Class 10: Halloween Mini-Project

Courtney Anderson (PID:A69038035)

What is in the dataset

```
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
```

```
flextable::flextable(head(candy))
```

chocolate	fruity	caramel	peanutyal-mondy	nougat	crispedrice-wafer	hard	bar	pluribus
1	0	1	0	0	1	0	1	0

chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus
1	0	0	0	1	0	0	1	0
0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0
0	1	0	0	0	0	0	0	0
1	0	0	1	0	0	0	1	0

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

```
library(dplyr)

candy |> nrow()
```

```
[1] 85
```

```
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

What is your favorite candy

```
candy|> select(winpercent)
```

```
                            winpercent
100 Grand                     66.97173
3 Musketeers                  67.60294
One dime                      32.26109
One quarter                   46.11650
```

Air Heads	52.34146
Almond Joy	50.34755
Baby Ruth	56.91455
Boston Baked Beans	23.41782
Candy Corn	38.01096
Caramel Apple Pops	34.51768
Charleston Chew	38.97504
Chewey Lemonhead Fruit Mix	36.01763
Chiclets	24.52499
Dots	42.27208
Dum Dums	39.46056
Fruit Chews	43.08892
Fun Dip	39.18550
Gobstopper	46.78335
Haribo Gold Bears	57.11974
Haribo Happy Cola	34.15896
Haribo Sour Bears	51.41243
Haribo Twin Snakes	42.17877
Hershey's Kisses	55.37545
Hershey's Krackel	62.28448
Hershey's Milk Chocolate	56.49050
Hershey's Special Dark	59.23612
Jawbusters	28.12744
Junior Mints	57.21925
Kit Kat	76.76860
Laffy Taffy	41.38956
Lemonhead	39.14106
Lifesavers big ring gummies	52.91139
Peanut butter M&M's	71.46505
M&M's	66.57458
Mike & Ike	46.41172
Milk Duds	55.06407
Milky Way	73.09956
Milky Way Midnight	60.80070
Milky Way Simply Caramel	64.35334
Mounds	47.82975
Mr Good Bar	54.52645
Nerds	55.35405
Nestle Butterfinger	70.73564
Nestle Crunch	66.47068
Nik L Nip	22.44534
Now & Later	39.44680
Payday	46.29660

```
Peanut M&Ms                   69.48379
Pixie Sticks                  37.72234
Pop Rocks                     41.26551
Red vines                     37.34852
Reese's Miniatures            81.86626
Reese's Peanut Butter cup     84.18029
Reese's pieces                73.43499
Reese's stuffed with pieces   72.88790
Ring pop                      35.29076
Rolo                          65.71629
Root Beer Barrels             29.70369
Runts                         42.84914
Sixlets                       34.72200
Skittles original             63.08514
Skittles wildberry            55.10370
Nestle Smarties               37.88719
Smarties candy                45.99583
Snickers                      76.67378
Snickers Crisper              59.52925
Sour Patch Kids               59.86400
Sour Patch Tricksters         52.82595
Starburst                     67.03763
Strawberry bon bons           34.57899
Sugar Babies                  33.43755
Sugar Daddy                   32.23100
Super Bubble                  27.30386
Swedish Fish                  54.86111
Tootsie Pop                   48.98265
Tootsie Roll Juniors          43.06890
Tootsie Roll Midgies          45.73675
Tootsie Roll Snack Bars       49.65350
Trolli Sour Bites             47.17323
Twix                          81.64291
Twizzlers                     45.46628
Warheads                      39.01190
Welch's Fruit Snacks          44.37552
Werther's Original Caramel    41.90431
Whoppers                      49.52411
```

```r
win<- candy$winpercent
win.mean <- mean(win)
round(win.mean)
```

```
[1] 50
```

```
candy%>% select(winpercent)
```

```
                            winpercent
100 Grand                     66.97173
3 Musketeers                  67.60294
One dime                      32.26109
One quarter                   46.11650
Air Heads                     52.34146
Almond Joy                    50.34755
Baby Ruth                     56.91455
Boston Baked Beans            23.41782
Candy Corn                    38.01096
Caramel Apple Pops            34.51768
Charleston Chew               38.97504
Chewey Lemonhead Fruit Mix    36.01763
Chiclets                      24.52499
Dots                          42.27208
Dum Dums                      39.46056
Fruit Chews                   43.08892
Fun Dip                       39.18550
Gobstopper                    46.78335
Haribo Gold Bears             57.11974
Haribo Happy Cola             34.15896
Haribo Sour Bears             51.41243
Haribo Twin Snakes            42.17877
Hershey's Kisses              55.37545
Hershey's Krackel             62.28448
Hershey's Milk Chocolate      56.49050
Hershey's Special Dark        59.23612
Jawbusters                    28.12744
Junior Mints                  57.21925
Kit Kat                       76.76860
Laffy Taffy                   41.38956
Lemonhead                     39.14106
Lifesavers big ring gummies   52.91139
Peanut butter M&M's           71.46505
M&M's                         66.57458
Mike & Ike                    46.41172
Milk Duds                     55.06407
Milky Way                     73.09956
Milky Way Midnight            60.80070
Milky Way Simply Caramel      64.35334
Mounds                        47.82975
Mr Good Bar                   54.52645
Nerds                         55.35405
Nestle Butterfinger           70.73564
Nestle Crunch                 66.47068
Nik L Nip                     22.44534
Now & Later                   39.44680
Payday                        46.29660
Peanut M&Ms                   69.48379
Pixie Sticks                  37.72234
Pop Rocks                     41.26551
Red vines                     37.34852
Reese's Miniatures            81.86626
Reese's Peanut Butter cup     84.18029
Reese's pieces                73.43499
Reese's stuffed with pieces   72.88790
Ring pop                      35.29076
Rolo                          65.71629
Root Beer Barrels             29.70369
Runts                         42.84914
Sixlets                       34.72200
Skittles original             63.08514
Skittles wildberry            55.10370
Nestle Smarties               37.88719
Smarties candy                45.99583
Snickers                      76.67378
Snickers Crisper              59.52925
Sour Patch Kids               59.86400
Sour Patch Tricksters         52.82595
Starburst                     67.03763
Strawberry bon bons           34.57899
Sugar Babies                  33.43755
Sugar Daddy                   32.23100
Super Bubble                  27.30386
Swedish Fish                  54.86111
Tootsie Pop                   48.98265
Tootsie Roll Juniors          43.06890
Tootsie Roll Midgies          45.73675
Tootsie Roll Snack Bars       49.65350
Trolli Sour Bites             47.17323
Twix                          81.64291
```

```
Twizzlers                     45.46628
Warheads                      39.01190
Welch's Fruit Snacks          44.37552
Werther's Original Caramel    41.90431
Whoppers                      49.52411
```

```r
candy |>
  select(winpercent)
```

```
                            winpercent
100 Grand                     66.97173
3 Musketeers                  67.60294
One dime                      32.26109
One quarter                   46.11650
Air Heads                     52.34146
Almond Joy                    50.34755
Baby Ruth                     56.91455
Boston Baked Beans            23.41782
Candy Corn                    38.01096
Caramel Apple Pops            34.51768
Charleston Chew               38.97504
Chewey Lemonhead Fruit Mix    36.01763
Chiclets                      24.52499
Dots                          42.27208
Dum Dums                      39.46056
Fruit Chews                   43.08892
Fun Dip                       39.18550
Gobstopper                    46.78335
Haribo Gold Bears             57.11974
Haribo Happy Cola             34.15896
Haribo Sour Bears             51.41243
Haribo Twin Snakes            42.17877
Hershey's Kisses              55.37545
Hershey's Krackel             62.28448
Hershey's Milk Chocolate      56.49050
Hershey's Special Dark        59.23612
Jawbusters                    28.12744
Junior Mints                  57.21925
Kit Kat                       76.76860
Laffy Taffy                   41.38956
Lemonhead                     39.14106
Lifesavers big ring gummies   52.91139
```

```
Peanut butter M&M's           71.46505
M&M's                         66.57458
Mike & Ike                    46.41172
Milk Duds                     55.06407
Milky Way                     73.09956
Milky Way Midnight            60.80070
Milky Way Simply Caramel      64.35334
Mounds                        47.82975
Mr Good Bar                   54.52645
Nerds                         55.35405
Nestle Butterfinger           70.73564
Nestle Crunch                 66.47068
Nik L Nip                     22.44534
Now & Later                   39.44680
Payday                        46.29660
Peanut M&Ms                   69.48379
Pixie Sticks                  37.72234
Pop Rocks                     41.26551
Red vines                     37.34852
Reese's Miniatures            81.86626
Reese's Peanut Butter cup     84.18029
Reese's pieces                73.43499
Reese's stuffed with pieces   72.88790
Ring pop                      35.29076
Rolo                          65.71629
Root Beer Barrels             29.70369
Runts                         42.84914
Sixlets                       34.72200
Skittles original             63.08514
Skittles wildberry            55.10370
Nestle Smarties               37.88719
Smarties candy                45.99583
Snickers                      76.67378
Snickers Crisper              59.52925
Sour Patch Kids               59.86400
Sour Patch Tricksters         52.82595
Starburst                     67.03763
Strawberry bon bons           34.57899
Sugar Babies                  33.43755
Sugar Daddy                   32.23100
Super Bubble                  27.30386
Swedish Fish                  54.86111
Tootsie Pop                   48.98265
Tootsie Roll Juniors          43.06890
Tootsie Roll Midgies          45.73675
Tootsie Roll Snack Bars       49.65350
Trolli Sour Bites             47.17323
Twix                          81.64291
Twizzlers                     45.46628
Warheads                      39.01190
Welch's Fruit Snacks          44.37552
Werther's Original Caramel    41.90431
Whoppers                      49.52411
```

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

```
[1] 81.64291
```

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

```r
candy["Gobstopper", ]$winpercent
```

```
[1] 46.78335
```

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

```r
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
```

```r
library("skimr")
skim(candy)
```

Table 2: Data summary

Name	candy
Number of rows	85
Number of columns	12
Column type frequency:	
numeric	12
Group variables	None

Variable type: numeric

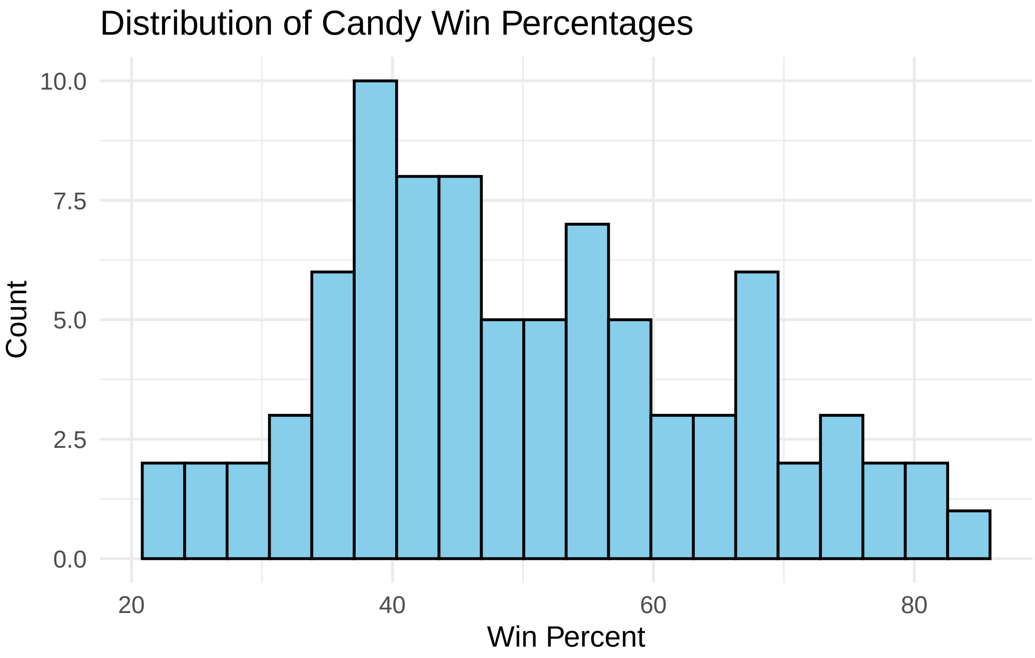
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.00	1.00	
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00	
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
nougat	0	1	0.08	0.28	0.00	0.00	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? The winpercent is on a 0 to 1 scale.

Q7. What do you think a zero and one represent for the candy$chocolate column? 0 means that the candy does not contain chocolate and 1 means that the candy does contain chocolate.

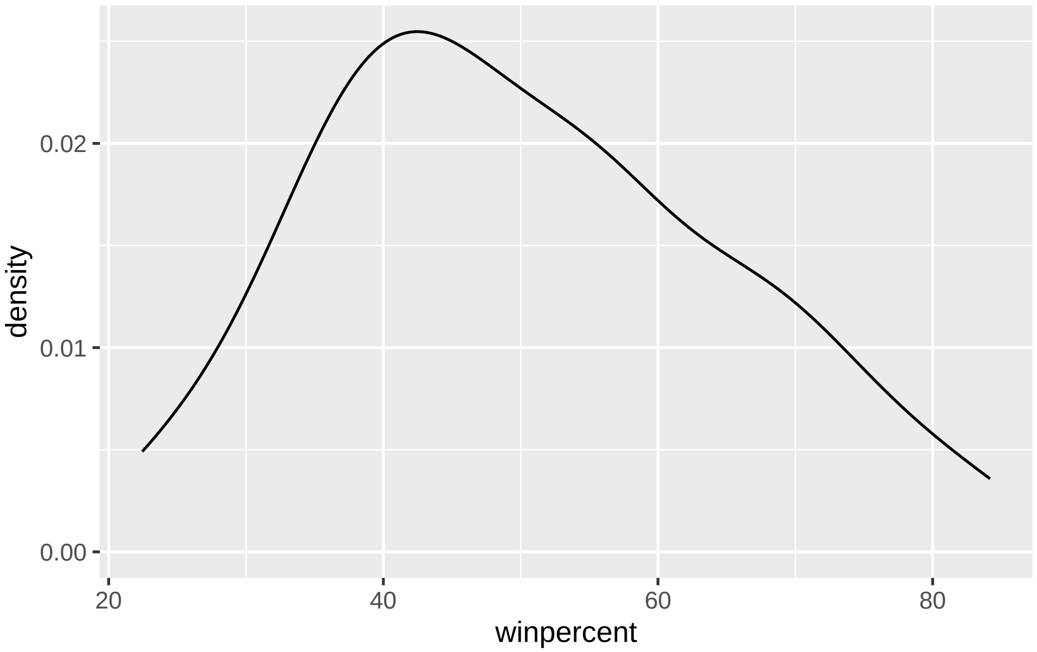
Q8. Plot a histogram of winpercent values

```r
library(ggplot2)

ggplot(candy, aes(x = winpercent)) +
  geom_histogram(bins=20, fill = "skyblue", color = "black") +
  labs(
    title = "Distribution of Candy Win Percentages",
    x = "Win Percent",
    y = "Count"
  ) +
  theme_minimal()
```



Q9. Is the distribution of winpercent values symmetrical?

```
ggplot(candy)+
aes(winpercent)+
geom_density()
```



No, the distribution of the winpercent is not symmetrical.

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

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

```
[1] 50.31676
```

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

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

The mean is above 50 and the median is below 50. For this data, I would go with the median.

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

```
library(dplyr)
#Find all chocolate candy in the data set
#Extract their winpercent values
#Find the mean of these values
candy |>
  group_by(chocolate) |>
  summarise(avg_winpercent = mean(winpercent))
```

```
# A tibble: 2 x 2
  chocolate avg_winpercent
      <int>          <dbl>
1         0           42.1
2         1           60.9
```

```
#Do the same for fruity candy
candy |>
  group_by(fruity) |>
  summarise(avg_winpercent = mean(winpercent))
```

```
# A tibble: 2 x 2
  fruity avg_winpercent
   <int>          <dbl>
1      0           55.3
2      1           44.1
```

```
# Now, which mean value is higher?
```

Chocolate candies have a higher winpercent 60.92%.

```
choc.inds<- as.logical(candy$chocolate)
```

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

```
[1] 60.92153
```

```r
fruity.inds<- as.logical(candy$fruity)

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

```
[1] 44.11974
```

Q12. Is this difference statistically significant?

```r
choc <- candy$winpercent[candy$chocolate == 1]
fruit <- candy$winpercent[candy$fruity == 1]

t.test(choc, fruit)
```

```

    Welch Two Sample t-test

data:  choc and fruit
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! It is statistically significant!

Overall Candy Rankings

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

```
                          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
Reese's pieces                    1      0       0              1      0
                          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
Reese's pieces                           0    0   0        1        0.406
                          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 pieces                   0.651   73.43499
```

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

```
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
```

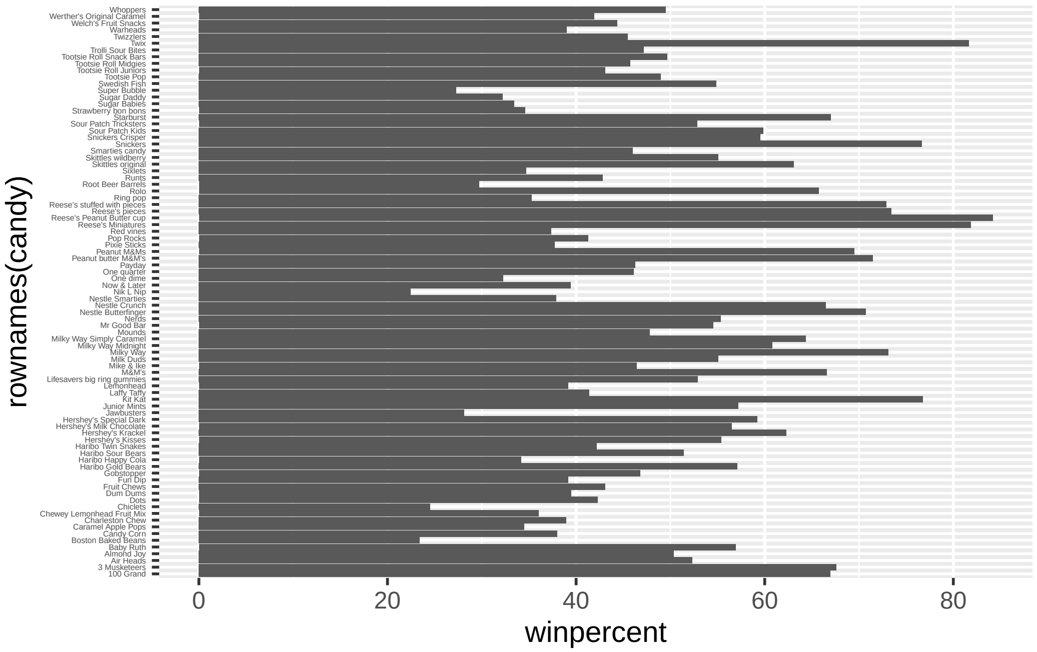
Q14. What are the top 5 all time favorite candy types out of this set?

```
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
```

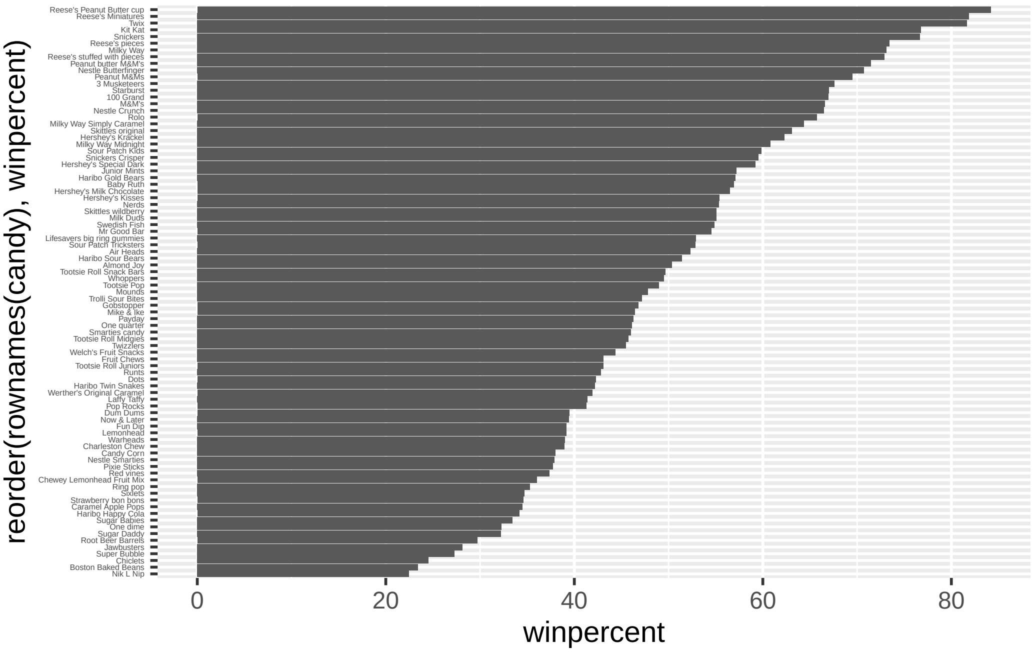
> Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_bar(stat = "identity")+
  theme(
    axis.text.y = element_text(size = 2.5),
  )
```



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

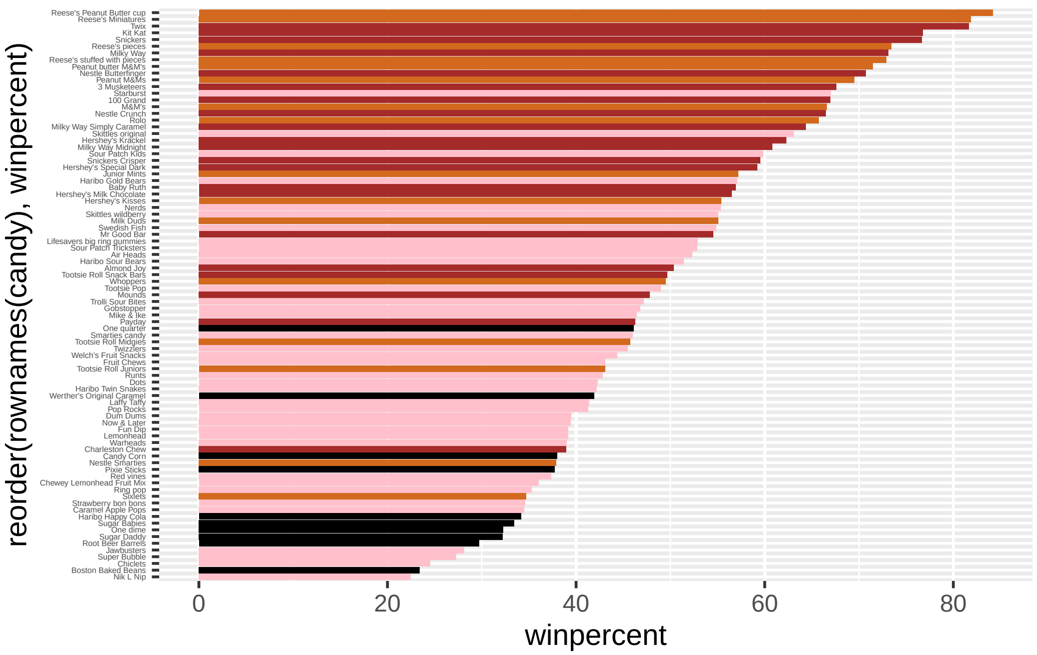
```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_bar(stat = "identity")+
  theme(
    axis.text.y = element_text(size = 2.5),
  )
```



Lets add some colors

```
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)+
  theme(
    axis.text.y = element_text(size = 2.5),
  )
```

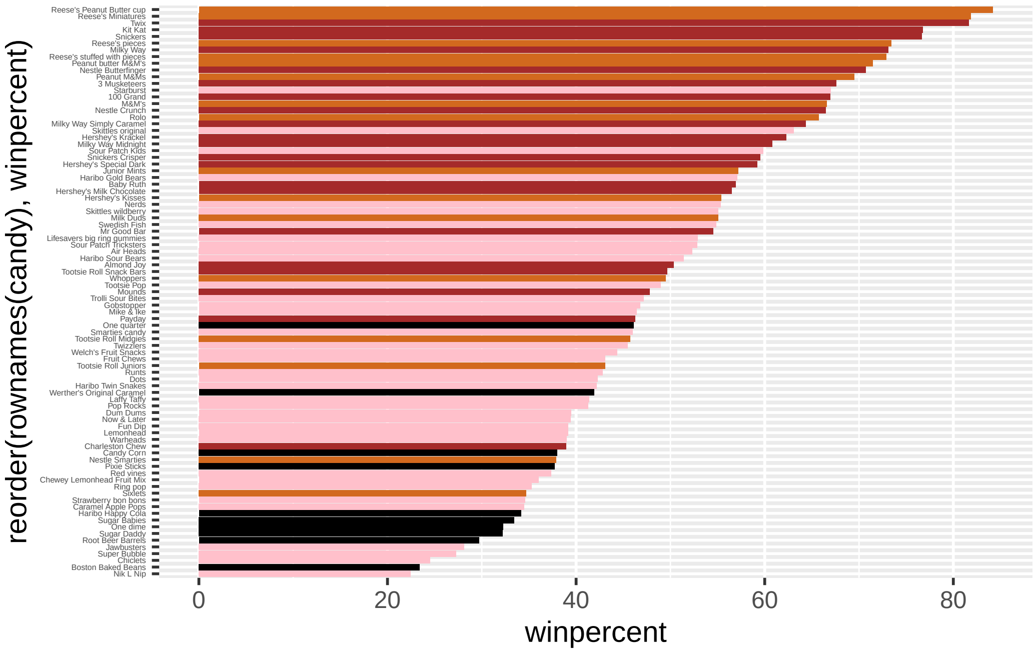


Lets add some different colors: (I did it again with Dr.Grant)

```
my_cols<- rep("black",nrow(candy))
my_cols[candy$chocolate==1]<-"chocolate"
my_cols[candy$bar==1]<-"brown"
my_cols[candy$fruity==1]<- "pink"
my_cols
```

```
 [1] "brown"     "brown"     "black"     "black"     "pink"      "brown"
 [7] "brown"     "black"     "black"     "pink"      "brown"     "pink"
[13] "pink"      "pink"      "pink"      "pink"      "pink"      "pink"
[19] "pink"      "black"     "pink"      "pink"      "chocolate" "brown"
[25] "brown"     "brown"     "pink"      "chocolate" "brown"     "pink"
[31] "pink"      "pink"      "chocolate" "chocolate" "pink"      "chocolate"
[37] "brown"     "brown"     "brown"     "brown"     "brown"     "pink"
[43] "brown"     "brown"     "pink"      "pink"      "brown"     "chocolate"
[49] "black"     "pink"      "pink"      "chocolate" "chocolate" "chocolate"
[55] "chocolate" "pink"      "chocolate" "black"     "pink"      "chocolate"
[61] "pink"      "pink"      "chocolate" "pink"      "brown"     "brown"
[67] "pink"      "pink"      "pink"      "pink"      "black"     "black"
[73] "pink"      "pink"      "pink"      "chocolate" "chocolate" "brown"
[79] "pink"      "brown"     "pink"      "pink"      "pink"      "black"
[85] "chocolate"
```

```r
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)+
  theme(
    axis.text.y = element_text(size = 2.5),
  )
```



Now, for the first time, using this plot we can answer questions like: > Q17. What is the worst ranked chocolate candy?

```r
library(dplyr)

candy |>
  filter(chocolate == 1) |>
  arrange(winpercent) |>
  head(1)
```

```
        chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
Sixlets         1      0       0              0      0                0    0
        bar pluribus sugarpercent pricepercent winpercent
Sixlets   0        1         0.22        0.081     34.722
```

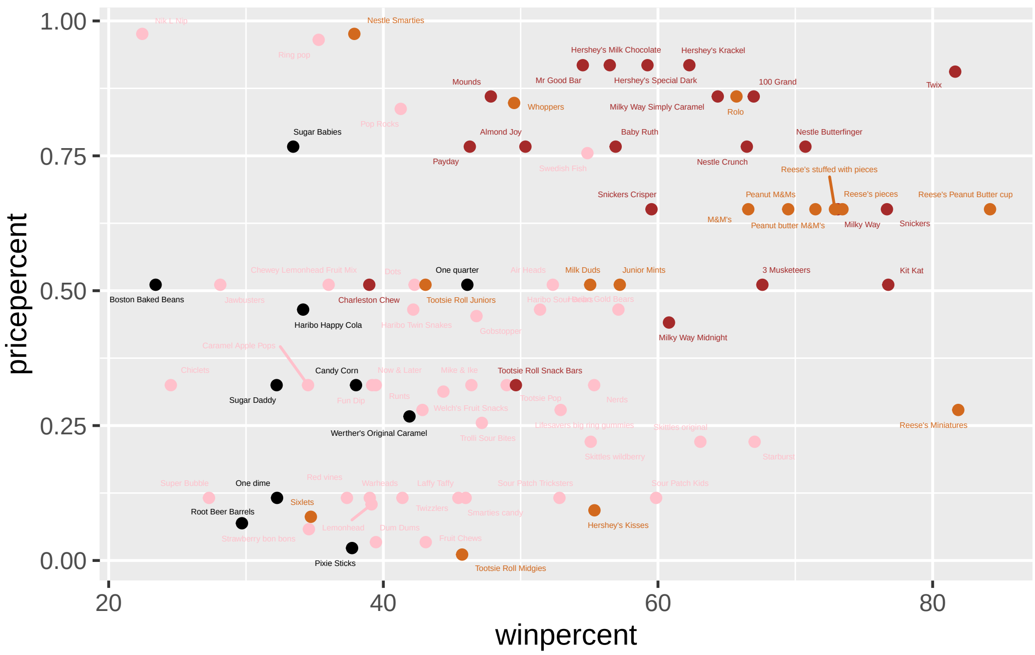
> Q18. What is the best ranked fruity candy?

```r
library(dplyr)

candy |>
  filter(fruity == 1) |>
  arrange(desc(winpercent)) |>
  head(1)
```

```
          chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
Starburst         0      1       0              0      0                0    0
          bar pluribus sugarpercent pricepercent winpercent
Starburst   0        1        0.151         0.22   67.03763
```

Taking a look at pricepercent

```r
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=1, max.overlaps = 12)
```



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?

```
library(dplyr)

candy |>
  mutate(bang_for_buck = winpercent / pricepercent) |>
  arrange(desc(bang_for_buck)) |>
  head(1)
```

```
                     chocolate fruity caramel peanutyalmondy nougat
Tootsie Roll Midgies         1      0       0              0      0
                     crispedricewafer hard bar pluribus sugarpercent
Tootsie Roll Midgies                0    0   0        1        0.174
                     pricepercent winpercent bang_for_buck
Tootsie Roll Midgies        0.011   45.73675      4157.886
```

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

```
library(dplyr)

# Step 1: Get the 5 most expensive candies
top5_expensive <- candy |>
  arrange(desc(pricepercent)) |>
  head(5)

top5_expensive
```

```
                         chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                        0      1       0              0      0
Nestle Smarties                  1      0       0              0      0
Ring pop                         0      1       0              0      0
Hershey's Krackel                1      0       0              0      0
Hershey's Milk Chocolate         1      0       0              0      0
                         crispedricewafer hard bar pluribus sugarpercent
Nik L Nip                               0    0   0        1        0.197
Nestle Smarties                         0    0   0        1        0.267
Ring pop                                0    1   0        0        0.732
Hershey's Krackel                       1    0   1        0        0.430
Hershey's Milk Chocolate                0    0   1        0        0.430
                         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
```

```
# Step 2: Among those 5, find the least popular
top5_expensive |>
  arrange(winpercent) |>
  head(1)
```

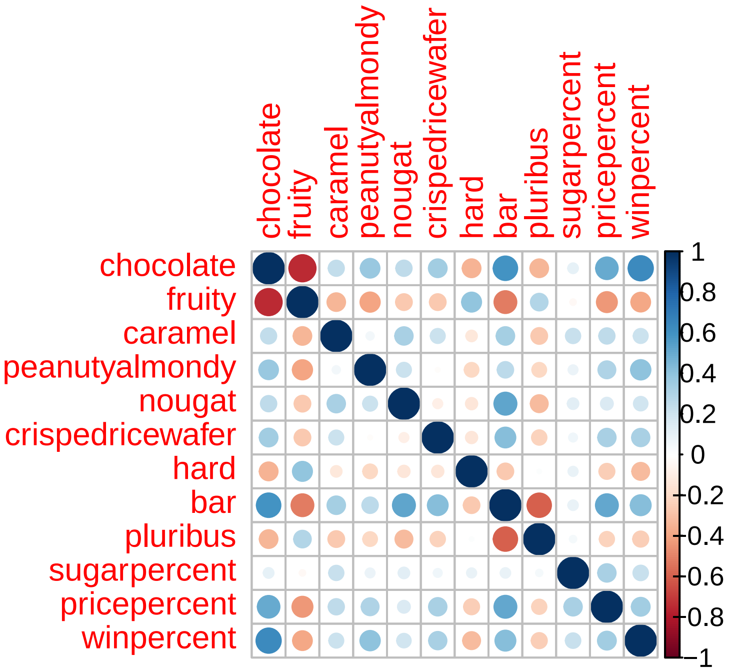
```
          chocolate fruity caramel peanutyalmondy nougat crispedricewafer hard
Nik L Nip         0      1       0              0      0                0    0
          bar pluribus sugarpercent pricepercent winpercent
Nik L Nip   0        1        0.197        0.976   22.44534
```

Exploring the correlation structure

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
# Select numeric columns only
numeric_candy <- candy[sapply(candy, is.numeric)]

# Compute correlation matrix
cij <- cor(numeric_candy)
corrplot(cij)
```



Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)? Fruity and chocolate are the most anti-correlated. Pluribus and bar are also anti-correlated. Also the things that are normally in a chocolate bar (caramel, peanut, nougat,crisp) are anticorrealted with fruit.

Q23. Similarly, what two variables are most positively correlated? The vairables are most correlated with themselves, but interms of differing variables. Winpercent and chocolate are the most positively correlated. Chocolate and bar are also positively correlated.

Principal Component Analysis

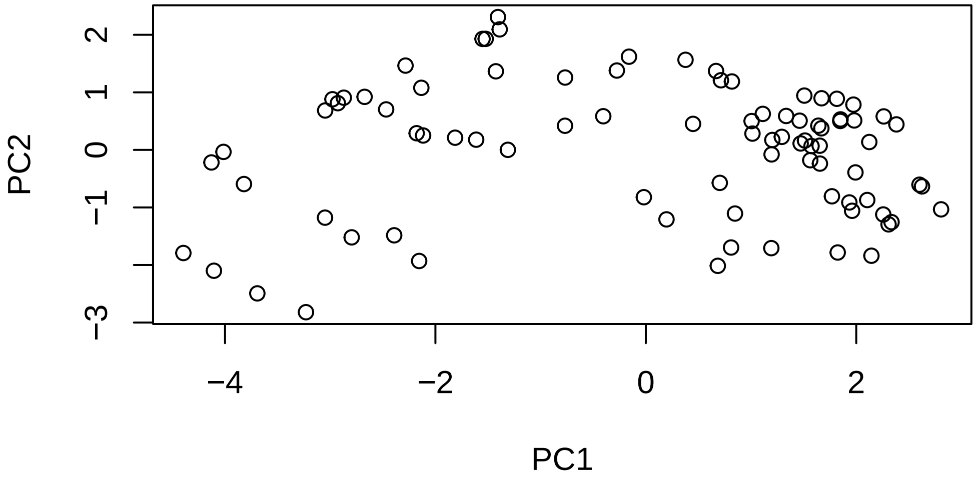
The main function for this is `prcomp()` and we want to set `scale=TRUE` here:

```
# Perform PCA on numeric columns only
numeric_candy <- candy[sapply(candy, is.numeric)]

pca <- prcomp(numeric_candy, scale. = TRUE)  # scale. = TRUE standardizes the variables
summary(pca)
```

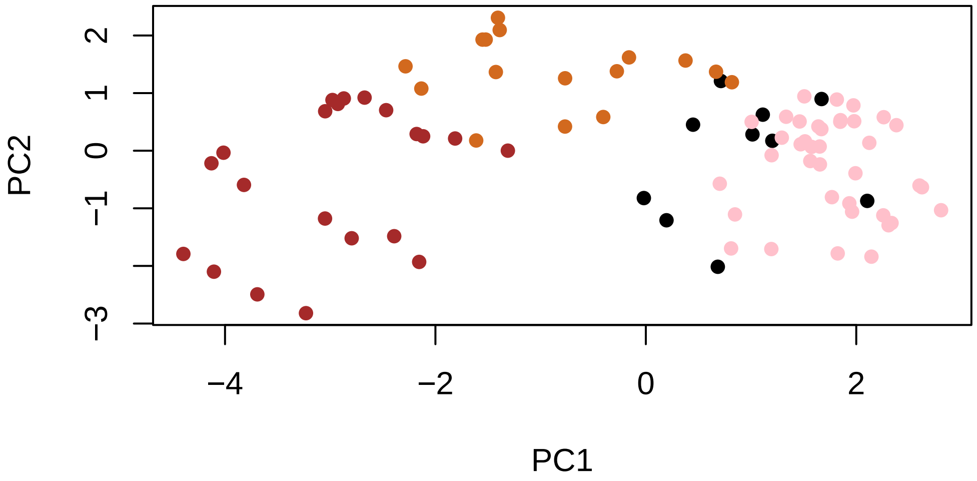
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
```

```
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 PC1 vs PC2
plot(pca$x[,1:2],)
```



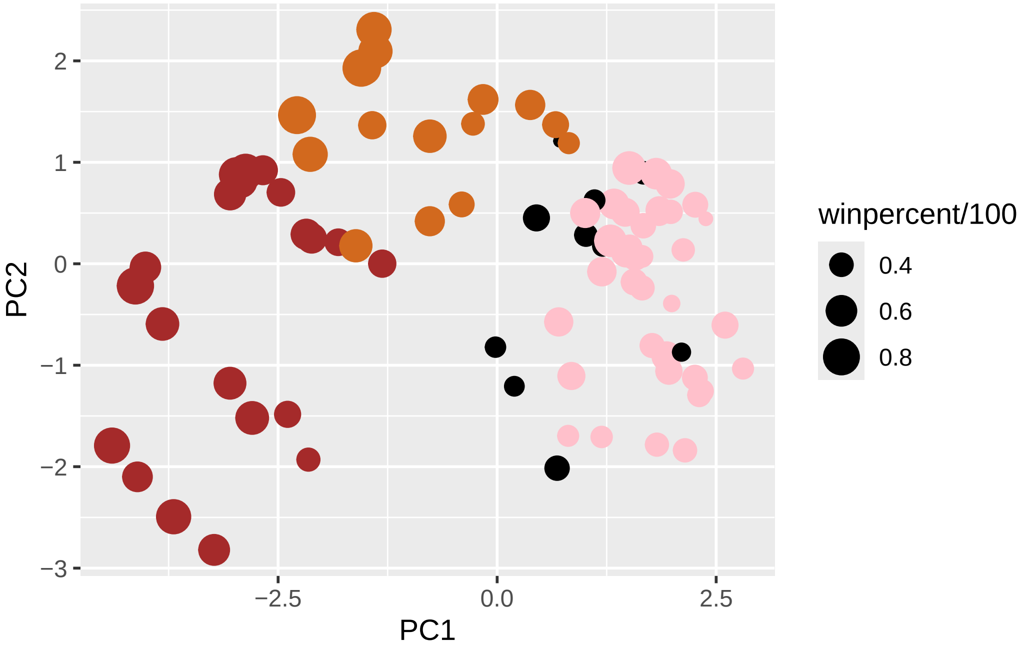
```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



Making a nicer plot using ggplot

```
my_data <- cbind(candy, pca$x[,1:3])
```

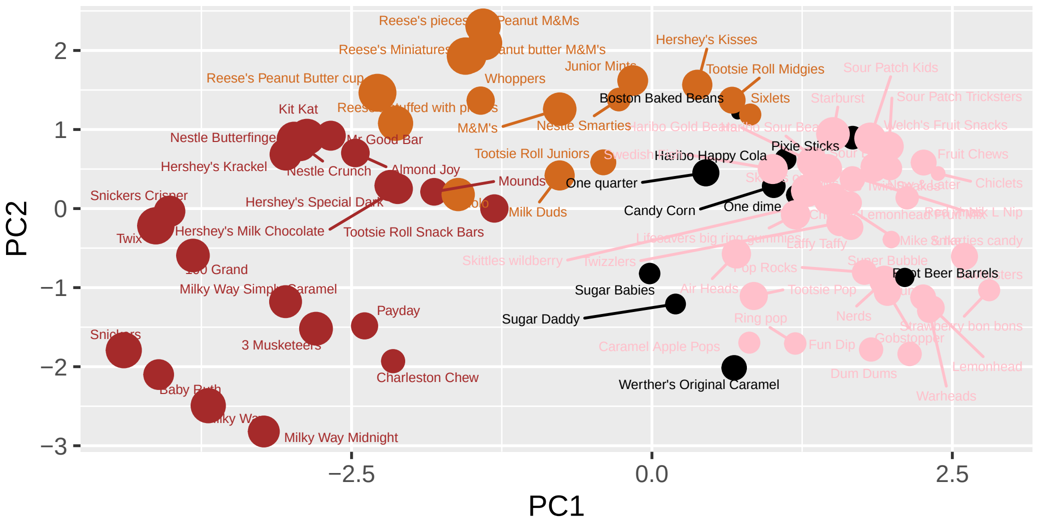
```
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



```r
library(ggrepel)

p + geom_text_repel(size=1.8, col=my_cols, max.overlaps = 30)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown),
       caption="Data from 538")
```

Halloween Candy PCA Space

Colored by type: chocolate bar (dark brown), chocolate other (light brown),



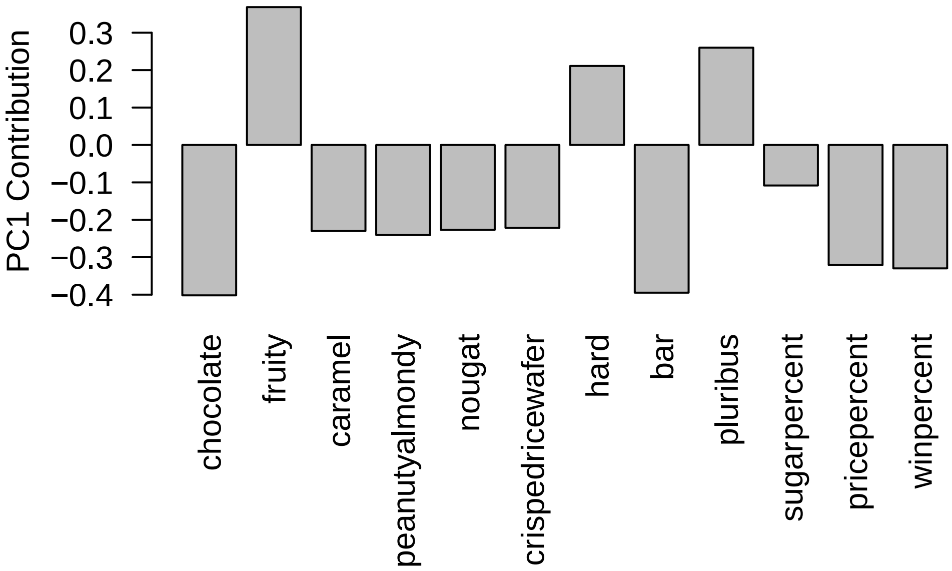
Data from 538

```r
library(plotly)
```

```

Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

Didn't include plotly because it was interfering with my pdf rendering.

```
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? The ones that are picked up strognly in the positive direction are fruity hard and pluribus. This makes sense to me because these variables are related. Most candies that are fruity are also hard and pluribus. Some examples are skittles, gobstoppers, and nerds.