Class 09: Candy Mini-Project

Renny Ng (A98061553)

Today we will analyze some data fro m538 about typical Halloween candy.

Our first job is t o go get the data and read it into R.

```
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
```

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

Check the number of rows with `nrow()`

```
nrow(candy)
```

```
[1] 85
```

```
ncol(candy)
```

```
[1] 12
```

There are 85 types of candy in this dataset.

Q. How many chocolate candy types are in this dataset?

Check by summing up the values in the "chocolate" column

```
sum(candy$chocolate)
```

```
[1] 37
```

There are 37 types of candy that are chocolate.

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

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

```
[1] 38
```

There are 38 types of candy that are fruity.

##Data Exploration There is a useful package that is helpful for first looking into a new dataset (skimr). We can see what it does to this dataset now. #install.packages("skimr")

```
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	

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

```
candy["Milky Way", "winpercent"]
```

```
[1] 73.09956
```

My favorite candy is Milky Way, and its winpercent value is 73.09956%

Q4. What is the winpercent value for "Kit Kat"?

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

```
[1] 76.7686
```

The winpercent value for Kit Kat is 76.7686%

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

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

```
[1] 49.6535
```

The winpercent value for Tootsie Rolls is 49.6535%

Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

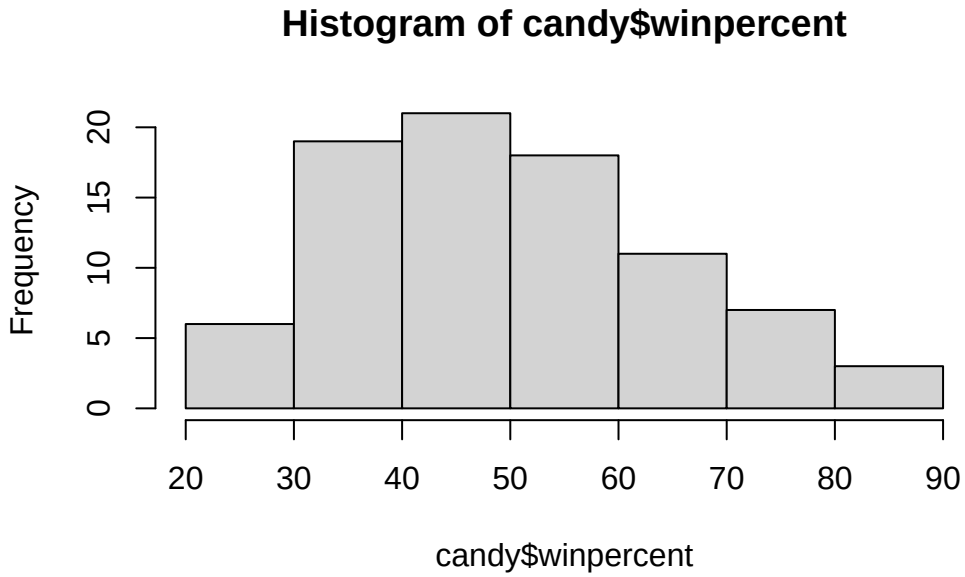
Yes, the "winpercent" variable is on a different scale from most of the other variables.

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

They represent whether the candy is (1) or is not (0) chocolate.

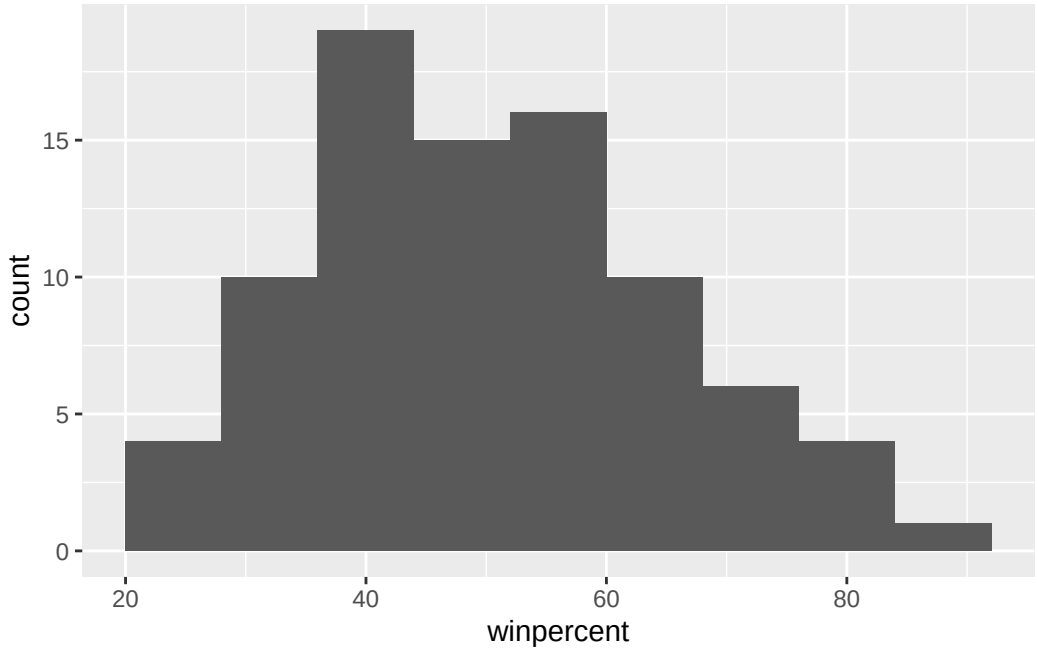
Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent)
```




```
library(ggplot2)

ggplot(candy) +
  aes (winpercent) +
  geom_histogram(binwidth=8)
```



Q9. Is the distribution of winpercent values symmetrical?

The distribution of winpercent values is not symmetrical; it is left-skewed towards the lower ($< 50\%$) values.

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

The center of the distribution is below 50%.

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

- First, find all chocolate candy (subset), so that the table is just chocolate
- Get their winpercent values
- Summarize these values into one metric

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

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

Chocolate candy is more popular at 60.9% than fruity candy at 44%.

Q12. Is this difference statistically significant?

We can use T-test on R.

```
t.test(candy[fruity.inds,]$winpercent, candy[choc.inds,]$winpercent)
```

```
    Welch Two Sample t-test

data:  candy[fruity.inds, ]$winpercent and candy[choc.inds, ]$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:
 -22.15795 -11.44563
sample estimates:
mean of x mean of y
 44.11974  60.92153
```

The p-value is 2.871e-08, indicating that these results are different in a statistically significant manner.

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

Try using `order`, which sorts on/reorders multiple variables. It also allows permutations to be applied to the data, so that all data is ordered by whatever parameter is selected.

```
inds <- order(candy$winpercent)
head(candy[inds, ])
```

	chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	sugarpercent	pricepercent
Nik L Nip	0	1	0	0	0	0	0	0	1	0.197	0.976
Boston Baked Beans	0	0	0	1	0	0	0	0	1	0.313	0.511
Chiclets	0	1	0	0	0	0	0	0	1	0.046	0.325
Super Bubble	0	1	0	0	0	0	0	0	0	0.162	0.116
Jawbusters	0	1	0	0	0	0	1	0	1	0.093	0.511
Root Beer Barrels	0	0	0	0	0	0	1	0	1	0.732	0.069

```
                   winpercent
Nik L Nip            22.44534
Boston Baked Beans   23.41782
Chiclets             24.52499
Super Bubble         27.30386
Jawbusters           28.12744
Root Beer Barrels    29.70369
```

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

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

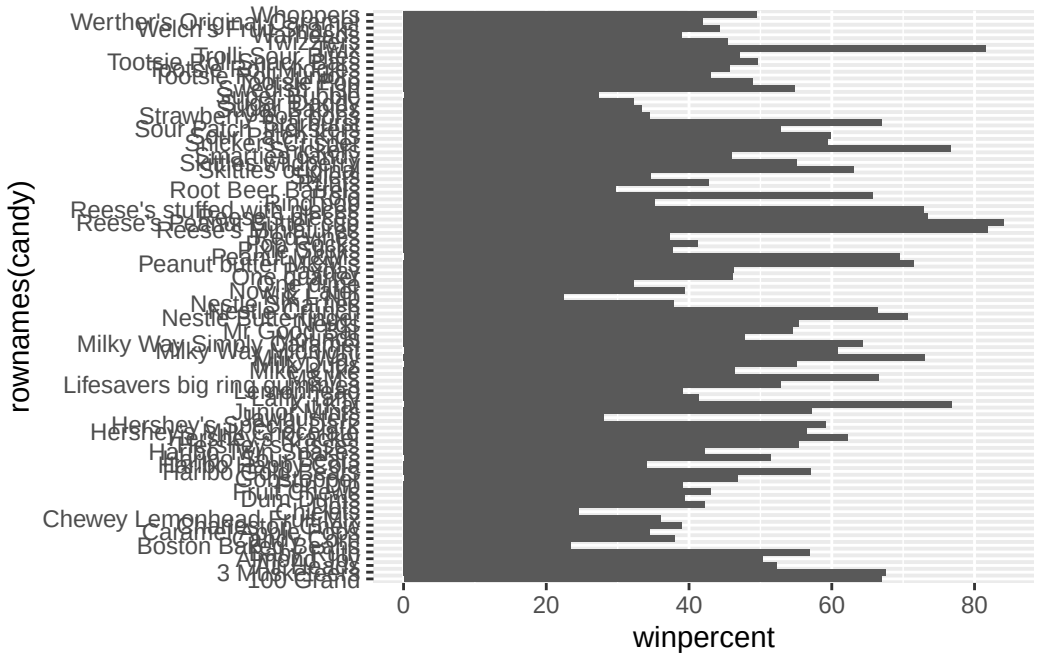
```
inds <- order(candy$winpercent)
tail(candy[inds, ])
```

```
                         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
```

Reese's PB cups, Reese's miniatures, Twix, Kit Kat, Snickers are the top 5.

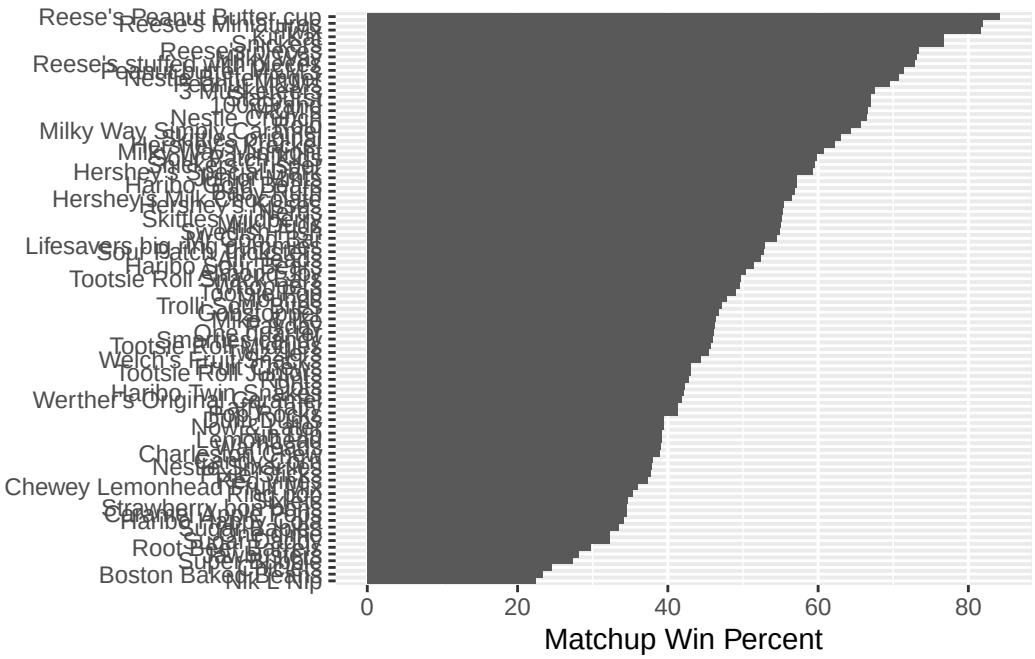
Q15. Make a first barplot of candy ranking based on winpercent values.

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



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

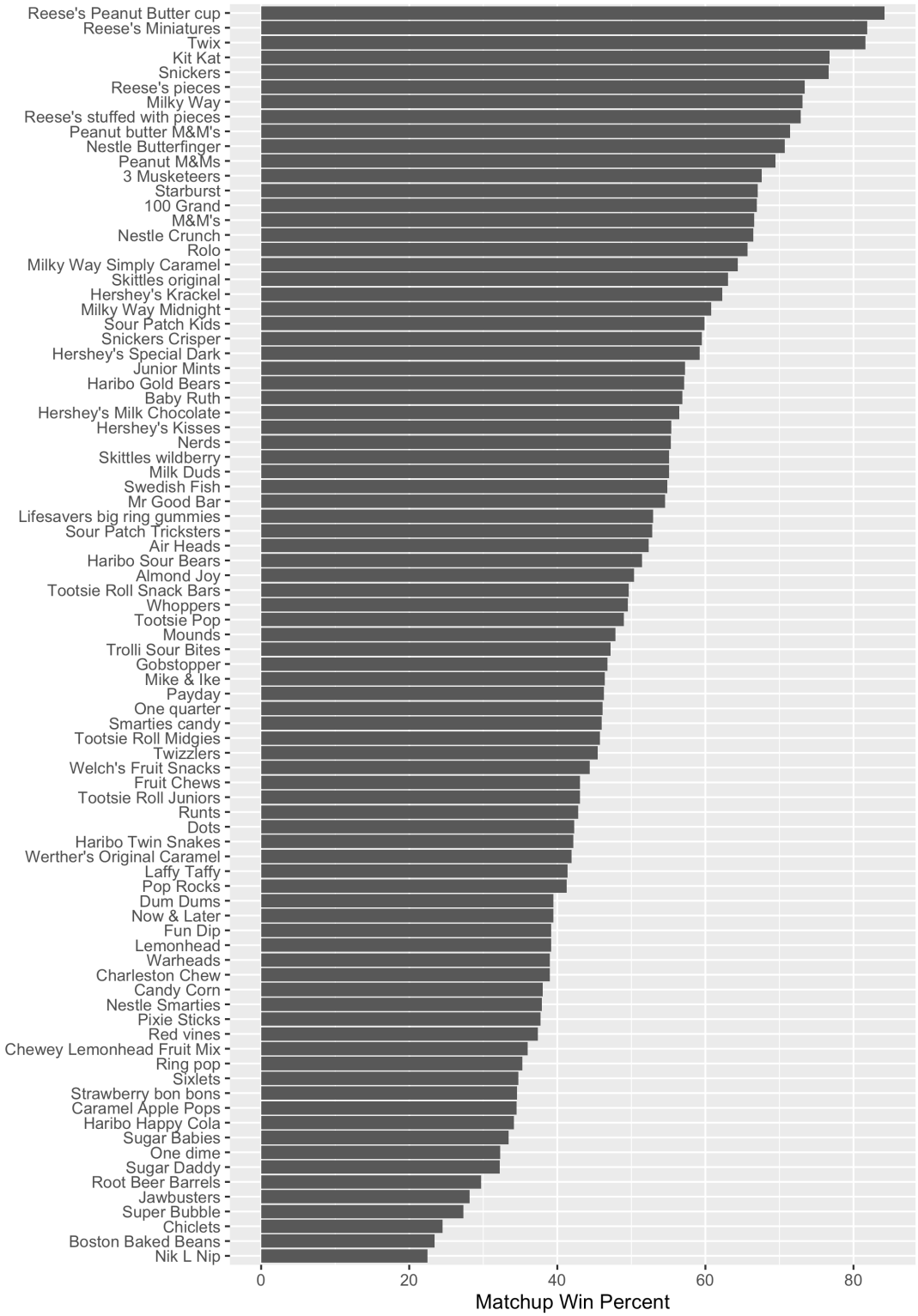
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col() +
  labs(x="Matchup Win Percent", y=NULL)
```




```
ggsave("barplot1.png", height=10, width=7)
```

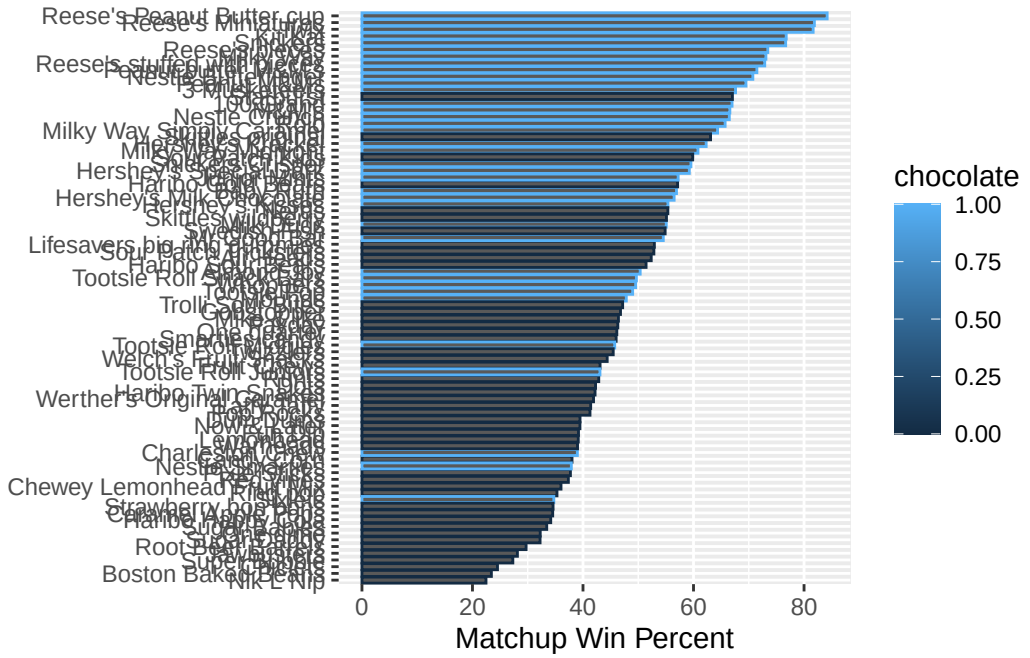
We can insert any image using markdown syntax. This is ! followed by square brackets and then normal brackets.

Put the file name in the normal brackets.


Reese's Peanut Butter cup
Reese's Miniatures
Twix
Kit Kat
Snickers
Reese's pieces
Milky Way
Reese's stuffed with pieces
Peanut butter M&M's
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&M's
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
Hershey's Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
Hershey's Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
Hershey's Milk Chocolate
Hershey's Kisses
Nerds
Skittles wildberry
Milk Duds
Swedish Fish
Mr Good Bar
Lifesavers big ring gummies
Sour Patch Tricksters
Air Heads
Haribo Sour Bears
Almond Joy
Tootsie Roll Snack Bars
Whoppers
Tootsie Pop
Mounds
Trolli Sour Bites
Gobstopper
Mike & Ike
Payday
One quarter
Smarties candy
Tootsie Roll Midgies
Twizzlers
Welch's Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
Werther's Original Caramel
Laffy Taffy
Pop Rocks
Dum Dums
Now & Later
Fun Dip
Lemonhead
Warheads
Charleston Chew
Candy Corn
Nestle Smarties
Pixie Sticks
Red vines
Chewey Lemonhead Fruit Mix
Ring pop
Sixlets
Strawberry bon bons
Caramel Apple Pops
Haribo Happy Cola
Sugar Babies
One dime
Sugar Daddy
Root Beer Barrels
Jawbusters
Super Bubble
Chiclets
Boston Baked Beans
Nik L Nip
0
20
40
60
80
Matchup Win Percent

We can also add some color.

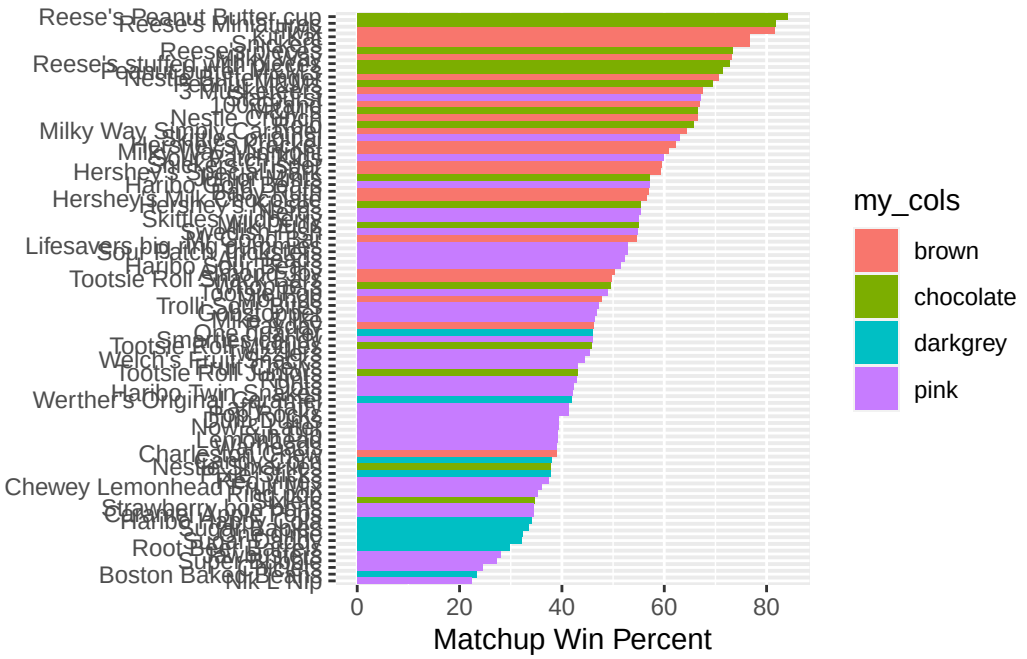
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent), color=chocolate) +
  geom_col() +
  labs(x="Matchup Win Percent", y=NULL)
```



We need to make our own color vector with the colors we like.

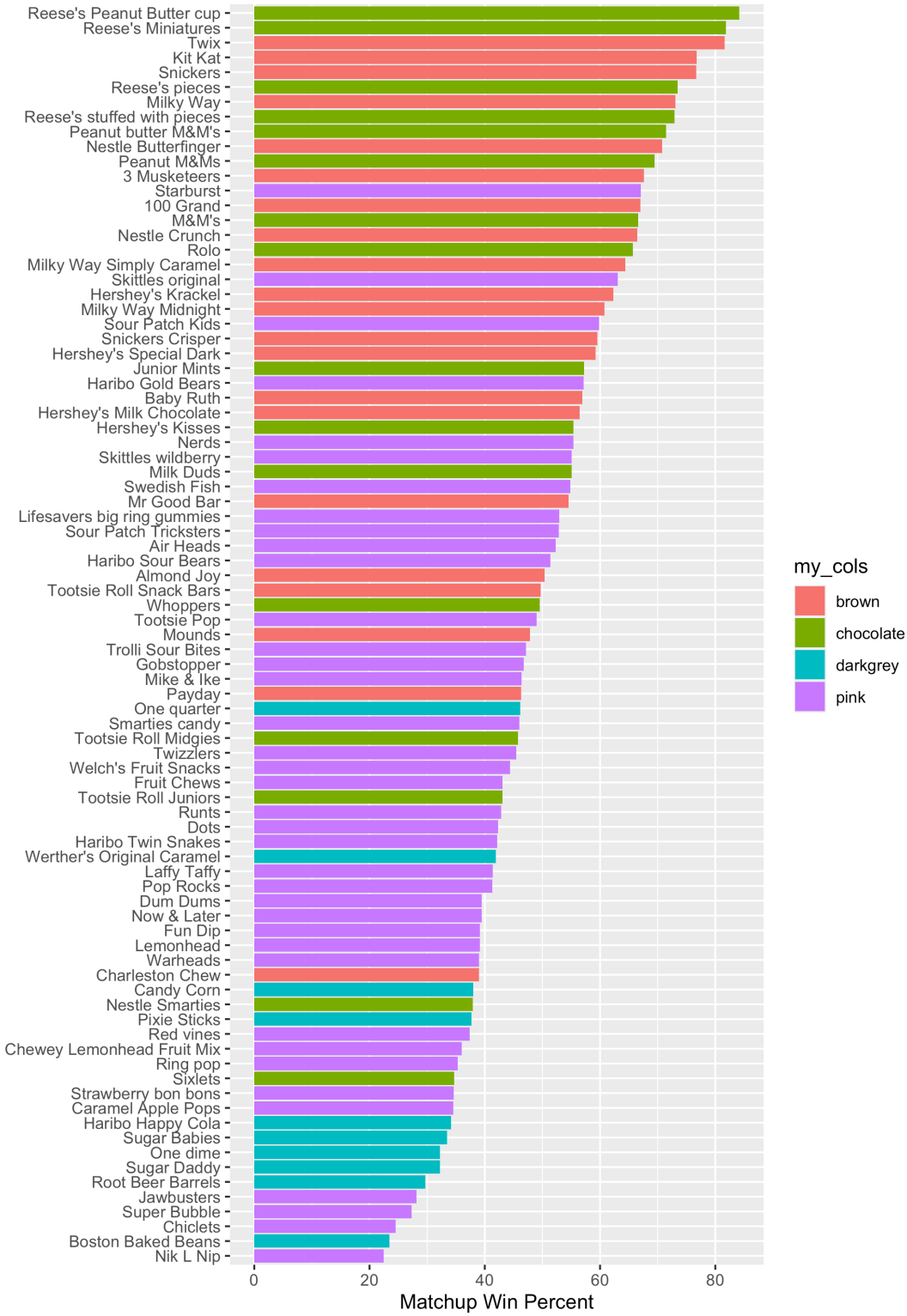
```
my_cols <- rep("darkgrey", nrow(candy))
my_cols[as.logical(candy$chocolate)] <- "chocolate"
my_cols[as.logical(candy$bar)] <- "brown"
my_cols[as.logical(candy$fruity)] <- "pink"
```

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent), fill=my_cols) +
  geom_col() +
  labs(x="Matchup Win Percent", y=NULL)
```




```
ggsave("barplot2.png", height=10, width=7)
```

Why did the graph turn red originally? Because aesthetic mapping from data is used, it'll try to match the data to the X and Y axes data. It's matching the first color on the palette (red). So, do not use the aesthetics to set the color; use it to the geometry.


Reese's Peanut Butter cup
Reese's Miniatures
Twix
Kit Kat
Snickers
Reese's pieces
Milky Way
Reese's stuffed with pieces
Peanut butter M&M's
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&M's
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
Hershey's Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
Hershey's Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
Hershey's Milk Chocolate
Hershey's Kisses
Nerds
Skittles wildberry
Milk Duds
Swedish Fish
Mr Good Bar
Lifesavers big ring gummies
Sour Patch Tricksters
Air Heads
Haribo Sour Bears
Almond Joy
Tootsie Roll Snack Bars
Whoppers
Tootsie Pop
Mounds
Trolli Sour Bites
Gobstopper
Mike & Ike
Payday
One quarter
Smarties candy
Tootsie Roll Midgies
Twizzlers
Welch's Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
Werther's Original Caramel
Laffy Taffy
Pop Rocks
Dum Dums
Now & Later
Fun Dip
Lemonhead
Warheads
Charleston Chew
Candy Corn
Nestle Smarties
Pixie Sticks
Red vines
Chewey Lemonhead Fruit Mix
Ring pop
Sixlets
Strawberry bon bons
Caramel Apple Pops
Haribo Happy Cola
Sugar Babies
One dime
Sugar Daddy
Root Beer Barrels
Jawbusters
Super Bubble
Chiclets
Boston Baked Beans
Nik L Nip
0
20
40
60
80
Matchup Win Percent
my_cols
brown
chocolate
darkgrey
pink

As

shown in {**?@fig-bar**} there are some ugly colors to pick from in R.

> Q17. What is the worst ranked chocolate candy?

Sixlets is the lowest ranked chocolate candy.

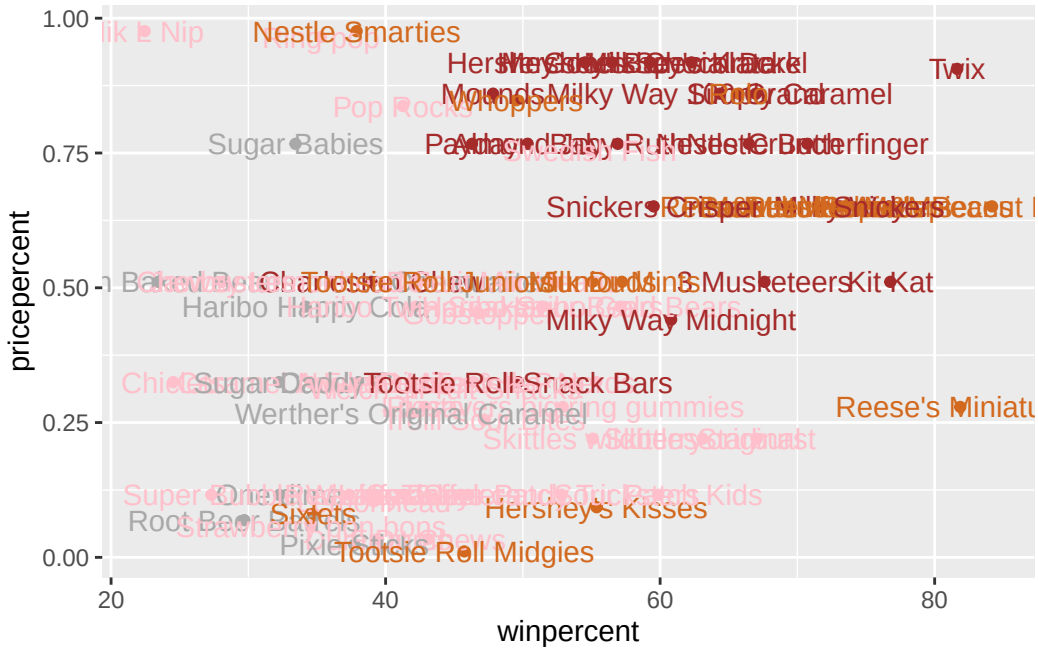
> Q18. What is the best ranked fruity candy?

Starbursts are the highest ranked fruity candy.

> 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?

Let's make a plot of winpercent vs pricepercent. The original idea with this 538 plot was to show you the best candy to get for your money.

```
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text(col=my_cols)
```



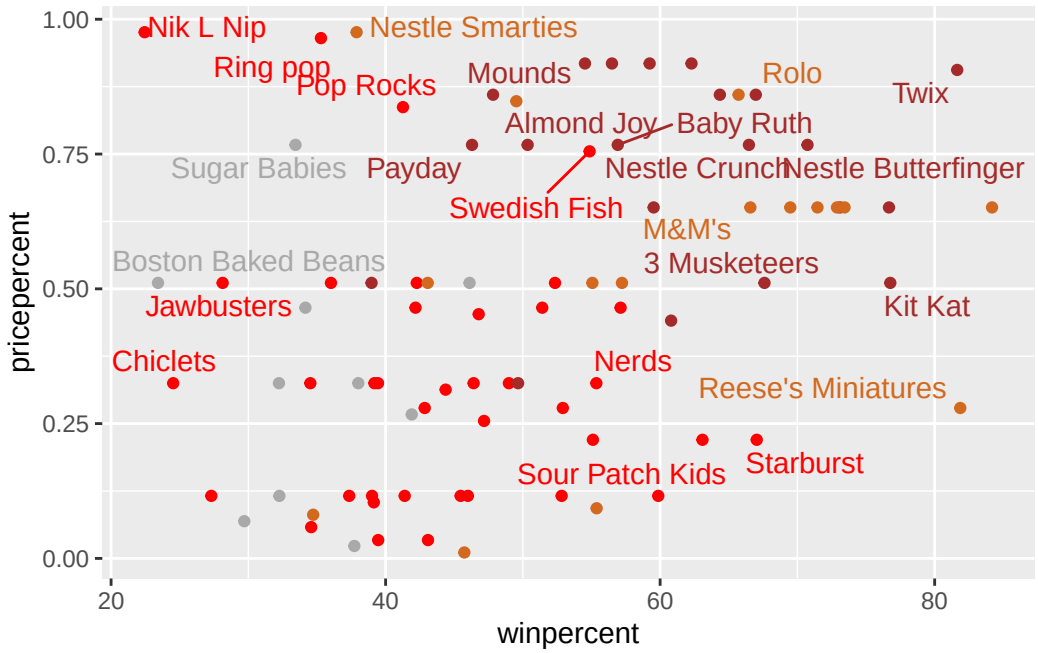
These labels are terrible. Get ggrepel package.

```
library(ggrepel)
```

```
my_cols[as.logical(candy$fruity)] <- "red"

ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, max.overlaps = 8)
```

```
Warning: ggrepel: 61 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



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

Nik L Nip, Ring Pop, Nestle Smarties, Mr Good Bar, and Hershey's Milk Chocolate are the most expensive. Nik L Nip is the least popular.

##Explore the correlation structure in candy data.

We will calcualte Pearson correlation values.

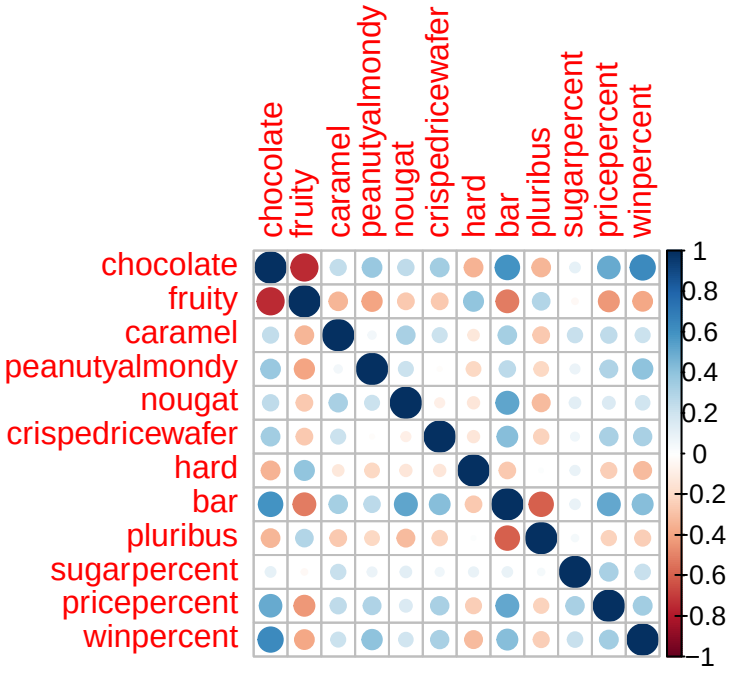
```
cij <- cor(candy)
#Shows the pairwise correlation between each parameter.
```

Let's install and load corrplot.

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

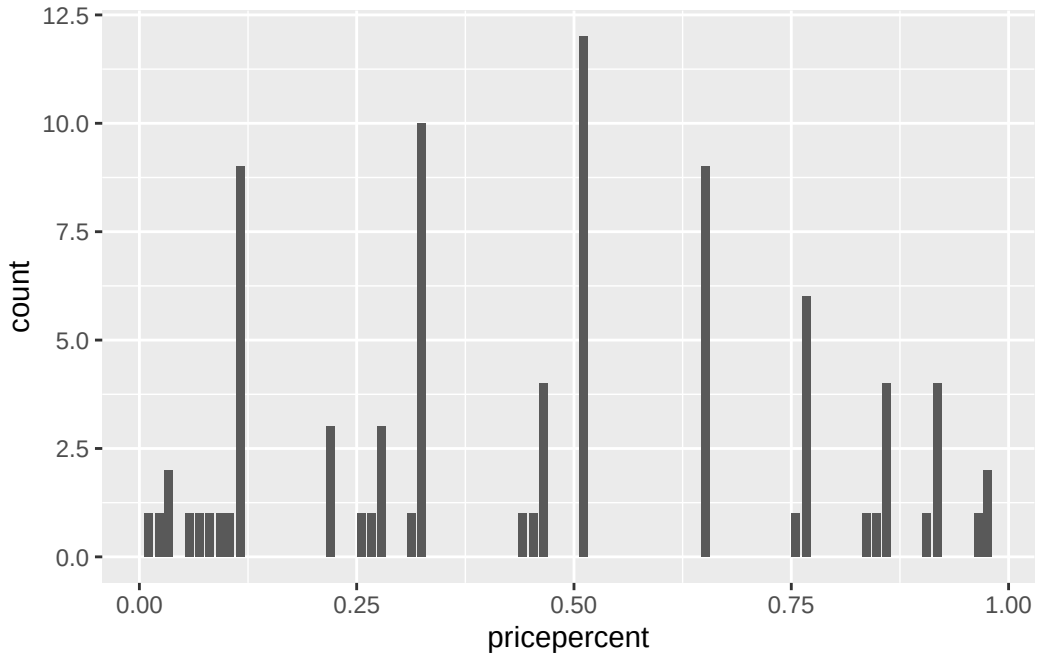
```
corrplot(cij)
```



This is the type of data that PCA takes advantage of. Let's see how PCA captures this correlation structure in an informative way.

> 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) +
  aes(pricepercent) +
  geom_bar()
```



##PCA The main function is called `prcomp`. We will need to scale. >Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

Chocolate and fruit are anti-correlated.

```
pca <- prcomp(candy, scale=TRUE)
```

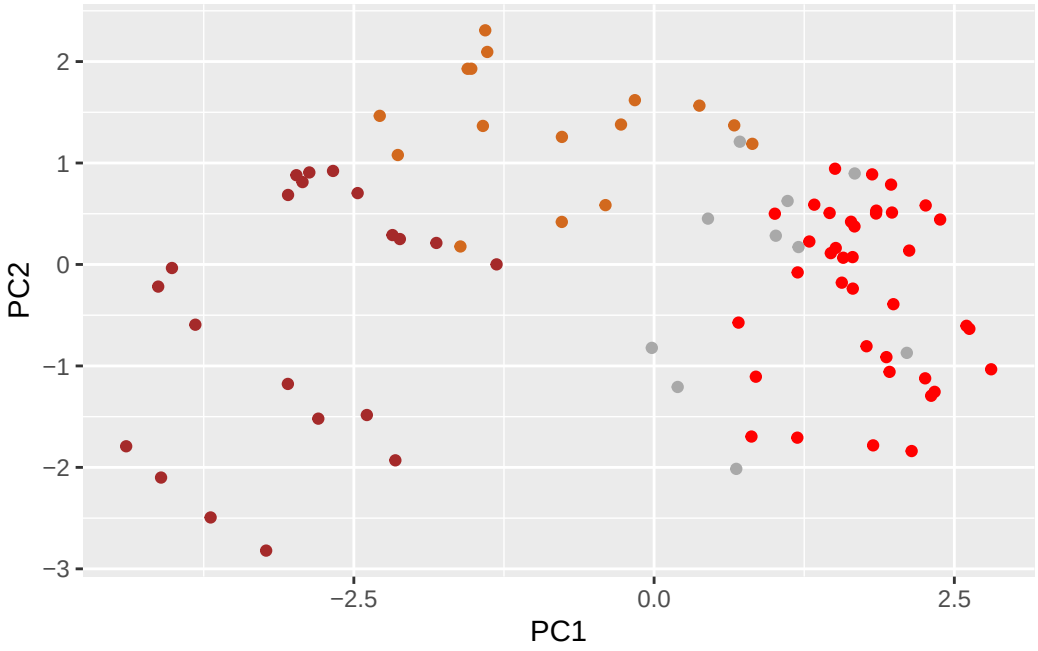
```
summary(pca)
```

```
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
```

```
pc.results <- as.data.frame(pca$x)
```

```
ggplot(pc.results) +
   aes(x=PC1, y=PC2) +
        geom_point(col=my_cols)
```

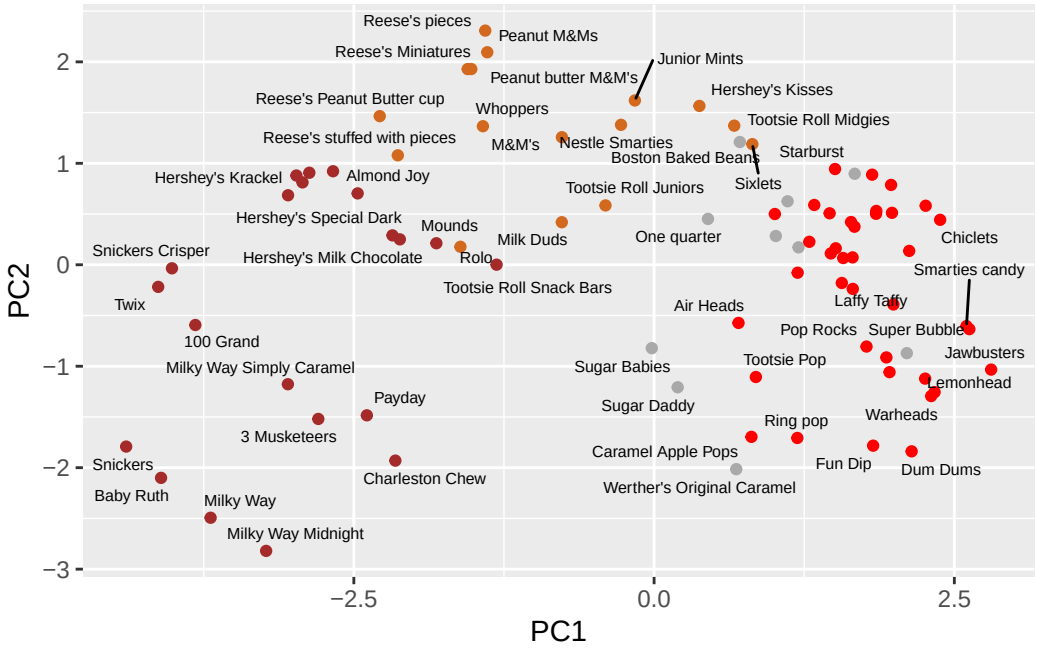


Note that chocolate is separated from fruit.

What does PC1 really caputure? Do a loadings plot.

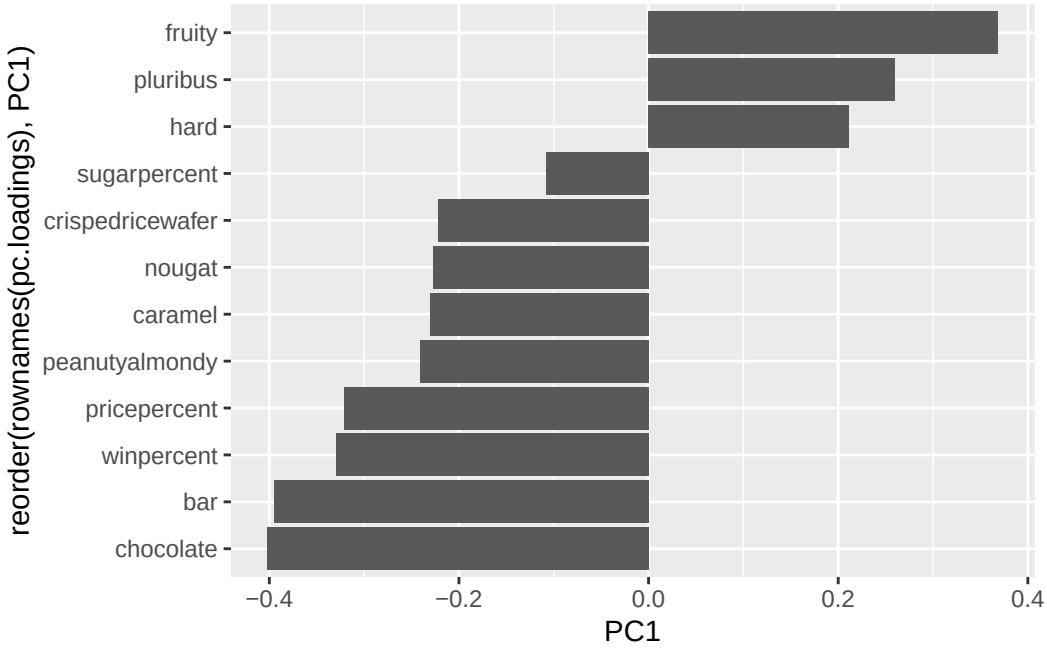
```
ggplot(pc.results) +
   aes(x=PC1, y=PC2, label=rownames(pc.results)) +
   geom_point(col=my_cols) +
  geom_text_repel(size = 2, max.overlaps = 8)
```

```
Warning: ggrepel: 32 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




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

ggplot(pc.loadings) +
  aes(PC1, reorder(rownames(pc.loadings), PC1)) +
  geom_col()
```



Explains why things are plotted the way they are; these are the things that factored and are reflected by PC1.

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

Fruit and pluribus are correlated.

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

"Fruit", "pluribus", and "hard" are the variables picked up strongly by PC1 in the positive direction.