class 8 halloween

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

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
```

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

```
nrow(candy_file)
```

```
[1] 85
```

Q2. How many fruity candy types are in the dataset? The functions dim(), nrow(), table() and sum() may be useful for answering the first 2 questions.

```
dim(candy_file)
```

```
[1] 85 12
```

```r
sum(candy_file$fruity)
```

```
[1] 38
```

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

```r
candy_file["Almond Joy",]$winpercent
```

```
[1] 50.34755
```

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

```r
candy_file["Kit Kat",]$winpercent
```

```
[1] 76.7686
```

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

```r
candy_file["Tootsie Roll Snack Bars",]$winpercent
```

```
[1] 49.6535
```

```r
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	

```r
skimr::skim(candy_file)
```

Table 3: 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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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?

hist

Q7. What do you think a zero and one represent for the candy$chocolate column? 0 - no chocolate 1- chocolate

```
candy_file$chocolate
```

```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

```
choc.ind <- as.logical(candy_file$chocolate)
fruit.ind <- as.logical(candy_file$fruity)
choc.win <- candy_file[choc.ind,]$winpercent
choc.win
```

```
 [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
```

```
mean(choc.ind)
```

```
[1] 0.4352941
```

```
mean(fruit.ind)
```

```
[1] 0.4470588
```

```
mean(candy_file$chocolate)
```

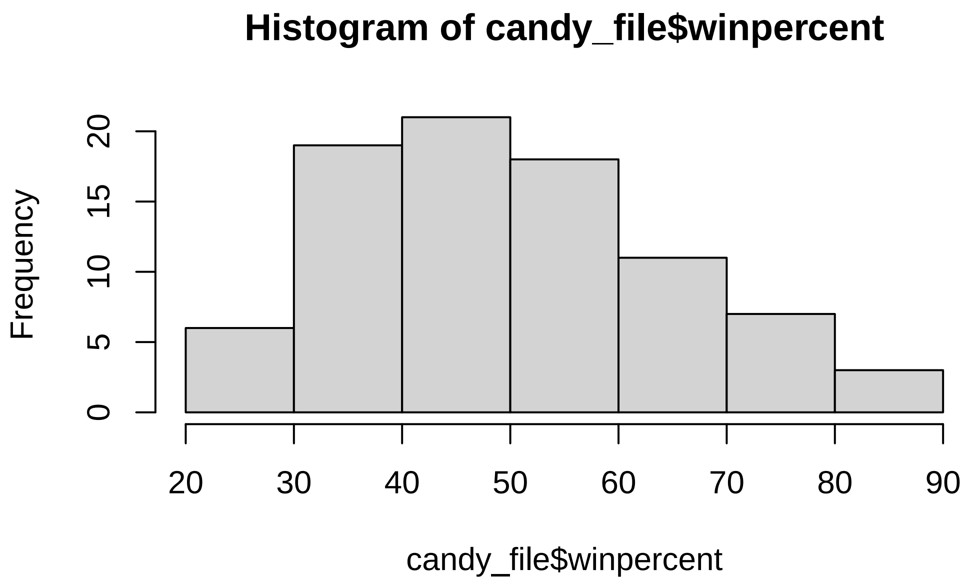
```
[1] 0.4352941
```

```
mean(candy_file$fruity)
```

```
[1] 0.4470588
```

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical? NO

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

```r
candy_file$winpercent[as.logical(candy_file$nougat)]
```

```
[1] 67.60294 56.91455 38.97504 73.09956 60.80070 46.29660 76.67378
```

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

```r
mean(candy_file$chocolate) >= mean(candy_file$fruity)
```

```
[1] FALSE
```

Q12. Is this difference statistically significant?

Welch Two Sample t-test

data: chocolate and fruity

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

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

```r
x <- c(5,2,3,6)
sort(x)
```

```
[1] 2 3 5 6
```

```
sort(x, decreasing=TRUE)
```

```
[1] 6 5 3 2
```

```
sort(x, decreasing =FALSE)
```

```
[1] 2 3 5 6
```

```
x
```

```
[1] 5 2 3 6
```

```
order(x)
```

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

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

```
[1] 2 3 5 6
```

```
y<-c("D", "A", "E")
order(y)
```

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

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

```
order.win <- order(candy_file$winpercent)
candy_file[order.win[1:5],1]
```

```
[1] 0 0 0 0 0
```

```r
tail(order.win)
```

```
[1] 54 65 29 80 52 53
```

```r
head(order.win)
```

```
[1] 45  8 13 73 27 58
```

```r
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
```

```r
tail(candy_file[order(candy_file$winpercent),], n=5)
```

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

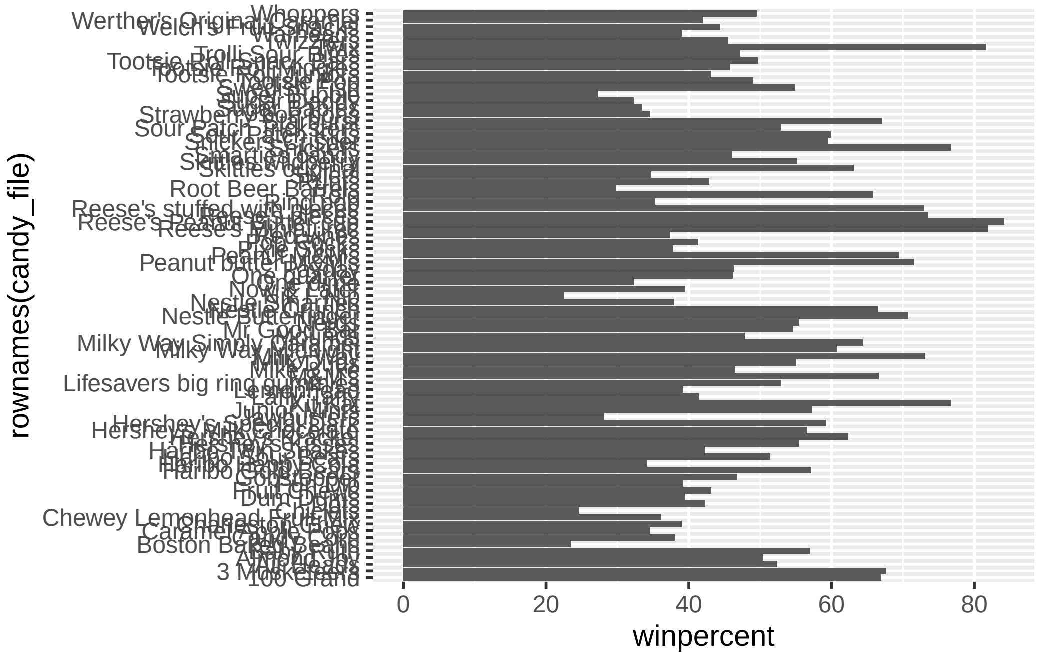
```
                          crispedricewafer hard bar pluribus sugarpercent
Snickers                                 0    0   1        0        0.546
Kit Kat                                  1    0   1        0        0.313
Twix                                     1    0   1        0        0.546
Reese's Miniatures                       0    0   0        0        0.034
Reese's Peanut Butter cup                0    0   0        0        0.720
                          pricepercent winpercent
Snickers                         0.651   76.67378
Kit Kat                          0.511   76.76860
Twix                             0.906   81.64291
Reese's Miniatures               0.279   81.86626
Reese's Peanut Butter cup        0.651   84.18029
```

```
x <- order(candy_file$winpercent)
order(candy_file$winpercent)
```

```
 [1] 45  8 13 73 27 58 72  3 71 20 10 70 60 56 12 51 49 63  9 11 82 31 17 46 15
[26] 50 30 84 22 14 59 76 16 83 81 77 64  4 47 35 18 79 40 75 85 78  6 21  5 68
[51] 32 41 74 36 62 42 23 25  7 19 28 26 66 67 38 24 61 39 57 44 34  1 69  2 48
[76] 43 33 55 37 54 65 29 80 52 53
```

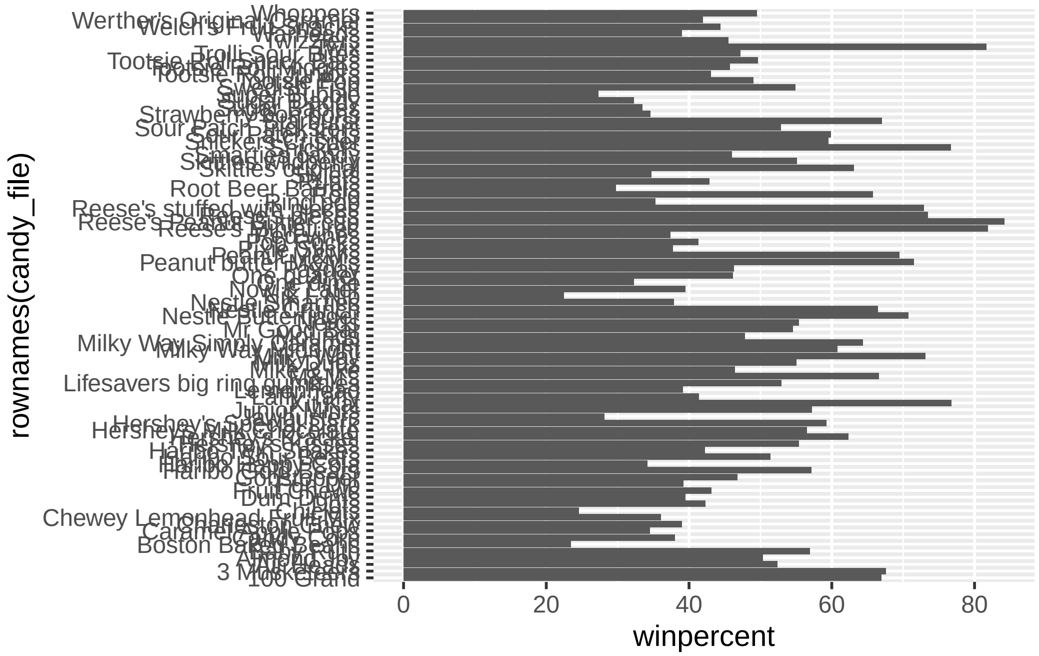
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)
ggplot(candy_file) +
  aes(winpercent, rownames(candy_file))+
  geom_col()
```

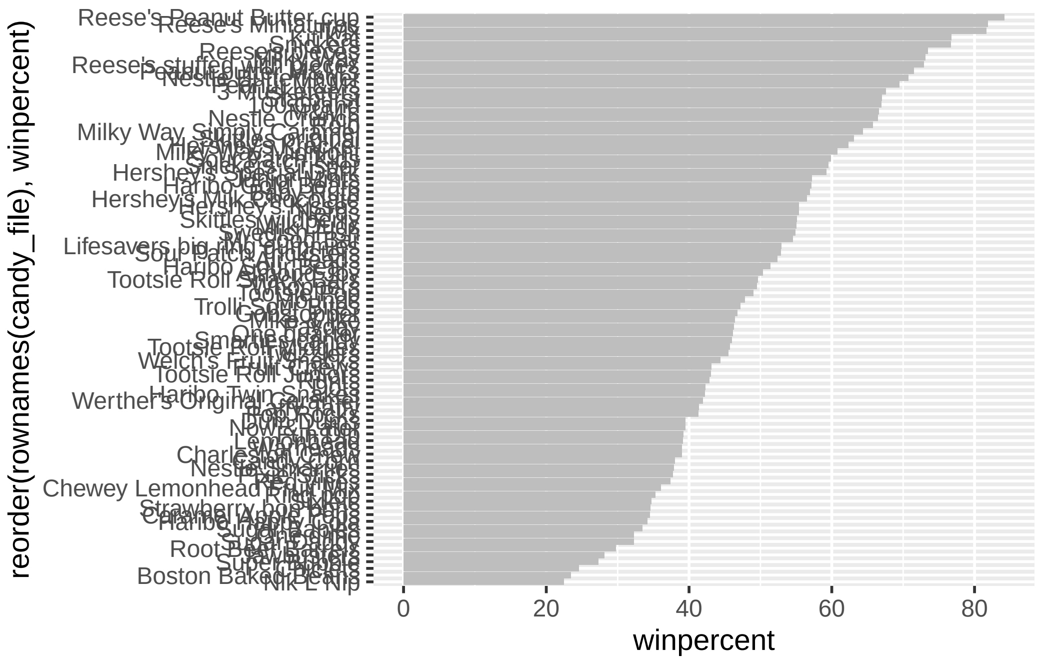


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

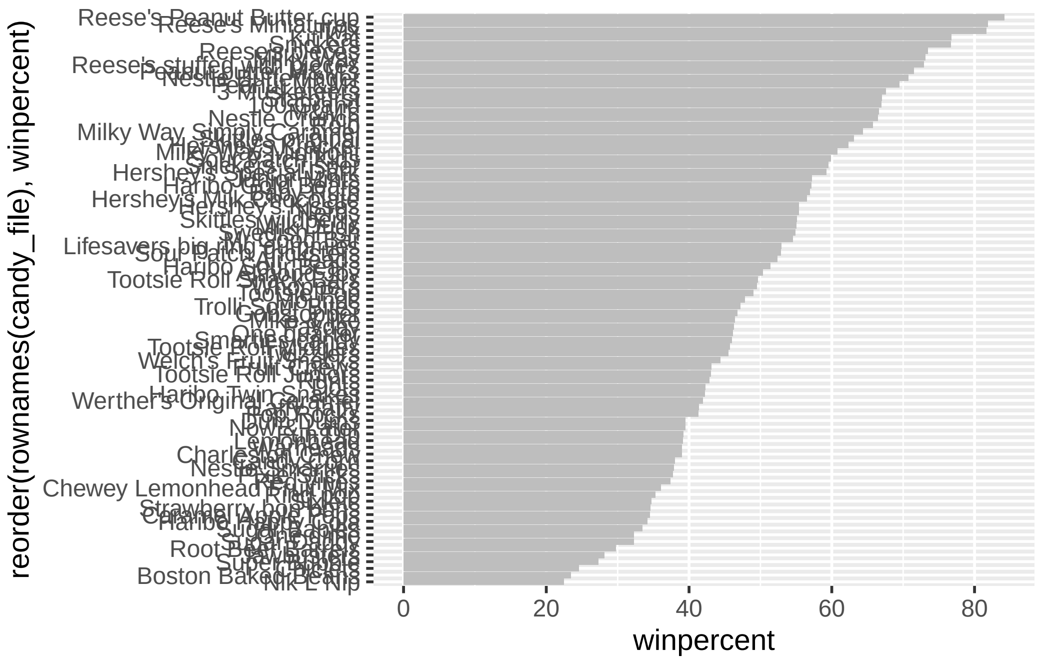
```
ggplot(candy_file) +
  aes(winpercent, rownames(candy_file), reorder(candy_file, FUN=mean(candy_file$winpercent
  geom_col()
```



```
mycols <- "gray"
ggplot(candy_file) +
  aes(winpercent, reorder(rownames(candy_file),winpercent)) +
  geom_col(bg=mycols)
```



```
mycols <- rep("gray",nrow(candy_file))

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

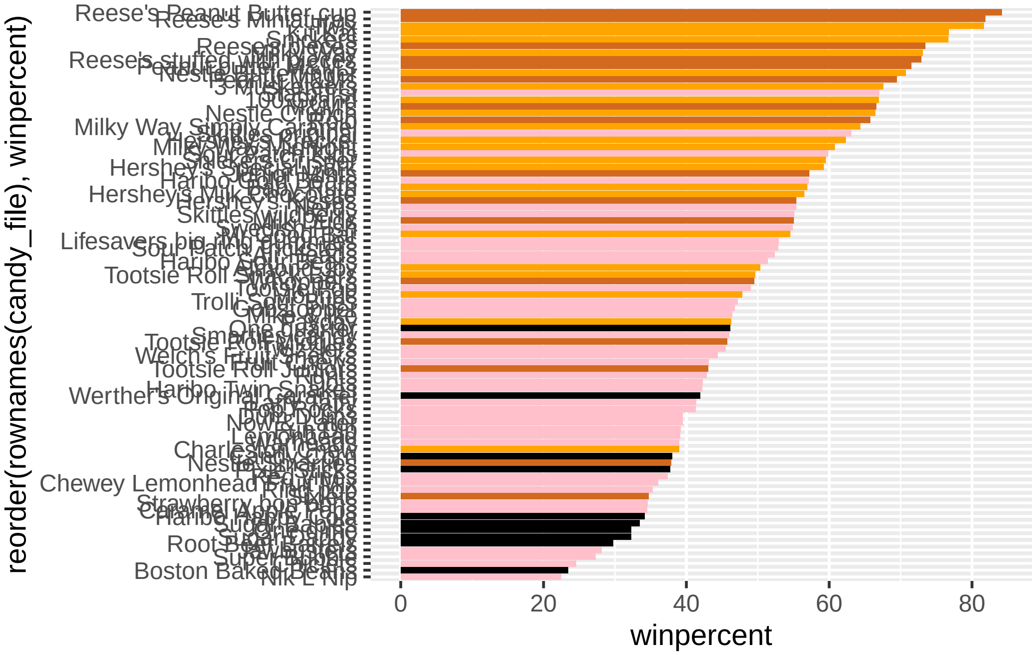


```r
mycols <- rep("gray", nrow(candy_file))
mycols[as.logical(candy_file$chocolate)] <- "brown"
mycols=rep("black", nrow(candy_file))
mycols[as.logical(candy_file$chocolate)] = "chocolate"
mycols[as.logical(candy_file$bar)] = "orange"
mycols[as.logical(candy_file$fruity)] = "pink"
mycols
```

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

```
[85] "chocolate"
```

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



Q17. What is the worst ranked chocolate candy? Resse's Peanut Butter cup

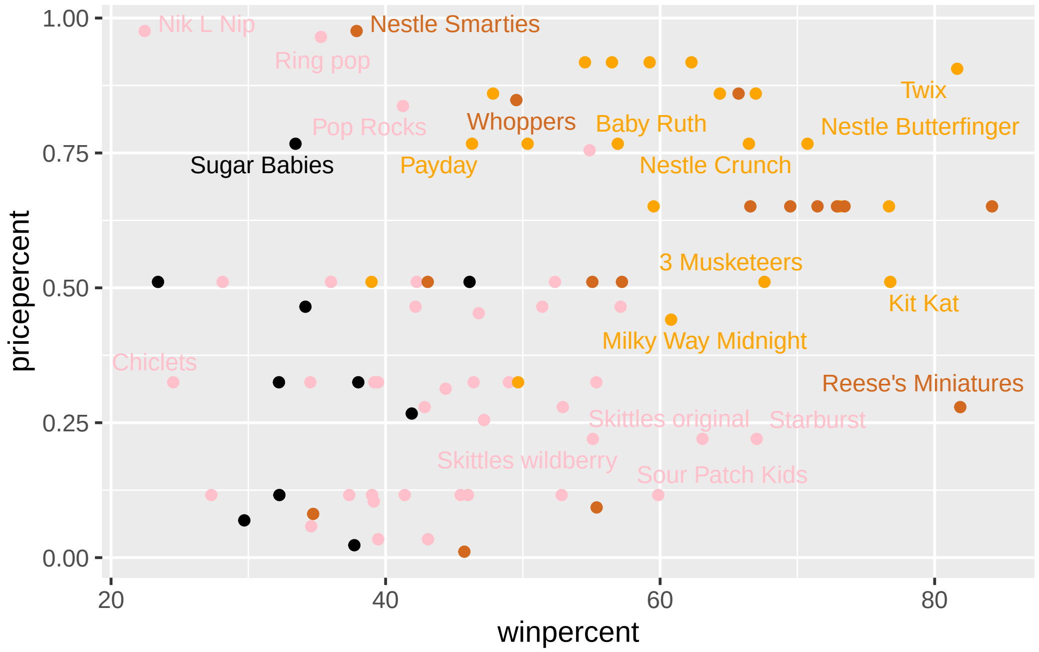
Q18. What is the best ranked fruity candy? Nik L Nip

How about a plot of price vs win

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(ggrepel)

ggplot(candy_file) +
  aes(winpercent, pricepercent, label=rownames(candy_file)) +
  geom_point(col=mycols) +
  geom_text_repel(col=mycols, size=3.3, max.overlaps = 5)
```

```
Warning: ggrepel: 65 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



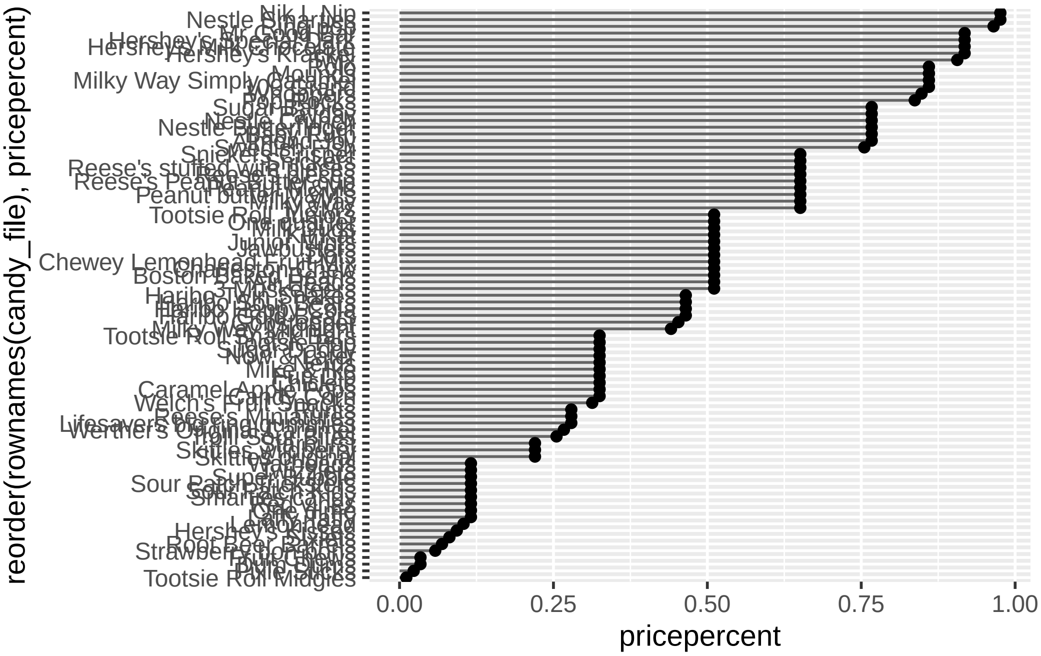
Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

```
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().

```r
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()
```

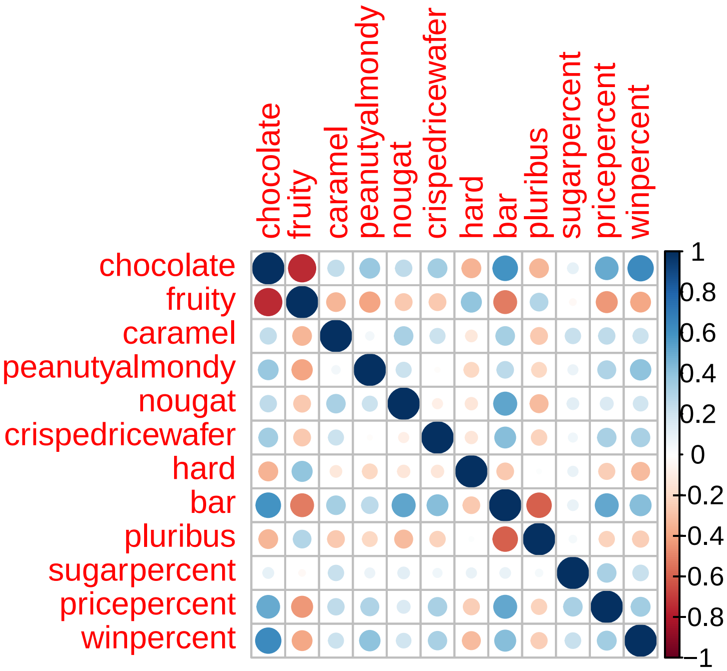


Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

```r
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



chocolate and fruity

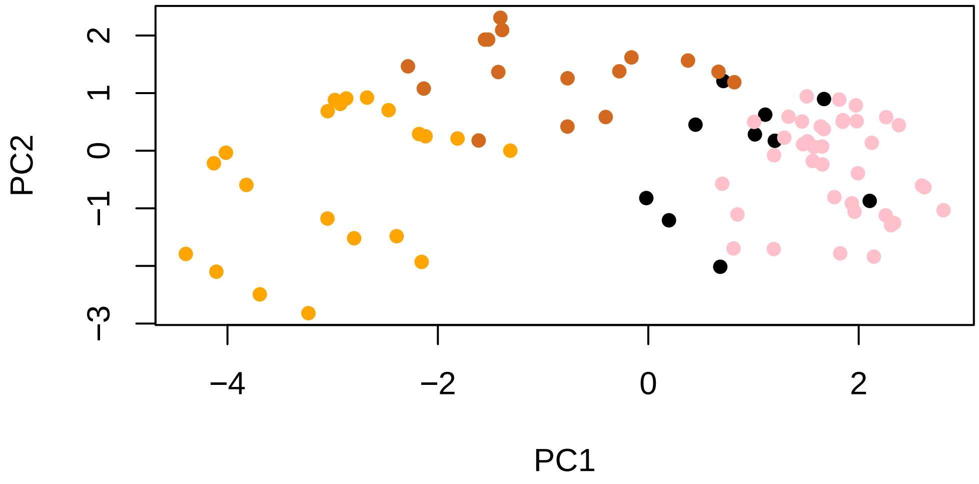
Q23. Similarly, what two variables are most positively correlated? chocolate and bar abd oricepercent and winpercent

#Principal Component Analysis

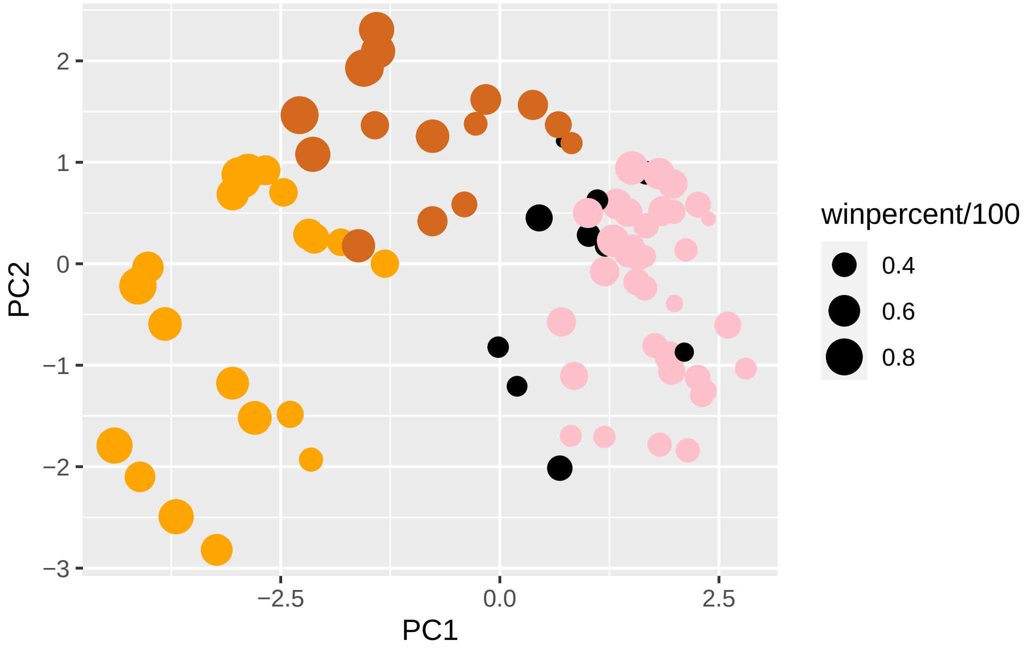
```
pca <- prcomp(candy_file, scale=TRUE)
summary(pca)
```

```
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[,1:2], col=mycols, pch=16)
```

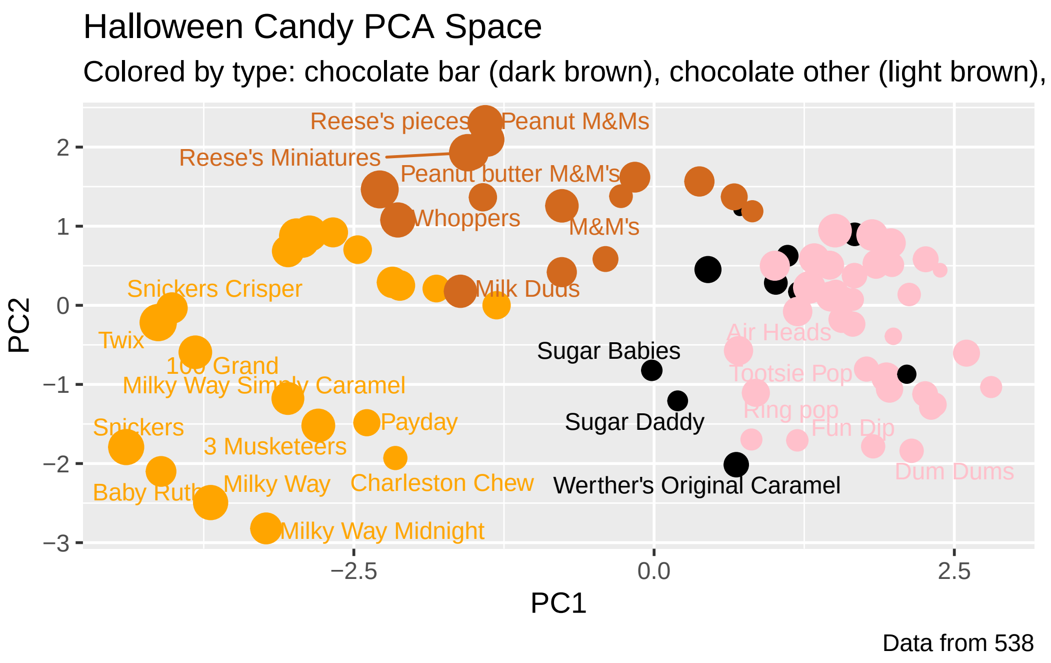


```
my_data <- cbind(candy_file, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=mycols)

p
```



```r
library(ggrepel)

p + geom_text_repel(size=3.3, col=mycols, 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
```



Halloween Candy PCA Space

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

Data from 538