class09: Spooky Days

Caliope Marin (PID: A13912583)

Today is

```
candy_file <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-p

head(candy_file)
```

```
             chocolate fruity caramel peanutyalmondy nougat crispedricewafer
100 Grand            1      0       1              0      0                 1
3 Musketeers         1      0       0              0      1                 0
One dime             0      0       0              0      0                 0
One quarter          0      0       0              0      0                 0
Air Heads            0      1       0              0      0                 0
Almond Joy           1      0       0              1      0                 0
             hard bar pluribus sugarpercent pricepercent winpercent
100 Grand       0   1        0        0.732        0.860   66.97173
3 Musketeers    0   1        0        0.604        0.511   67.60294
One dime        0   0        0        0.011        0.116   32.26109
One quarter     0   0        0        0.011        0.511   46.11650
Air Heads       0   0        0        0.906        0.511   52.34146
Almond Joy      0   1        0        0.465        0.767   50.34755
```

q.1 How many different candy types are in this dataset?

```
nrow(candy_file)
```

```
[1] 85
```

#There are 85 candy types in this dataset

```
table(candy_file["fruity"])
```

```
fruity
 0  1
47 38
```

```
sum(candy_file["fruity"])
```

```
[1] 38
```

Q2. there are 38 candies that are fruity

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

```
[1] NA
```

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

```
candy_file["Air Heads", ] $winpercent
```

```
[1] 52.34146
```

```
library(dplyr)
```

```
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
candy_file |>
  filter(rownames(candy_file)%in% c("Dum Dums", "Twix")) |>
  select(winpercent)
```

```
          winpercent
Dum Dums   39.46056
Twix       81.64291
```

Q4. What is the winpercent value for "Kit Kat"? Q5. What is the winpercent value for "Tootsie Roll Snack Bars"?

```
candy_file |>
  filter(rownames(candy_file)%in% c("Kit Kat", "Tootsie Roll Snack Bars")) |>
  select(winpercent)
```

```
                        winpercent
Kit Kat                    76.7686
Tootsie Roll Snack Bars    49.6535
```

```
#The win percent for Kit Kat is 76.76% and Tootsie Roll is # 49.65%
```

The %in% operator is useful for checking the intersection of two vectors

```
c("barry","liz", "chandra") %in% c("paul", "alice", "liz")
```

```
[1] FALSE  TRUE FALSE
```

```
candy_file |>
  filter(winpercent > 75) |>
  filter(pricepercent < 0.5)
```

```
                  chocolate fruity caramel peanutyalmondy nougat
Reese's Miniatures         1      0       0              1      0
                  crispedricewafer hard bar pluribus sugarpercent pricepercent
Reese's Miniatures                0    0   0        0        0.034        0.279
                  winpercent
Reese's Miniatures   81.86626
```

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_file)
```

Table 1: Data summary

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

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.00	1.00	
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00	
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
nougat	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00	
crispedricewafer	0	1	0.08	0.28	0.00	0.00	0.00	0.00	1.00	
hard	0	1	0.18	0.38	0.00	0.00	0.00	0.00	1.00	
bar	0	1	0.25	0.43	0.00	0.00	0.00	0.00	1.00	
pluribus	0	1	0.52	0.50	0.00	0.00	1.00	1.00	1.00	
sugarpercent	0	1	0.48	0.28	0.01	0.22	0.47	0.73	0.99	
pricepercent	0	1	0.47	0.29	0.01	0.26	0.47	0.65	0.98	
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18	

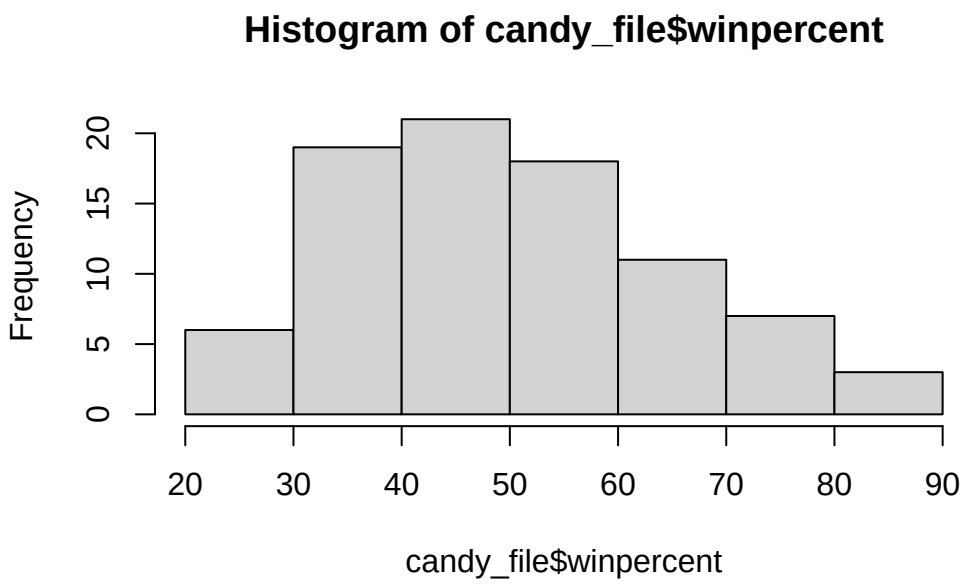
```
# Q6. The column that is on a different scale is the #winpercent because it is in percentage
```

Q7. What do you think a zero and one represent for the candy$chocolate column?
The "0" and "1" are binary for "False" and "true", respectively.

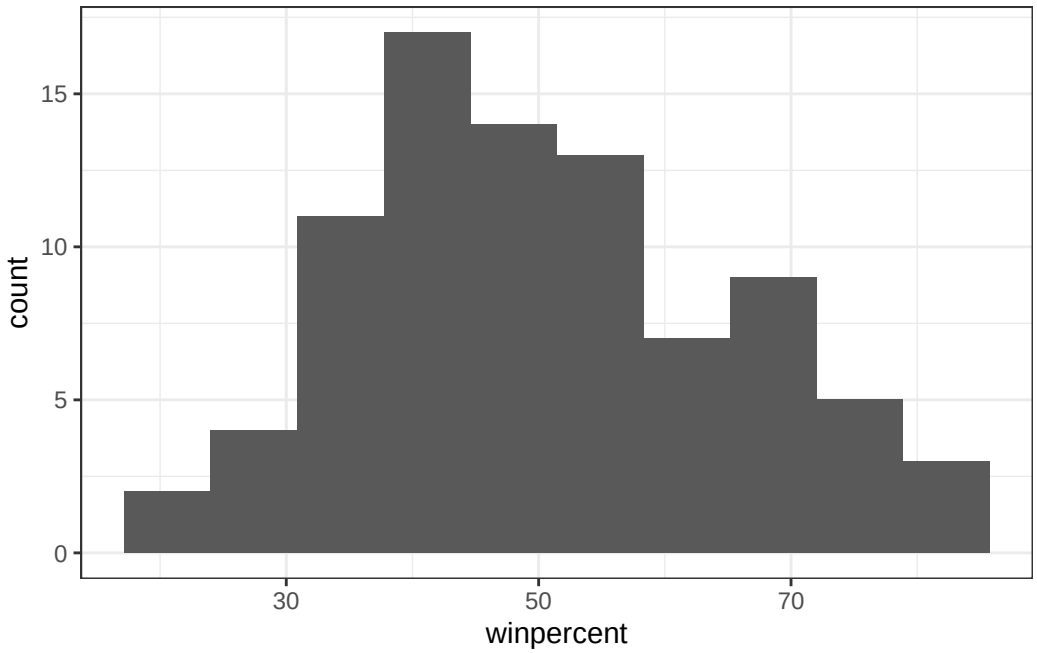
Q8. Plot a histogram of winpercent values

We can do this a few different ways: with base R `hist()`or with `ggplot()`

```
hist(candy_file$winpercent)
```




```
library(ggplot2)
ggplot(candy_file)+
  aes(winpercent)+
  geom_histogram(bins=10)+
  theme_bw()
```



Q9. Is the distribution of winpercent values symmetrical? Q10. Is the center of the distribution above or below 50%?

#Q.9 no the distribution is not symmetrical #Q.10 the center of the distribution below 50%

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

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

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

```
inds <-as.logical(candy_file$chocolate)
candy_file[inds,]$winpercent
```

```
 [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
```

```
candy_file$chocolate==1
```

```
 [1]  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE FALSE
[13] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE  TRUE
[25]  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE  TRUE FALSE  TRUE
[37]  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE  TRUE
[49] FALSE FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE  TRUE
[61] FALSE FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
[73] FALSE FALSE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE FALSE FALSE FALSE
[85]  TRUE
```

```
candy_file[inds,]$winpercent
```

```
 [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
```

```
candy_file |>
  filter(chocolate==1)|>
  select(winpercent)
```

	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
Hershey's Krackel	62.28448
Hershey's Milk Chocolate	56.49050
Hershey's Special Dark	59.23612
Junior Mints	57.21925
Kit Kat	76.76860
Peanut butter M&M's	71.46505
M&M's	66.57458
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
Nestle Butterfinger           70.73564
Nestle Crunch                 66.47068
Peanut M&Ms                   69.48379
Reese's Miniatures            81.86626
Reese's Peanut Butter cup     84.18029
Reese's pieces                73.43499
Reese's stuffed with pieces   72.88790
Rolo                          65.71629
Sixlets                       34.72200
Nestle Smarties               37.88719
Snickers                      76.67378
Snickers Crisper              59.52925
Tootsie Pop                   48.98265
Tootsie Roll Juniors          43.06890
Tootsie Roll Midgies          45.73675
Tootsie Roll Snack Bars       49.65350
Twix                          81.64291
Whoppers                      49.52411
```

```r
candy_file |>
  filter(fruity==1)|>
  select(winpercent)
```

```
                          winpercent
Air Heads                   52.34146
Caramel Apple Pops          34.51768
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 Sour Bears           51.41243
Haribo Twin Snakes          42.17877
Jawbusters                  28.12744
Laffy Taffy                 41.38956
Lemonhead                   39.14106
```

```
Lifesavers big ring gummies   52.91139
Mike & Ike                    46.41172
Nerds                         55.35405
Nik L Nip                     22.44534
Now & Later                   39.44680
Pop Rocks                     41.26551
Red vines                     37.34852
Ring pop                      35.29076
Runts                         42.84914
Skittles original             63.08514
Skittles wildberry            55.10370
Smarties candy                45.99583
Sour Patch Kids               59.86400
Sour Patch Tricksters         52.82595
Starburst                     67.03763
Strawberry bon bons           34.57899
Super Bubble                  27.30386
Swedish Fish                  54.86111
Tootsie Pop                   48.98265
Trolli Sour Bites             47.17323
Twizzlers                     45.46628
Warheads                      39.01190
Welch's Fruit Snacks          44.37552
```

#Q.11 Yes chocolate has a higher preference than fruity candies

```
inds<-candy_file$chocolate==1
choc.win <-candy_file[inds,]$winpercent

inds<-candy_file$fruity==1
fruit.win <-candy_file[inds,]$winpercent
#Then I could compare these
summary(choc.win)
```

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

```
summary(fruit.win)
```

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

```
t.test(choc.win, fruit.win)
```

```
    Welch Two Sample t-test

data:  choc.win and fruit.win
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
```

Q12. Is this difference statistically significant?

#Yes the pvalue is really low -> 2.871-08 and chocolate #is more preferred than fruity candy. #

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

There are two related functions that are useful here `sort()` and `order()`

```
play <- c(2,1,5,3)
sort(play)
```

```
[1] 1 2 3 5
```

```
order(play)
```

```
[1] 2 1 4 3
```

```
#Its giving you the number in the row of each variable not the actual character
```

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

```
[1] 1 2 3 5
```

```
l<-c("c","a","b")
sort(l)
```

```
[1] "a" "b" "c"
```

```
order(l)
```

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

```
n <-c("d","a")
n[order(n)]
```

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

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

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

```
                  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
Root Beer Barrels         0      0       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
Root Beer Barrels                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
```

```r
#these are the least fave candies below
```

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

```r
#inds means the bucket of a vector
inds <- order(candy_file$winpercent, decreasing= T)
head(candy_file[inds,],5)
```

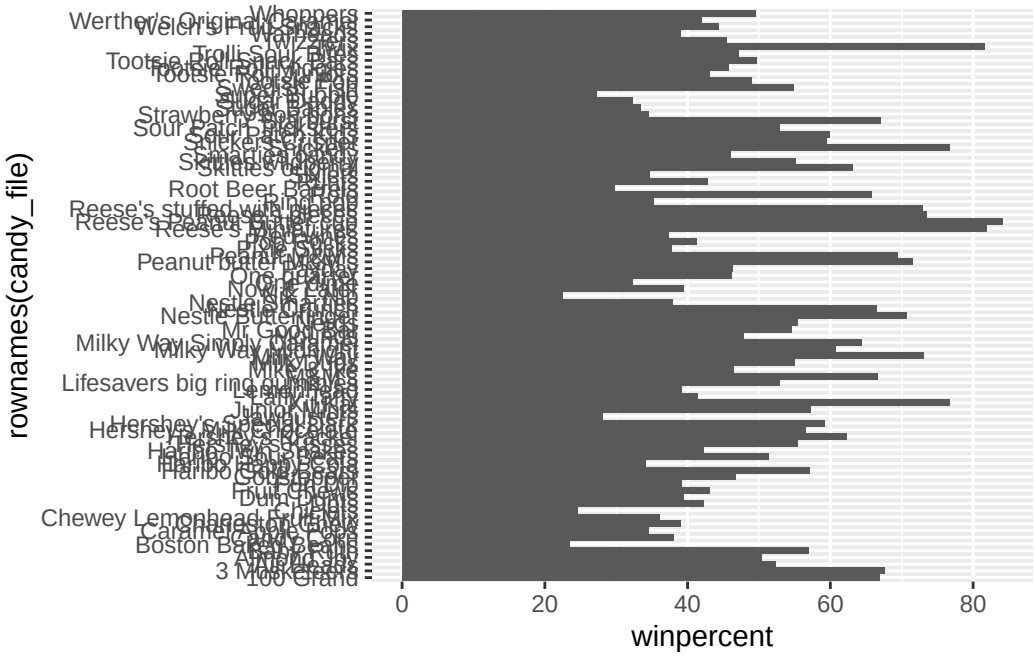
```
                        chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           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
```

```
#these are the fave candies above
```

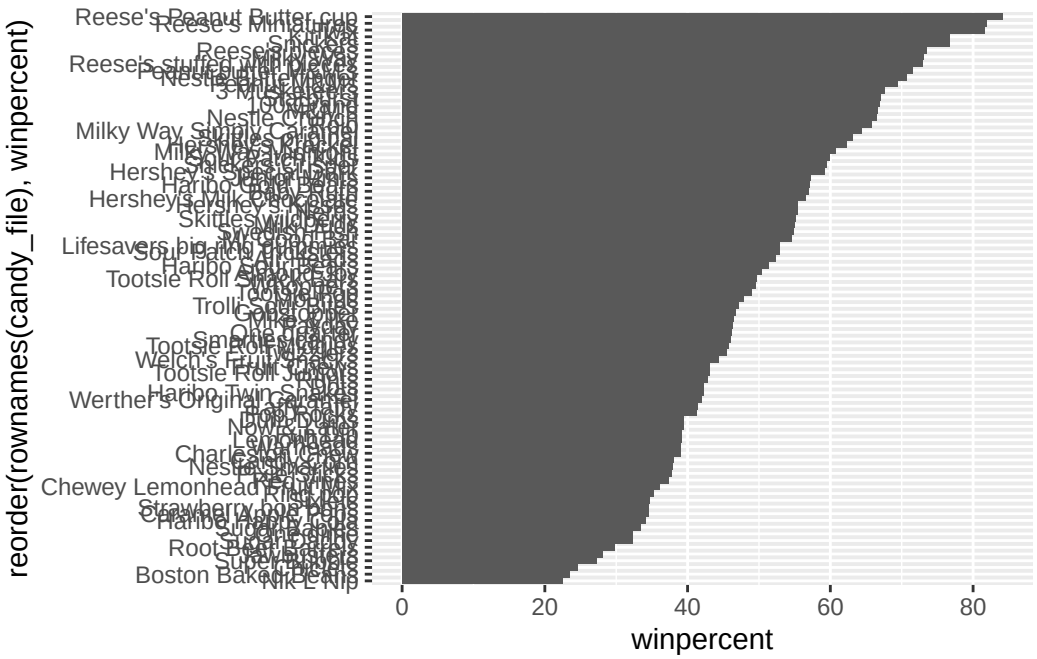
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

ggplot(candy_file) +
  aes(y=rownames(candy_file), x=winpercent ) +
  geom_col()
```



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

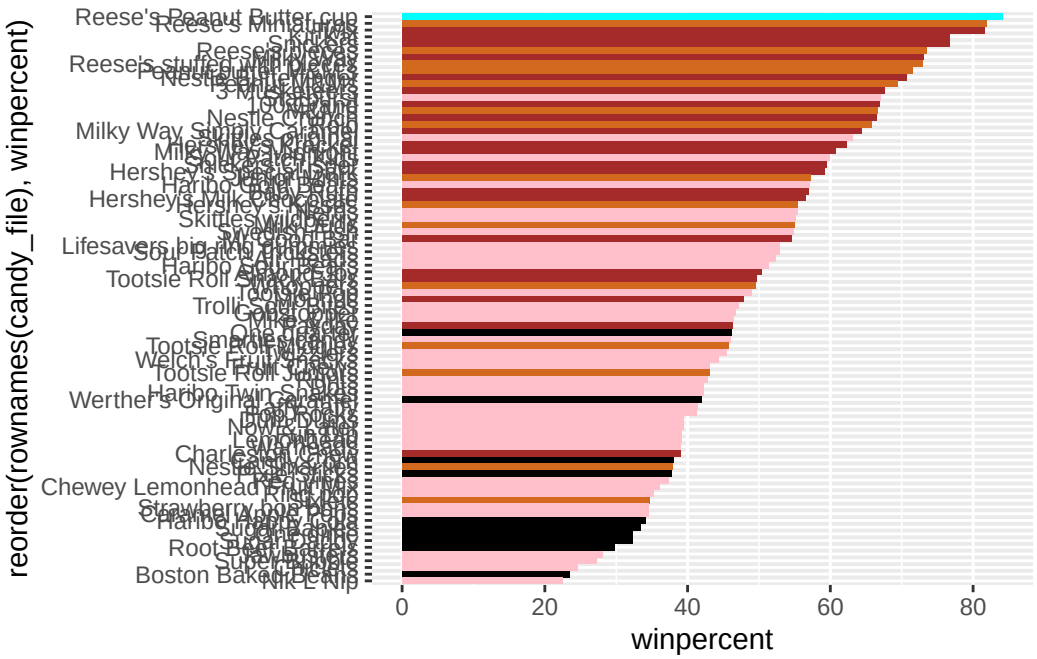
```
ggplot(candy_file) +
  aes(winpercent, reorder(rownames(candy_file),winpercent)) +
  geom_col()
```




```
#my fave candy is `Reese's Peanut Butter cup`
my_cols=rep("black", nrow(candy_file))

my_cols[as.logical(candy_file$chocolate)] = "chocolate"
my_cols[as.logical(candy_file$bar)] = "brown"
my_cols[as.logical(candy_file$fruity)] = "pink"
my_cols [rownames(candy_file)=="Reese's Peanut Butter cup"] <- "cyan"
```

Q. Color your favorite candy

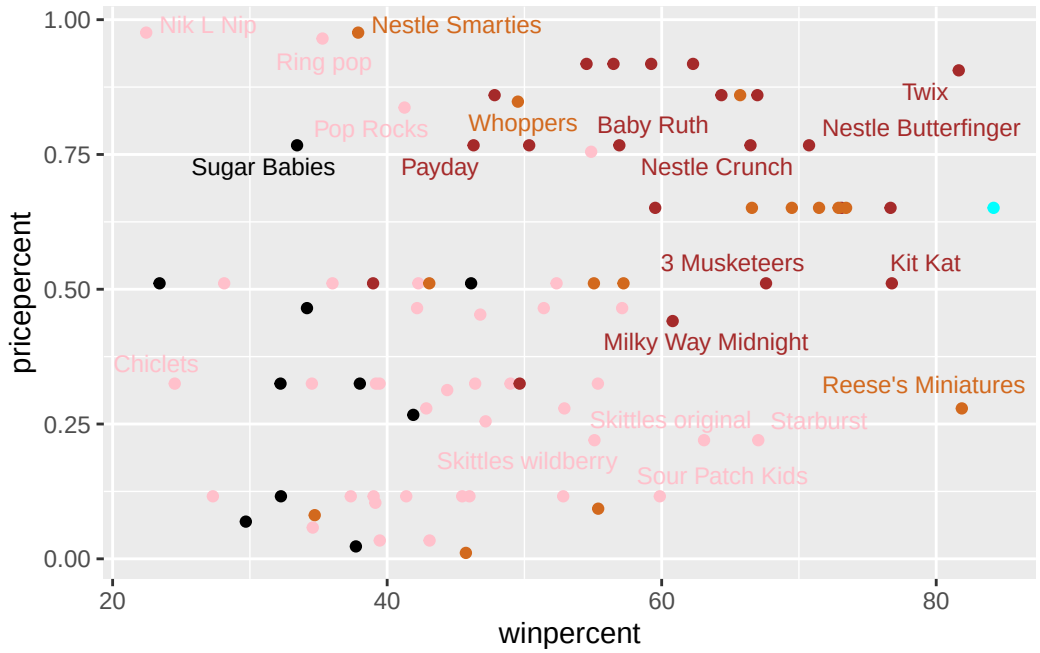
```
ggplot(candy_file) +
  aes(x=winpercent, y=reorder(rownames(candy_file),winpercent)) +
  geom_col(fill=my_cols)
```



Q17. What is the worst ranked chocolate candy? #Sixlets Q18. What is the best ranked fruity candy? #Starbursts

```
library(ggrepel)

# How about a plot of price vs win
ggplot(candy_file) +
  aes(winpercent, pricepercent, label=rownames(candy_file)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 5)
```

```
Warning: ggrepel: 65 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's is the biggest bag for your buck

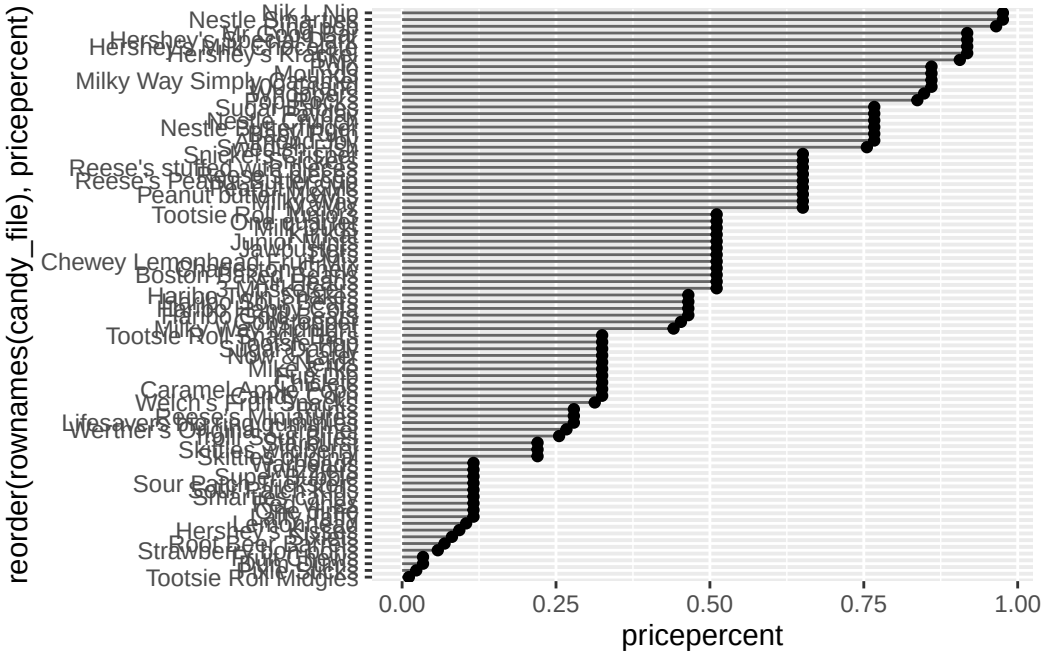
Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular? #Nik L Nip is the most expensive and least popular

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

```
                      pricepercent winpercent
Nik L Nip                    0.976   22.44534
Nestle Smarties              0.976   37.88719
Ring pop                     0.965   35.29076
Hershey's Krackel            0.918   62.28448
Hershey's Milk Chocolate     0.918   56.49050
```

Q21. Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping geom_col() for geom_point() + geom_segment().

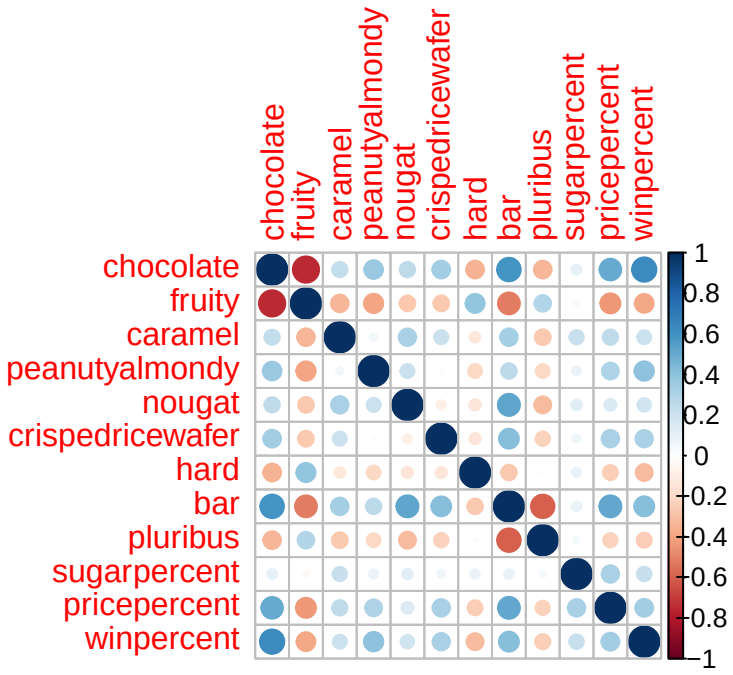
```
ggplot(candy_file) +
  aes(pricepercent, reorder(rownames(candy_file), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy_file), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```




```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
cij <- cor(candy_file)
corrplot(cij)
```



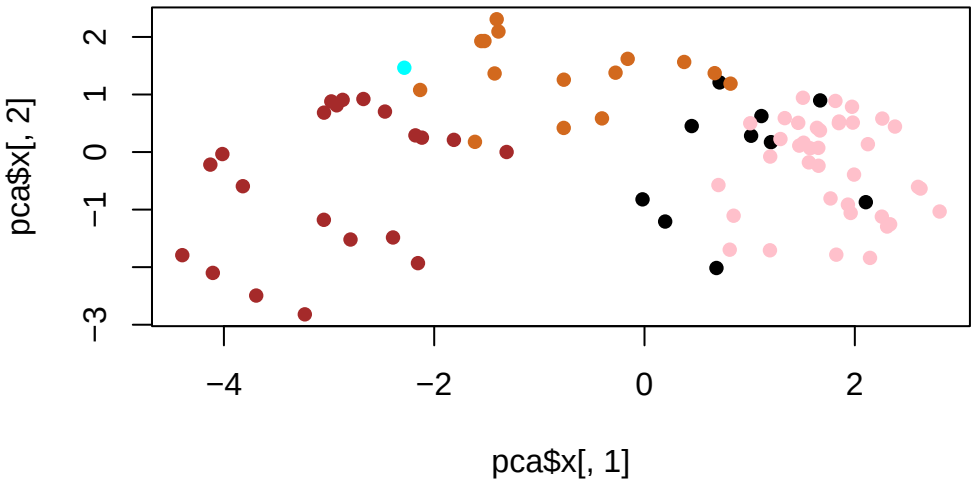
Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)? #The anti-correlated variables are fruity and chocolate Q23. Similarly, what two variables are most positively correlated? The highly correlated variables are chocolate and bar.

##Principal Component Analysis##

```
pca <- prcomp(candy_file, scale=TRUE)
summary(pca)
```

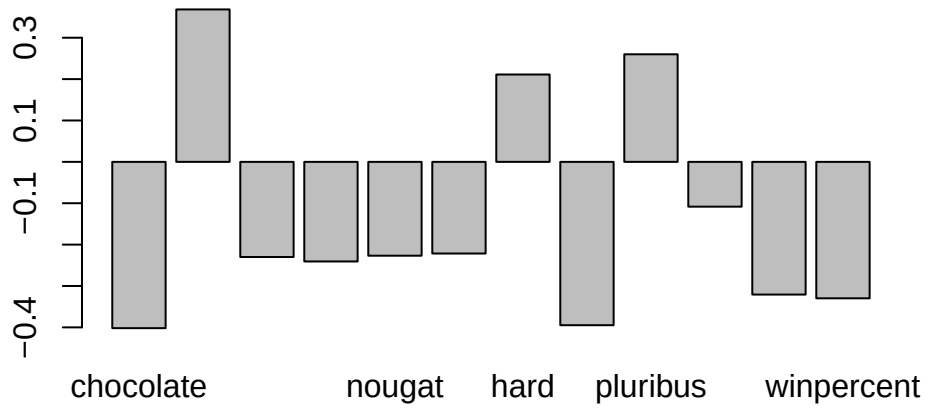
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 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
```

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

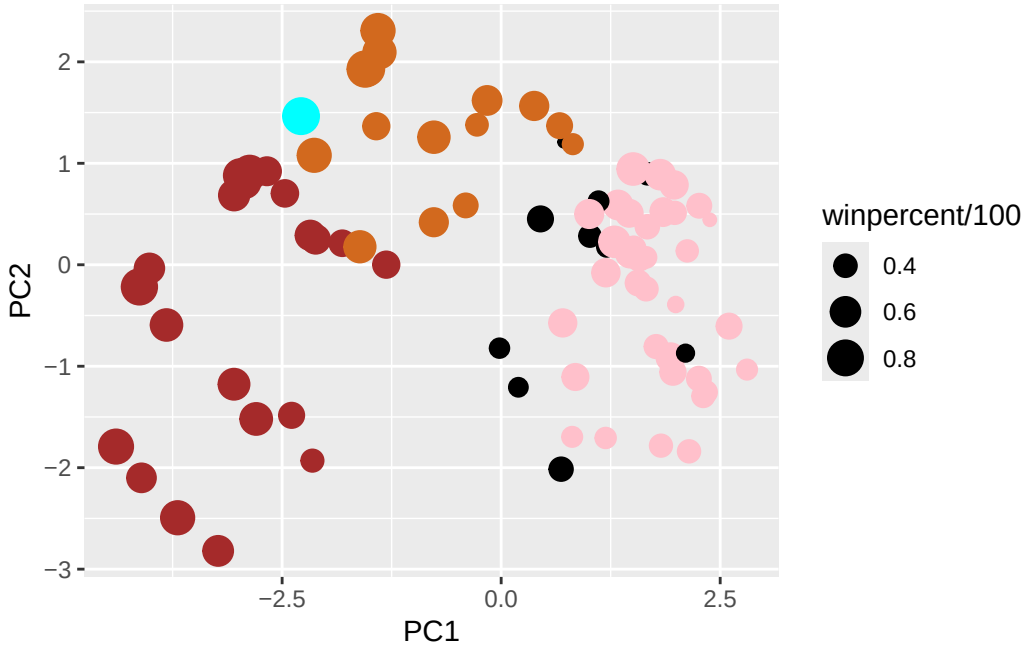



```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy_file, pca$x[,1:3])
```

```
barplot(pca$rotation[,1])
```

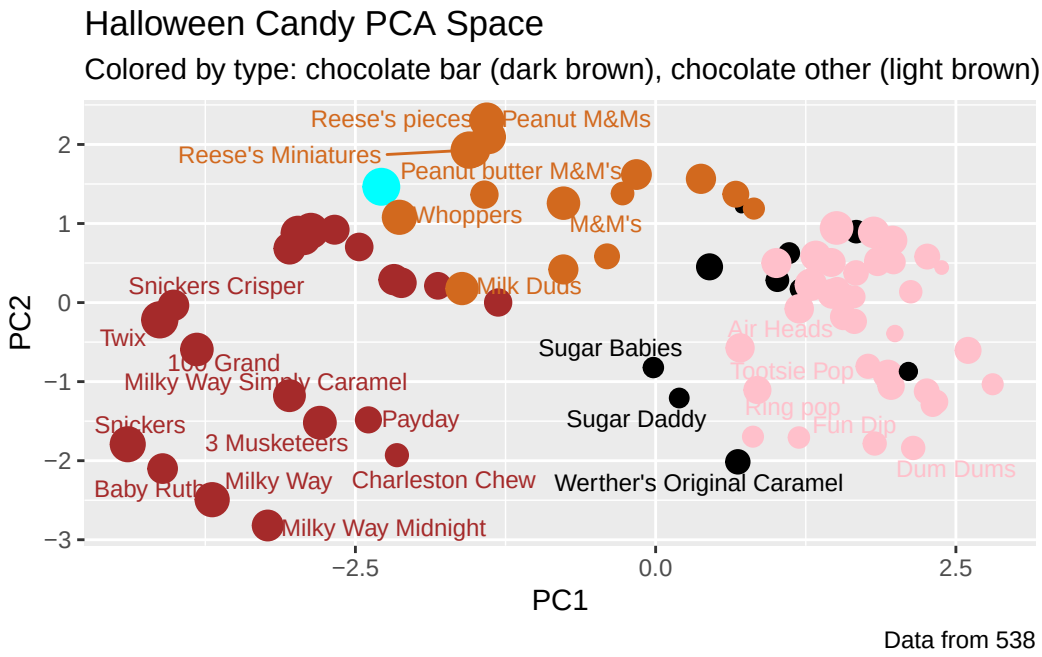



```
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



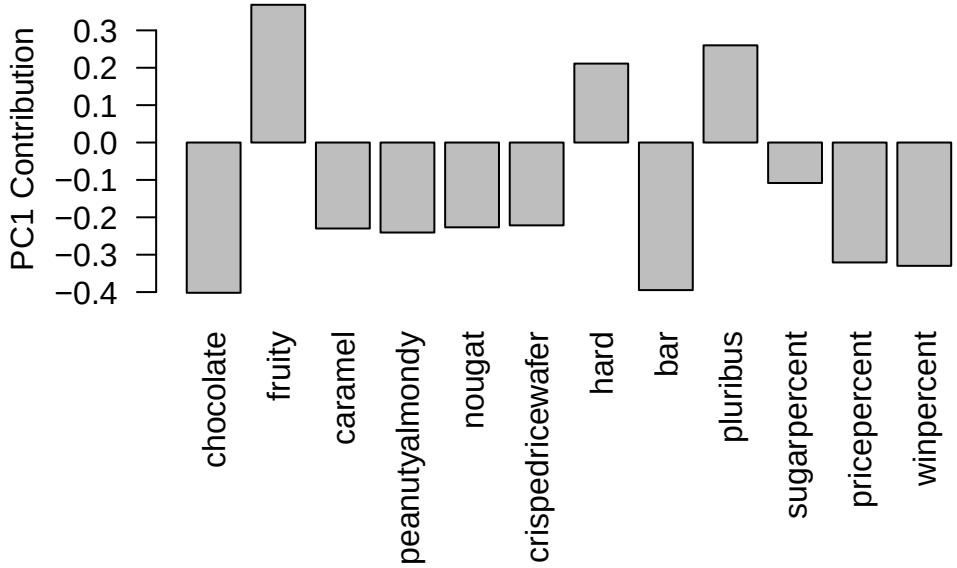

```
library(ggrepel)

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)
#ggplotly(p)
```

```
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 variables are fruity, hard, and pluribus because those are anticorrelated with preference.