Class 10: Halloween Mini-Project

Derek Chang (PID: 16942232)

1. Importing Candy Data

```
candy_file <- read.csv("candy-data.csv", row.names = 1)
head(candy_file)
```

```
             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_file)
```

```
[1] 85
```

There are 85 different candy types

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

```
sum(candy_file$fruity)
```

```
[1] 38
```

There are 38 fruity candy types.

##2. What is your favorite candy?

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

```
[1] 81.64291
```

Q. Findy Fruity candy with a winpercent above 50

```
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_file %>% filter(fruity == 1) %>% filter(winpercent > 50)
```

	chocolate	fruity	caramel	peanutyalmondy	nougat
Air Heads	0	1	0	0	0
Haribo Gold Bears	0	1	0	0	0
Haribo Sour Bears	0	1	0	0	0
Lifesavers big ring gummies	0	1	0	0	0
Nerds	0	1	0	0	0
Skittles original	0	1	0	0	0
Skittles wildberry	0	1	0	0	0
Sour Patch Kids	0	1	0	0	0
Sour Patch Tricksters	0	1	0	0	0

```
Starburst                         0      1       0              0      0
Swedish Fish                      0      1       0              0      0
                          crispedricewafer hard bar pluribus sugarpercent
Air Heads                                 0    0   0        0        0.906
Haribo Gold Bears                         0    0   0        1        0.465
Haribo Sour Bears                         0    0   0        1        0.465
Lifesavers big ring gummies               0    0   0        0        0.267
Nerds                                     0    1   0        1        0.848
Skittles original                         0    0   0        1        0.941
Skittles wildberry                        0    0   0        1        0.941
Sour Patch Kids                           0    0   0        1        0.069
Sour Patch Tricksters                     0    0   0        1        0.069
Starburst                                 0    0   0        1        0.151
Swedish Fish                              0    0   0        1        0.604
                          pricepercent winpercent
Air Heads                        0.511   52.34146
Haribo Gold Bears                0.465   57.11974
Haribo Sour Bears                0.465   51.41243
Lifesavers big ring gummies      0.279   52.91139
Nerds                            0.325   55.35405
Skittles original                0.220   63.08514
Skittles wildberry               0.220   55.10370
Sour Patch Kids                  0.116   59.86400
Sour Patch Tricksters            0.116   52.82595
Starburst                        0.220   67.03763
Swedish Fish                     0.755   54.86111
```

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

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

```
[1] 76.7686
```

My favorite andy is Kit Kats and the winpercent vlaue is 76.7686.

Q4. What is the win percent value for Kit Kat?

The win percent value for kit kat is 76.7686.

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

The win percent for tootsie rolls is 49.6535.

To get a quick insight into a new dataset some folks like using the skimer package and its `skimr` function

```
library("skimr")
skimr::skim(candy_file)
```

Table 1: Data summary

Name	candy_file
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?

Looks like the `winpercent` variable or column looks to be on a different scale to the other columns in the dataset, as the other scales are out of 1, and this seems to be out of 100. I will need to scale my data before using PCA or other analysis.

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

```
candy_file$chocolate
```

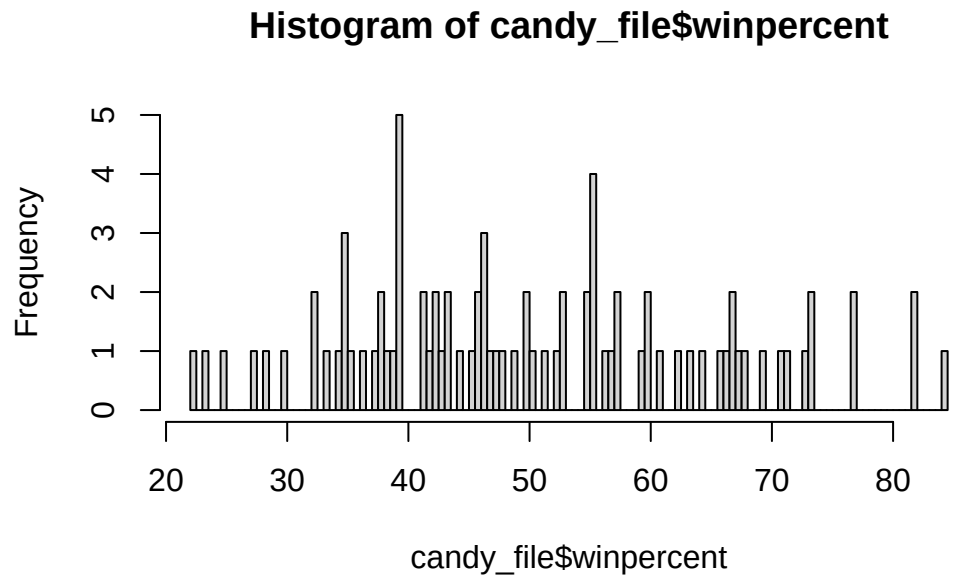
```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

A zero in the `candy$chocolate` column would indicate that the candy is not a chocolate, while a 1 would indicate that the candy is a chocolate.

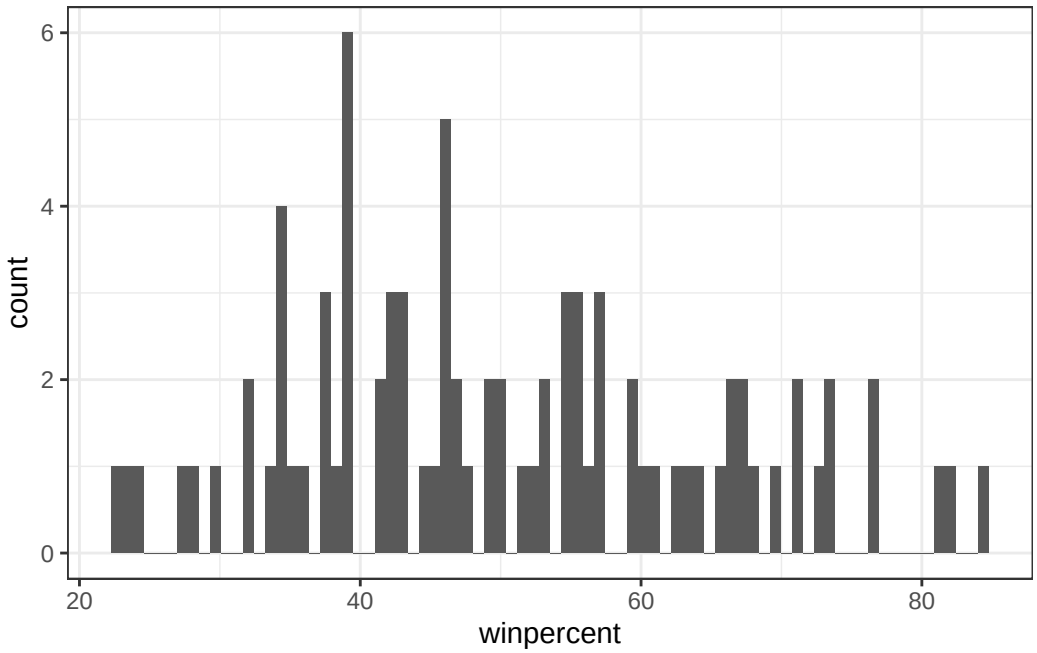
Q8. Plot a histogram of winpercent values

We can do this a few ways, e.g. the "base" R `hist()` function or with `ggplot()`

```
library(ggplot2)
hist(candy_file$winpercent, breaks = 100)
```



```
ggplot(candy_file, aes(winpercent)) +
  geom_histogram(bins = 80) +
  theme_bw()
```



Q9. Is the distribution of winpercent values symmetrical?

No, the distribution of winpercent values are not symmetrical, as the histogram is not a perfect bell shaped graph.

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

```
summary(candy_file$winpercent)
```

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

Since the median is below 50% at 47.83,the center of distribution is below. We are using the median since there are outliers.

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

```
fruity.candy <- candy_file %>% filter(fruity == 1)
summary(fruity.candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.45   39.04   42.97   44.12   52.11   67.04
```

```
chocolate.candy <- candy_file %>% filter(chocolate == 1)
summary(chocolate.candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  34.72   50.35   60.80   60.92   70.74   84.18
```

On average chocolate candy is higher ranked than fruity candy, with a higher median and mean. Chocolate seems to win more often.

Q12. Is this difference statistically significant?

```
t.test(chocolate.candy$winpercent, fruity.candy$winpercent)
```

```
    Welch Two Sample t-test

data:  chocolate.candy$winpercent and fruity.candy$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
```

With a very small p-value of 2.87e-08, the difference between chocolate and fruity is statistically significant. Chocolate is statistically better than fruit.

3. Overall Candy Rankings

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

Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters.

```
head(candy_file[order( candy_file$winpercent),], 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
```

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

```
play <- c("d","a","c")
sort(play)
```

```
[1] "a" "c" "d"
```

```
order(play)
```

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

```
play[ order(play)]
```

```
[1] "a" "c" "d"
```

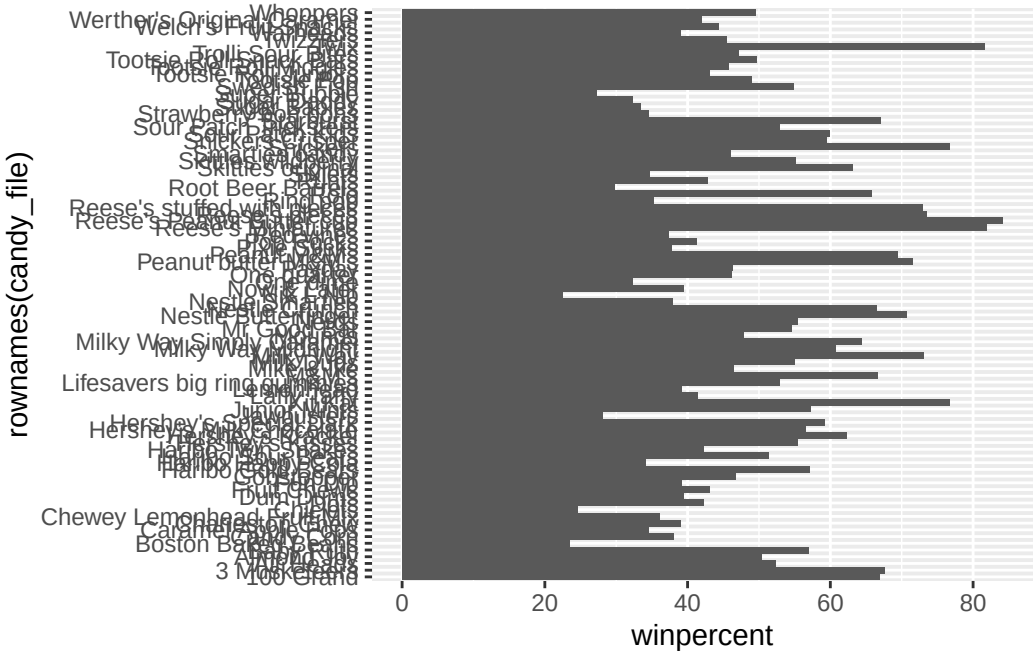
```
tail(candy_file[order( candy_file$winpercent),], 5)
```

```
                           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
```

Snickers, Kit Kat, Twix, Reese's Miniatures, Reese's Peanut Butter cup. are the 5 most popular.

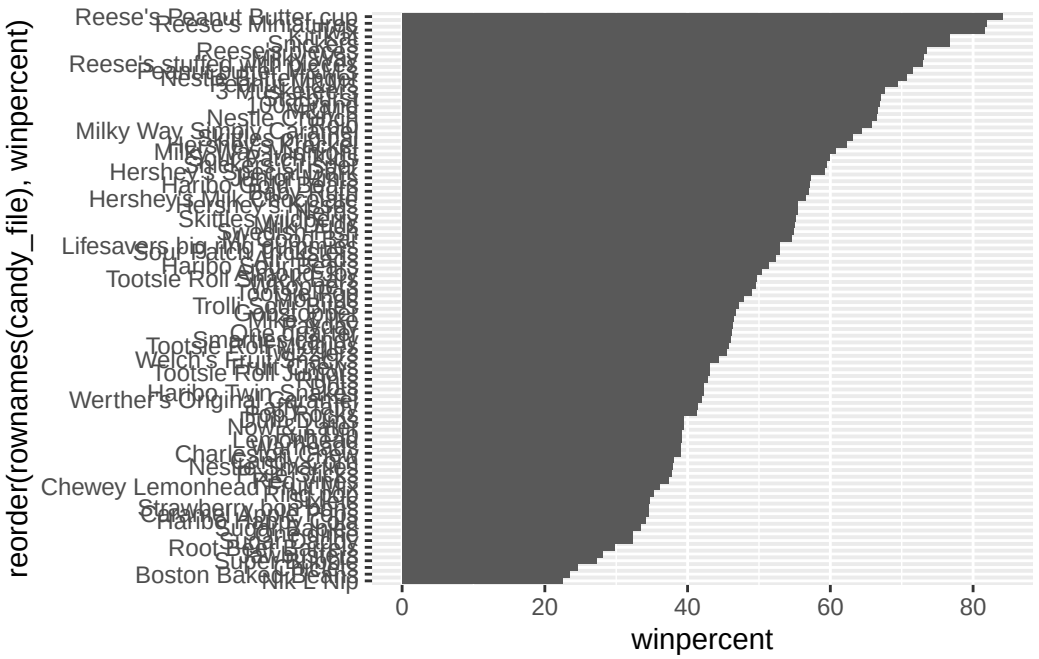
Q15. Make a first barplot of candy ranking based on winpercent values.

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



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

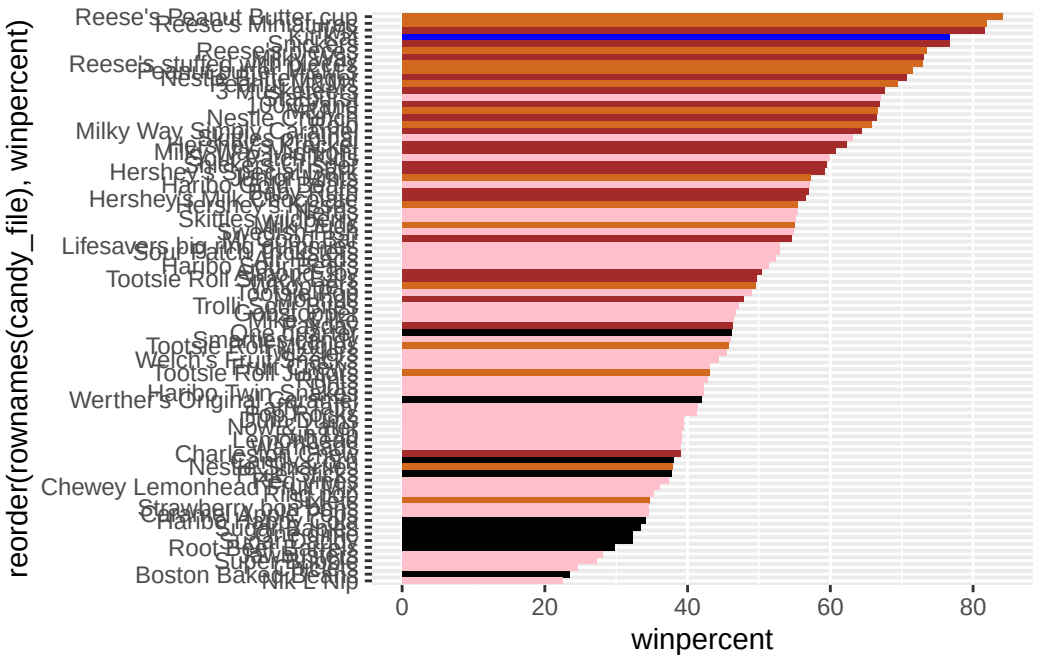
```
ggplot(candy_file, aes(winpercent, reorder(rownames(candy_file),winpercent))) +
  geom_col()
```



I want a more custom color scheme where I can see different categories of candy on the same graph. To do this we can roll our own color vector.

```
# Place Holder Color Vector
mycols <- rep("black", nrow(candy_file))
mycols[as.logical(candy_file$chocolate)] <- "chocolate"
mycols[as.logical(candy_file$bar)] <- "brown"
mycols[as.logical(candy_file$fruity)] <- "pink"
mycols[(row.names(candy_file) == "Kit Kat")] <- "blue"
```

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



Q17. What is the worst ranked chocolate candy?

Sixlet is the worst ranked chocolate candy

Q18. What is the best ranked fruity candy?

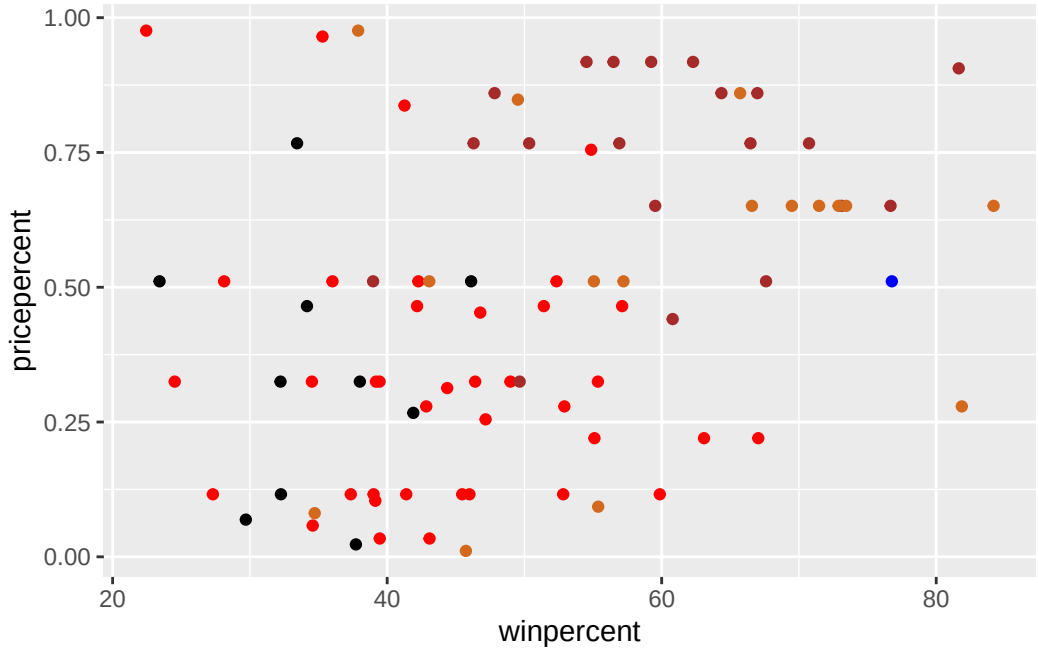
Starbursts are the best ranked fruity candy.

4. Taking a look at price percent

Plot of winpercent vs price percent to see what the best candy to buy is.

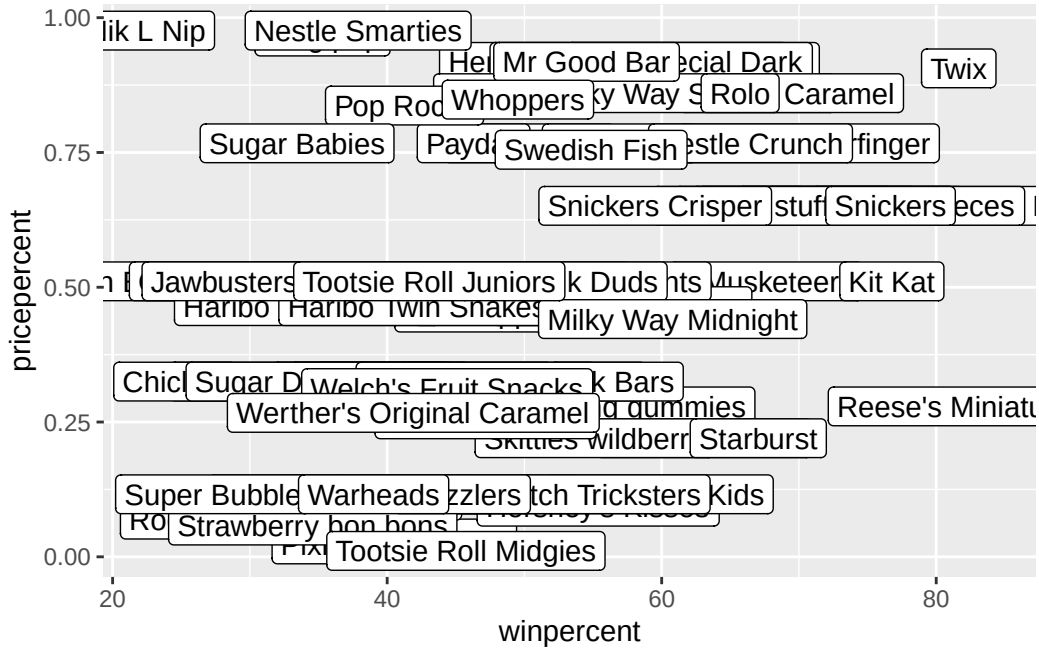
```
mycols[as.logical(candy_file$fruity)] <- "red"
```

```
ggplot(candy_file) + aes(winpercent, pricepercent) +
  geom_point(col = mycols)
```



Adding Labels

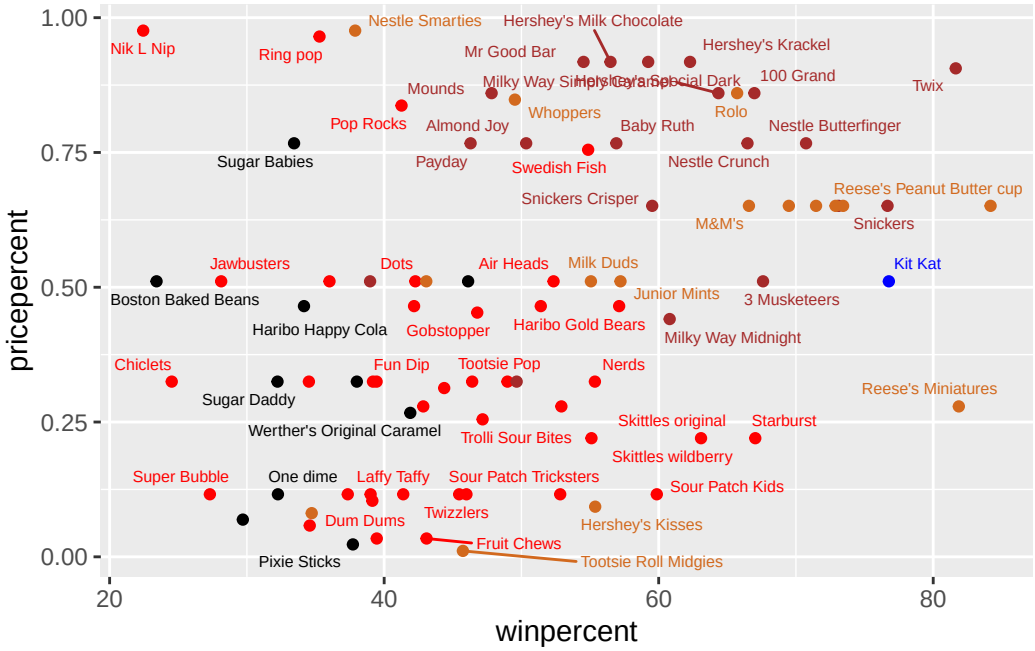
```
library(ggrepel)
ggplot(candy_file) + aes(winpercent, pricepercent, label = rownames(candy_file)) +
  geom_point(col = mycols) +
  geom_label()
```



Make the labels non-overlapping

```
library(ggrepel)
ggplot(candy_file) + aes(winpercent, pricepercent, label = rownames(candy_file)) +
  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
```



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?

The Reese's Miniatures are the highest ranked in terms of winpercent for the least money.

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

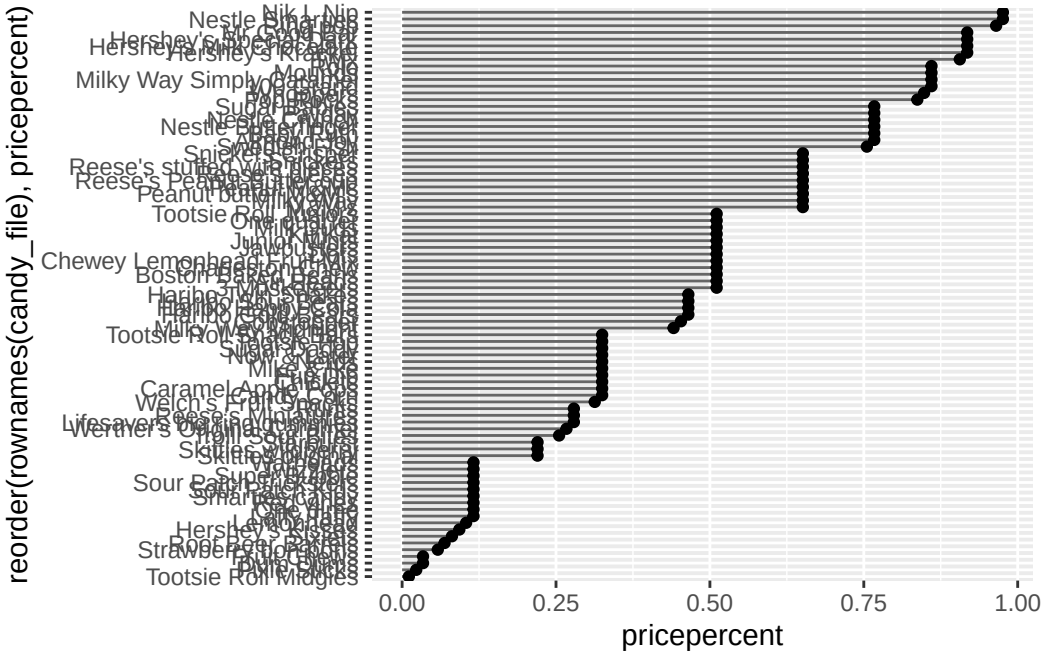
```
ord <- order(candy_file$pricepercent, decreasing = TRUE)
head( candy_file[ord,c(11,12)], n=5 )
```

```
                      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 are Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate. Nik L Nip is 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_file) +
  aes(pricepercent, reorder(rownames(candy_file), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy_file), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```

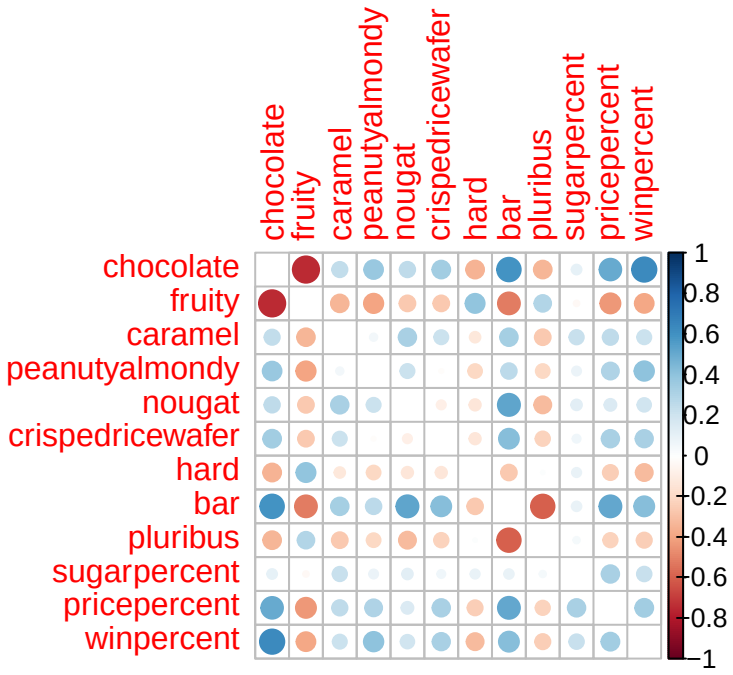


5. Exploring the correlation structure.

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
cij <- cor(candy_file)
corrplot(cij, diag = F)
```



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

Fruity and chocolate are two strong anti-correlated variables

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

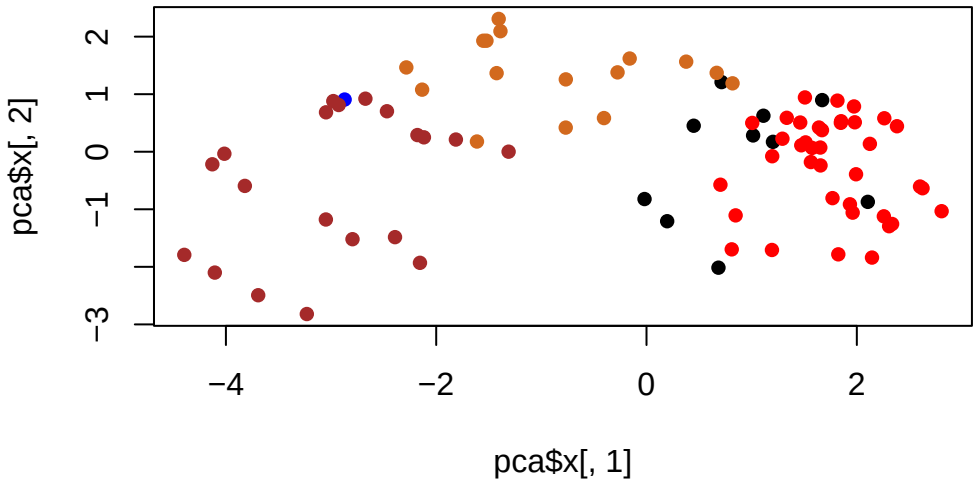
Win Percent and Chocolate are the most positively correlated variables.

#Principal COmponent Analysis

```
pca <-  prcomp(candy_file, 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
```

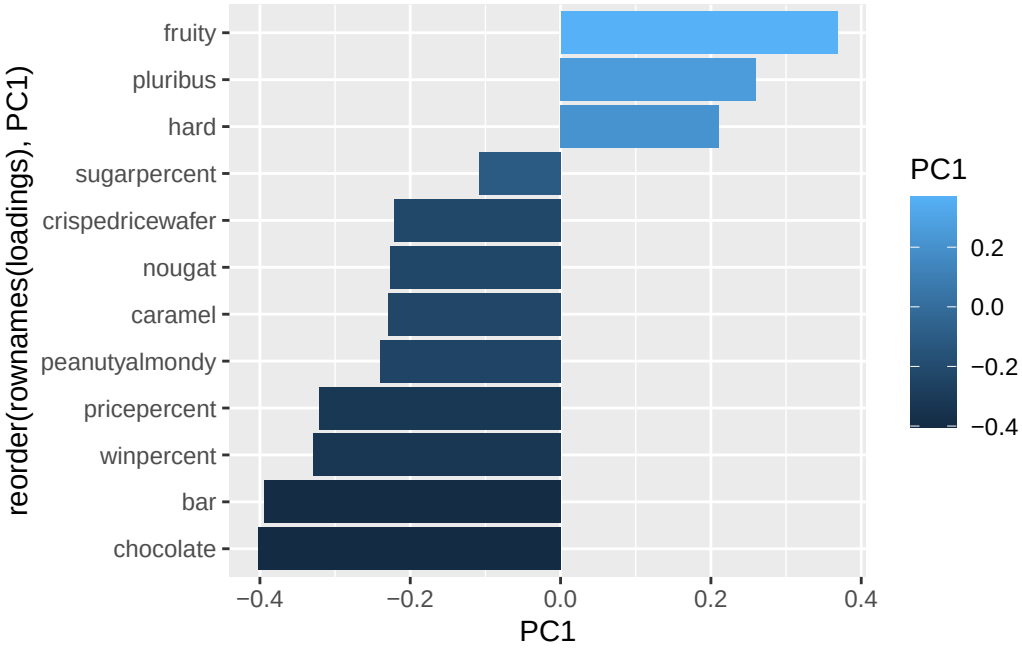
```
plot(pca$x[,1], pca$x[,2], col = mycols, pch = 16)
```



How do the original variables (columns) contribute to the new PCs. I will look at PC1 here

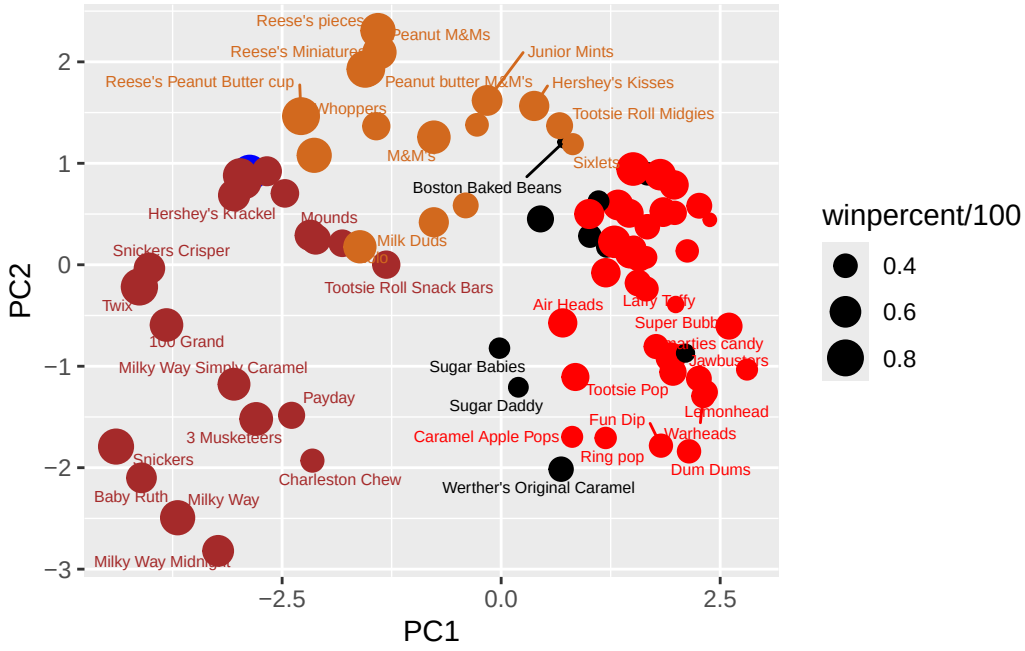
```
loadings <- as.data.frame(pca$rotation)

ggplot(loadings, aes(PC1, reorder(rownames(loadings), PC1), fill = PC1)) + geom_col()
```




```
my_data <- cbind(candy_file, pca$x[,1:3])
ggplot(my_data, aes(PC1, PC2, size = winpercent/100, text = rownames(my_data), label = rownar
  geom_text_repel(size = 2, col = mycols, max.overlaps = 8)
```

```
Warning: ggrepel: 42 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



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

The original variables of fruity, hard, and pluribus, It makes sense because these variables were positively correlated together compared to the chocolate, which were correlated together.