Class 9: Halloween Candy Mini-Project

Yoonjin Lim (PID:A16850635)

Table of contents

Today we will take a wee step back to some data we can taste and explore the correlation structure and principal components of some Halloween candy.

Data Import

```r
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?

```
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

What is your favorite candy?

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

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

```
[1] 81.64291
```

```
candy["Almond Joy",]$winpercent
```

```
[1] 50.34755
```

```
candy["M&M's",]$winpercent
```

```
[1] 66.57458
```

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”?

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

```
[1] 49.6535
```

Exploratory Analysis

We can use the **skimr** package to get a quick overview of a given dataset. This can be useful for the first time you encounter a new dataset.

```r
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?

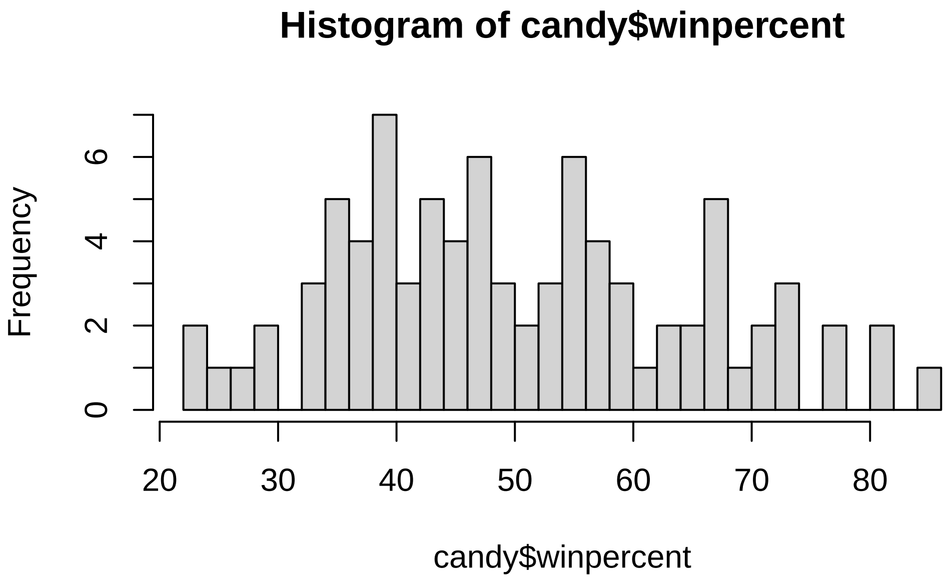
It looks like the last column `candy$winpercent` is on a different scale to all others.

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

Based on the table above in `candy` dataset, I believe that a chocolate column includes several zero and one. As the information on chocolate column is encoded as integer, it is likely for a zero to represent FALSE and a one to represent TRUE. Therefore, a one likely represents that the candy contains chocolate, and a zero refers that it does not contain chocolate in the candy$chocolate column.

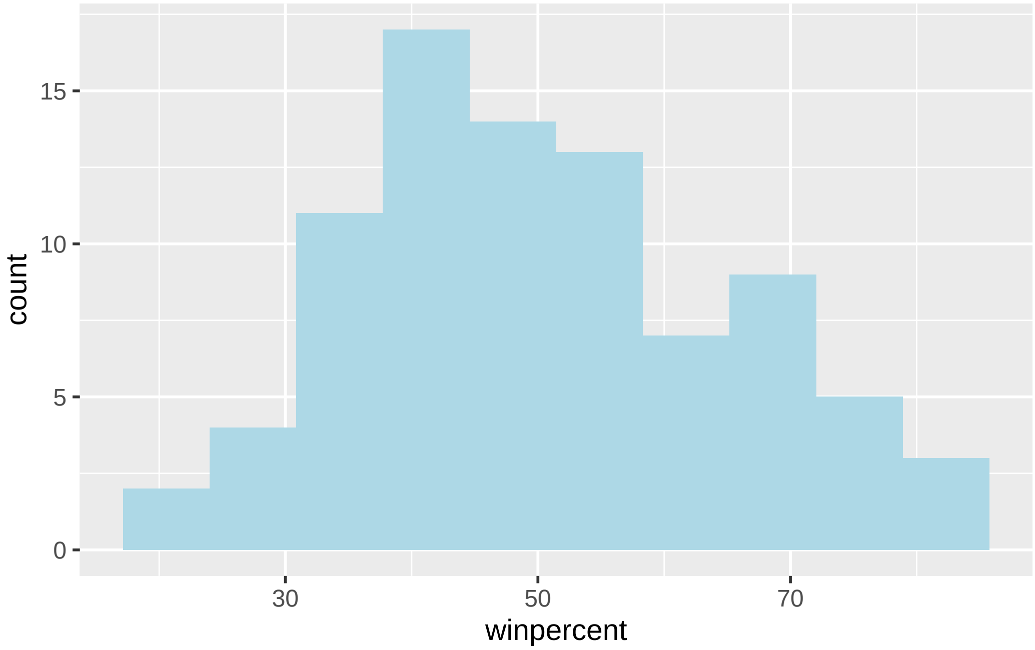
> Q8. Plot a histogram of winpercent values

```r
hist(candy$winpercent, breaks=30)
```



```r
library(ggplot2)

ggplot(candy) +
  aes(winpercent) +
  geom_histogram(bins=10, fill="lightblue")
```



> Q9. Is the distribution of winpercent values symmetrical?

Based on the observation of results above, the distribution of winpercent values is not symmetrical.

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

Determining the center of the distribution depends on which parameters we choose, such as median or mean. However, as I chose the median as a parameter, the center of distribution is below 50% since the median is below 50% (47.83%).

```
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.14   47.83   50.32   59.86   84.18 
```

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

```
choc.inds <- candy$chocolate==1
choc.candy <- candy[choc.inds,]
choc.win <- choc.candy$winpercent
mean(choc.win)
```

```
[1] 60.92153
```

On average, the chocolate candy is higher ranked than fruity candy since the mean for winpercent of chocolate candy is over 50%. The below result is the ranking of fruity candy by measuring the mean for winpercent of fruity candy: 44.11974%.

```
fruit.win1 <- candy[as.logical(candy$fruity), ]$winpercent
mean(fruit.win1)
```

```
[1] 44.11974
```

```
fruit.inds <- candy$fruity==1
fruit.candy <- candy[fruit.inds,]
fruit.win <- fruit.candy$winpercent
mean(fruit.win)
```

```
[1] 44.11974
```

Q12. Is this difference statistically significant?

```
ans <- t.test(choc.win, fruit.win)
ans
```

```

    Welch Two Sample t-test

data:  choc.win and fruit.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 
```

Yes, with a P-value of 2.8713778×10^{-8}.

```
ans$p.value
```

```
[1] 2.871378e-08
```

Notes: Based on the result above, I believe that this difference is statistically significant between these two candies (chocolate candy and fruity candy) because the p-value is significantly low as 2.871378e-08.

3. Overall Candy Rankings

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

There are two related functions that can help here, one is the classic `sort()` and `order()`.

```r
x <- c(5,10,1,4)
sort(x, decreasing=T)
```

```
[1] 10  5  4  1
```

```r
order(x)
```

```
[1] 3 4 1 2
```

```r
inds <- order(candy$winpercent)
head (candy[inds,], 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, and Jawbusters are the five least liked candy types in this set.

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

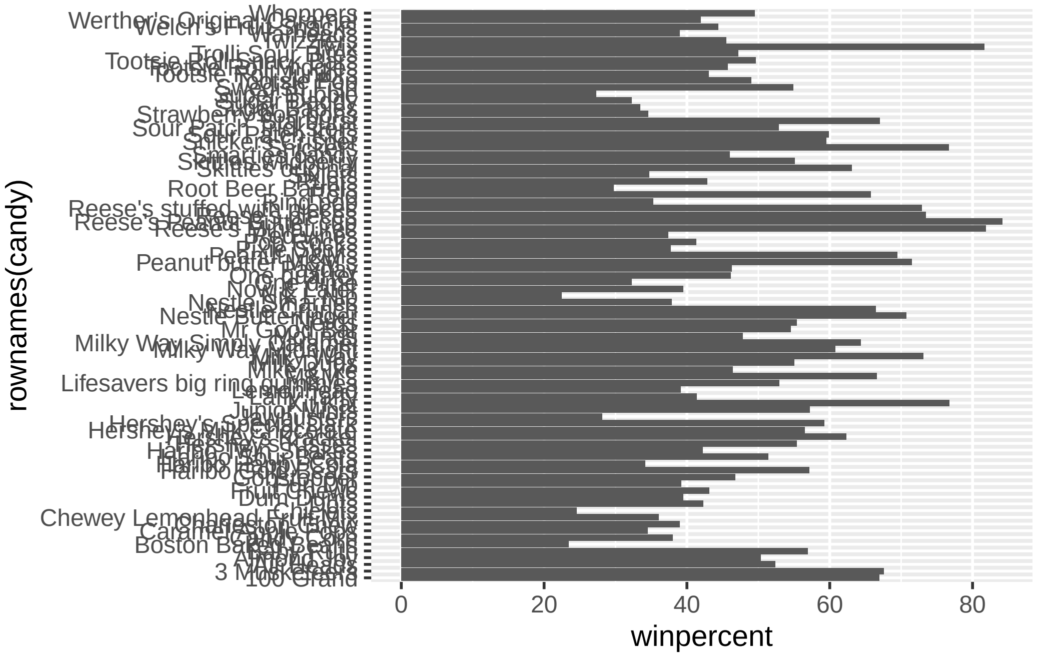
```
inds <- order(candy$winpercent, decreasing =T)
head(candy[inds,], 5)
```

```
                         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
```

Based on the result above, the top 5 all time favorite candy types out of this set are Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, and Snickers.

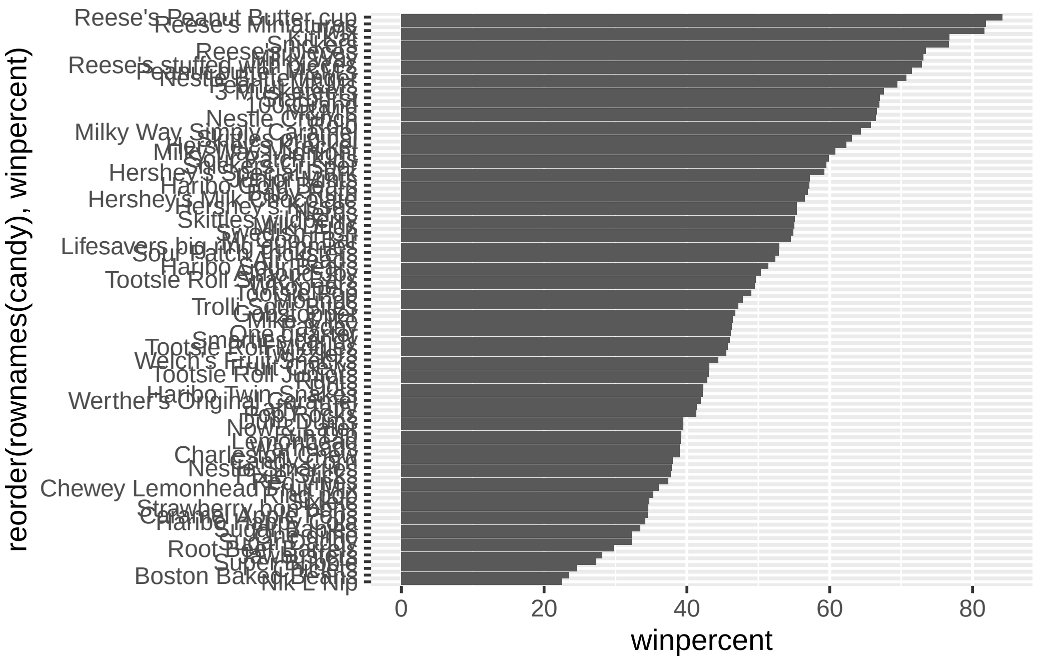
> Q15. Make a bar plot and order it by winpercent values.

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

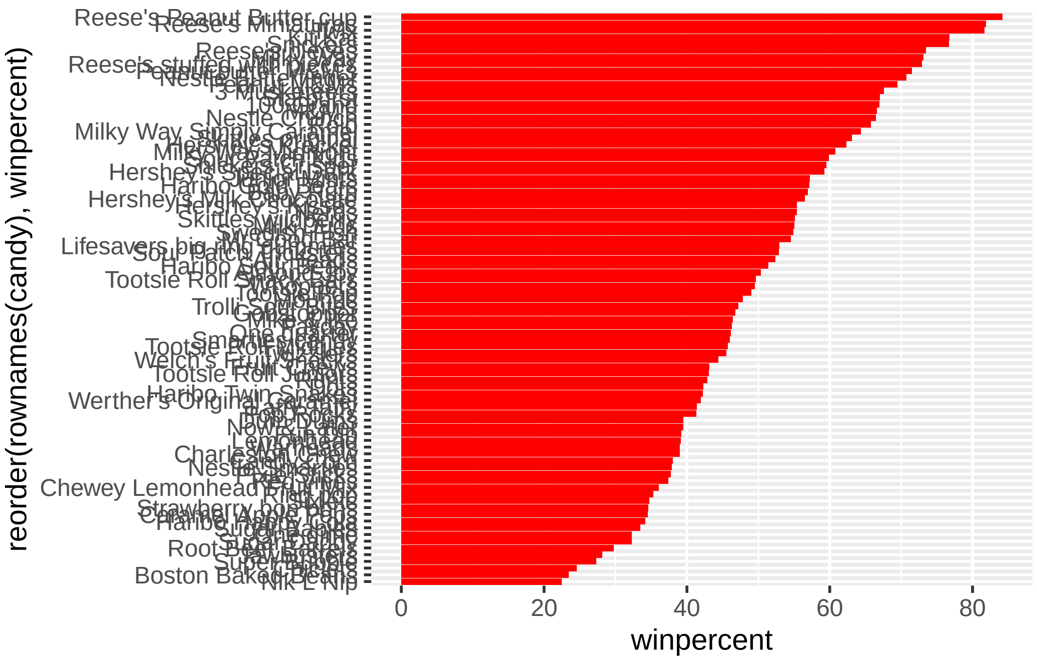


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

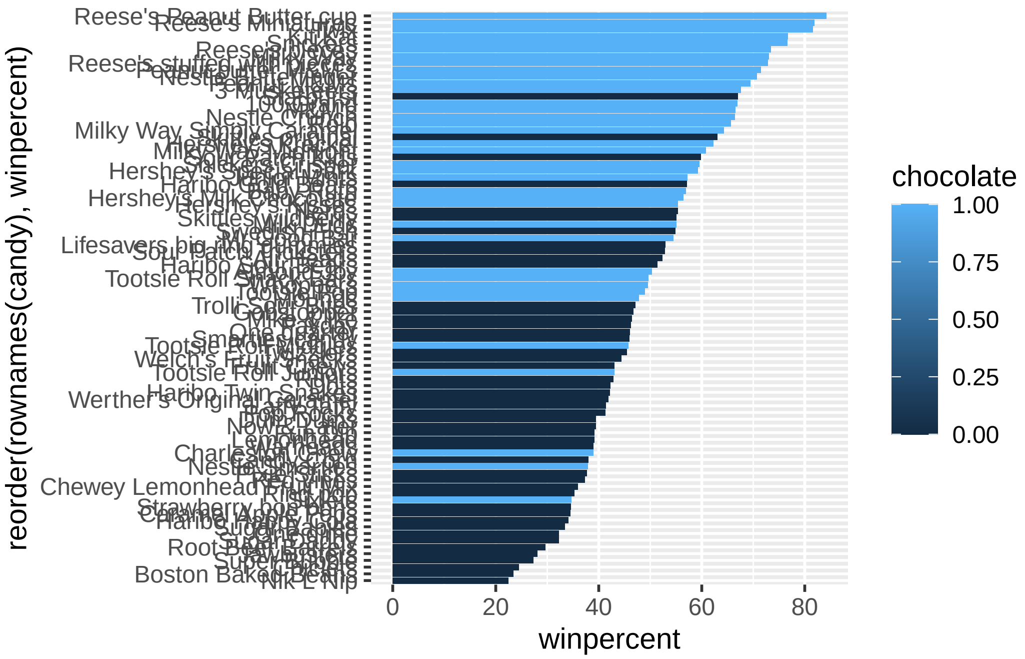
```
ggplot(candy) +
  aes(winpercent,reorder(rownames(candy), winpercent)) +
  geom_col()
```



```r
ggplot(candy) +
  aes(winpercent,reorder(rownames(candy), winpercent)) +
  geom_col(fill="red")
```



```
ggplot(candy) +
  aes(x= winpercent,
      y= reorder(rownames(candy), winpercent),
      fill=chocolate)+
  geom_col()
```

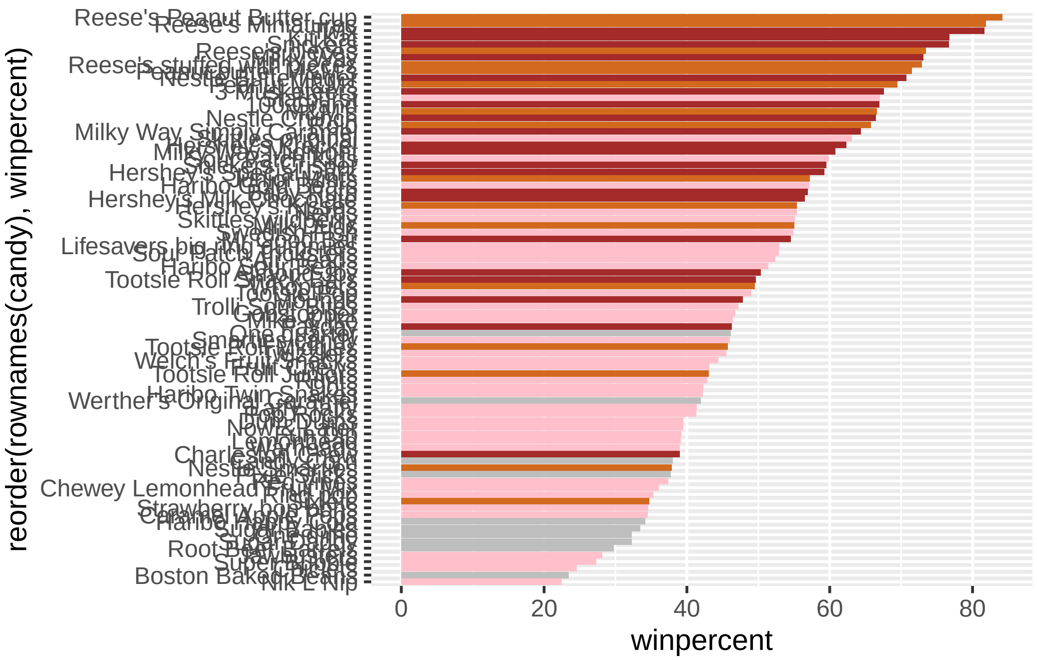


Here we want a custom color vector to color each bar the way we want - with `chocolate` and `fruity` candy together with whether it is a `bar` or not.

```
mycols <- rep ("gray", nrow(candy))
mycols[as.logical(candy$chocolate)] <- "chocolate"
mycols[as.logical(candy$fruity)] <- "pink"
mycols[as.logical(candy$bar)] <- "brown"

#mycols

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



```
ggsave("mybarplot.png", width=3, height=6)
```

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

> 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.

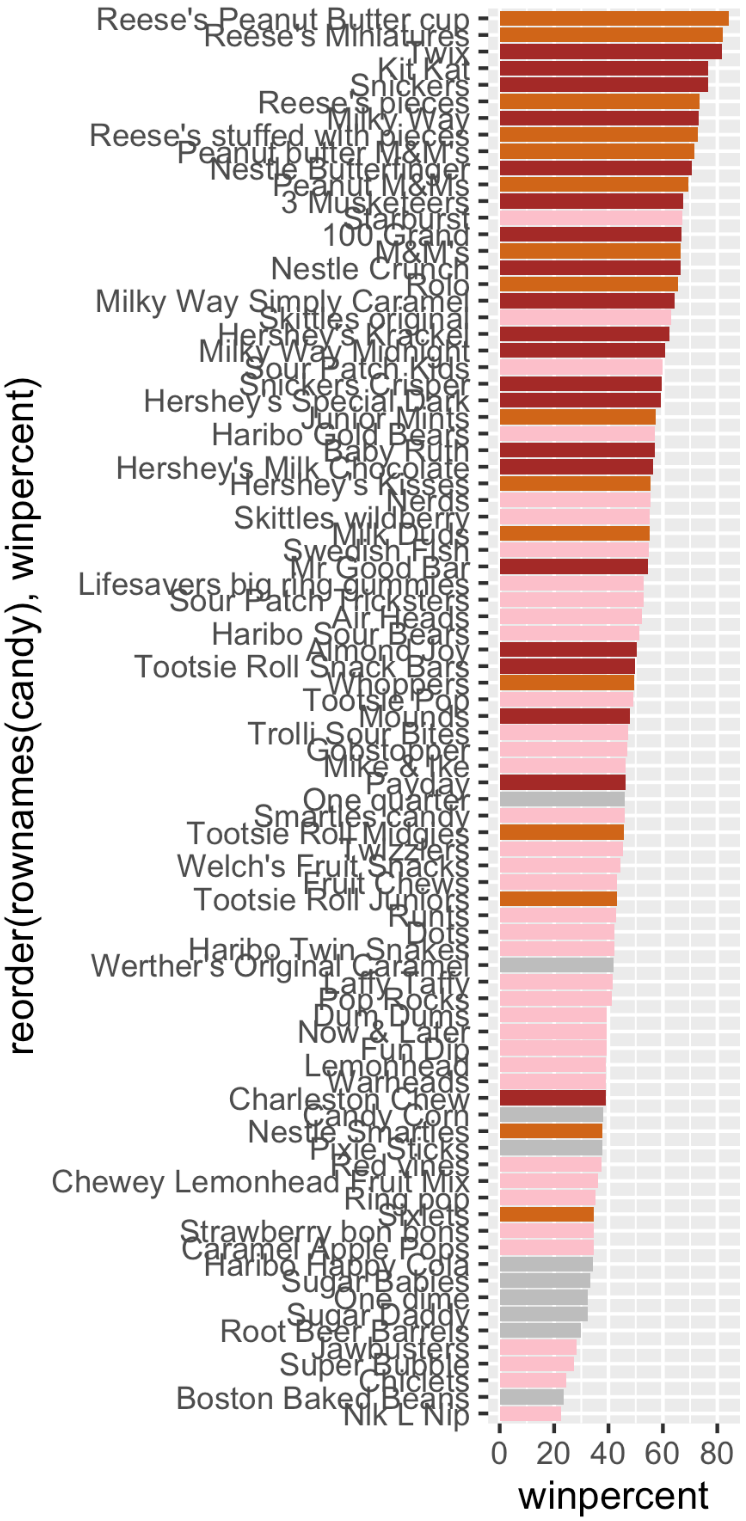


Figure 1: My silly barplot image

Winpercent vs. Pricepercent

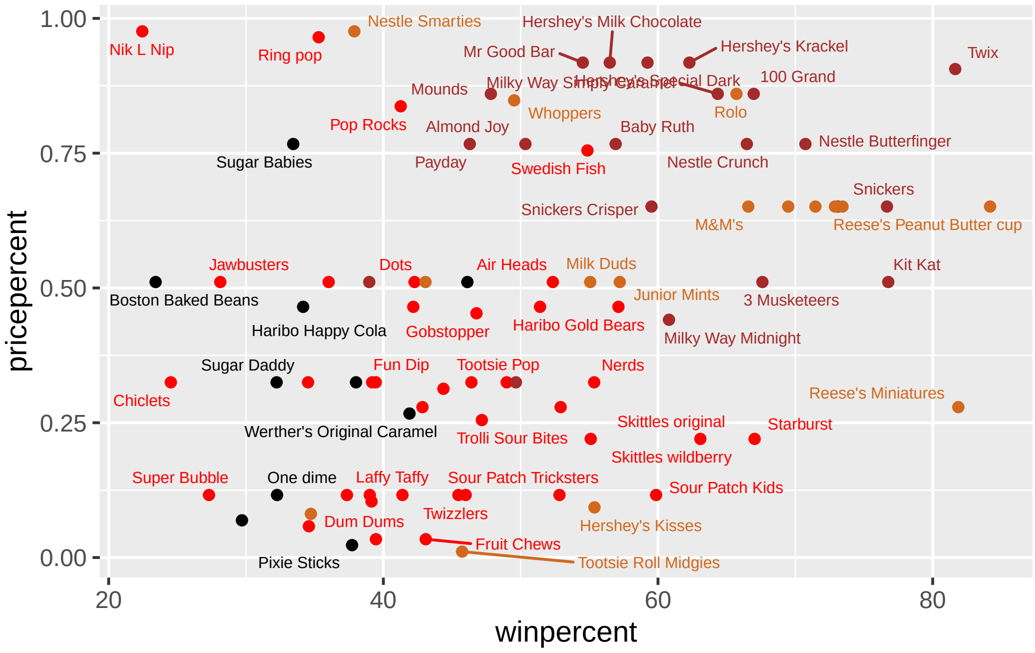
```
#Pink and grey is too light, let's change to red and black

mycols <- rep ("black", nrow(candy))
mycols[as.logical(candy$chocolate)] <- "chocolate"
mycols[as.logical(candy$fruity)] <- "red"
mycols[as.logical(candy$bar)] <- "brown"


library(ggrepel)

# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=mycols) +
  geom_text_repel(col=mycols, size=2, max.overlaps = 8)
```

```
Warning: ggrepel: 26 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



5. Correlation Structure

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

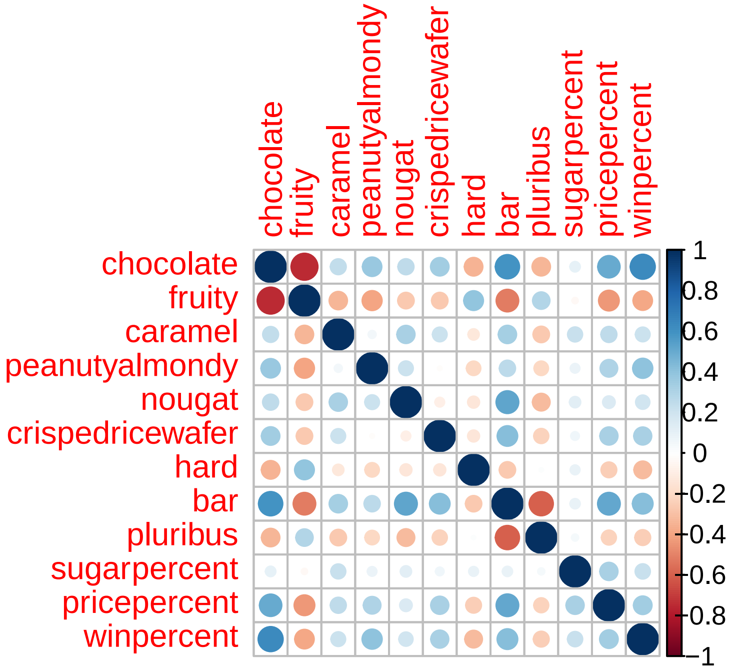
```
                 chocolate      fruity     caramel peanutyalmondy      nougat
chocolate        1.0000000 -0.74172106  0.24987535     0.37782357  0.25489183
fruity          -0.7417211  1.00000000 -0.33548538    -0.39928014 -0.26936712
caramel          0.2498753 -0.33548538  1.00000000     0.05935614  0.32849280
peanutyalmondy   0.3778236 -0.39928014  0.05935614     1.00000000  0.21311310
nougat           0.2548918 -0.26936712  0.32849280     0.21311310  1.00000000
crispedricewafer 0.3412098 -0.26936712  0.21311310    -0.01764631 -0.08974359
hard            -0.3441769  0.39067750 -0.12235513    -0.20555661 -0.13867505
bar              0.5974211 -0.51506558  0.33396002     0.26041960  0.52297636
pluribus        -0.3396752  0.29972522 -0.26958501    -0.20610932 -0.31033884
sugarpercent     0.1041691 -0.03439296  0.22193335     0.08788927  0.12308135
pricepercent     0.5046754 -0.43096853  0.25432709     0.30915323  0.15319643
winpercent       0.6365167 -0.38093814  0.21341630     0.40619220  0.19937530
                crispedricewafer        hard         bar    pluribus
chocolate             0.34120978 -0.34417691  0.59742114 -0.33967519
fruity               -0.26936712  0.39067750 -0.51506558  0.29972522
caramel               0.21311310 -0.12235513  0.33396002 -0.26958501
peanutyalmondy       -0.01764631 -0.20555661  0.26041960 -0.20610932
nougat               -0.08974359 -0.13867505  0.52297636 -0.31033884
crispedricewafer      1.00000000 -0.13867505  0.42375093 -0.22469338
hard                 -0.13867505  1.00000000 -0.26516504  0.01453172
bar                   0.42375093 -0.26516504  1.00000000 -0.59340892
pluribus             -0.22469338  0.01453172 -0.59340892  1.00000000
sugarpercent          0.06994969  0.09180975  0.09998516  0.04552282
pricepercent          0.32826539 -0.24436534  0.51840654 -0.22079363
winpercent            0.32467965 -0.31038158  0.42992933 -0.24744787
                sugarpercent pricepercent winpercent
chocolate         0.10416906    0.5046754  0.6365167
fruity           -0.03439296   -0.4309685 -0.3809381
caramel           0.22193335    0.2543271  0.2134163
peanutyalmondy    0.08788927    0.3091532  0.4061922
nougat            0.12308135    0.1531964  0.1993753
crispedricewafer  0.06994969    0.3282654  0.3246797
hard              0.09180975   -0.2443653 -0.3103816
bar               0.09998516    0.5184065  0.4299293
pluribus          0.04552282   -0.2207936 -0.2474479
```

```
sugarpercent      1.00000000   0.3297064  0.2291507
pricepercent      0.32970639   1.0000000  0.3453254
winpercent        0.22915066   0.3453254  1.0000000
```

```
library (corrplot)
```

```
corrplot 0.95 loaded
```

```
corrplot(cij)
```



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

Based on the plot above, I believe that chocolate and fruity candy are negatively correlated because their correlation value are the most negative: -0.74. It is also shown in the plot as biggest size of red circle representing as negative.

```
round(cij["chocolate", "fruity"], 2)
```

```
[1] -0.74
```

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

I believe that bar candy and chocolate candy are the most positively correlated.

Principal Component Analysis (PCA)

We need to be sure to scale our input `candy` data before PCA as we have the `winpercent` column on a different scale to all others in the dataset.

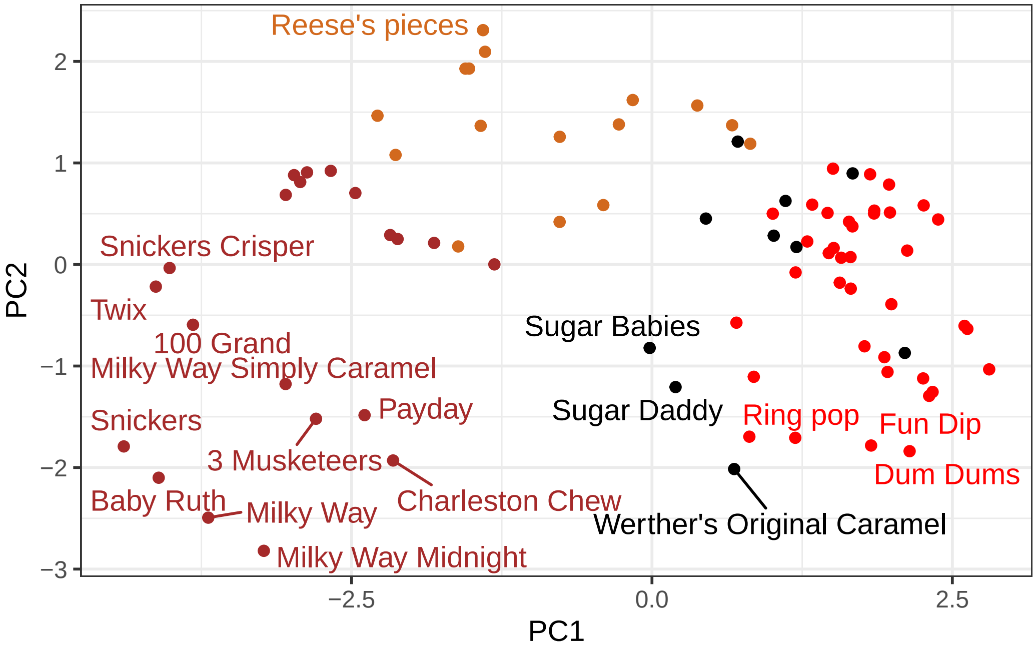
```
pca <- prcomp(candy, scale=T)
summary (pca)
```

```
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
```

First main result figure is my "PCA plot"

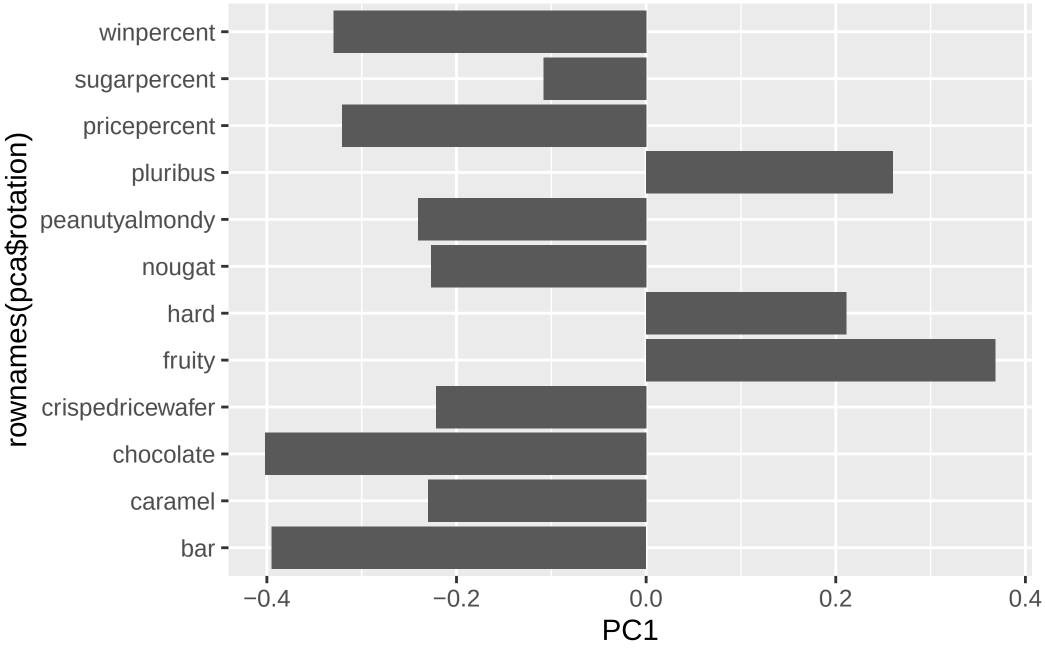
```
#pca$x
ggplot(pca$x) +
  aes (PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=mycols) +
  geom_text_repel(max.overlaps=6, col=mycols) +
  theme_bw()
```

```
Warning: ggrepel: 67 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

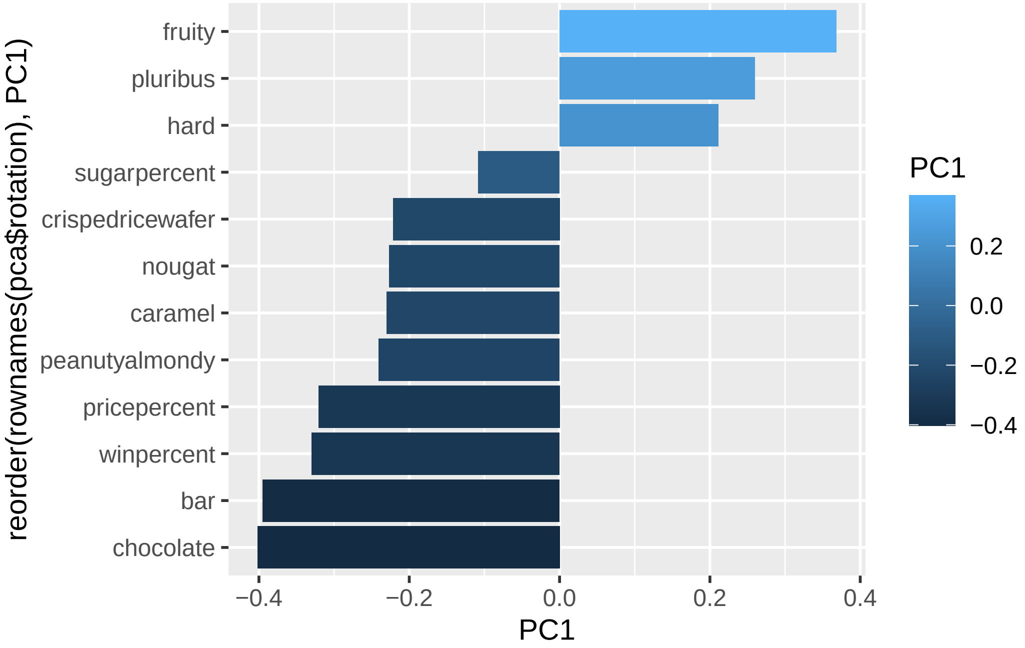


The second main PCA result is in the `pca$rotation` we can plot this to generate a so-called "loagings plot".

```
#pca$rotation
ggplot(pca$rotation) +
  aes(PC1, rownames(pca$rotation)) +
  geom_col()
```



```
ggplot(pca$rotation) +
  aes(PC1, reorder(rownames(pca$rotation), PC1), fill=PC1) +
  geom_col()
```



> Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

Based on the plots above, I believe that the fruity, pluribus meaning the candy coming in a bag or box of multiple candies, and hard candies are picked up strongly by PC1 in the positive direction. This result makes sense pretty well because fruity and chocolate are negatively correlated. The fruity candies are not likely to contain the chocolate, and similarly, the chocolate candies are not likely to include any fruity taste. Things related to chocolate is shown in the opposite direction of fruity/pluribus/hard candies' direction.