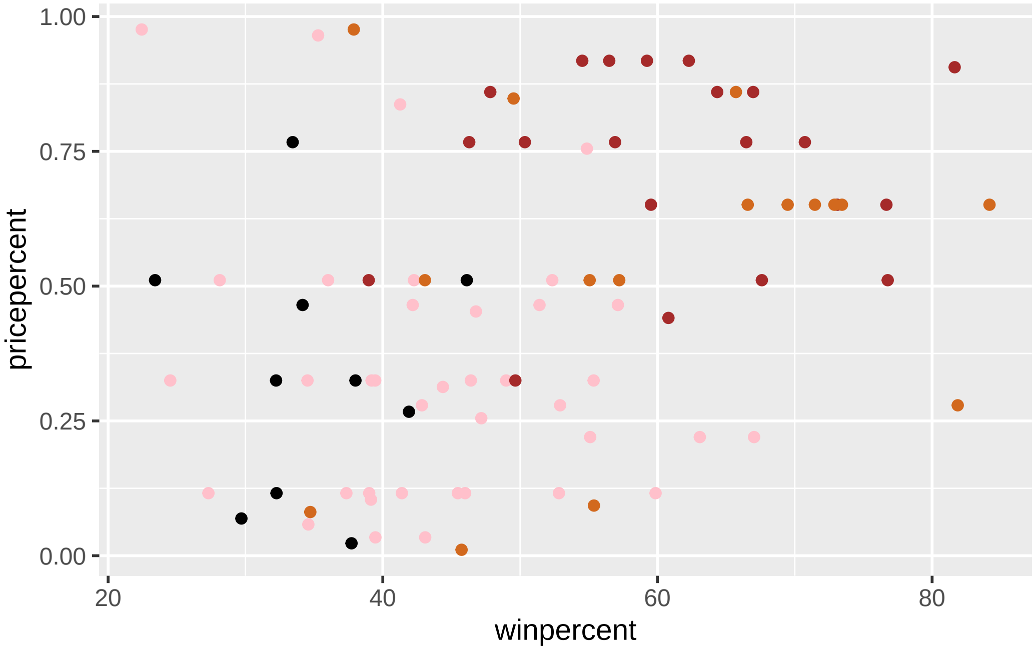
Q18: What is the best ranked fruity candy?

Starburst

Exploring pricepercent...

```
#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 = 3.3, max.overlaps = 5)
```

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?

```
#ordering pricepercent to see most/least expensive candy
ord <- order(candy$pricepercent, decreasing = FALSE)
head( candy[ord, c(11,12)], n = 5)
```

```
                     pricepercent winpercent
Tootsie Roll Midgies        0.011   45.73675
```

```
Pixie Sticks                0.023   37.72234
Dum Dums                    0.034   39.46056
Fruit Chews                 0.034   43.08892
Strawberry bon bons         0.058   34.57899
```

Tootsie Roll Midgies

> 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
```

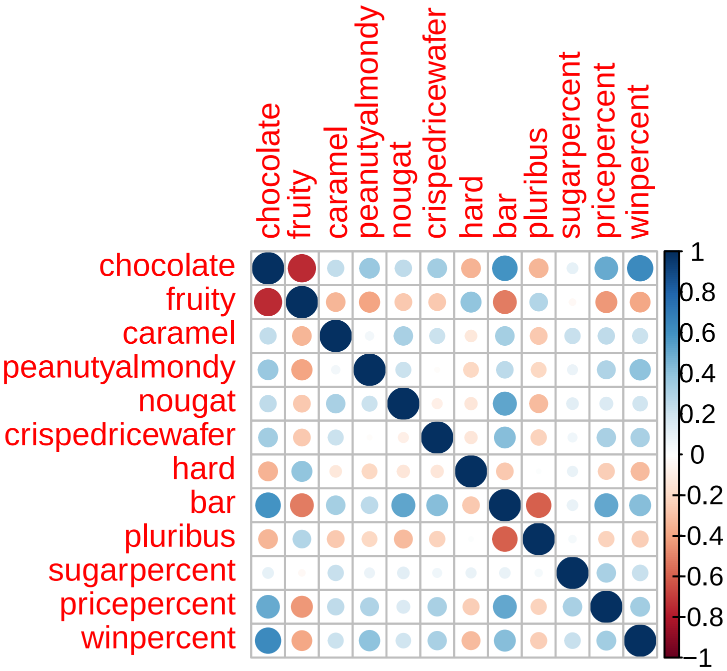
Nik L Nip is the most expensive and least popular of the dataset.

Correlation Plot

```
library(corrplot)
```

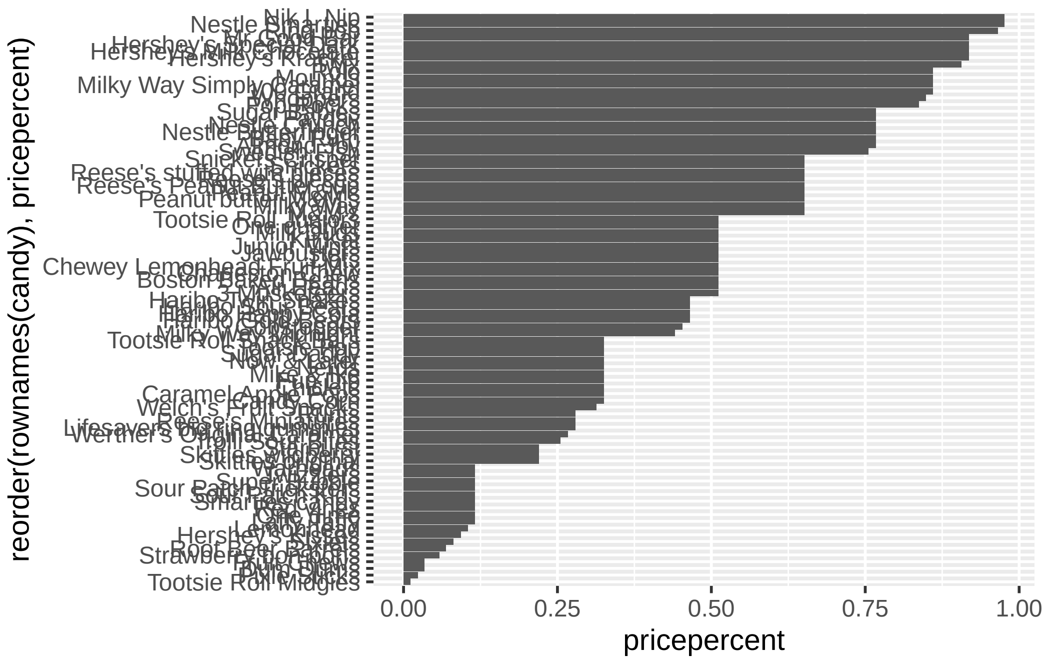
```
corrplot 0.92 loaded
```

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

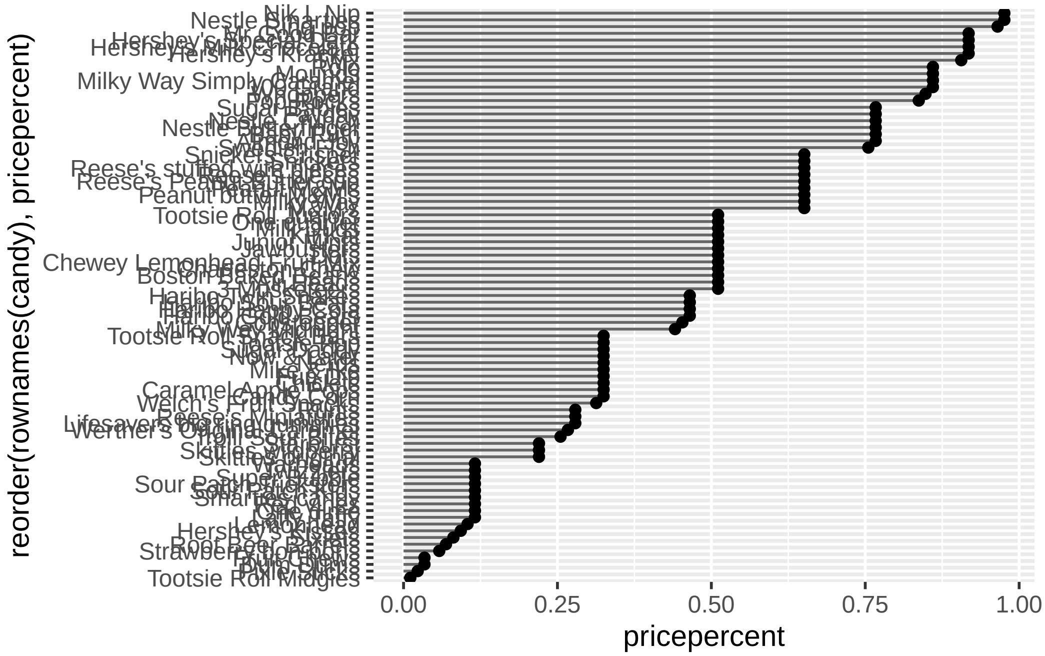


Q21: Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping geom_col() for geom_point() + geom_segment().

```
#barplot of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_col()
```



```
#making it prettier: make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```



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

chocolate and fruity

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

chocolate and winpercent

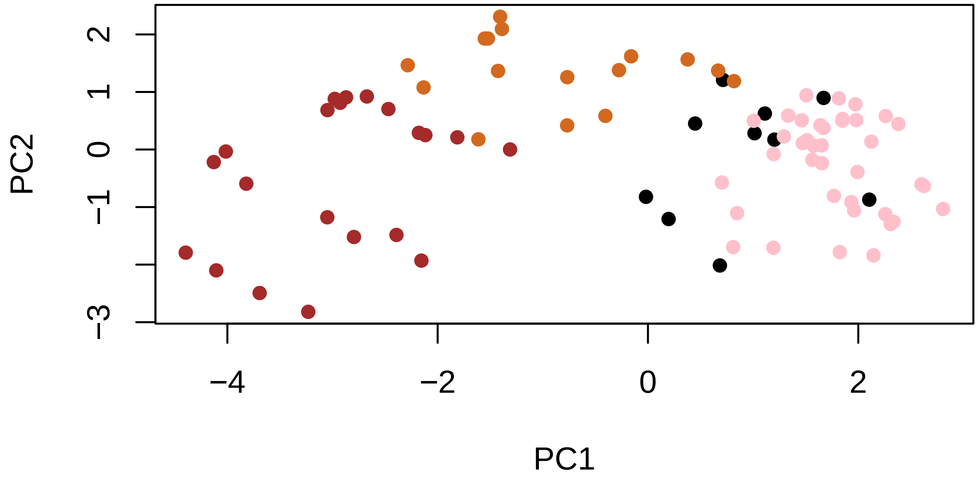
PCA

```
#setting up PCA object
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
```

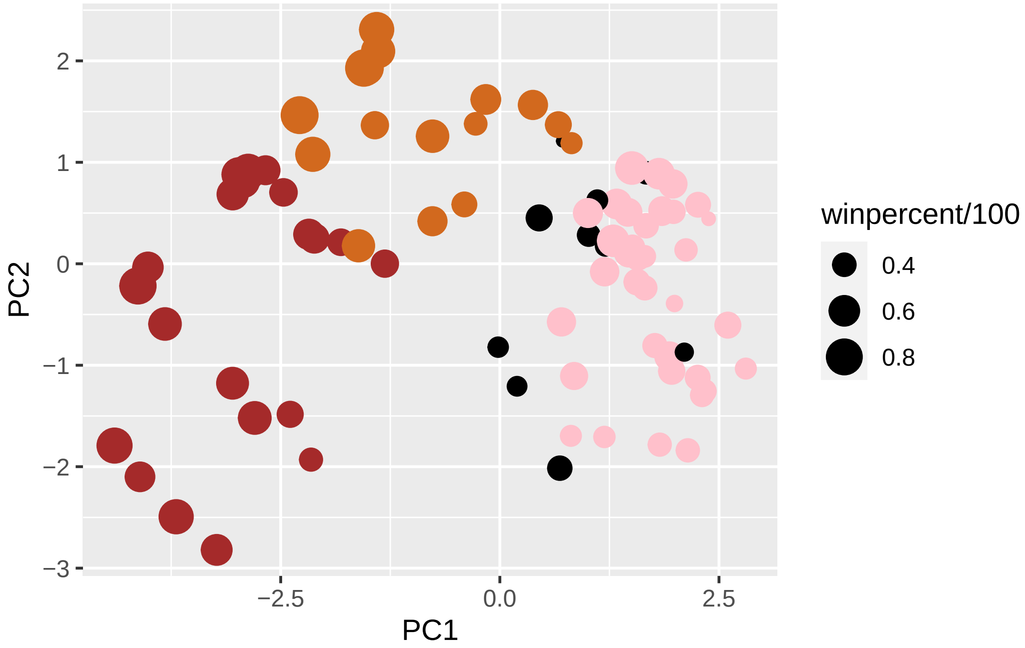
```
plot(pca$x[, 1:2], col = my_cols, pch = 16)
```



```
# 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
```



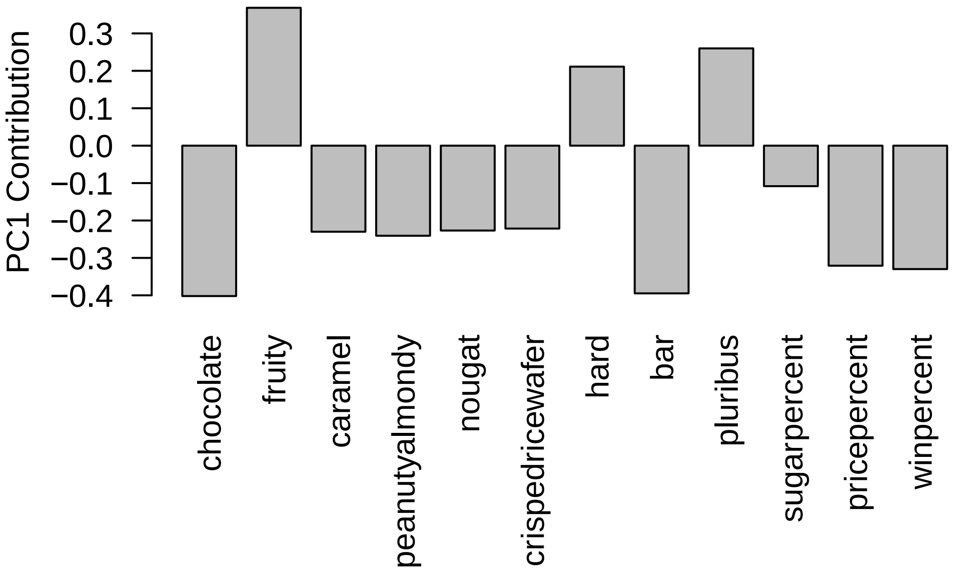
```
#putting labels on each of the dots; making it prettier
#library(ggrepel)

#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 brow
       #caption="Data from 538")

#passing plot 'p' to plotly to get interactive plot (removes overlap)
#library(plotly)
#ggplotly(p)
```

PC Loadings

```
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, pluribus - yes, it makes sense (they normally come in this form)