Halloween Candy!

Daira

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

```
             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
```

what is in the dataset?

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

```
nrow(candy.data)
```

```
[1] 85
```

A: There are 85 candy types in this dataset

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

```
fruitycandy <- sum(candy.data$fruity)
fruitycandy
```

```
[1] 38
```

A: There are 38 fruity candies

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

```
fave <- candy.data["Nerds",]$winpercent
fave
```

```
[1] 55.35405
```

A: My favorite candy is Nerds and it has a 55.4% winpercent value.

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

```
kitkat <- candy.data["Kit Kat",]$winpercent
kitkat
```

```
[1] 76.7686
```

A: The winpercent value for Kit Kats is 76.8%

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

```
tootsie <- candy.data["Tootsie Roll Snack Bars",]$winpercent
tootsie
```

```
[1] 49.6535
```

A: The win percent value for tootsie rolls is 49.7%

skimr

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

Table 1: Data summary

Name	candy.data
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?

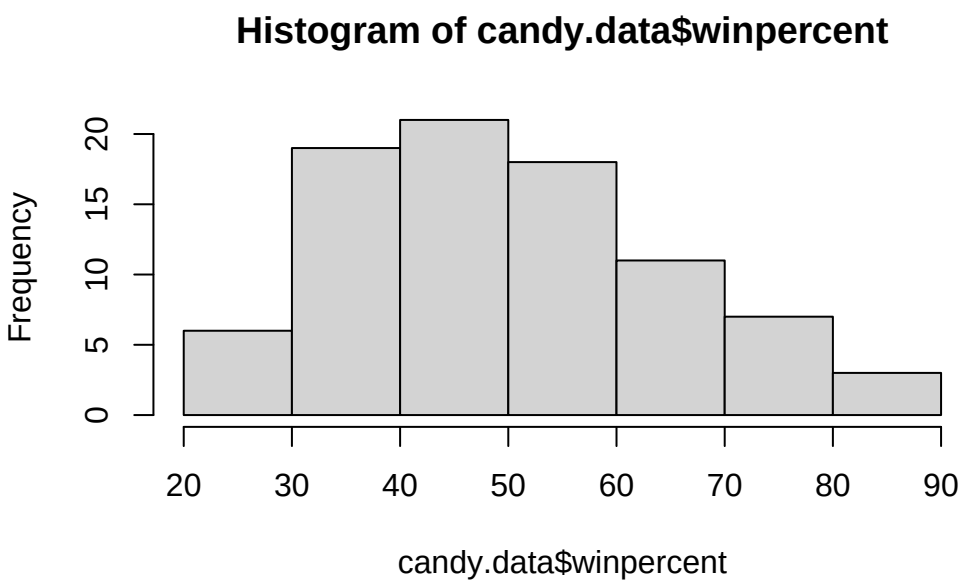
A: Based on the skim, the win percent values seem on a different scale than the others (ie the mean, the sd..ete are larger than in a range from 0.00 to 1.00).

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

A: binary, for Boolean values so TRUE and FALSE, if its chocolate or not.

Q8. Plot a histogram of winpercent values

```
hist(candy.data$winpercent)
```



Q9. Is the distribution of winpercent values symmetrical?

A: Not it is not symmetrical.

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

```
center <- median(candy.data$winpercent)
center
```

```
[1] 47.82975
```

A: It appears Below 50%

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

```
win.choc <- candy.data[as.logical(candy.data$chocolate), "winpercent"]
win.choc
```

```
 [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
```

```
win.fruit <- candy.data[as.logical(candy.data$fruity), "winpercent"]
win.fruit
```

```
 [1] 52.34146 34.51768 36.01763 24.52499 42.27208 39.46056 43.08892 39.18550
 [9] 46.78335 57.11974 51.41243 42.17877 28.12744 41.38956 39.14106 52.91139
[17] 46.41172 55.35405 22.44534 39.44680 41.26551 37.34852 35.29076 42.84914
[25] 63.08514 55.10370 45.99583 59.86400 52.82595 67.03763 34.57899 27.30386
[33] 54.86111 48.98265 47.17323 45.46628 39.01190 44.37552
```

now compare averages

```
mean(win.choc)
```

```
[1] 60.92153
```

```
mean(win.fruit)
```

```
[1] 44.11974
```

A: On average, chocolate candy is higher ranked than fruity.

Q12. Is this difference statistically significant? do a sample t-test

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

```
    Welch Two Sample t-test

data:  win.choc and win.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
```

A: Yes, based on our t-test it looks like there is a significant difference between the average means (small P-value)

```
sort(c(5,4,1,2))
```

```
[1] 1 2 4 5
```

```
order(c(5,4,1,2))
```

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

Overall Candy Ranking

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

```
order.ind <- order(candy.data$winpercent)
head(candy.data[order.ind,], 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
```

A: The five least liked are: Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters.

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

```
order.new <-order.ind <- order(candy.data$winpercent, decreasing= TRUE)
head(candy.data[order.new,], 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
```

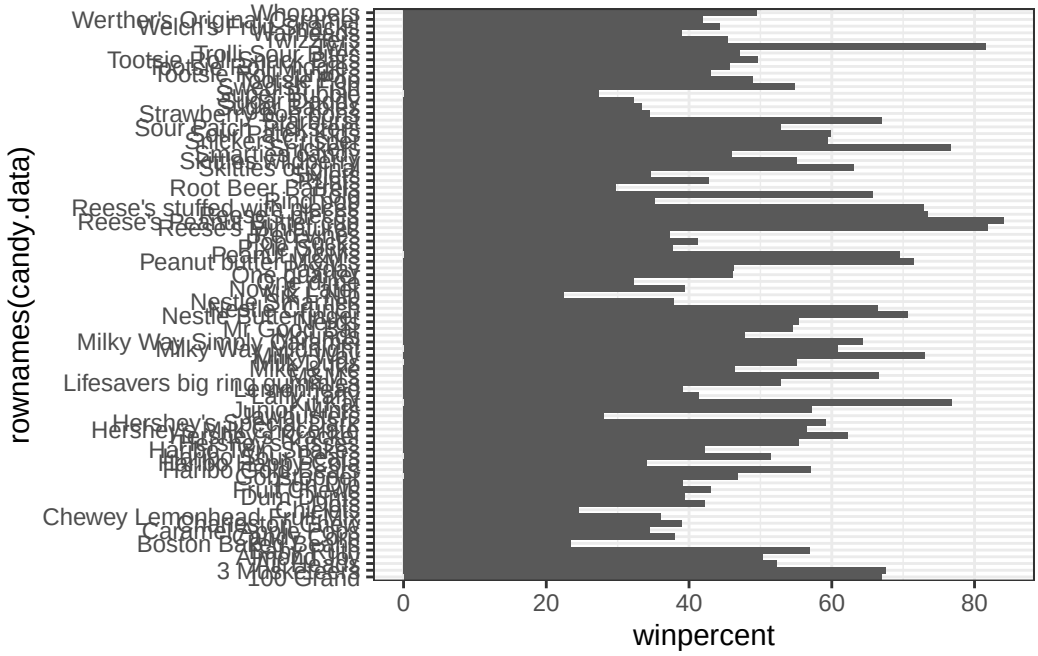
A: The top five are: ReesePeanut Butter Cup, Reese Miniatures, Twix, Kit Kat, and Snickers.

##now use ggplot2

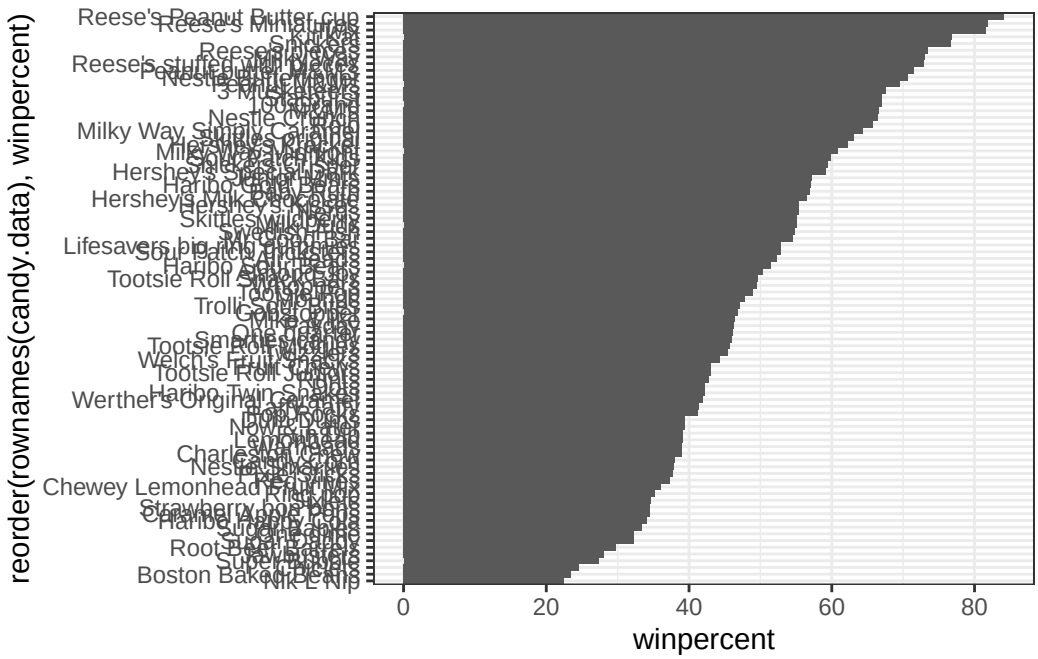
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

ggplot(candy.data) +
  aes(x = winpercent, rownames(candy.data)) +
```

```
  geom_col() +
  theme_bw ()
```



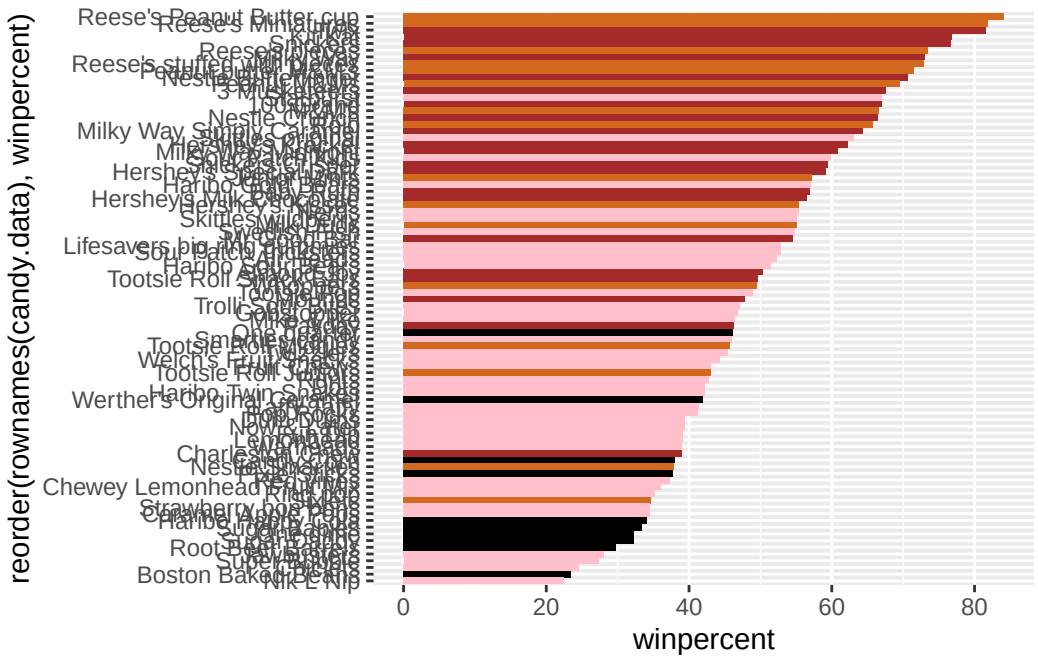

```
ggplot(candy.data) +
  aes(winpercent, reorder(rownames(candy.data),winpercent)) +
  geom_col() +
  theme_bw ()
```



#can add some color

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

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



very colorful now we can answer some questions

Q17. What is the worst ranked chocolate candy?

A: The worst ranked chocolate candy are Sixlets

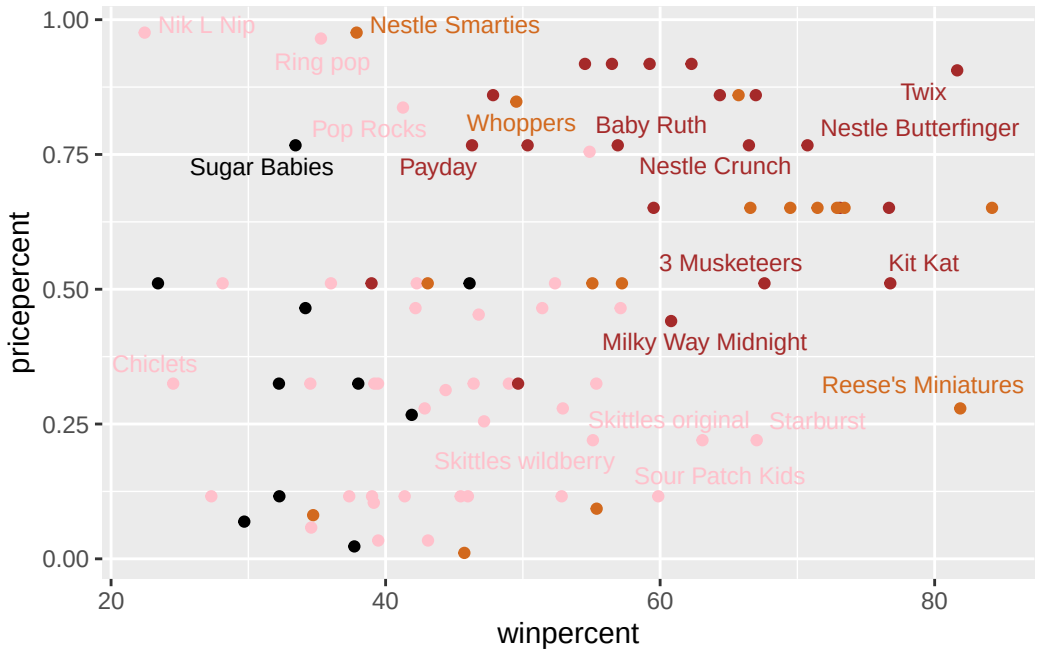
Q18. What is the best ranked fruity candy?

A: The best ranked fruity candy are Starbursts

Price Percent

```
library(ggrepel)

# How about a plot of price vs win
ggplot(candy.data) +
  aes(winpercent, pricepercent, label=rownames(candy.data)) +
  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?

A: It appears that Reese's Miniatures are the highest rank for winpercent(80) and lower for $

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

```
ord <- order(candy.data$pricepercent, decreasing = TRUE)
head( candy.data[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
```

A: The most expensive and worst is the Nik L Nip!

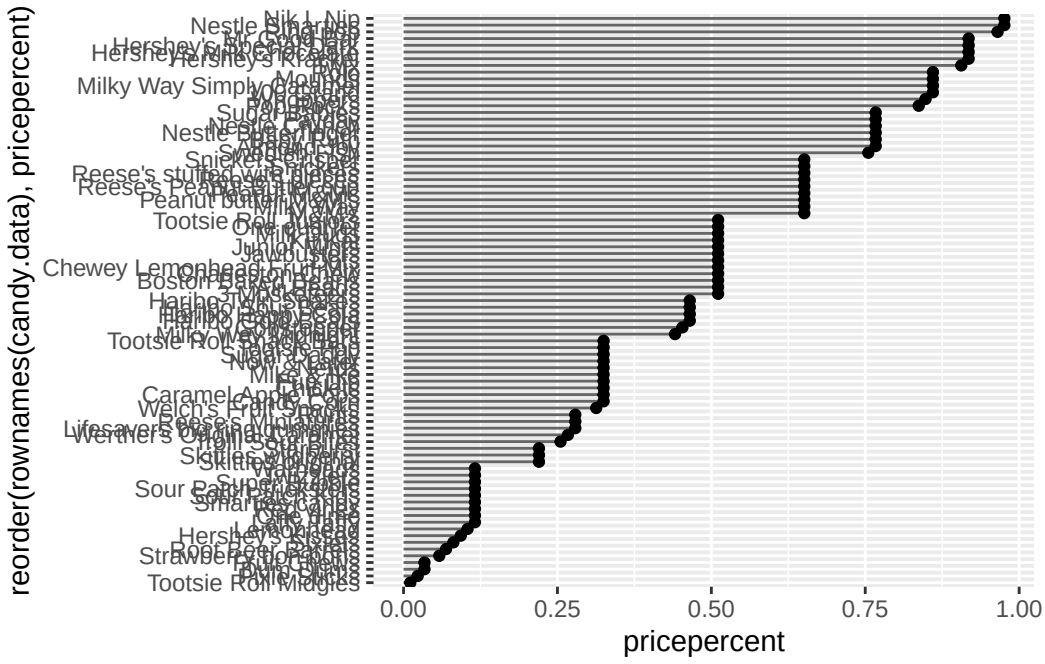
Q21. Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a

so called “dot chat” or “lollipop” chart by swapping geom_col() for geom_point() + geom_segment().

A: OPTIONAL

Make a Lollipop Chart of Price percent

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

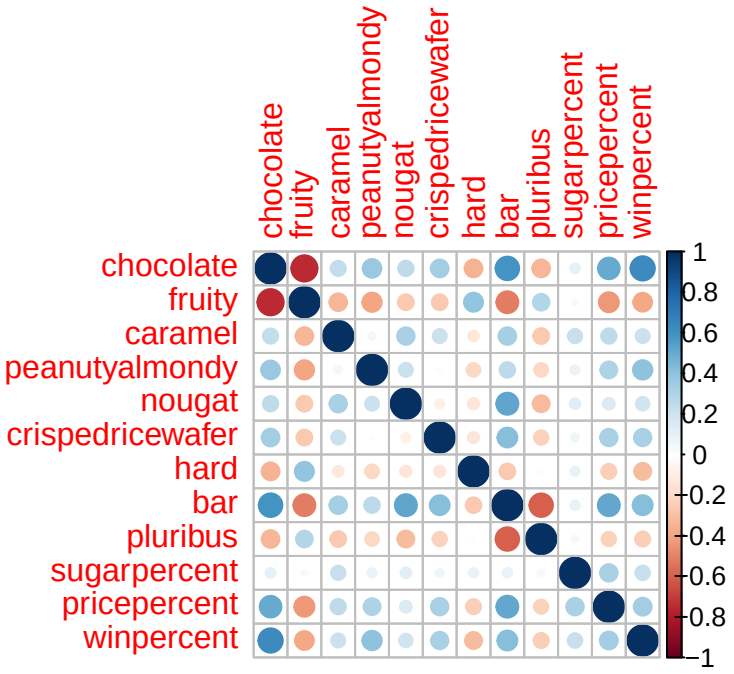


Exploring Correlation

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

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



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

A: The two variables most anti-correlated, chocolate and fruity

Q23. Similarly, what two variables are most positively correlated?

A: The most positively correlated are chocolate and win percent (and if we weren't looking at percents and other qualities of candy then it is most positively correlated with bar which makes sense because chocolate bars are delicious)

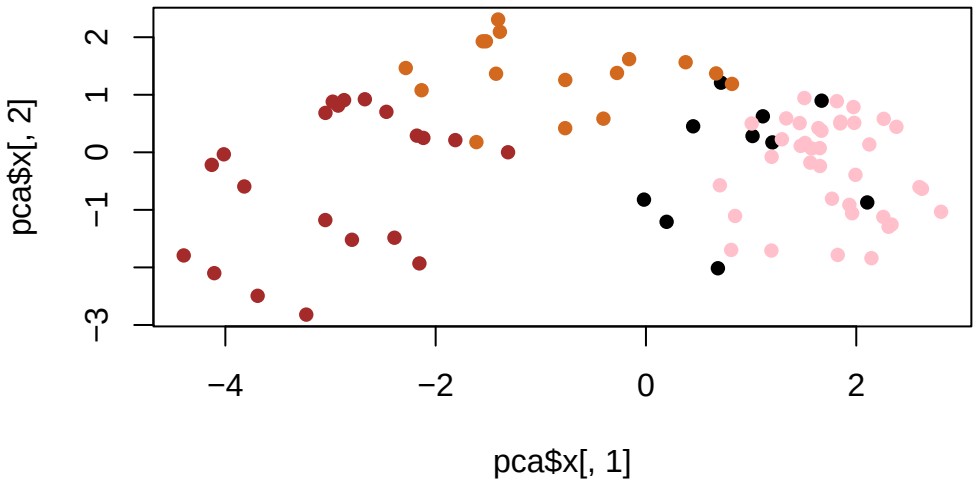
now time for PCA

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

```
Importance of components:
```

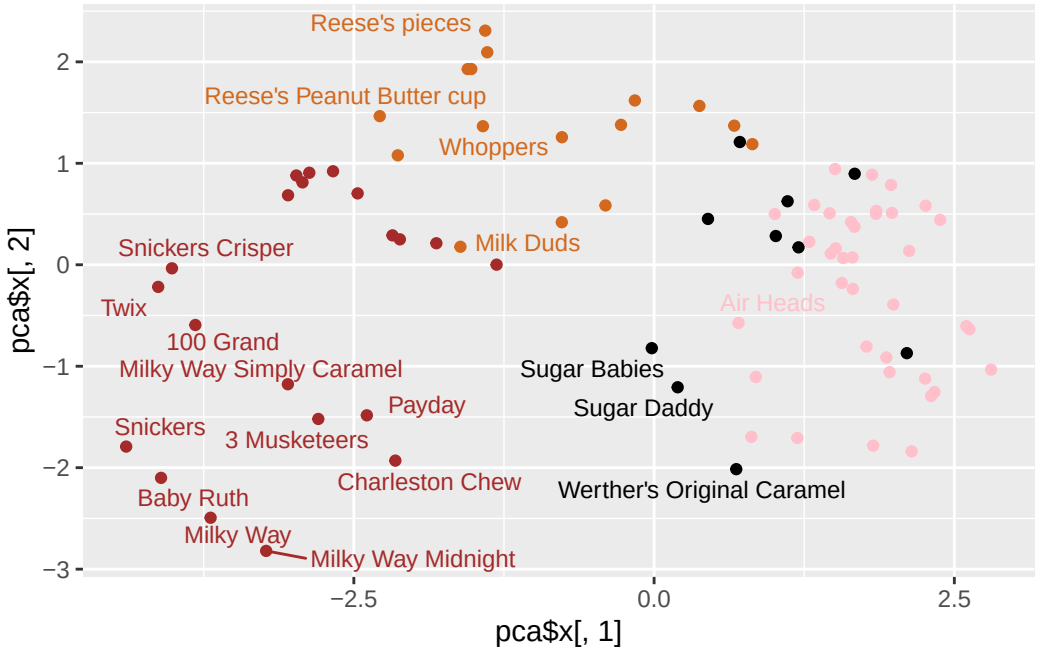
```
                           PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 0.7424 0.79830 0.85369
                           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
```

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

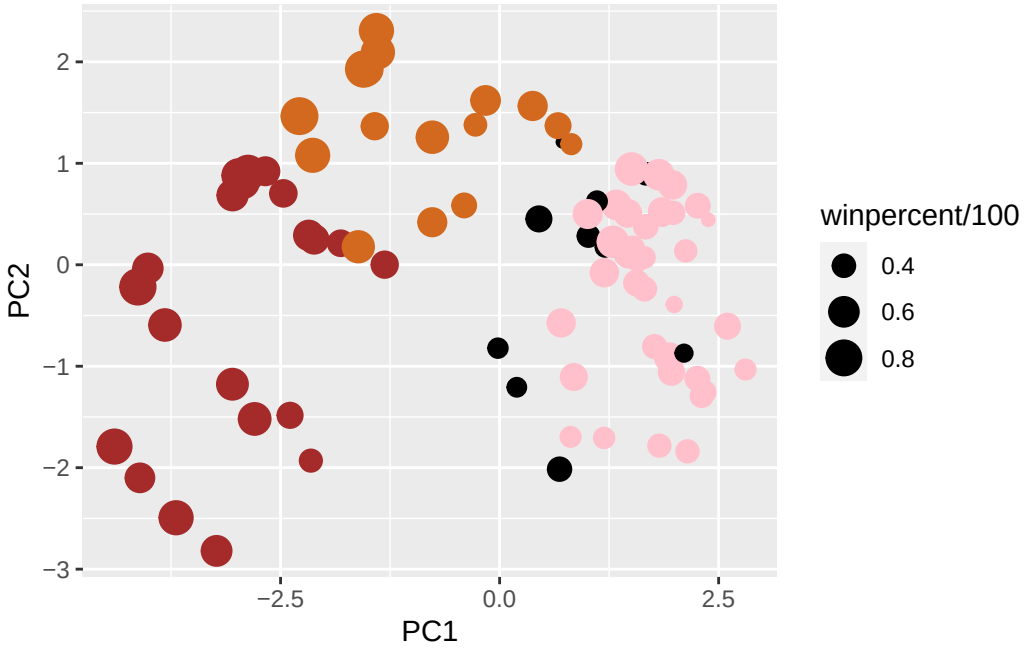



```r
ggplot(candy.data) +
aes(pca$x[,1], pca$x[,2], label=rownames(candy.data)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 5)
```

```
Warning: ggrepel: 66 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

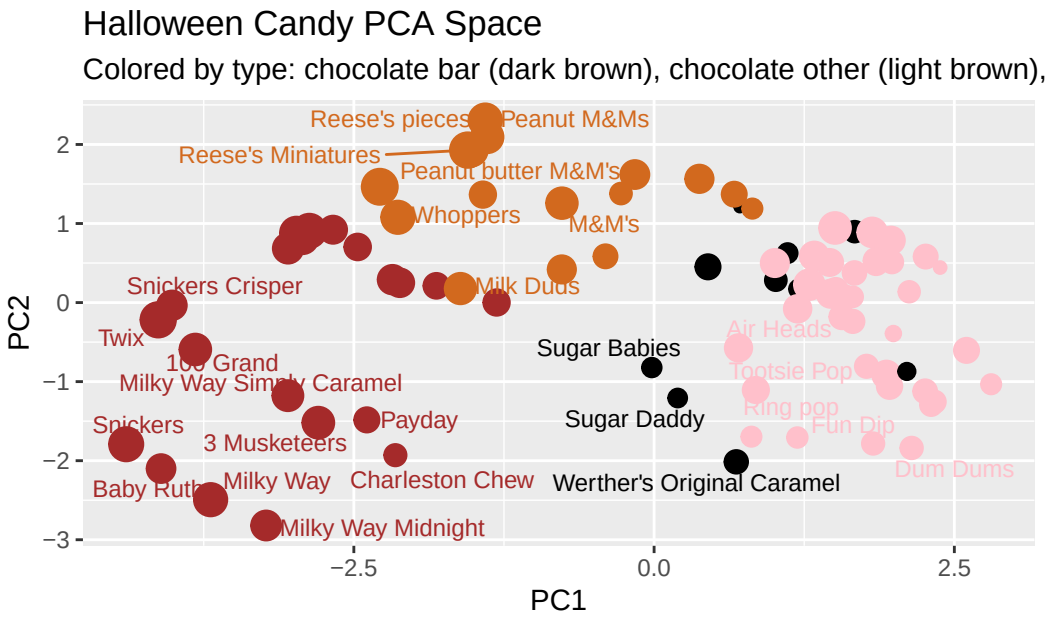



```
my_data <- cbind(candy.data, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



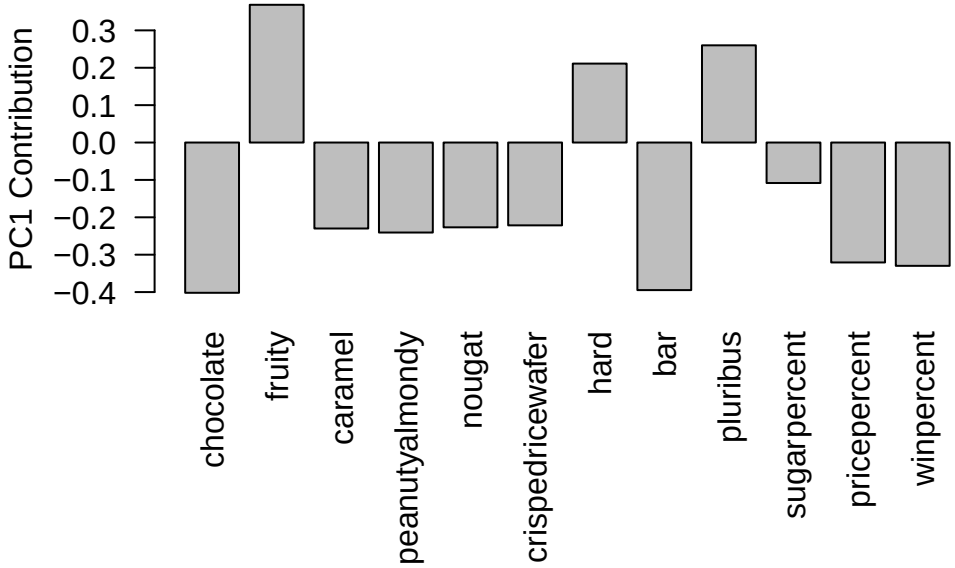

```
#| eval: false
#| echo: false

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)
```

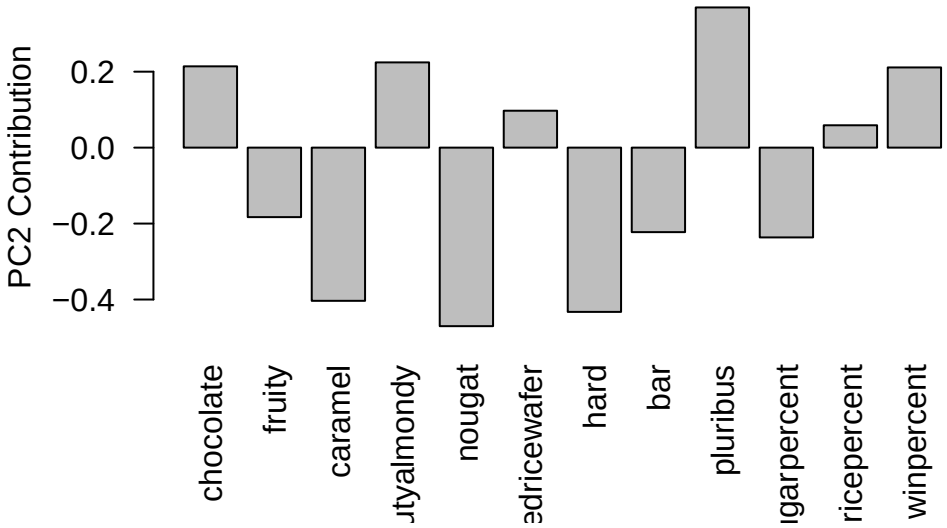
```
par(mar=c(8,4,2,2))
barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```



Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

A: The variables that are picked up by PC1 in the positive are fruity, hard, and pluribus (comes in multiple). this makes sense as fruity drives the difference in correlation of the data and msot fuirty candies are hard and come in bags of many whereas chocolate are usually in bars and in one.

```
barplot(pca$rotation[,2], las=2, ylab="PC2 Contribution")
```


PC2 Contribution
0.2
0.0
−0.2
−0.4
chocolate
fruity
caramel
ɹtyalmondy
nougat
ɐdricewafer
hard
bar
pluribus
ıgarpercent
ricepercent
winpercent