Halloween_Candy

AUTHOR
Qianqian Tao

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

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

```r
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

```r
#using table
table(candy$fruity)
```

```

 0  1 
47 38 
```

> Q2.5 What are these fruity candy?

```r
indices <- which(candy$fruity==1)

print_indices <- function(indices, dataset){
  for(i in indices){
    return(dataset[indices,])
  }
}
#print_indices(indices, candy)

#Another easier way to do this
rownames(candy[candy$fruity==1,])
```

```
 [1] "Air Heads"                   "Caramel Apple Pops"         
 [3] "Chewey Lemonhead Fruit Mix"  "Chiclets"                   
 [5] "Dots"                        "Dum Dums"                   
 [7] "Fruit Chews"                "Fun Dip"
 [9] "Gobstopper"                 "Haribo Gold Bears"
[11] "Haribo Sour Bears"          "Haribo Twin Snakes"
[13] "Jawbusters"                 "Laffy Taffy"
[15] "Lemonhead"                  "Lifesavers big ring gummies"
[17] "Mike & Ike"                 "Nerds"
[19] "Nik L Nip"                  "Now & Later"
[21] "Pop Rocks"                  "Red vines"
[23] "Ring pop"                   "Runts"
[25] "Skittles original"          "Skittles wildberry"
[27] "Smarties candy"             "Sour Patch Kids"
[29] "Sour Patch Tricksters"      "Starburst"
[31] "Strawberry bon bons"        "Super Bubble"
[33] "Swedish Fish"               "Tootsie Pop"
[35] "Trolli Sour Bites"          "Twizzlers"
[37] "Warheads"                   "Welch's Fruit Snacks"
```

How often does my favorite candy win

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

```
candy["Hershey's Krackel", ]$winpercent
```

```
[1] 62.28448
```

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

```
[1] 66.57458
```

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

```
[1] 81.64291
```

> 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”?

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

```
[1] 49.6535
```

```
library("skimr")
#skimr::skim(candy)--> using only one function from the package without loading the whole
skim(candy)
```

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?

Yeap, the `winpercent` column is on a 0:100 scale and all other appear to be 0: 1 scale.

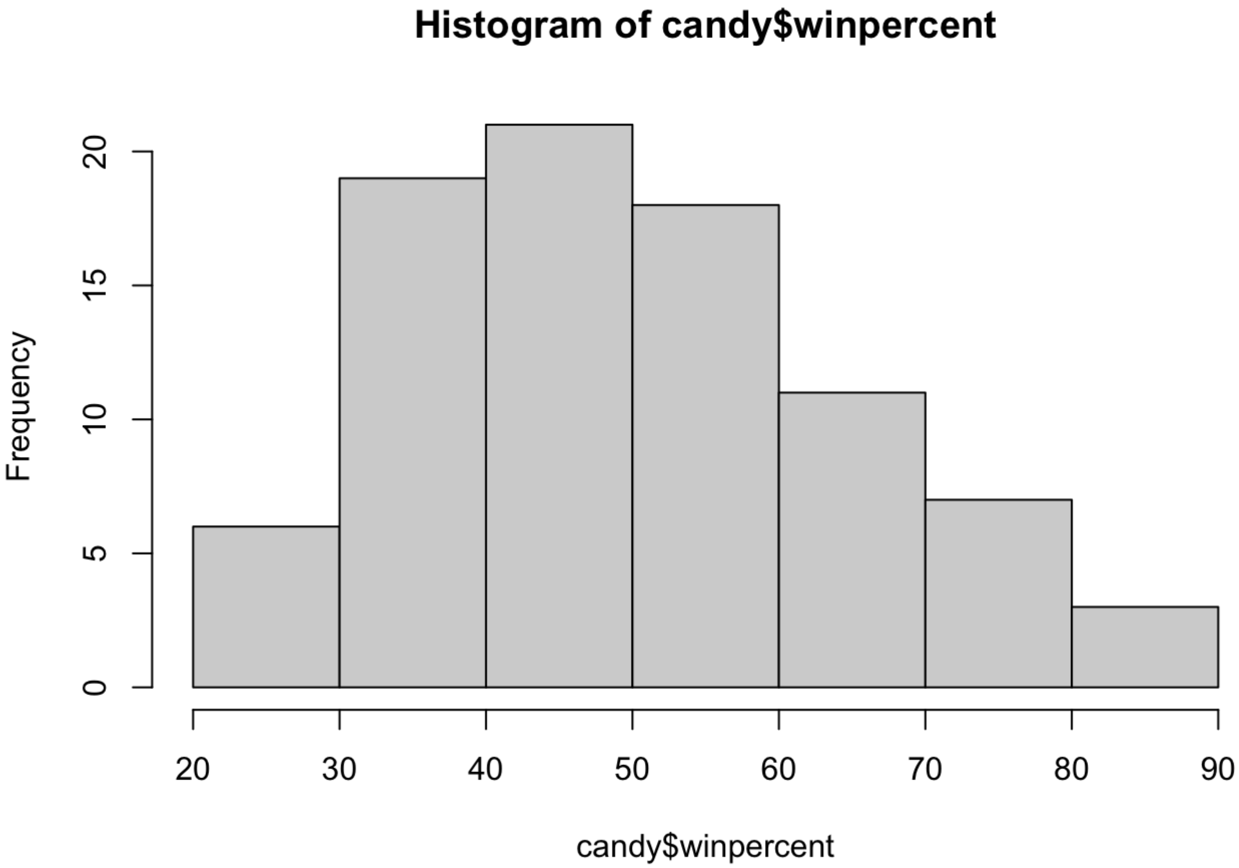
> Q7. What do you think a zero and one represent for the candy$chocolate column?

A 0 means this candy is not classified as containing chocolate and 1 means this candy is classified as containing chocolate.

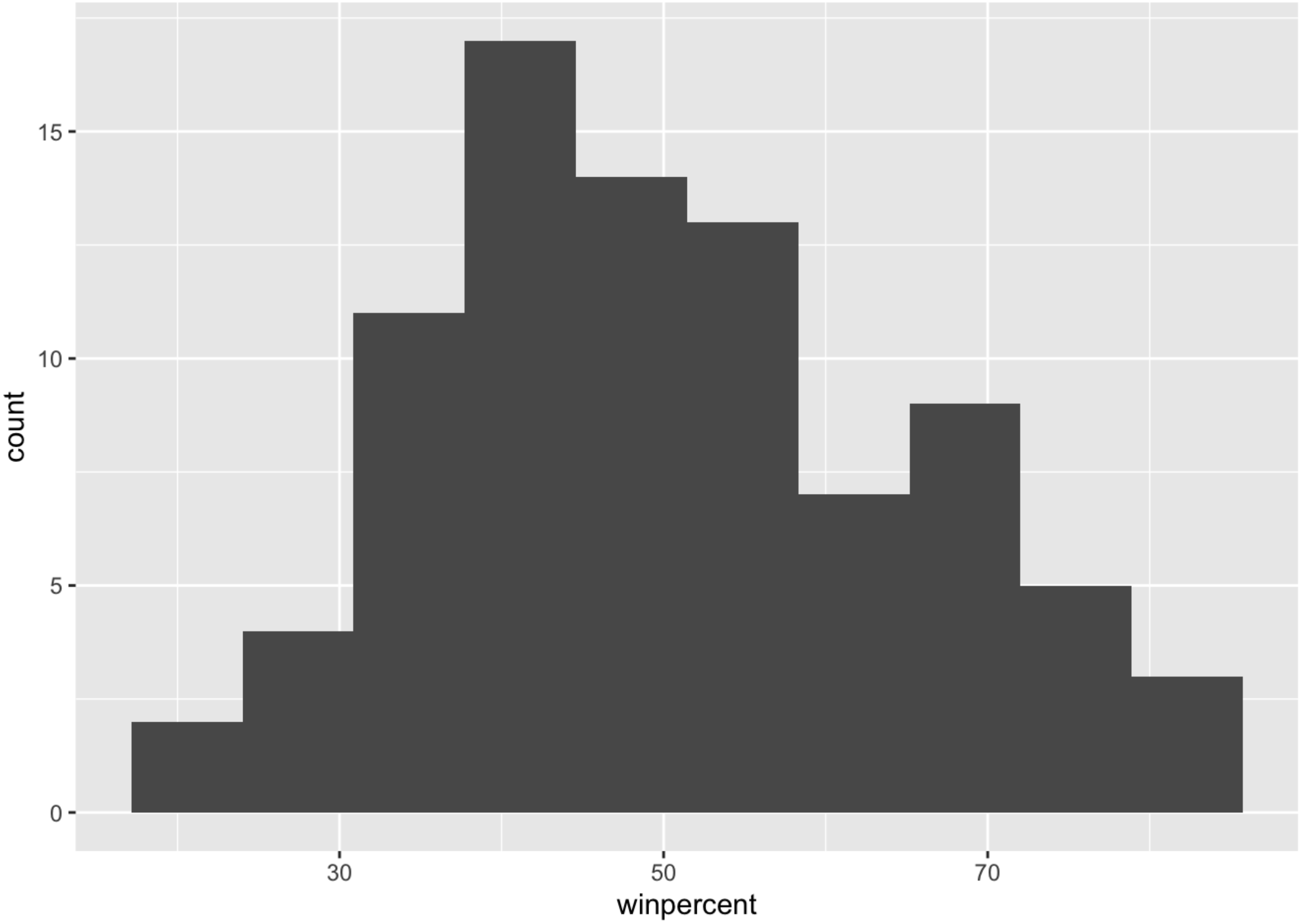
> Q8. Plot a histogram of winpercent values

```
#one way to make a histogram in base R graphics:
hist(candy$winpercent)

#use ggplot
library(ggplot2)
```



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



> Q9. Is the distribution of winpercent values symmetrical?

No. It is left skewed

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

It is below 50% with a mean:

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

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

To answer this question I will need to :

-“subset” (a.k.a “select”, “filter”) the candy dataset to just chocolate candy, - get their winpercent values, -calculate the mean of these Then do the same for fruity candy and compate.

```
col_function <- function(dataset, col1, variable){
  mean(dataset[,col1][dataset[,variable]==1])
}
col_function(candy, "winpercent", "chocolate")
```

```
[1] 60.92153
```

```
chocolate_candy <- candy[candy$chocolate==1,]
choco_mean_win <- mean(chocolate_candy$winpercent)


#Professor's method
#Filter/select/subset to just chocolate rows
chocolate.candy <- candy[as.logical(candy$chocolate),]
#Get their winpercent values
chocolate.winpercent <- chocolate.candy$winpercent
#Calculate the mean value
mean(chocolate.winpercent)
```

```
[1] 60.92153
```

```
#Then do the same thing for fruit
#Filter/select/subset to just chocolate rows
fruity.candy <- candy[as.logical(candy$fruity),]
#Get their winpercent values
fruity.winpercent <- fruity.candy$winpercent
#Calculate the mean value
mean(fruity.winpercent)
```

```
[1] 44.11974
```

Chocolate wins!

> 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 
```

```r
t.test(chocolate.candy, fruity.candy)
```

```
	Welch Two Sample t-test

data:  chocolate.candy and fruity.candy
t = 1.4907, df = 808.56, p-value = 0.1364
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4733867  3.4619260
sample estimates:
mean of x mean of y 
  5.41088   3.91661 
```

super low p-value–> there is significant difference between chocolate and fruit. >Q13. What are the five least liked candy types in this set?

```r
x <- c(5,2,10)
#use sort
sort(x)
```

```
[1]  2  5 10
```

```r
#use order
order(x)
```

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

```r
x[order(x)]
```

```
[1]  2  5 10
```

```r
#I can order by winpercent
ord <- order(candy$winpercent)
head(candy[ord,],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
```

```r
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
```

```r
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
```

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

```r
#usual way
ord_2 <- order(candy$winpercent, decreasing=TRUE)
head(candy[ord_2,],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
```

```r
#one easy one
tail(candy[ord,],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
```

```r
col_function <- function(dataset, col1, variable){
  mean(dataset[,col1][dataset[,variable]==1])
}
col_function(candy, "winpercent", "chocolate")
```

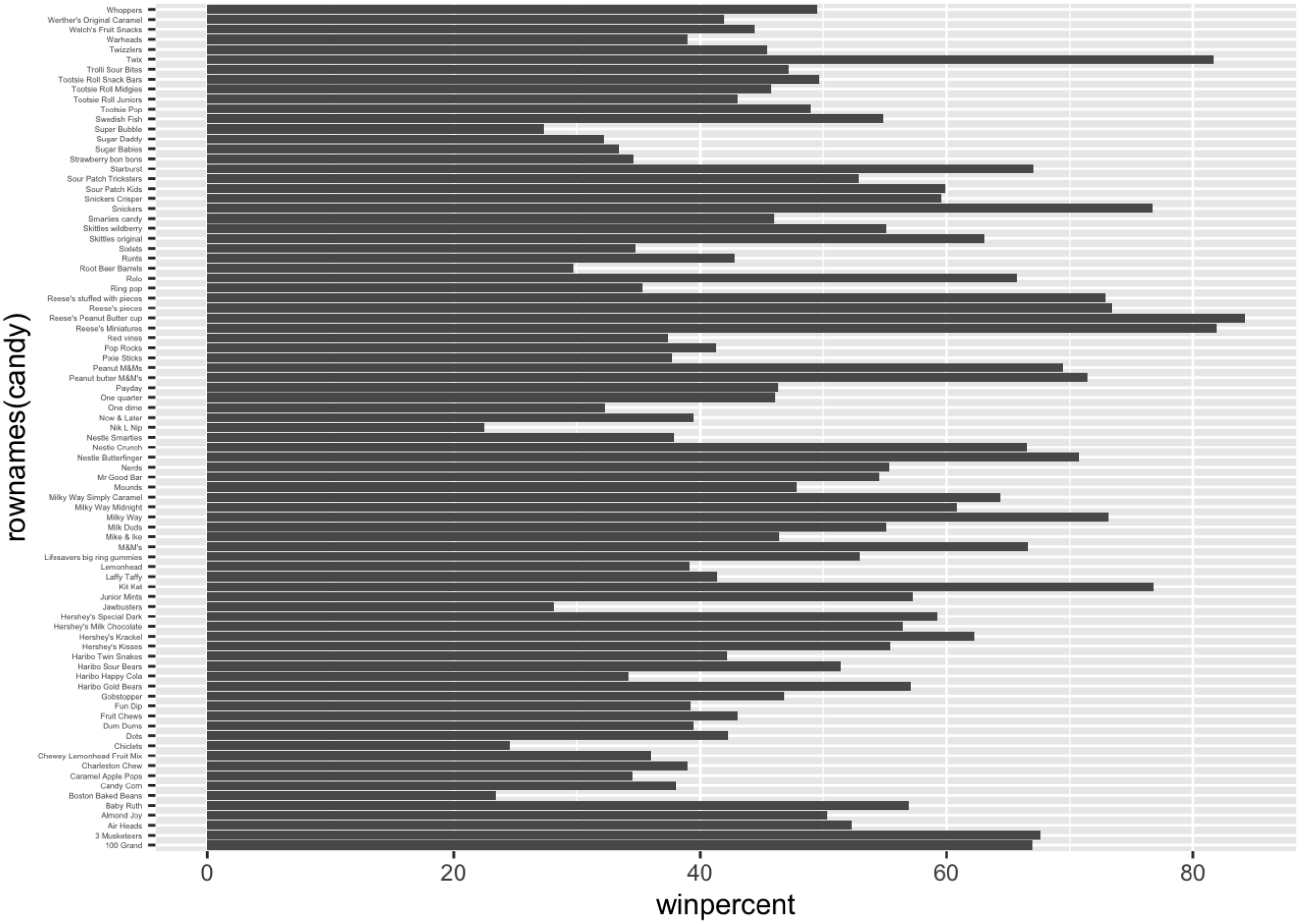
```
[1] 60.92153
```

```r
candy[,"winpercent"][candy[,"chocolate"]==1]
```

```
 [1] 66.97173 67.60294 50.34755 56.91455 38.97504 55.37545 62.28448 56.49050
 [9] 59.23612 57.21925 76.76860 71.46505 66.57458 55.06407 73.09956 60.80070
[17] 64.35334 47.82975 54.52645 70.73564 66.47068 69.48379 81.86626 84.18029
[25] 73.43499 72.88790 65.71629 34.72200 37.88719 76.67378 59.52925 48.98265
[33] 43.06890 45.73675 49.65350 81.64291 49.52411
```

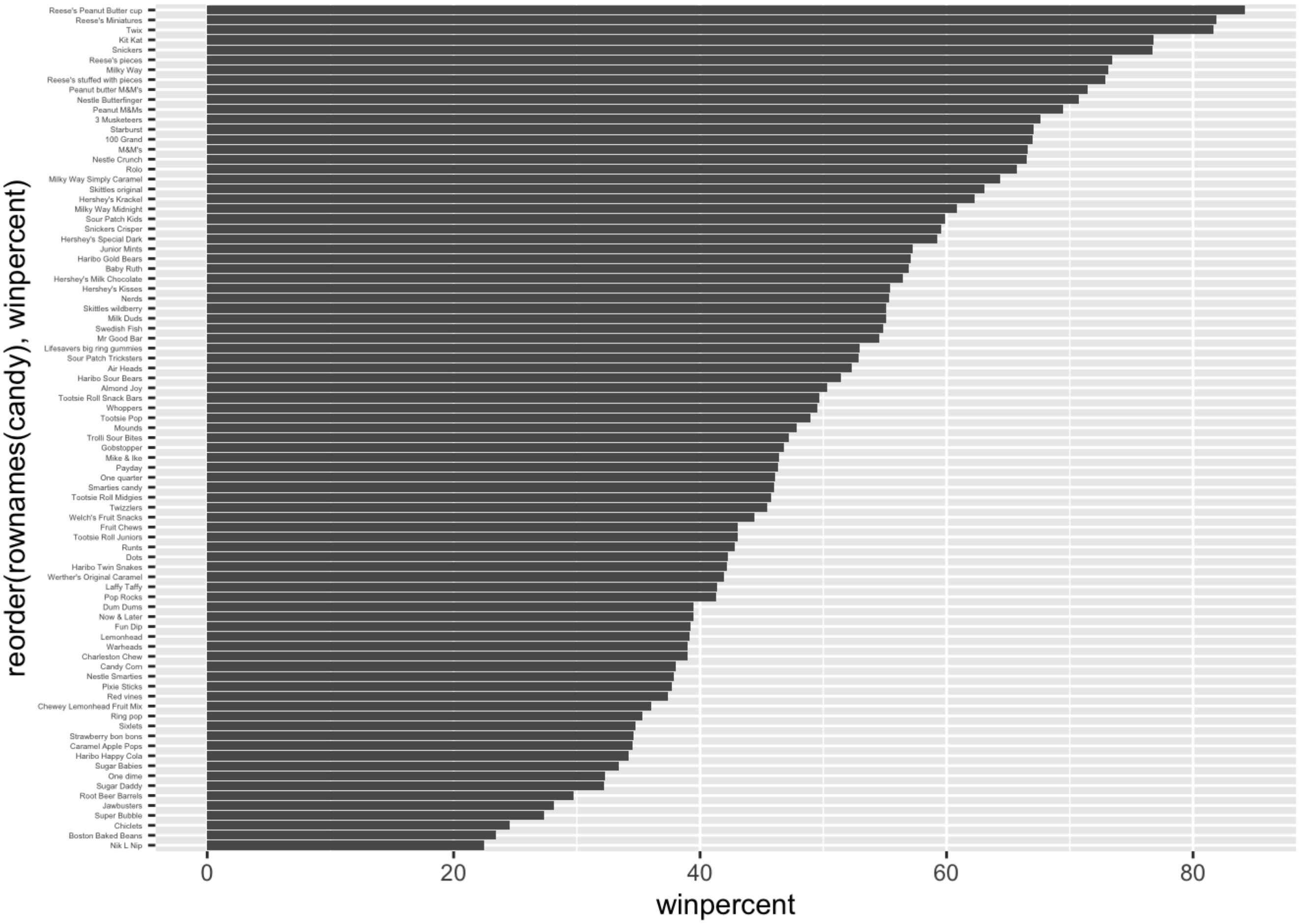
> Q15. Make a first barplot of candy ranking based on winpercent values.

```
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col()+
  theme(axis.text.y = element_text(
          size = 3))
```

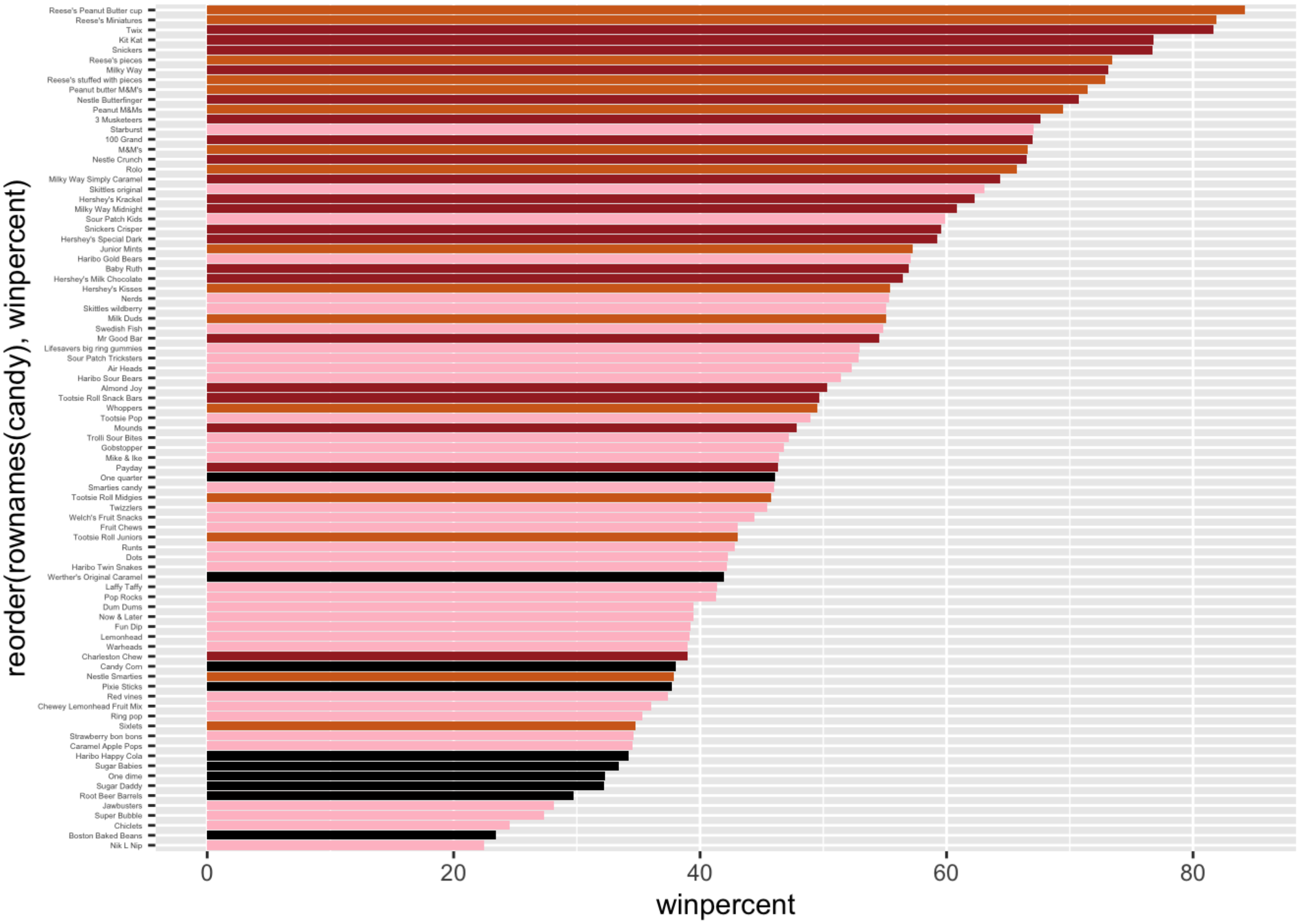


> Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent))+
  geom_col()+
  theme(axis.text.y = element_text(
          size = 3))
```



```
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"
# set color to my_cols only make the frame to be that color; set fill to my_cols make the
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols) +
  theme(axis.text.y = element_text(
          size = 3))
```



> Q17. What is the worst ranked chocolate candy?

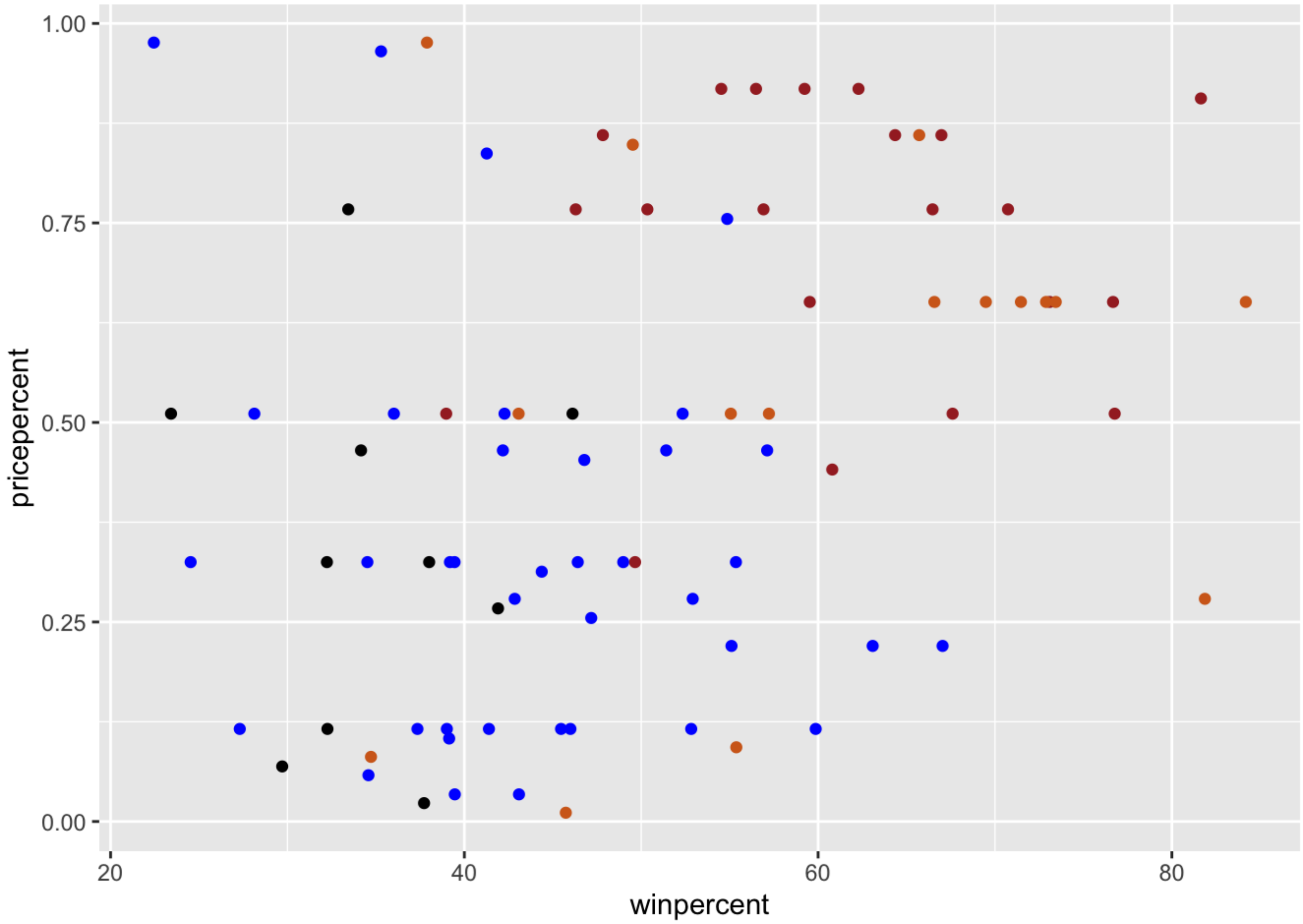
Sixlets

> Q18. What is the best ranked fruity candy?

Starburst

> Q. What is the best candy

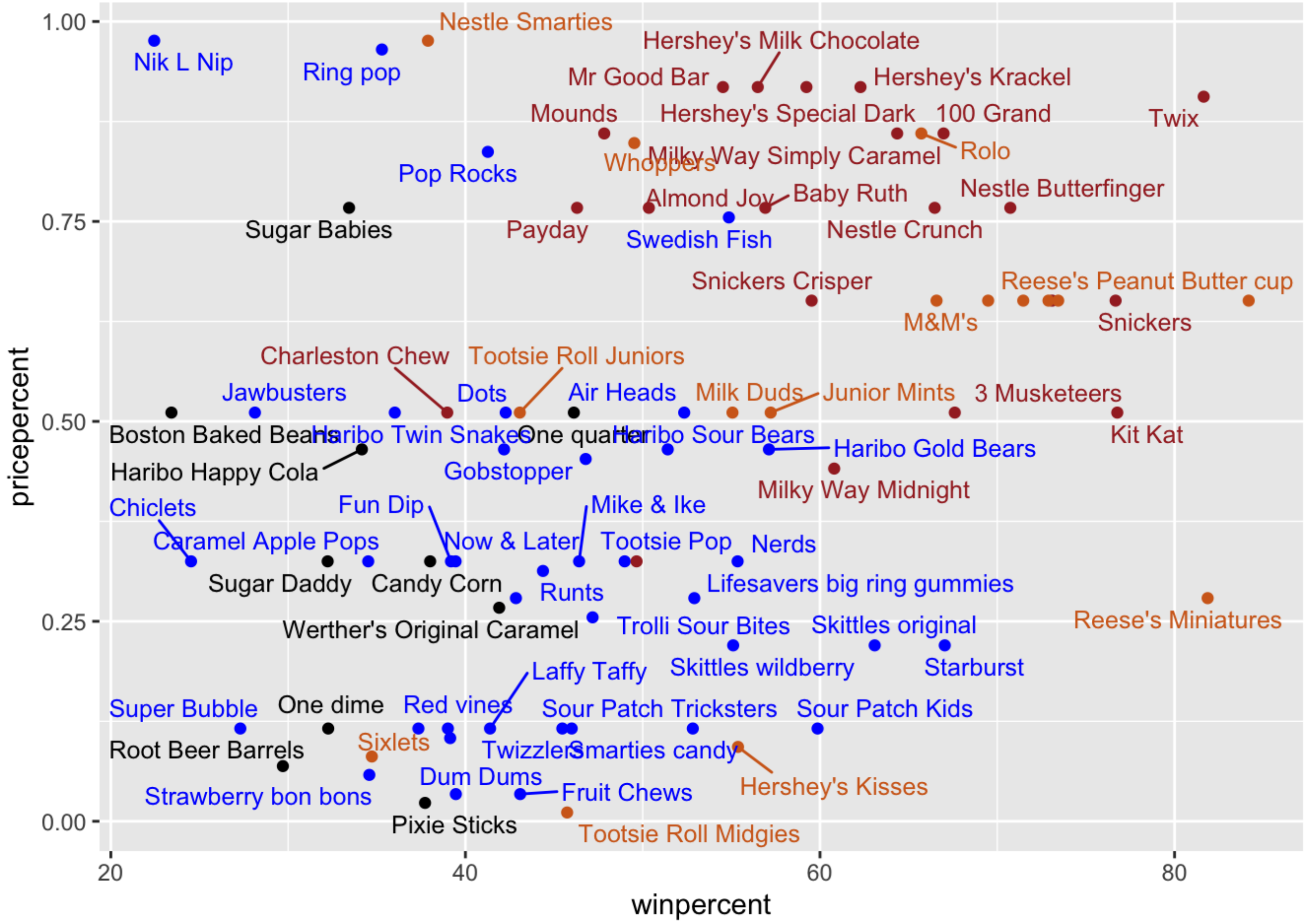
```
library(ggrepel)
my_cols[as.logical(candy$fruity)] = "blue"
ggplot(candy, aes(winpercent, pricepercent))+
  geom_point(col=my_cols)
```



Add some labels

```
ggplot(candy, aes(winpercent, pricepercent, label=rownames(candy)))+
  geom_point(col=my_cols)+
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 10)
```

```
Warning: ggrepel: 10 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?

Reese’s Miniature or Reese’s Peanut Butter cup (chocolate)

```
order_win <- order(candy$winpercent)
tail(candy[order_win,])
```

```
                          chocolate fruity caramel peanutyalmondy nougat
Reese's pieces                    1      0       0              1      0
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
Reese's pieces                           0    0   0        1        0.406
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
```

```
Reese's pieces                  0.651   73.43499
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
```

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

```
order_price <- order(candy$pricepercent)
tail(candy[order_price,])
```

```
                         chocolate fruity caramel peanutyalmondy nougat
Hershey's Milk Chocolate         1      0       0              0      0
Hershey's Special Dark           1      0       0              0      0
Mr Good Bar                      1      0       0              1      0
Ring pop                         0      1       0              0      0
Nik L Nip                        0      1       0              0      0
Nestle Smarties                  1      0       0              0      0
                         crispedricewafer hard bar pluribus sugarpercent
Hershey's Milk Chocolate                0    0   1        0        0.430
Hershey's Special Dark                  0    0   1        0        0.430
Mr Good Bar                             0    0   1        0        0.313
Ring pop                                0    1   0        0        0.732
Nik L Nip                               0    0   0        1        0.197
Nestle Smarties                         0    0   0        1        0.267
                         pricepercent winpercent
Hershey's Milk Chocolate        0.918   56.49050
Hershey's Special Dark          0.918   59.23612
Mr Good Bar                     0.918   54.52645
Ring pop                        0.965   35.29076
Nik L Nip                       0.976   22.44534
Nestle Smarties                 0.976   37.88719
```

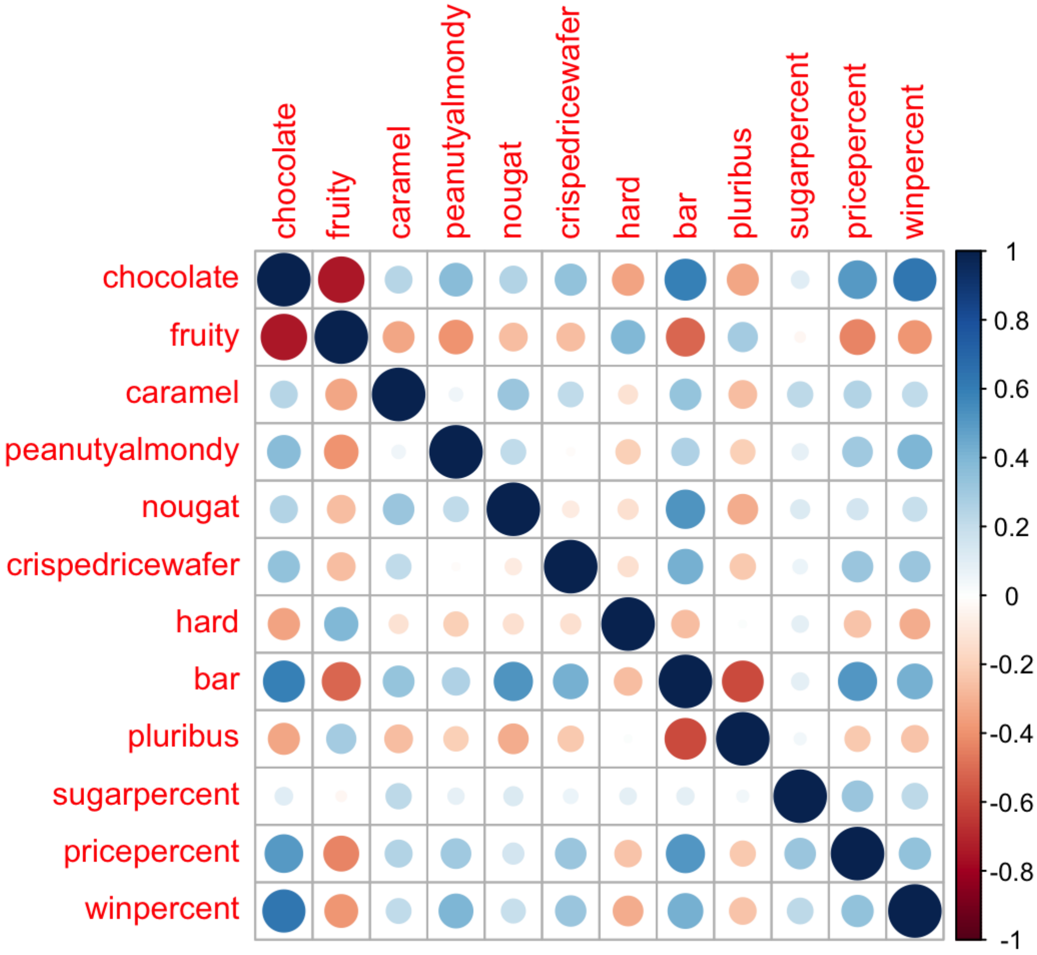
5 Exploring th ecorrelations structure

Pearson correlation goes between -1 and +1 with zero indicating no correlation and values close to 1 being highly correlated.

```
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)?

fruit and chocolate are anti-correlated >Q23. Similarly, what two variables are most positively correlated?

chocolate and winpercent or chocolate and bar are most positively correlated

#Principal Component Analysis

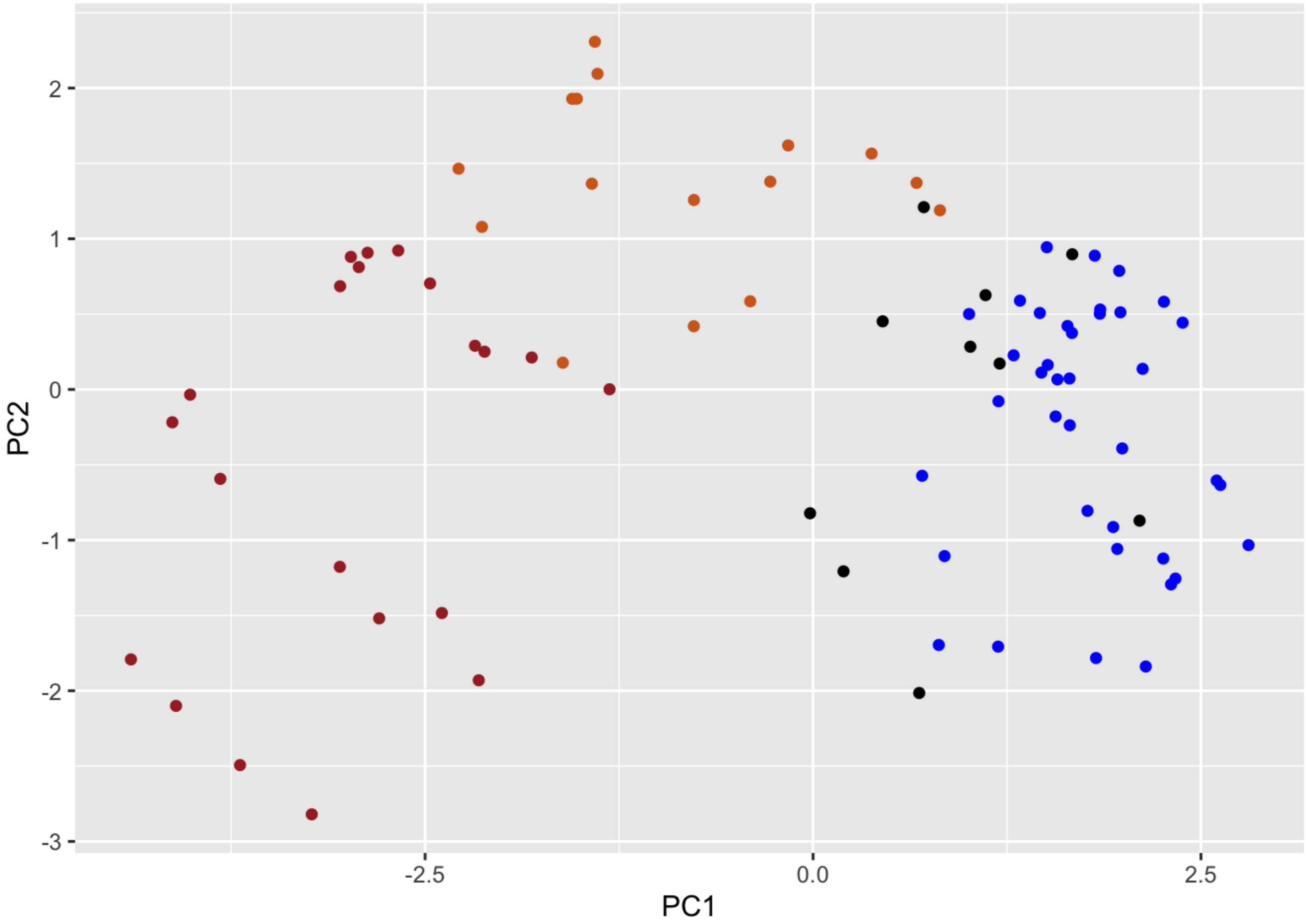
The base R function for PCA is called `pcromp()` and we can set “scale=TRUE/FALSE”

```
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
```

The main result of PCA - i.e. the new PC plot (projection of candy on our new PC axis) is contained in `pca$x`

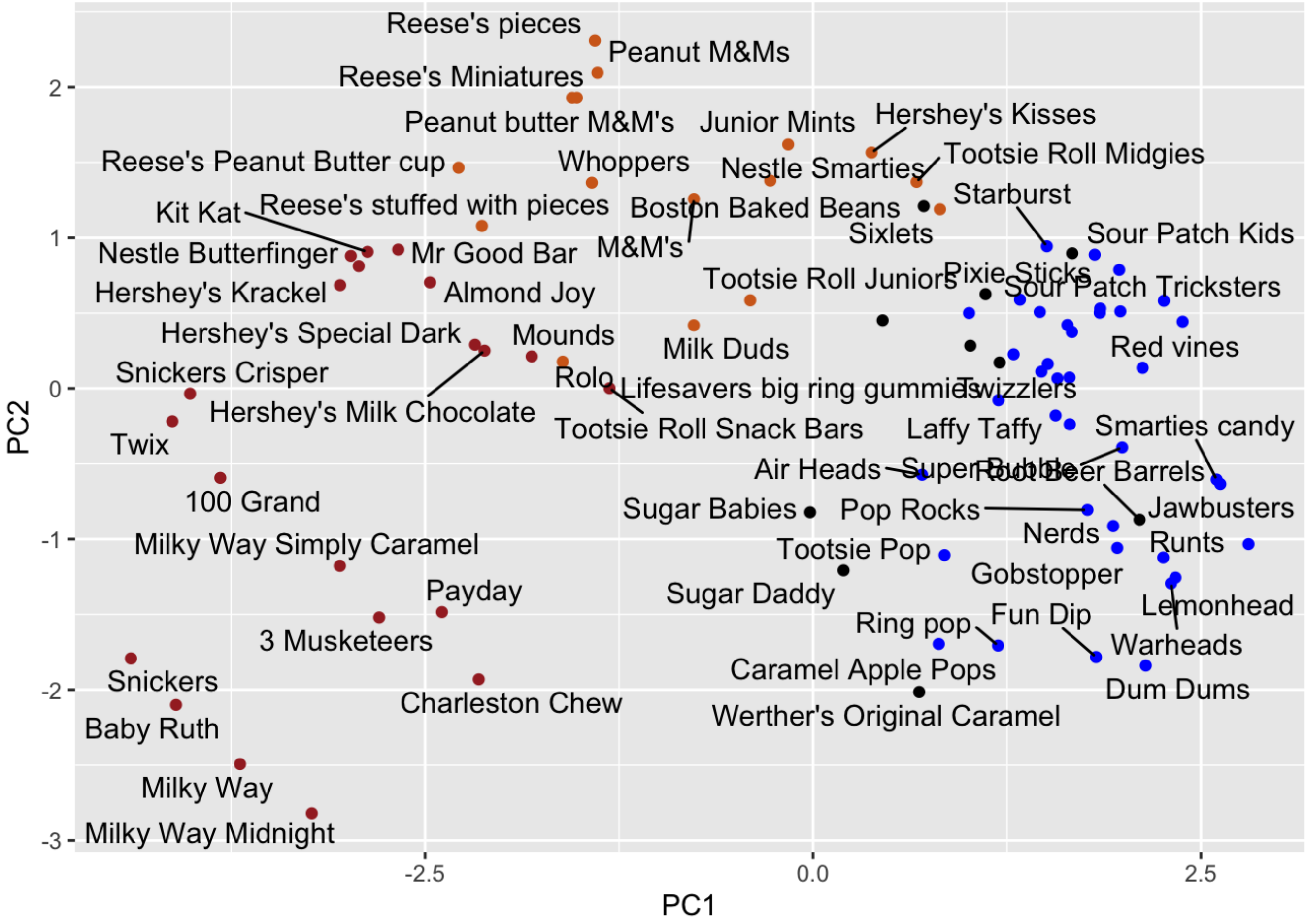
```
pc <- as.data.frame(pca$x)
ggplot(pc)+
  aes(PC1, PC2, label=rownames(pc))+
  geom_point(col=my_cols)
```



```
 # geom_text_repel(max.overlaps = 10)
```

```
ggplot(pc)+
  aes(PC1, PC2, label=rownames(pc))+
  geom_point(col=my_cols)+
  geom_text_repel(max.overlaps = 10)
```

```
Warning: ggrepel: 21 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?

Fruity, hard and pluribus

PC 1 captures correlation structure. If a candy is fruity, hard and pluribus, it will be on the positive side of the axis.