Class10: Halloween Mini Project

Jessica PID A15647602

1. Importing Candy Data

```
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
```

Q1. How many different candy types are in this dataset? A. There are 85 different candy types.

```
nrow(candy)
```

```
[1] 85
```

Q2. How many fruity candy types are in the dataset? A. There are 38 fruity candy types.

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

```
[1] 38
```

2. What is your favorite candy?

> Q3. What is your favorite candy in the dataset and what is it's winpercent value? A. Snickers; 76.67378

```
candy["Snickers", ]$winpercent
```

```
[1] 76.67378
```

> Q4. What is the winpercent value for "Kit Kat"? A. 76.7686 Q5. What is the winpercent value for "Tootsie Roll Snack Bars"? A. 49.6535

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

```
[1] 76.7686
```

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

```
[1] 49.6535
```

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

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

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

Table 1: Data summary

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

Variable type: numeric

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

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

Q7. What do you think a zero and one represent for the candy$chocolate column?
A. The 1 means that the candy has chocolate in it.

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

inds <- candy$fruity== 1
fruit.win <- candy[inds,]$winpercent

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)
```

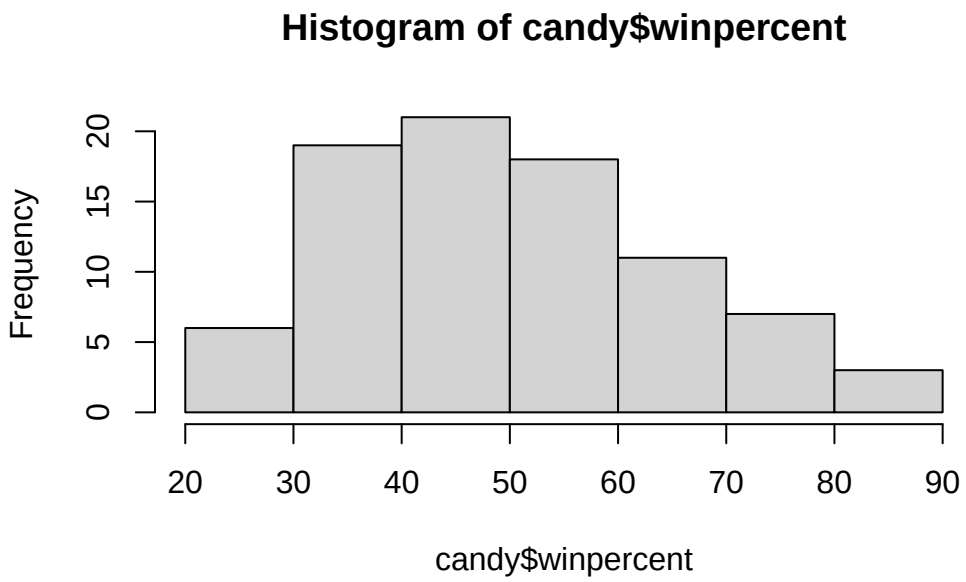
```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  22.45   39.04   42.97   44.12   52.11   67.04
```

```
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
```

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical? A. No it is not symmetrical.

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

Q11. On average is chocolate candy higher or lower ranked than fruit candy? A. The chocolate (60) is ranked higher than the candy (44).

Q12. Is this difference statistically significant? A. Yes, the difference is statistically significant.

```
mean(candy$winpercent[as.logical(candy$chocolate)])
```

```
[1] 60.92153
```

```
mean(candy$winpercent[as.logical(candy$fruity)])
```

```
[1] 44.11974
```

3. Overall Candy Rankings

Q13. What are the five least liked candy types in this set? A. Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters

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

Q14. What are the top 5 all time favorite candy types out of this set? A. Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, and Snickers.

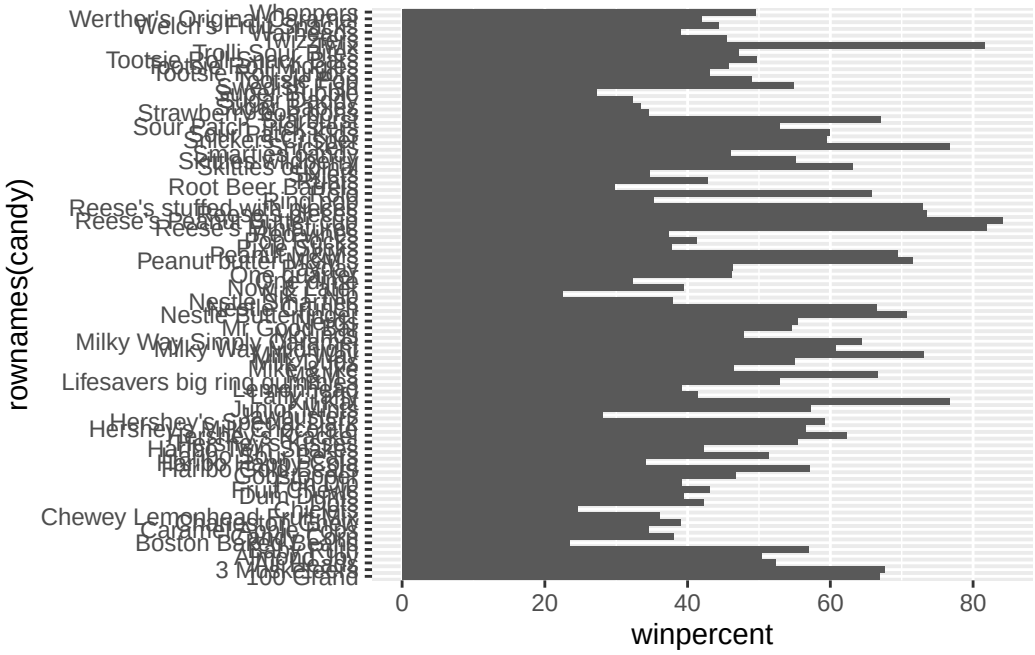
```
inds2 <- order(desc(candy$winpercent))
head(candy[inds2,], n= 5)
```

```
                        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
```

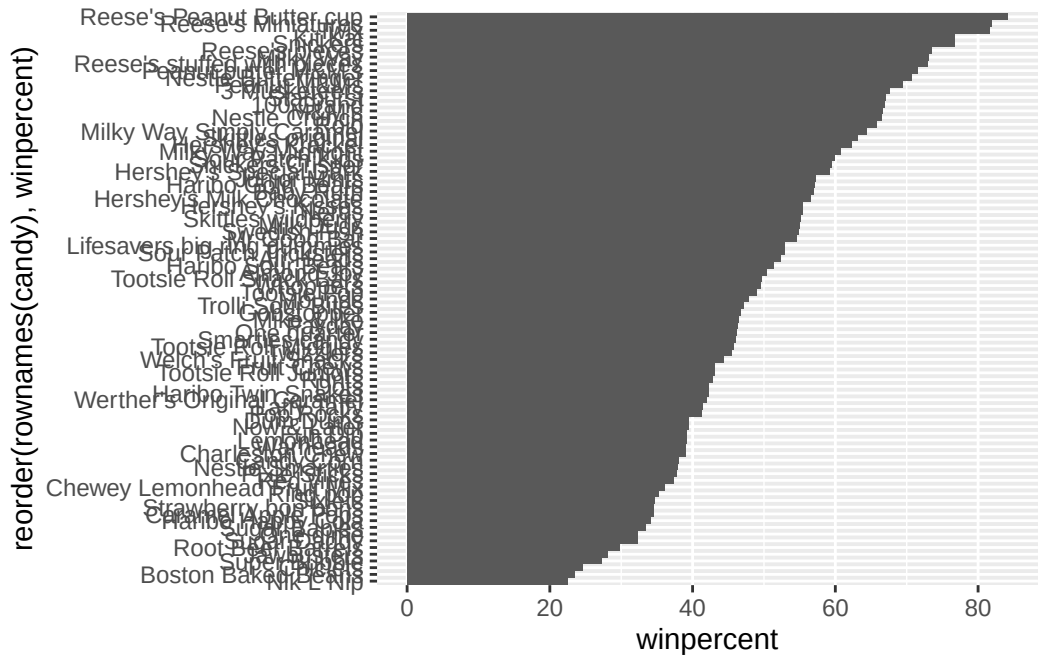
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

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

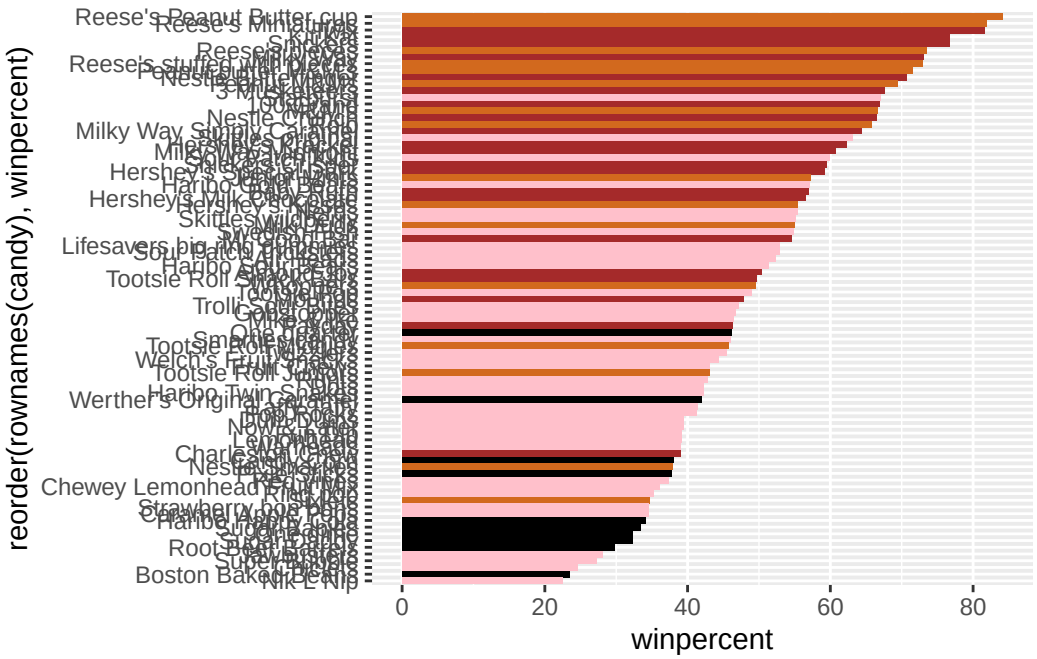


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
library(ggplot2)

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



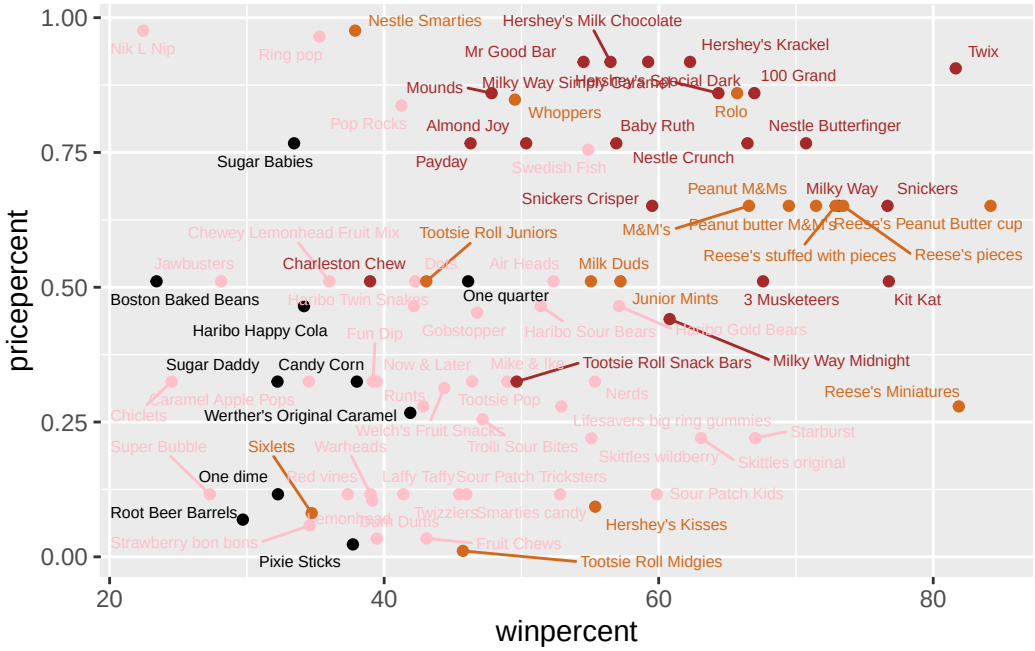

```
#Adding useful colors
my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "brown"
my_cols[as.logical(candy$fruity)] = "pink"

ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)
```



Q17. What is the worst ranked chocolate candy? A. The worst ranked chocolate candy is Sixlets. Q18. What is the best ranked fruity candy? A. The best ranked fruity candy is Starburst.

4. Taking a look at pricepercent

```
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=2, max.overlaps = 80)
```



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? A. Tootsie Roll Midgies

```
order <- order((candy$winpercent/candy$pricepercent), decreasing = TRUE)
head(candy[order,c(11,12)], n=1 )
```

```
                  pricepercent winpercent
Tootsie Roll Midgies        0.011   45.73675
```

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

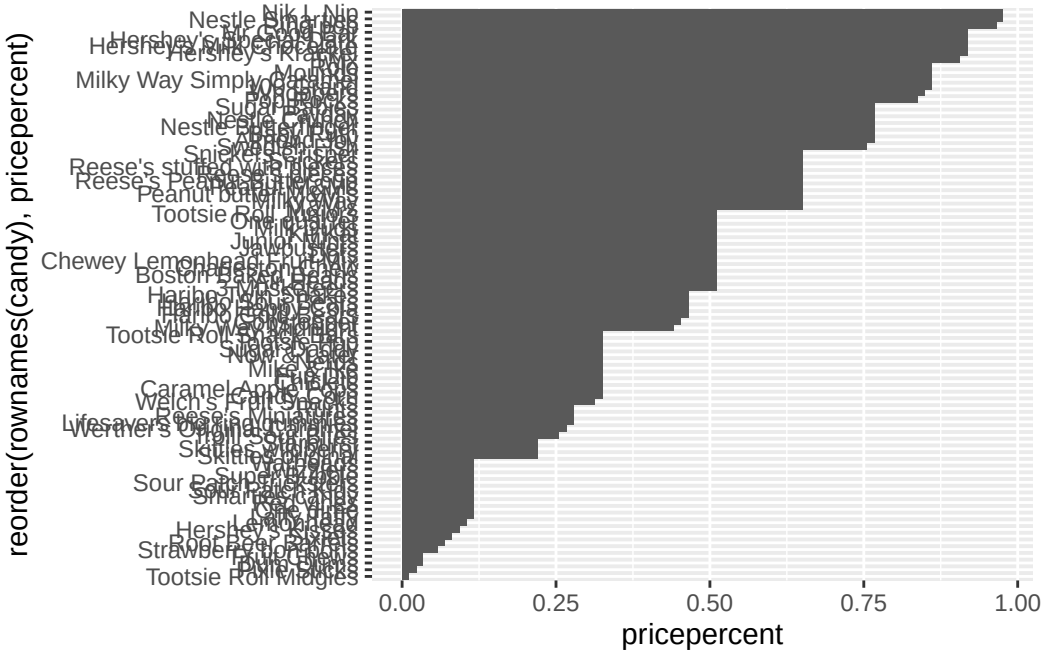
A. The 5 most expensive candies are Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate. The least popular is Nik L Nip.

```
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
```

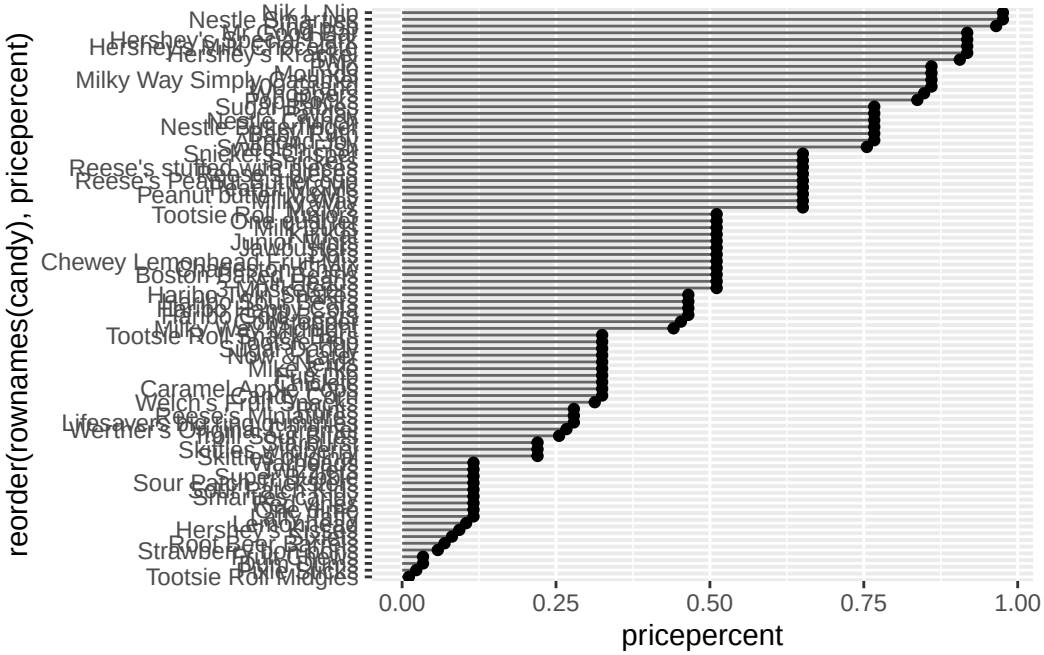
#Q21. Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping geom_col() for geom_point() + geom_segment().

```
library(ggplot2)

ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_col()
```



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()
```

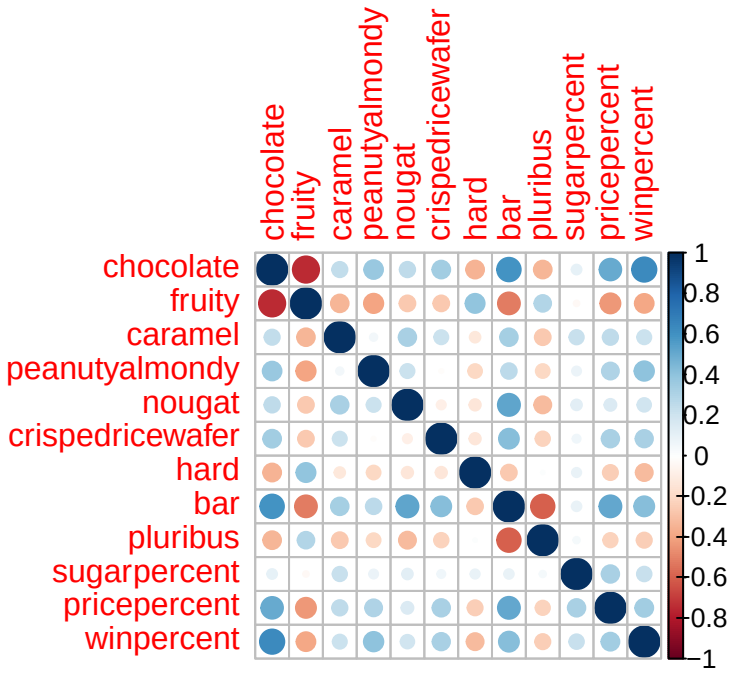


5. Exploring the Correlation structure

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



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

6. Principal Component Analysis

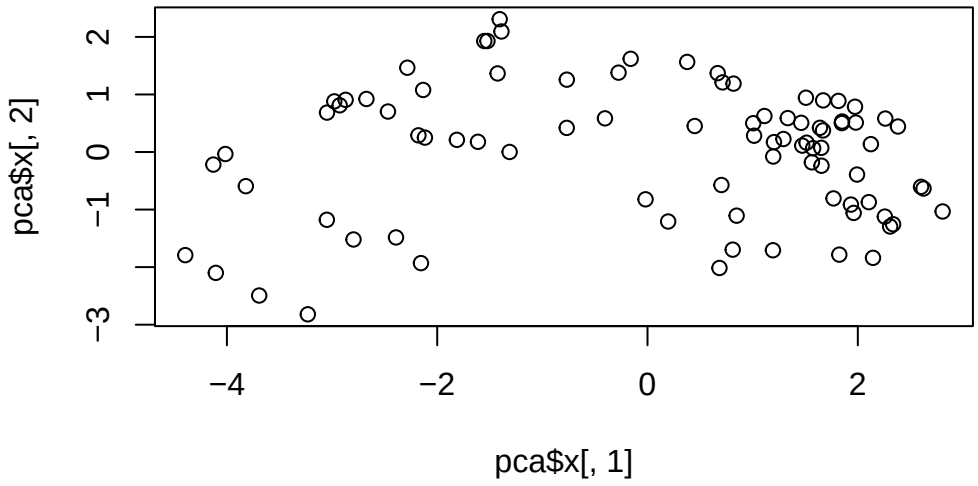
```
pca <- prcomp(candy, scale. = TRUE)
summary(pca)
```

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

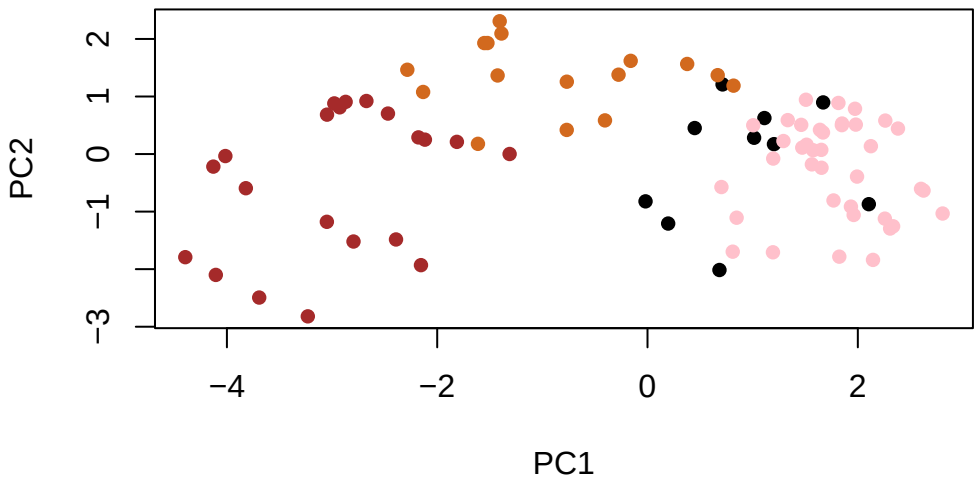
```
pca$rotation[,1]
```

```
      chocolate           fruity          caramel   peanutyalmondy
     -0.4019466        0.3683883       -0.2299709       -0.2407155
         nougat crispedricewafer             hard              bar
     -0.2268102       -0.2215182        0.2111587       -0.3947433
       pluribus     sugarpercent     pricepercent       winpercent
      0.2600041       -0.1083088       -0.3207361       -0.3298035
```

```
#Now we can plot our main PCA score plot of PC1 vs PC2.
plot(pca$x[, 1], pca$x[, 2])
```



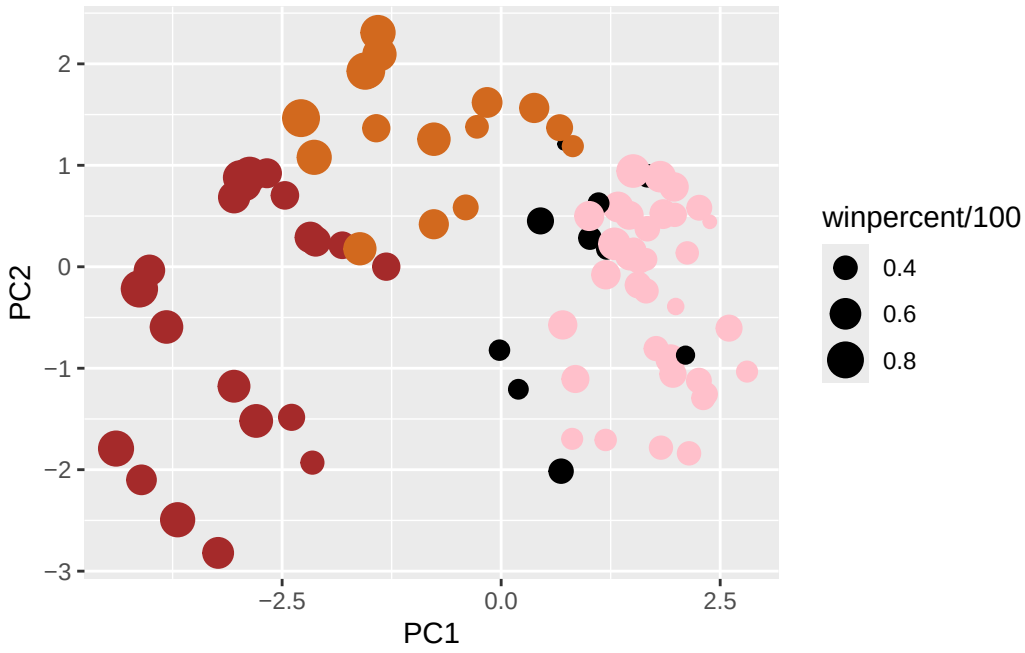

```
#We can change the plotting character and add some color:
plot(pca$x[,1:2], col=my_cols, pch=16)
```



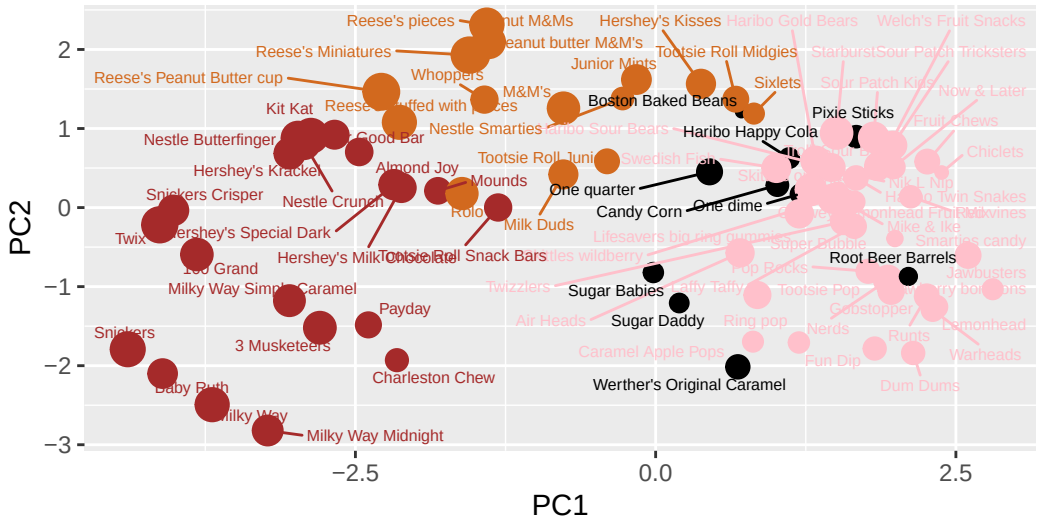
We can make a much nicer plot with the ggplot2

```
# 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
```



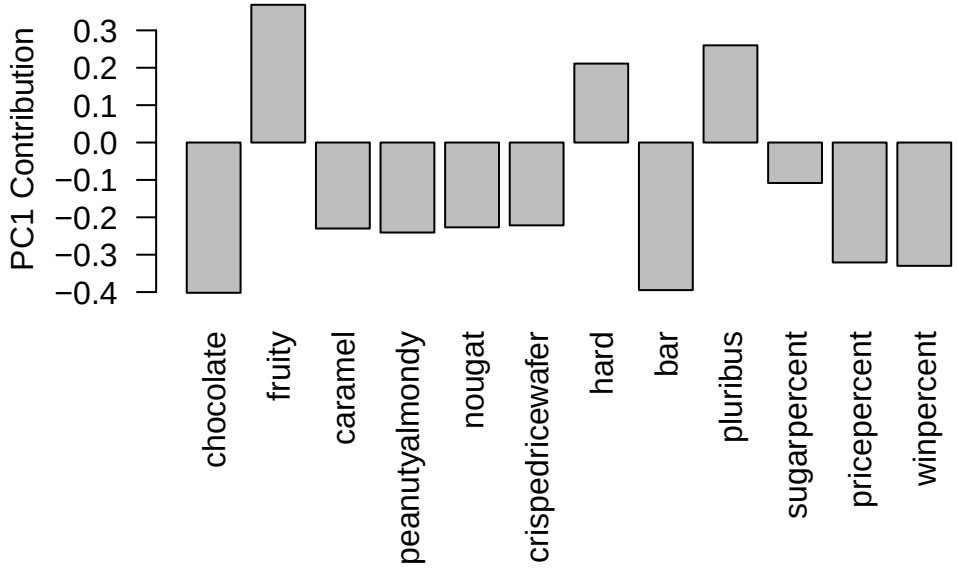

```
library(ggrepel)

p + geom_text_repel(size= 2, col=my_cols, max.overlaps = 60)  +
  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")
```



Data from 538

```
#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? A. Fruity candy, hard candy, and pluribus candy(coming from a bag or box of multiple candy). Yes these variable make sense.