ClassLab10: Halloween Candy Mini-Project

Nathaniel Nono (PID:A16782656)

Background

Candy voting tradition use as the basis for the dataset

Importing Candy Data

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

The functions `dim()`, `nrow()`, `table()` and `sum()` may be useful for answering the first 2 questions.

```
candy_file <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranl

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
```

```
nrow(candy)
```

```
[1] 85
```

There are 85 different candy types in the data set

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

```
# Shows a T/F (1 or 0) index
table(candy$fruity)
```

```
 0  1
47 38
```

There are 38 fruity candy types in the data

What is your favorate candy?

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

```
rownames(candy)
```

```
 [1] "100 Grand"                 "3 Musketeers"
 [3] "One dime"                  "One quarter"
 [5] "Air Heads"                 "Almond Joy"
 [7] "Baby Ruth"                 "Boston Baked Beans"
 [9] "Candy Corn"                "Caramel Apple Pops"
[11] "Charleston Chew"           "Chewey Lemonhead Fruit Mix"
[13] "Chiclets"                  "Dots"
[15] "Dum Dums"                  "Fruit Chews"
[17] "Fun Dip"                   "Gobstopper"
[19] "Haribo Gold Bears"         "Haribo Happy Cola"
[21] "Haribo Sour Bears"         "Haribo Twin Snakes"
[23] "Hershey's Kisses"          "Hershey's Krackel"
[25] "Hershey's Milk Chocolate"  "Hershey's Special Dark"
```

```
 [27] "Jawbusters"                   "Junior Mints"
 [29] "Kit Kat"                      "Laffy Taffy"
 [31] "Lemonhead"                    "Lifesavers big ring gummies"
 [33] "Peanut butter M&M's"          "M&M's"
 [35] "Mike & Ike"                   "Milk Duds"
 [37] "Milky Way"                    "Milky Way Midnight"
 [39] "Milky Way Simply Caramel"     "Mounds"
 [41] "Mr Good Bar"                  "Nerds"
 [43] "Nestle Butterfinger"          "Nestle Crunch"
 [45] "Nik L Nip"                    "Now & Later"
 [47] "Payday"                       "Peanut M&Ms"
 [49] "Pixie Sticks"                 "Pop Rocks"
 [51] "Red vines"                    "Reese's Miniatures"
 [53] "Reese's Peanut Butter cup"    "Reese's pieces"
 [55] "Reese's stuffed with pieces" "Ring pop"
 [57] "Rolo"                         "Root Beer Barrels"
 [59] "Runts"                        "Sixlets"
 [61] "Skittles original"            "Skittles wildberry"
 [63] "Nestle Smarties"              "Smarties candy"
 [65] "Snickers"                     "Snickers Crisper"
 [67] "Sour Patch Kids"              "Sour Patch Tricksters"
 [69] "Starburst"                    "Strawberry bon bons"
 [71] "Sugar Babies"                 "Sugar Daddy"
 [73] "Super Bubble"                 "Swedish Fish"
 [75] "Tootsie Pop"                  "Tootsie Roll Juniors"
 [77] "Tootsie Roll Midgies"         "Tootsie Roll Snack Bars"
 [79] "Trolli Sour Bites"            "Twix"
 [81] "Twizzlers"                    "Warheads"
 [83] "Welch's Fruit Snacks"         "Werther's Original Caramel"
 [85] "Whoppers"
```

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

```
[1] 76.7686
```

My favorite candy "Reese's pieces" has a win percent of 76.7686%

Find fruity candy with a winpercent above 50%

Dplyr package approach: Filter the data

```
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
```

```
fruit_win <- candy |>
  filter(winpercent > 50) |>
  filter(fruity == 1)

head(fruit_win)
```

```
                           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
                           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
                           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
```

Base R approach: Index the data

```
top.candy <- candy[candy$winpercent > 50,]
head(top.candy[top.candy$fruity == 1,])
```

```
                           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
                           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
                           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
```

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

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

```
[1] 76.7686
```

Kit kat has a win percent value of 76.7686%

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

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

```
[1] 49.6535
```

Tootsie Roll Snack Bars have a win percent value of 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?

```
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

Looks like the `winpercent` variable or column is measured on a different scale than everything else. Need to scale my data before doing any analysis like PCA

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

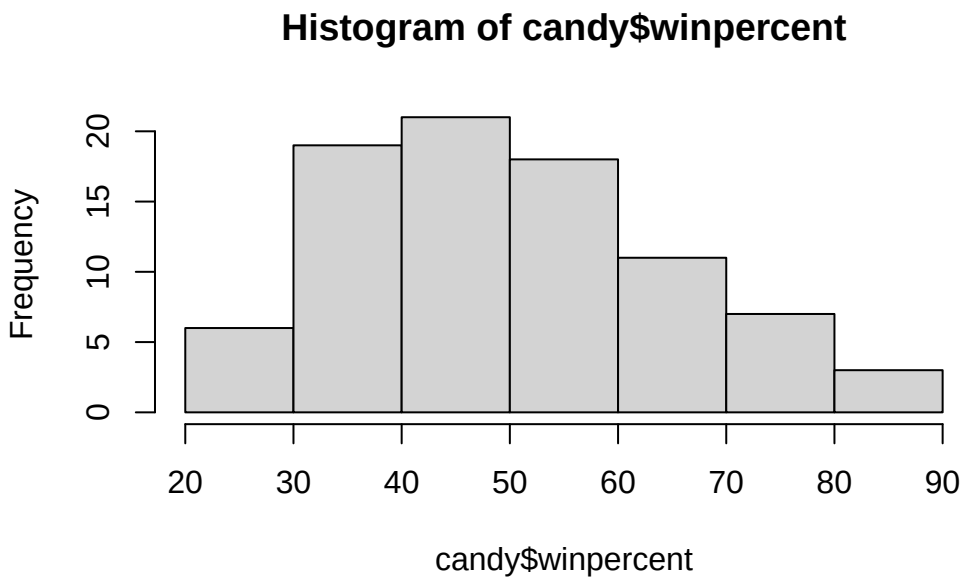
A zero represents that the candy is not a chocolate (FALSE logical) while a one represents that the candy is a chocolate (TRUE logical)

##Q8. Plot a histogram of winpercent values

We can do this a few ways (use base R or ggplot)

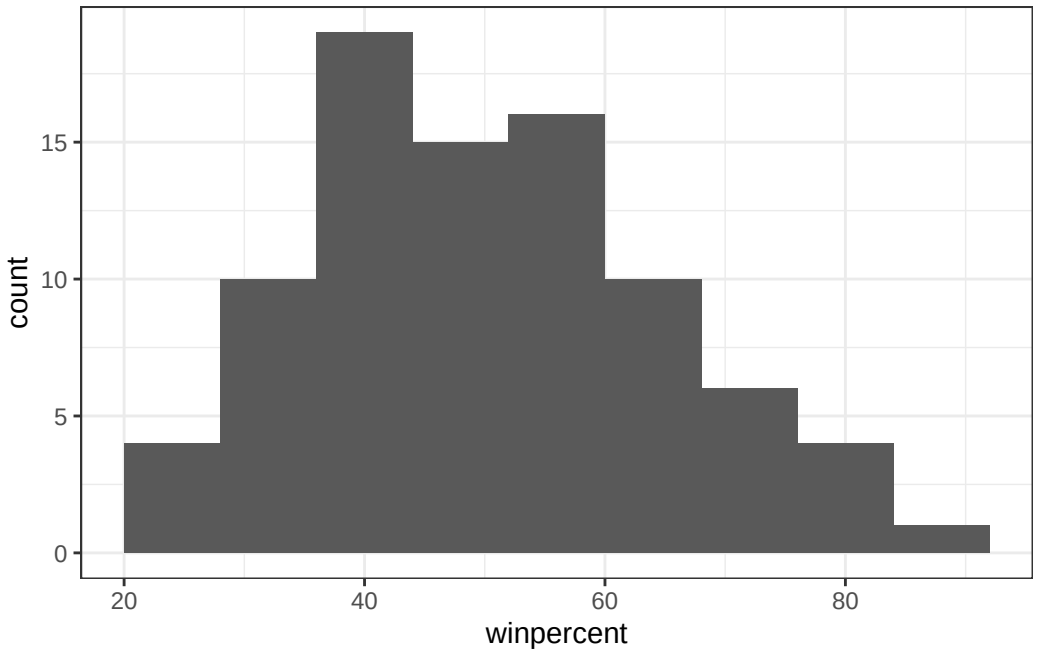
Base R approach: `hist()`

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



ggplot approach

```
library(ggplot2)

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



##Q9. Is the distribution of winpercent values symmetrical?

No it appears to be skewed

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

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

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

Because the distribution of the winpercent is not symmetrical (not normal) we should use the median to determine the center. The center of the distrbution is 47.83 which is below 50%

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

```
chocolate_candy <- candy |>
  filter(chocolate==1)

head(chocolate_candy)
```

```
                  chocolate fruity caramel peanutyalmondy nougat
100 Grand                 1      0       1              0      0
3 Musketeers              1      0       0              0      1
Almond Joy                1      0       0              1      0
Baby Ruth                 1      0       1              1      1
Charleston Chew           1      0       0              0      1
Hershey's Kisses          1      0       0              0      0
                  crispedricewafer hard bar pluribus sugarpercent pricepercent
100 Grand                        1    0   1        0        0.732        0.860
3 Musketeers                     0    0   1        0        0.604        0.511
Almond Joy                       0    0   1        0        0.465        0.767
Baby Ruth                        0    0   1        0        0.604        0.767
Charleston Chew                  0    0   1        0        0.604        0.511
Hershey's Kisses                 0    0   0        1        0.127        0.093
                  winpercent
100 Grand           66.97173
3 Musketeers        67.60294
Almond Joy          50.34755
Baby Ruth           56.91455
Charleston Chew     38.97504
Hershey's Kisses    55.37545
```

```
fruity_candy <- candy |>
  filter(fruity==1)

head(fruity_candy)
```

```
                          chocolate fruity caramel peanutyalmondy nougat
Air Heads                         0      1       0              0      0
Caramel Apple Pops                0      1       1              0      0
Chewey Lemonhead Fruit Mix        0      1       0              0      0
Chiclets                          0      1       0              0      0
Dots                              0      1       0              0      0
Dum Dums                          0      1       0              0      0
                          crispedricewafer hard bar pluribus sugarpercent
Air Heads                                0    0   0        0        0.906
Caramel Apple Pops                       0    0   0        0        0.604
Chewey Lemonhead Fruit Mix               0    0   0        1        0.732
Chiclets                                 0    0   0        1        0.046
Dots                                     0    0   0        1        0.732
Dum Dums                                 0    1   0        0        0.732
                          pricepercent winpercent
```

```
Air Heads                          0.511   52.34146
Caramel Apple Pops                 0.325   34.51768
Chewey Lemonhead Fruit Mix         0.511   36.01763
Chiclets                           0.325   24.52499
Dots                               0.511   42.27208
Dum Dums                           0.034   39.46056
```

```r
summary(chocolate_candy$winpercent)
```

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

```r
summary(fruity_candy$winpercent)
```

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

On average, the chocolate candy has a higher rank in win percent (60.92%) compared to fruity candy (44.12%)

##Q12. Is this difference statistically significant?

```r
# Perform a t-test for signficance
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 really small p-value we can conclude that there is a statistically significant difference between the win percent of chocolate and fruit candy

Overall Candy Rankings

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

Base R Approach

```
head(candy[order(candy$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
```

Dplyr Approach

```
least_candy <- candy |>
  arrange((winpercent))

head(least_candy, 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
```

The five least liked candy types in this data are "Nik L Nip", "Boston Baked Beans", "Chiclets", "Super Bubble", "Jawbusters" (Least -> Highest)

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

Base R Approach

```
head(candy[order(candy$winpercent, decreasing = T),], 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
```

Dplyr Approach

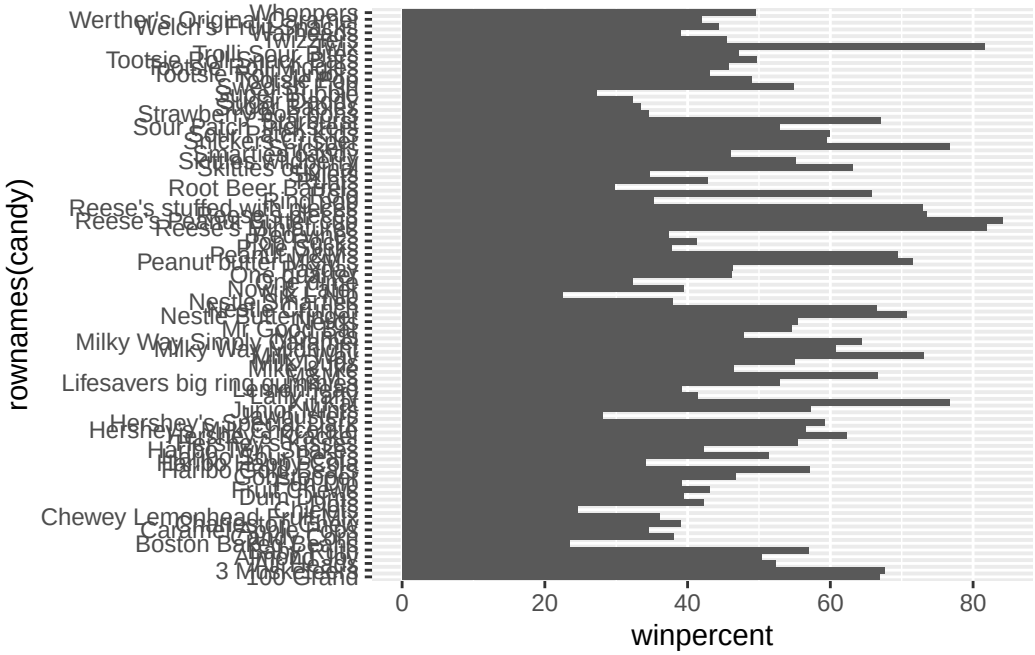
```
highest_candy <- candy |>
  arrange(desc(winpercent))

head(highest_candy, 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
```

The top five favorite candy types in this data set are "Reese's Peanut Butter cup", "Reese's Miniatures", "Twix", "Kit Kat", "Snickers" (Highest -> Lowest)

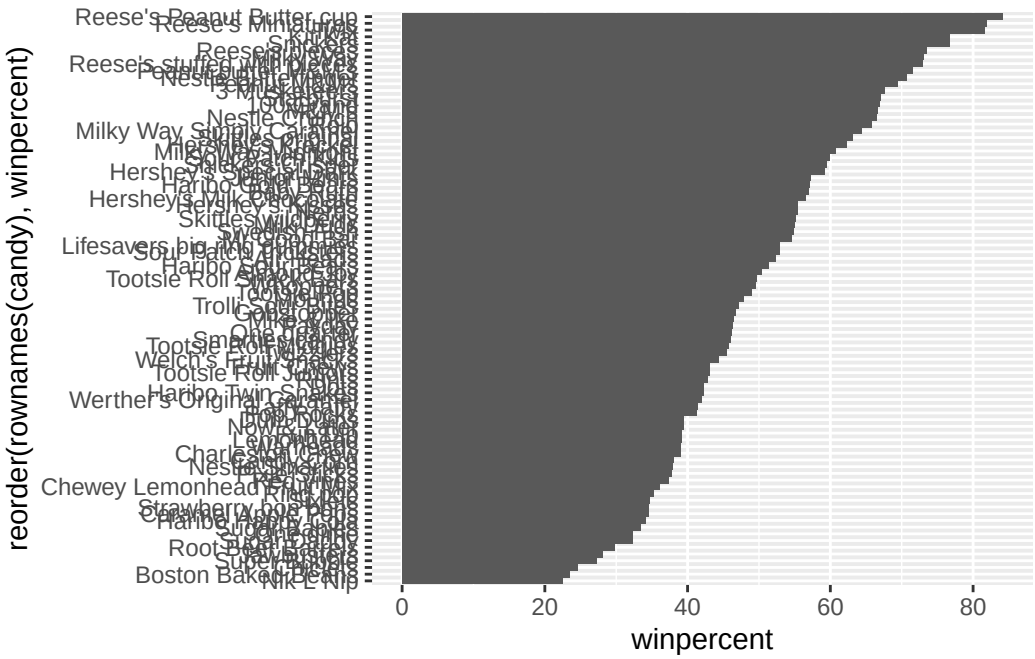
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

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 function to put the highest win percent on top
      reorder(rownames(candy), winpercent)) +
  geom_col()
```



Q17. What is the worst ranked chocolate candy?

I want a more specialized color scheme where I can see both chocolate and bar and fruity, etc. all from one plot.

- Roll our own color vector
- Add that color vector onto ggplot aes layer `fill =`

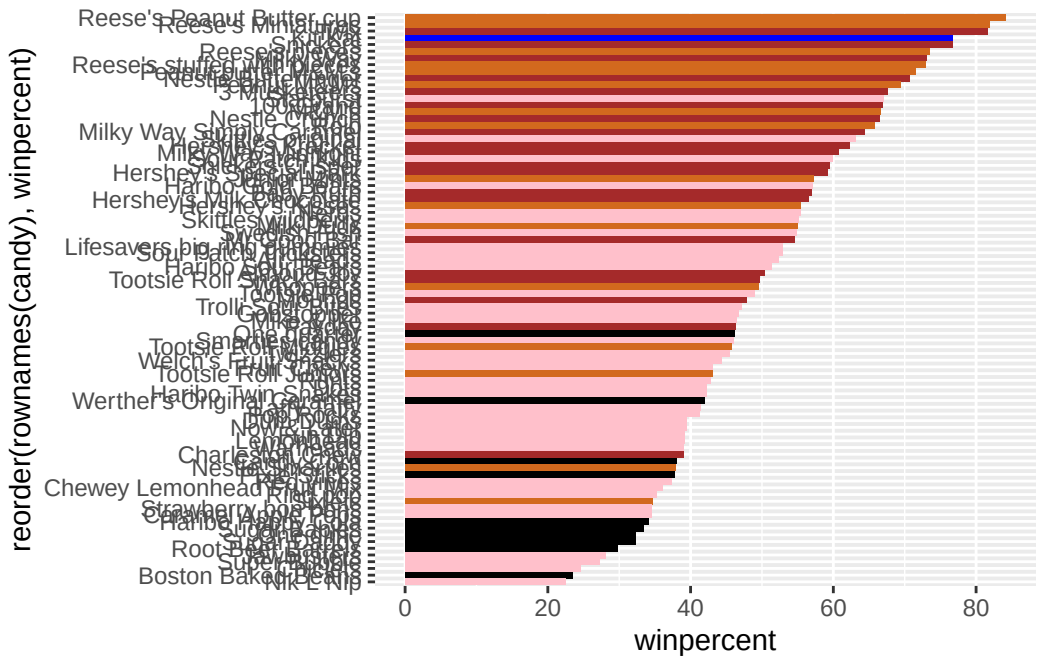
```
# Place holder color vector
my_cols <- rep('black', nrow(candy))
# Will override based on the index
my_cols[as.logical(candy$chocolate)] <- 'chocolate'
my_cols[as.logical(candy$bar)] <- 'brown'
my_cols[as.logical(candy$fruity)] <- 'pink'
```

Use blue for favorite candy = Kit Kat

- Get all the rownames in candy
- Find the point where it is the conditional is true
- Index that value and turn that to a blue color

```
# Use blue for favorite candy = Kit Kat
my_cols[rownames(candy) == 'Kit Kat'] <- 'blue'
```

```
ggplot(candy) +
  aes(winpercent,
      reorder(rownames(candy), winpercent)) +
  # Want the color in the geom layer
  geom_col(fill=my_cols)
```



The worst ranked chocolate candy are sixlits

Q18. What is the best ranked fruity candy?

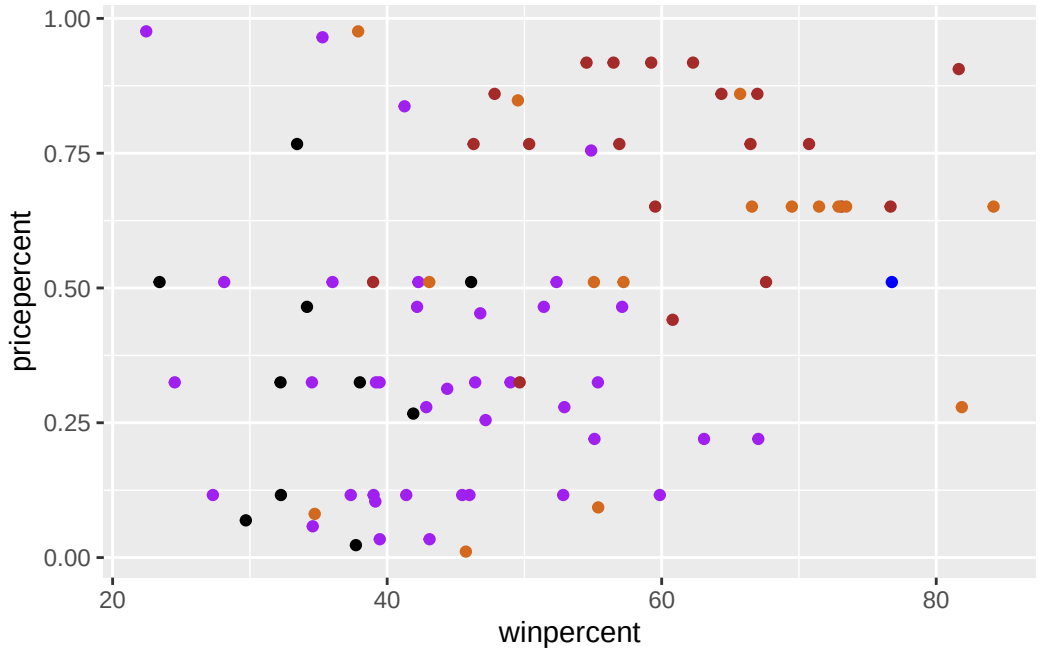
The best ranked fruity candy are starbursts

Takin a look at pricepercent

Plot of winpercent vs pricepercent to see what would be the ebst candy to buy

```
my_cols[as.logical(candy$fruity)] <- 'purple'
```

```
ggplot(candy) +
  aes(x = winpercent,
      y = pricepercent) +
  geom_point(col = my_cols)
```

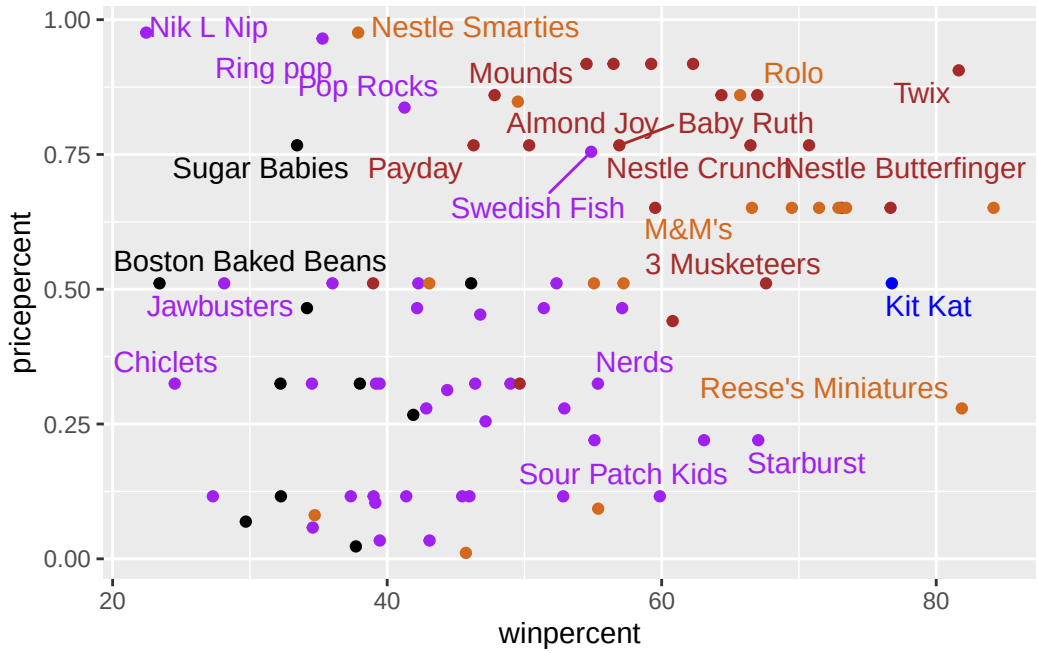


Add labels

- `label=rownames` to label
- `geom_text_repel` to prevent the

```
library(ggrepel)

ggplot(candy) +
  aes(x = winpercent,
      y = 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
```



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 highest ranked of candy in terms of winpercent for the least money are Reese's miniatures

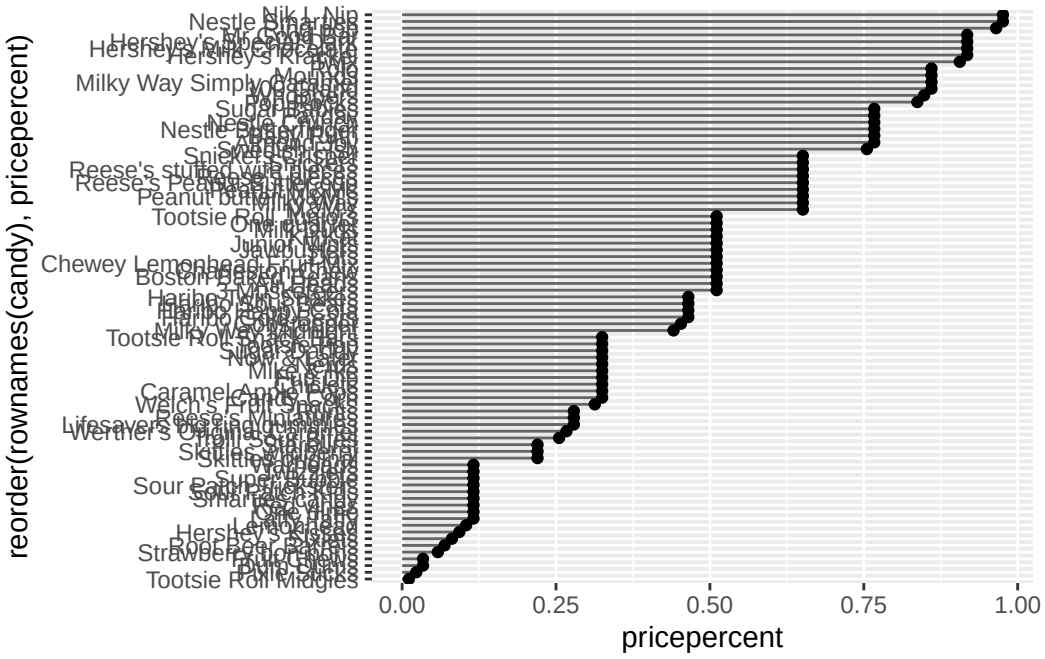
##Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[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 top 5 most expensive candy types are "Hershey's Milk Chocolate", Ring pop", "Nestle Smarties" and the least popular are "Nik L Nip"

```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```



Exploring the correlation structure

```
library(corrplot)
```

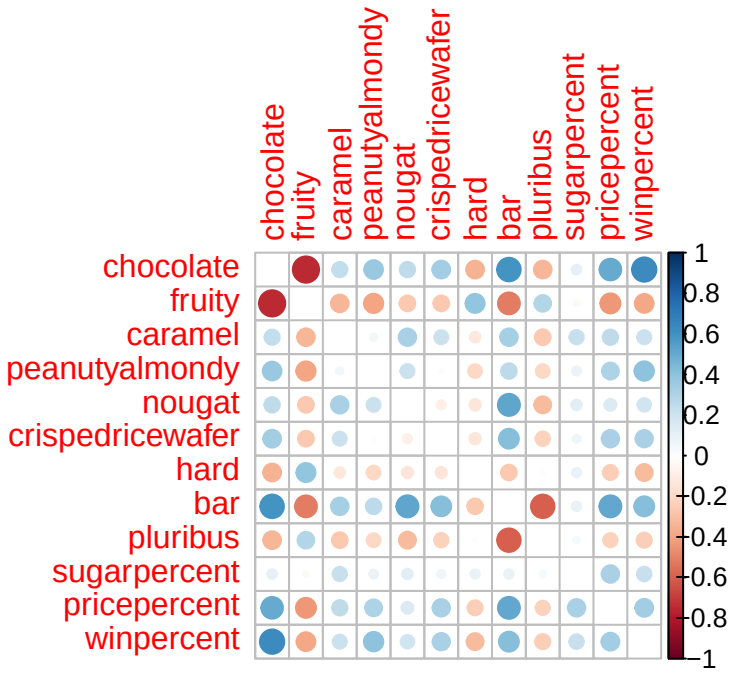
```
corrplot 0.95 loaded
```

```
# Correlation between i and j
cij <- cor(candy)
cij
```

```
                  chocolate      fruity     caramel peanutyalmondy      nougat
```

```
chocolate         1.0000000 -0.74172106  0.24987535     0.37782357  0.25489183
fruity           -0.7417211  1.00000000 -0.33548538    -0.39928014 -0.26936712
caramel           0.2498753 -0.33548538  1.00000000     0.05935614  0.32849280
peanutyalmondy    0.3778236 -0.39928014  0.05935614     1.00000000  0.21311310
nougat            0.2548918 -0.26936712  0.32849280     0.21311310  1.00000000
crispedricewafer  0.3412098 -0.26936712  0.21311310    -0.01764631 -0.08974359
hard             -0.3441769  0.39067750 -0.12235513    -0.20555661 -0.13867505
bar               0.5974211 -0.51506558  0.33396002     0.26041960  0.52297636
pluribus         -0.3396752  0.29972522 -0.26958501    -0.20610932 -0.31033884
sugarpercent      0.1041691 -0.03439296  0.22193335     0.08788927  0.12308135
pricepercent      0.5046754 -0.43096853  0.25432709     0.30915323  0.15319643
winpercent        0.6365167 -0.38093814  0.21341630     0.40619220  0.19937530
                  crispedricewafer        hard         bar    pluribus
chocolate              0.34120978 -0.34417691  0.59742114 -0.33967519
fruity                -0.26936712  0.39067750 -0.51506558  0.29972522
caramel                0.21311310 -0.12235513  0.33396002 -0.26958501
peanutyalmondy        -0.01764631 -0.20555661  0.26041960 -0.20610932
nougat                -0.08974359 -0.13867505  0.52297636 -0.31033884
crispedricewafer       1.00000000 -0.13867505  0.42375093 -0.22469338
hard                  -0.13867505  1.00000000 -0.26516504  0.01453172
bar                    0.42375093 -0.26516504  1.00000000 -0.59340892
pluribus              -0.22469338  0.01453172 -0.59340892  1.00000000
sugarpercent           0.06994969  0.09180975  0.09998516  0.04552282
pricepercent           0.32826539 -0.24436534  0.51840654 -0.22079363
winpercent             0.32467965 -0.31038158  0.42992933 -0.24744787
                  sugarpercent pricepercent winpercent
chocolate           0.10416906    0.5046754  0.6365167
fruity             -0.03439296   -0.4309685 -0.3809381
caramel             0.22193335    0.2543271  0.2134163
peanutyalmondy      0.08788927    0.3091532  0.4061922
nougat              0.12308135    0.1531964  0.1993753
crispedricewafer    0.06994969    0.3282654  0.3246797
hard                0.09180975   -0.2443653 -0.3103816
bar                 0.09998516    0.5184065  0.4299293
pluribus            0.04552282   -0.2207936 -0.2474479
sugarpercent        1.00000000    0.3297064  0.2291507
pricepercent        0.32970639    1.0000000  0.3453254
winpercent          0.22915066    0.3453254  1.0000000
```

```
corrplot(cij, diag = F)
```



How to read:

- (-1) and 1 = Perfectly negatively/positively correlated
- 0 = Perfectly uncorrelated
- `diag = F` = Turn off correlating valeus against itself (Redundant)
- `type = lower` = Turns off the upper one

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

The two variables that are anti-correlated are chocolate and fruity

Q23. Similarly, what two variables are most positively correlated

The two variables that are most positively correlated are chocolate and bar

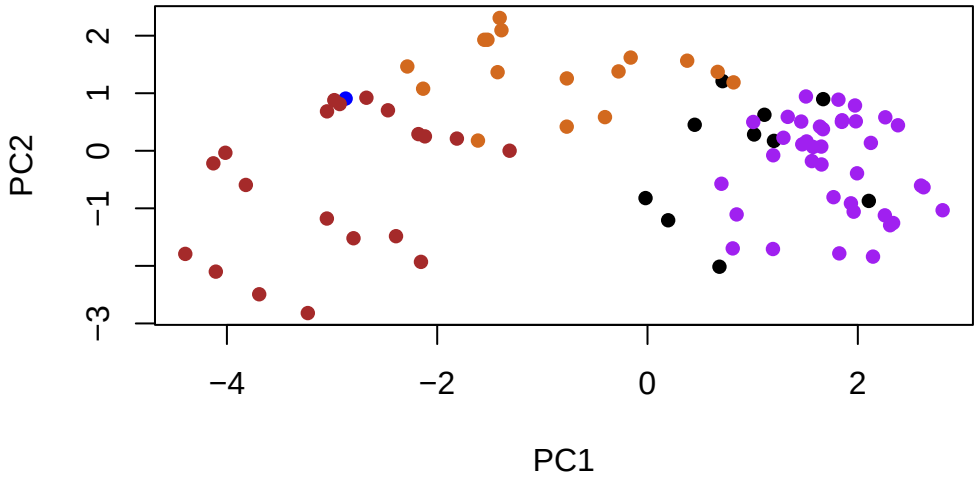
Principal Component Analysis

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

```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 0.7424 0.79830 0.85369
                           PC8     PC9    PC10    PC11    PC12
Standard deviation     0.74530 0.67824 0.62349 0.43974 0.39760
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

Base R approach

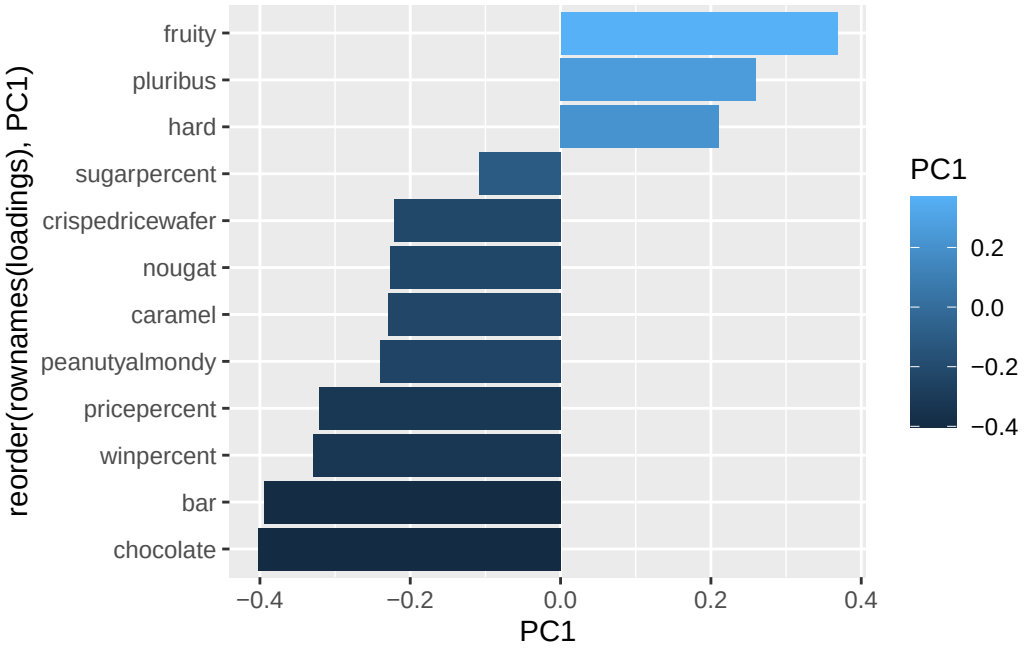
```
plot(pca$x[,1], pca$x[,2], col = my_cols, pch = 16,
     xlab='PC1',
     ylab='PC2')
```



ggplot approach

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

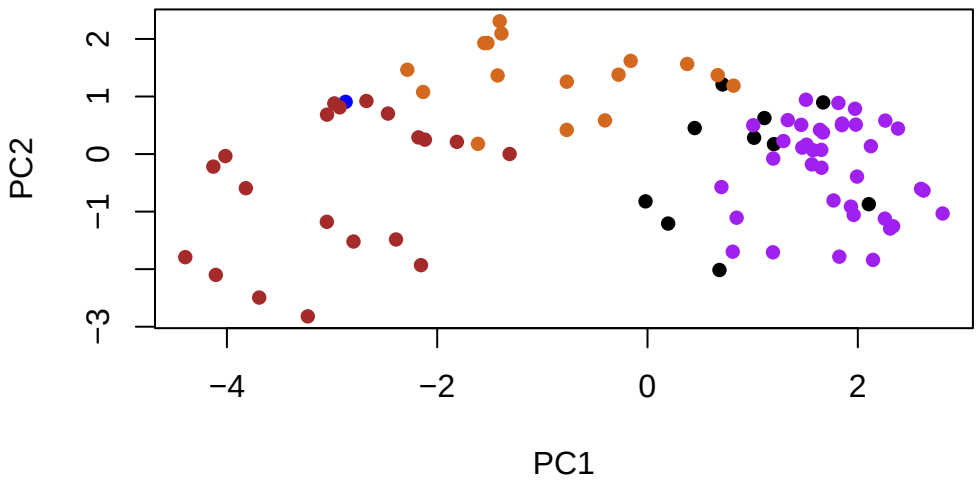
```
# Putting the candy data with the rotation data
loadings <- as.data.frame(pca$rotation)

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



Interpretation

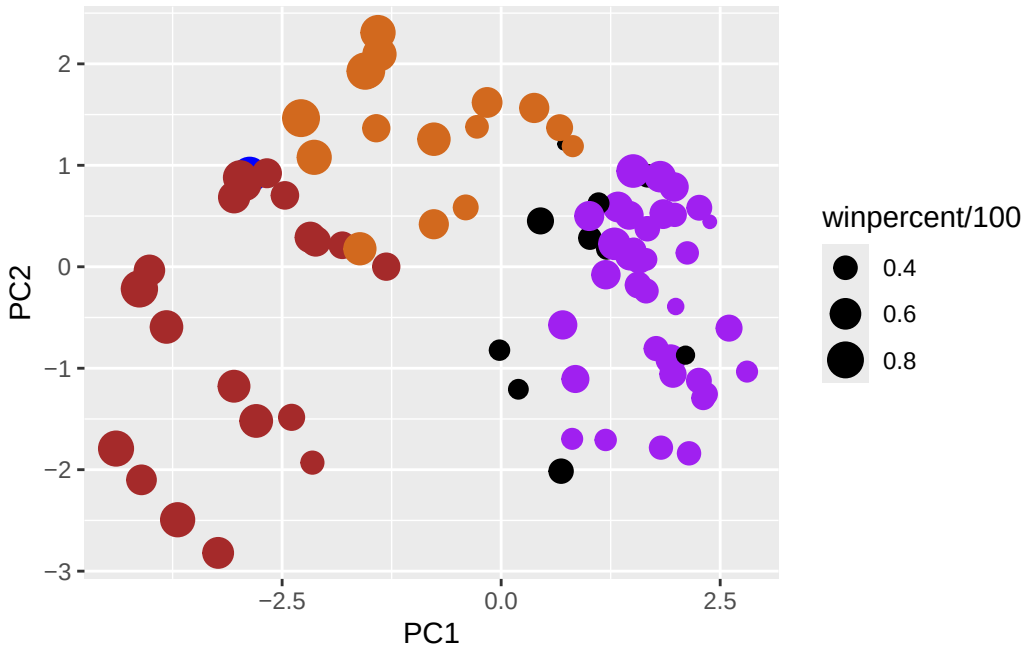
- Anything on the positive side is fruity and hard candy
- Anything on the negative side is chocolate and bar candy

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



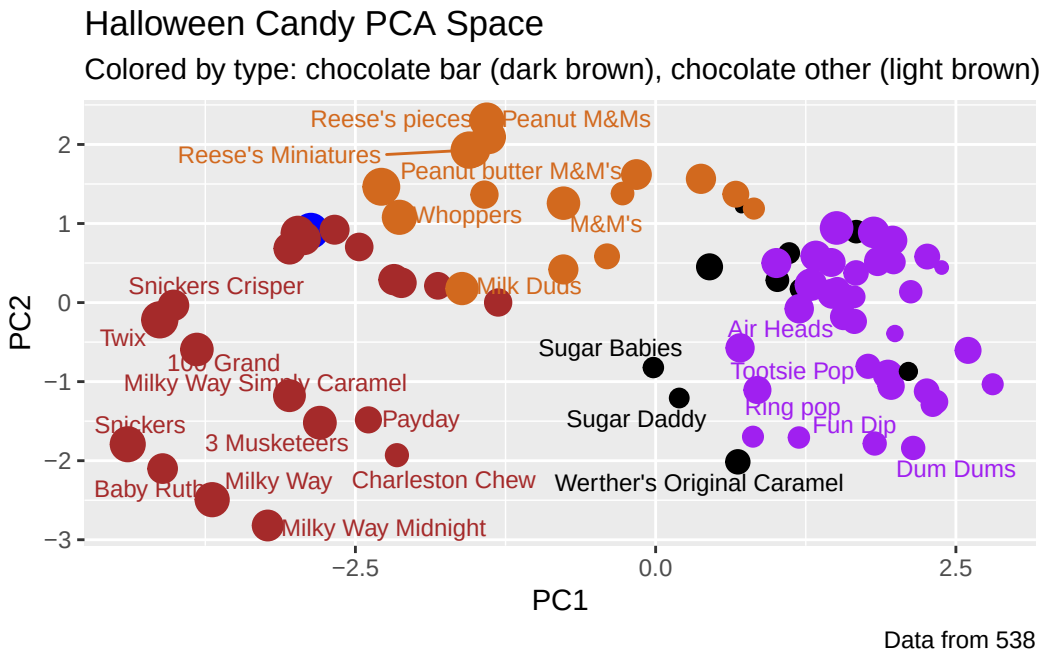

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

```
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
```




```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
  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")
```

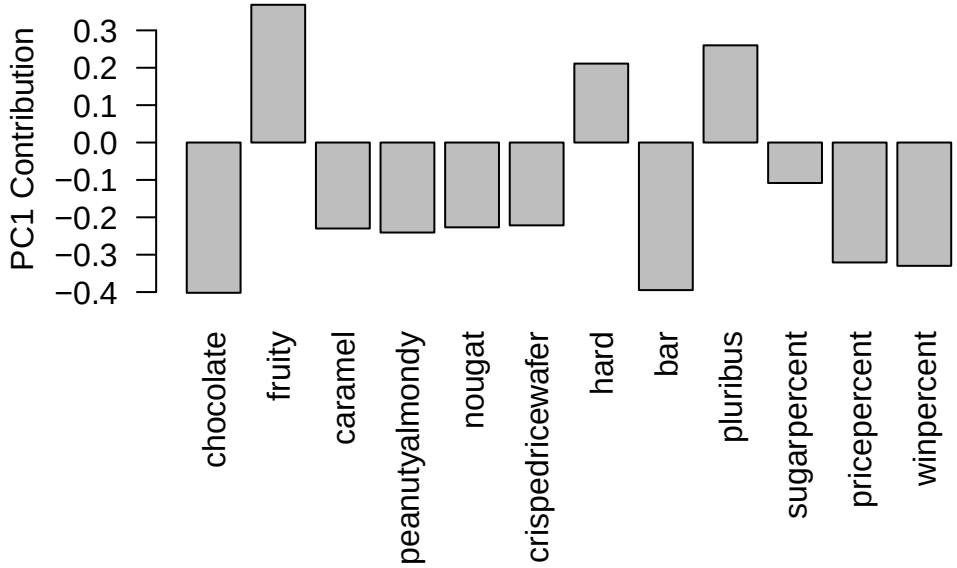
```
Warning: ggrepel: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




```
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
```

```
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 base R plot and the ggplot both pick up the same variables from PC1. It get the fruity, hard, and pluribus variables in the positive direction. Still having a little trouble doing it from scratch but it makes sense interpreting the results