Halloween Mini Project Class 9

Matthew White

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

```
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 in dataset? 85 Q2: How many fruity candy types are in dataset? 38

```
nrow(candy)
```

```
[1] 85
```

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

```
fruity
 0  1
47 38
```

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

```
[1] 38
```

Q3: What is your favorite candy in the dataset and its `winpercent` value? 50.35 Q4: What is the `winpercent` value for "Kit Kat"? 76.76 Q5: What is the `winpercent` value for "Tootsie Roll Snack Bars"? 49.65

```
#rownames(candy)
candy["Almond Joy", "winpercent"]
```

```
[1] 50.34755
```

```
candy["Almond Joy",]$winpercent
```

```
[1] 50.34755
```

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

```
[1] 76.7686
```

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

```
[1] 49.6535
```

```
#Could also do this all at once in pipe syntax, once load dplyr package
#The %in% function is asking which of the following info is within the previous vector/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
```

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

```
                      winpercent
Almond Joy              50.34755
Kit Kat                 76.76860
Tootsie Roll Snack Bars 49.65350
```

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 to the rest of the columns.

Q7. What do you think a zero and one represent for the candy$chocolate column? Whether a candy has chocolate or not.

```
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

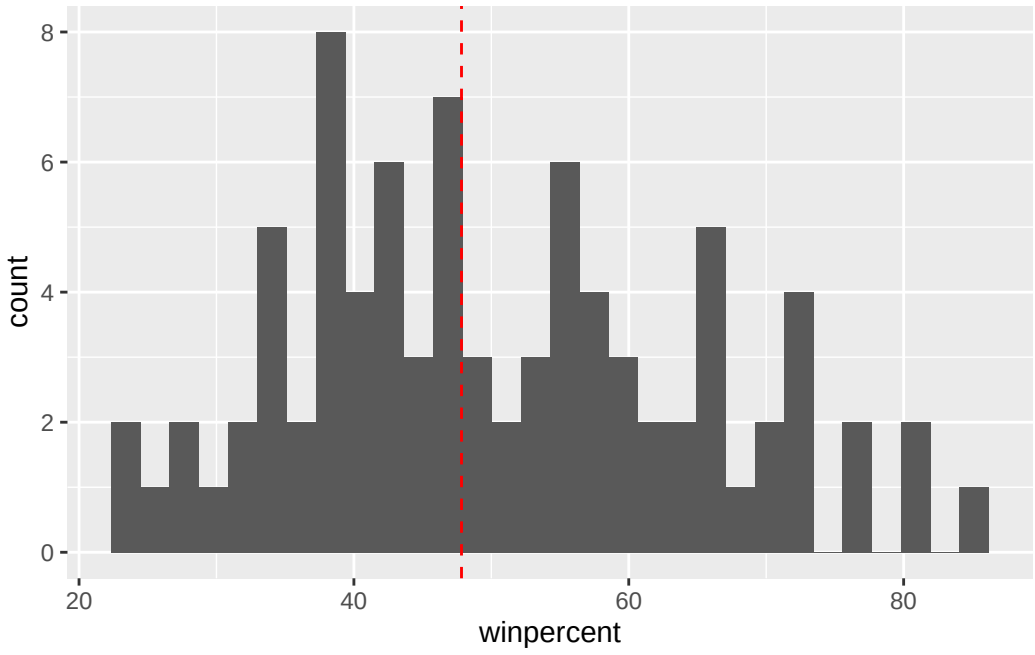
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.00	1.00	
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.00	1.00	
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.00	1.00	
nougat	0	1	0.08	0.28	0.00	0.00	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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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	

Q8. Plot a histogram of winpercent values see code Q9. Is the distribution of winpercent values symmetrical? distribution is not symmetrical Q10. Is the center of the distribution above or below 50%? center is below 50%, at 47% Q11. On average is chocolate candy higher or lower ranked than fruit candy? Chocoalte is higher on avg Q12. Is this difference statistically significant?

```
library("ggplot2")

#Q8, 9, 10
ggplot(candy, aes(winpercent)) + geom_histogram() + geom_vline(aes(xintercept = median(winpe
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
median(candy$winpercent)
```

```
[1] 47.82975
```

```
#Q11, 12
chocolate_avg_rank <- candy %>%
    filter(chocolate == 1) %>%
    select(winpercent)

fruity_avg_rank <- candy %>%
    filter(fruity == 1) %>%
    select(winpercent)

mean(chocolate_avg_rank$winpercent) > mean(fruity_avg_rank$winpercent)
```

```
[1] TRUE
```

```
t.test(chocolate_avg_rank$winpercent, fruity_avg_rank$winpercent)
```

```
    Welch Two Sample t-test

data:  chocolate_avg_rank$winpercent and fruity_avg_rank$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
```

Q13. What are the five least liked candy types in this set? 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? Reeses PB Cup, Reeses Mini, Twix, Kit Kat, Snickers

```
#tell the candy number (row number) with the lowest to highest winpercent order
inds <- order(candy$winpercent)
head(candy[inds,], 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
```

```
(tail(candy[inds,], 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
```

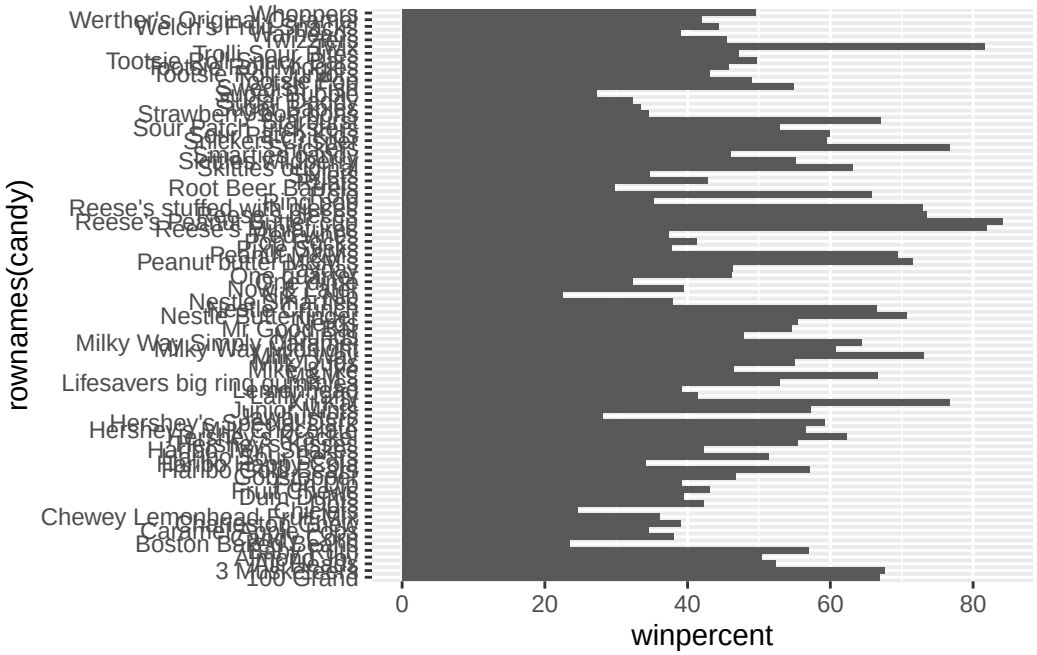
```
#can also use decreasing = T argument inside order function or rev() function to reverse the
```

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

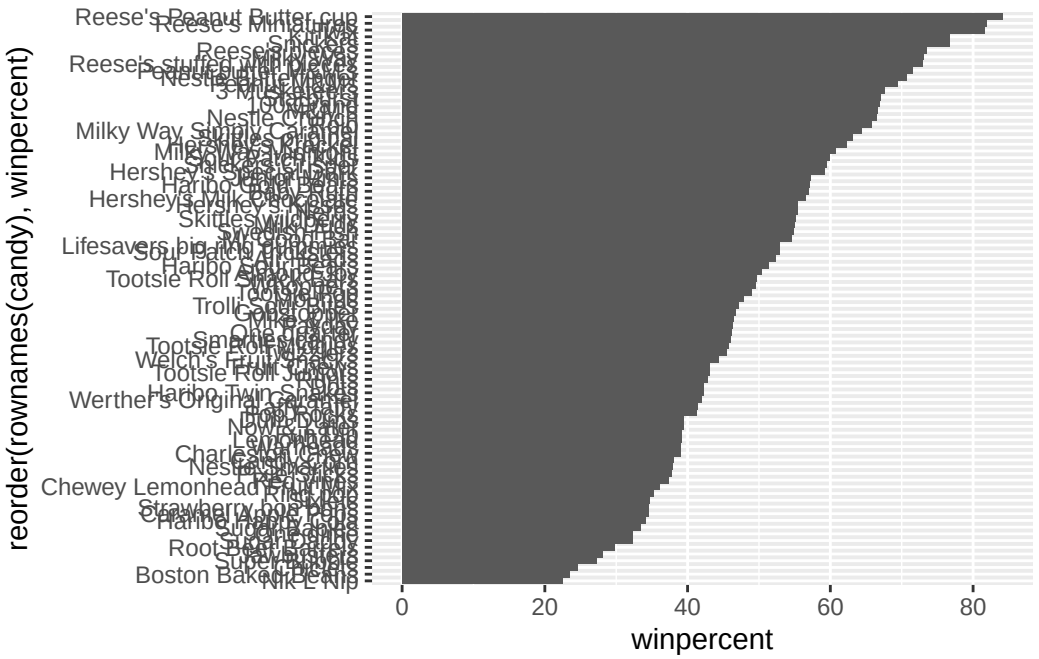
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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

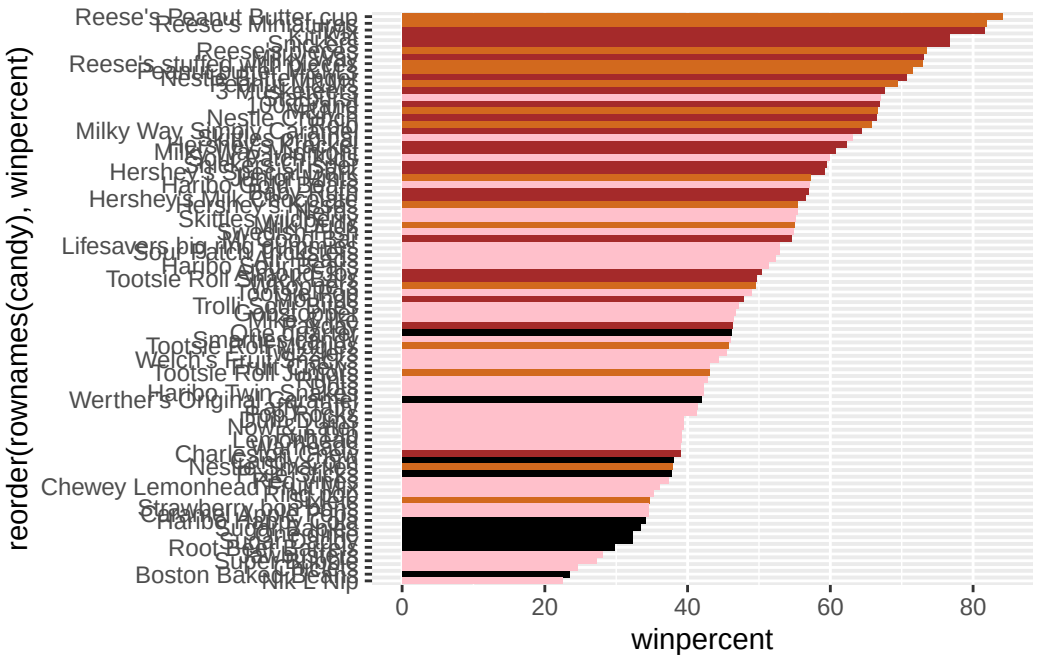
```
ggplot(candy, aes(winpercent, rownames(candy))) + geom_col()
```




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

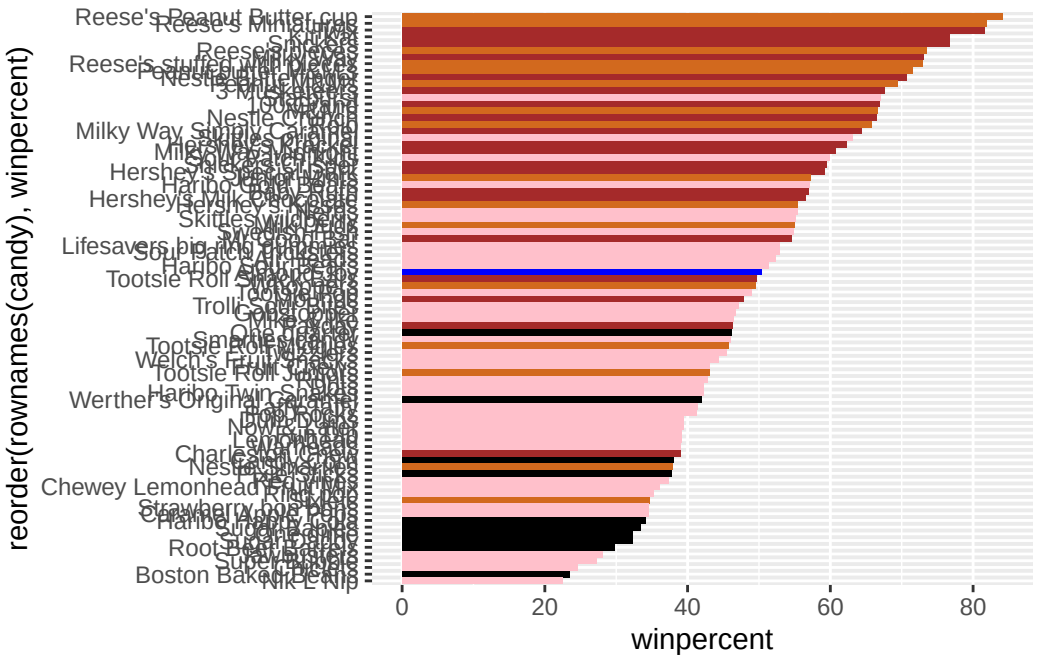



```
#rep function replicates first argument for specified number of times
my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "brown"
my_cols[as.logical(candy$fruity)] = "pink"

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



Extra: Color favorite candy by favorite color

```
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"
my_cols[as.logical(rownames(candy) == "Almond Joy")] <- "blue"

# ?Could use this syntax: candy[,"Almond Joy"] if the candy names were still themselves a col

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

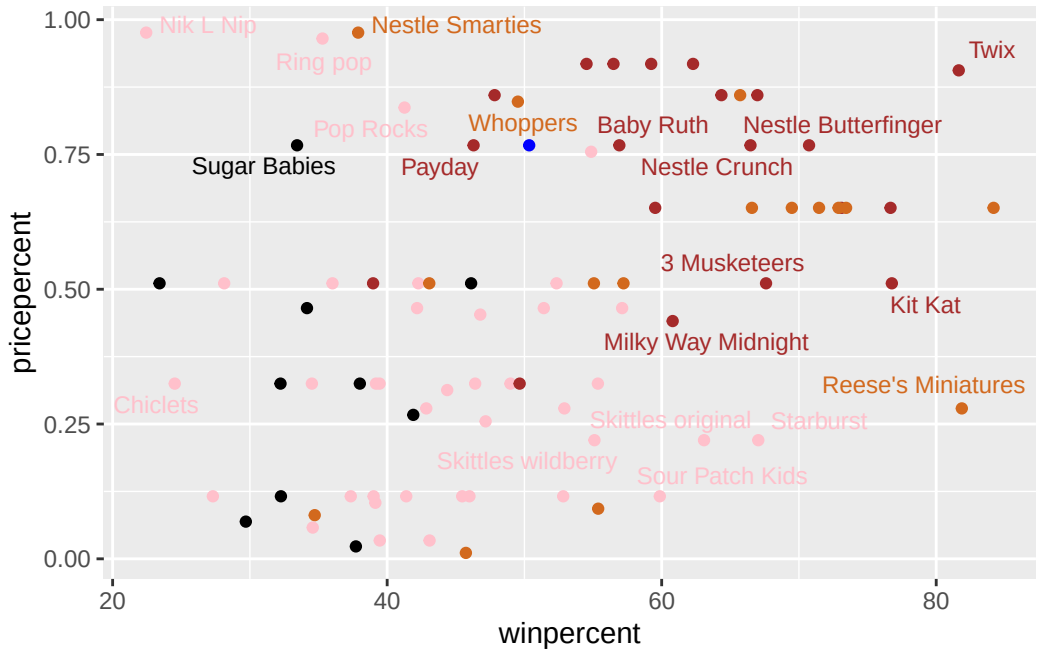


Q19. Which candy type is the highest ranked in terms of winpercent for the least money - i.e. offers the most bang for your buck? Reese's mini have highest winpercent at lowest pricepercent

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

```
library(ggrepel)

ggplot(candy, aes(winpercent, pricepercent, label = rownames(candy))) + geom_point(col = my_
```

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

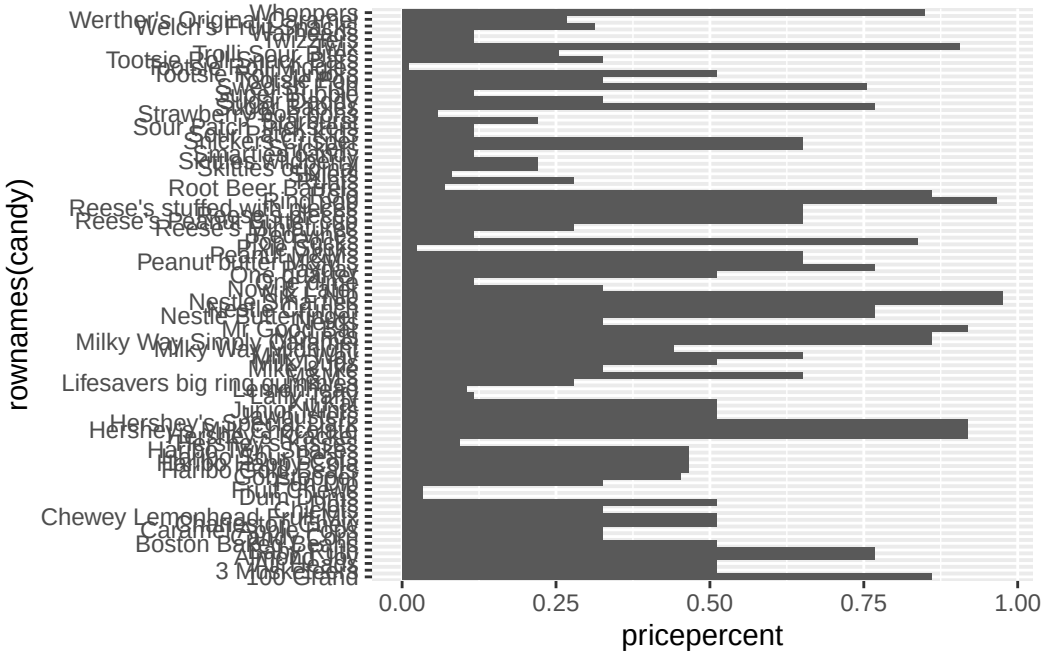



```
#c(11,12) gives us only the price and winpercent columns of the ordered rows we ask for.
candy_price_order <- order(candy$pricepercent, decreasing = TRUE)
head(candy[candy_price_order,c(11,12)], 5)
```

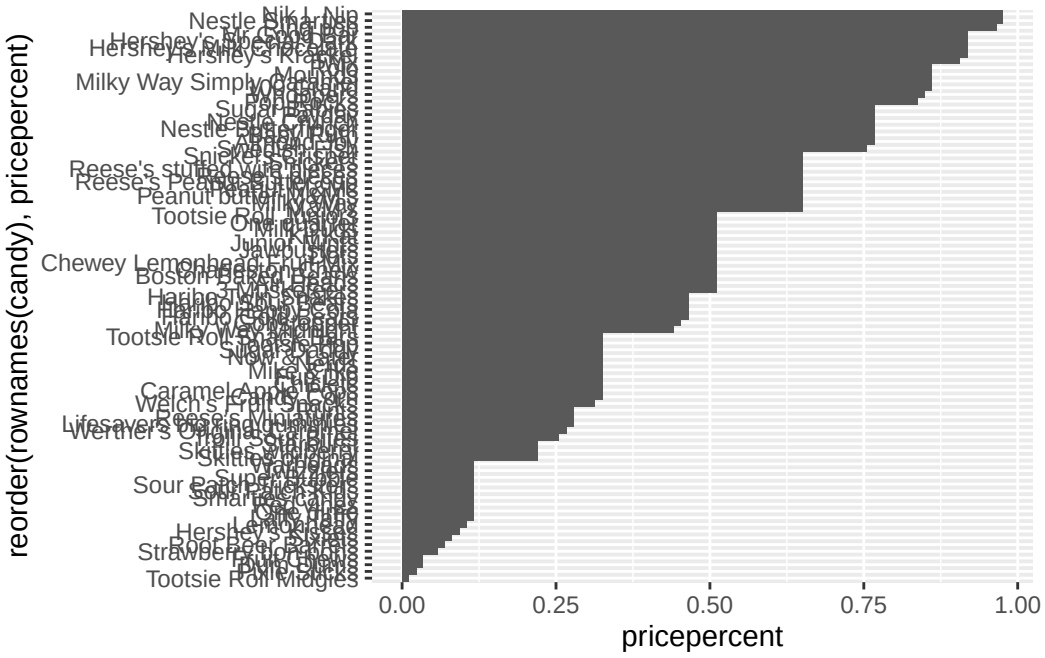
```
                      pricepercent winpercent
Nik L Nip                    0.976   22.44534
Nestle Smarties              0.976   37.88719
Ring pop                     0.965   35.29076
Hershey's Krackel            0.918   62.28448
Hershey's Milk Chocolate     0.918   56.49050
```

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().

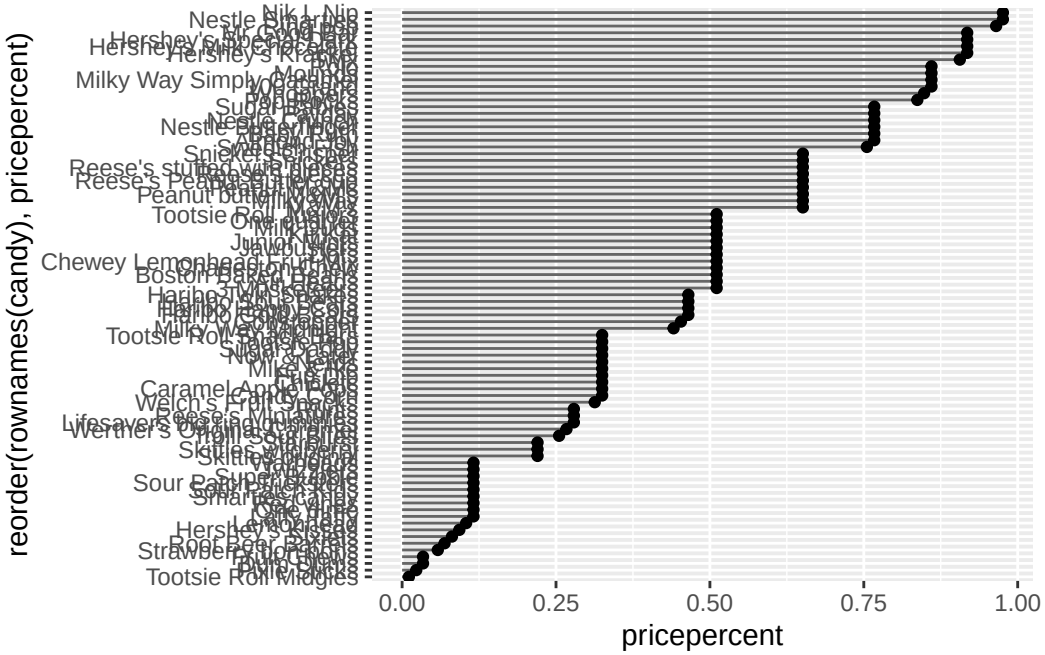
```
ggplot(candy, aes(pricepercent, rownames(candy))) + geom_col()
```




```
#order x-axis by value
ggplot(candy, aes(pricepercent, reorder(rownames(candy), pricepercent))) + geom_col()
```



```
#make a dot/lollipop chart
ggplot(candy, aes(pricepercent, reorder(rownames(candy), pricepercent))) + geom_segment(aes(y
                    xend = 0), col="gray40") + geom_point()
```



##Explore the correlation structure

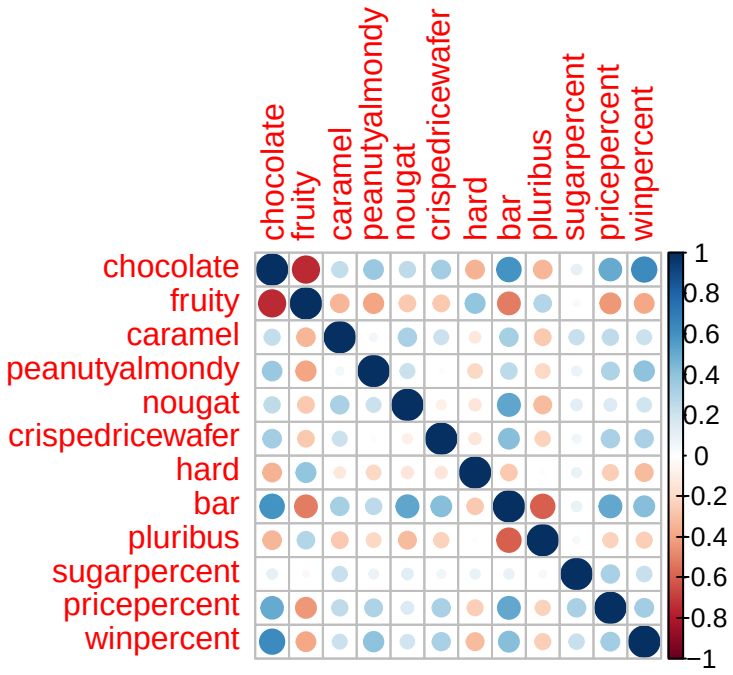
Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)? Fruity and chocolate candies are anti-correlated. Bar and pluribus candies too, understandably.

Q23. Similarly, what two variables are most positively correlated? Chocolate is highly correlated with winpercent!

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



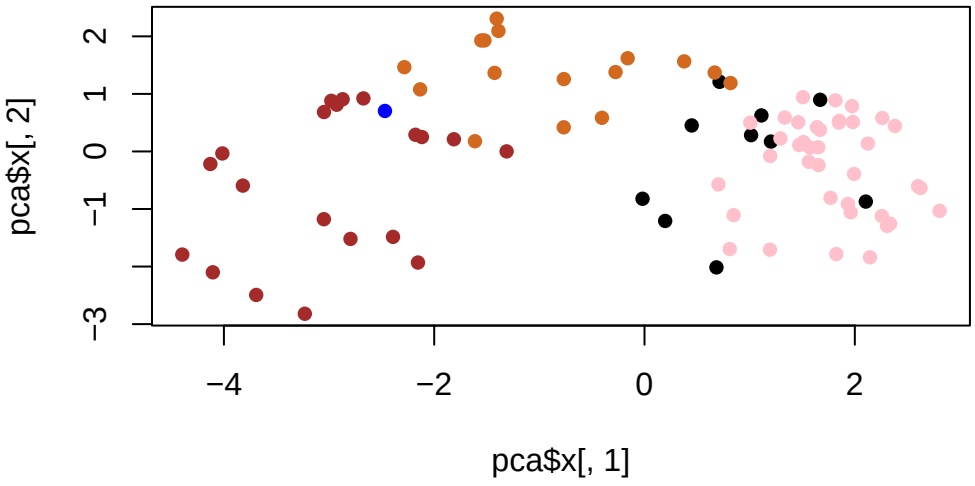

```
#I dig this plot
```

Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you? Fruity, hard, and pluribus variables. These variables are highly anticorrelated with other variables such as chocolate and bar and winpercent that also have high magnitudes in PC1 but in the opposite direction.

```
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
```

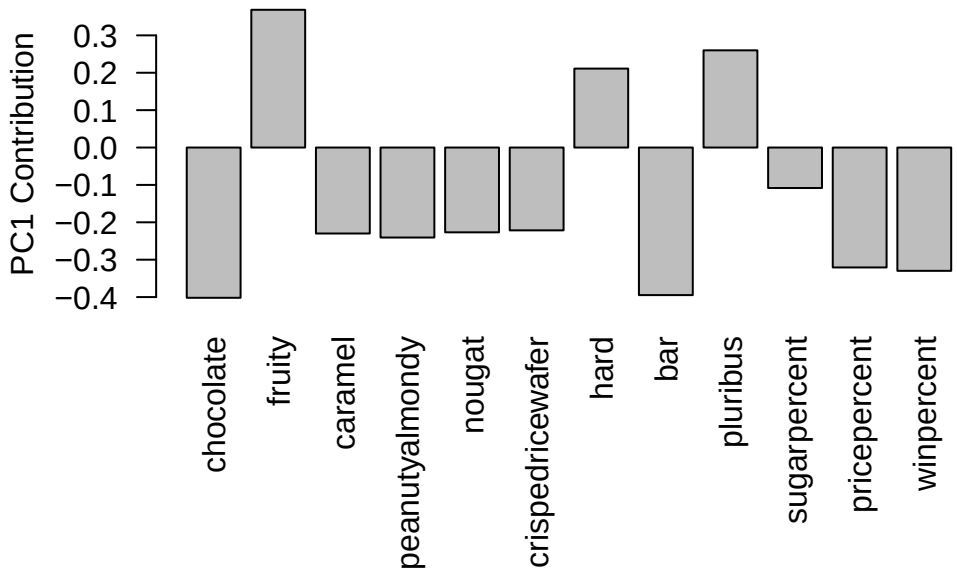
```
#recall that x column contains the pca coordinates
plot(pca$x[,1], pca$x[,2], col = my_cols, pch =16)
```



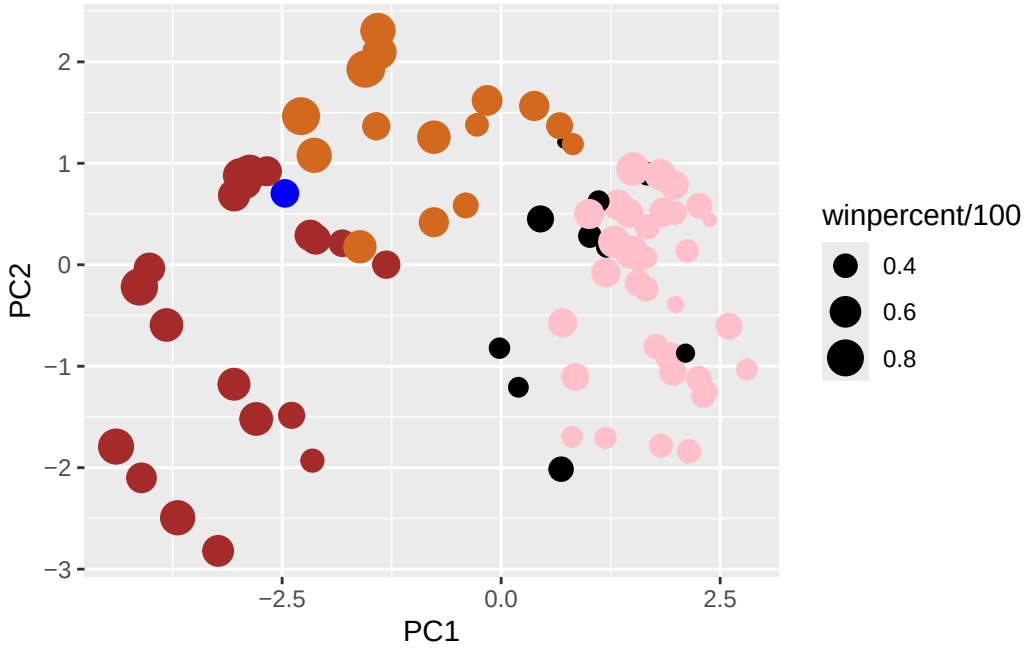

```
#tell where each type of candy lies on the pca pc1 column
pca$rotation[,1]
```

```
     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
```

```
#barplot of the pca1 coordinates for our candies, par function sets the margians
par(mar=c(8,4,2,2))
barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```

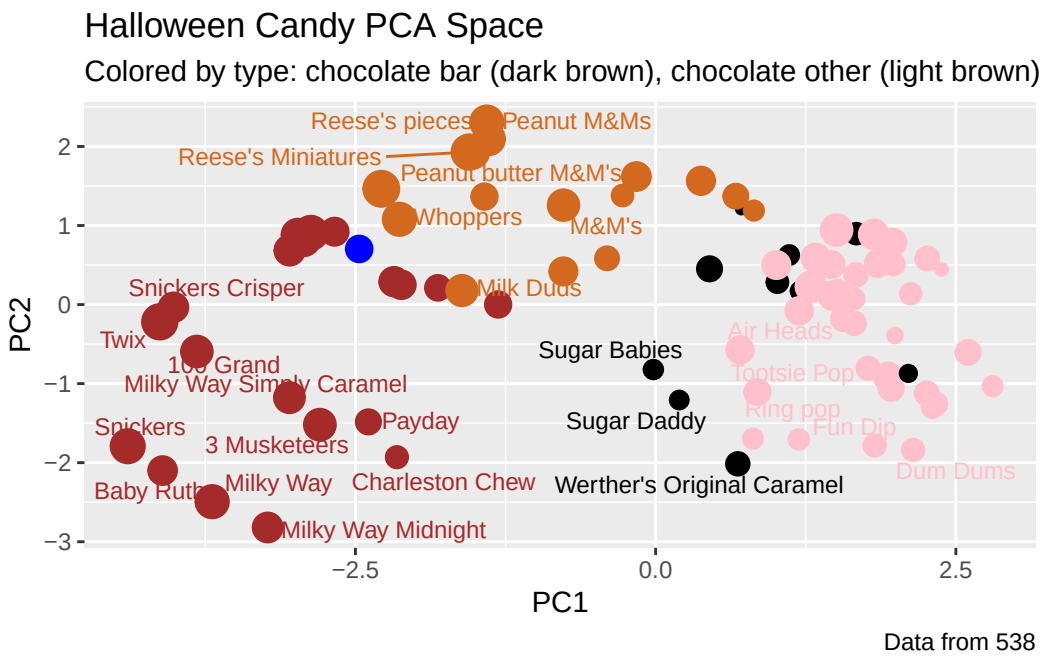



```
#putting pca results for pc1,2, and 3 into our candy data frame to help with ggplot functiona
my_data <- cbind(candy, pca$x[,1:3])

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
```




```
#use the ggrepel function
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
```


Halloween Candy PCA Space
Colored by type: chocolate bar (dark brown), chocolate other (light brown),
Reese's pieces
Peanut M&Ms
Reese's Miniatures
Peanut butter M&M's
Whoppers
M&M's
Snickers Crisper
Milk Duds
Twix
Sugar Babies
Air Heads
Tootsie Pop
Ring pop
Fun Dip
Dum Dums
Milky Way Simply Caramel
Payday
Snickers
Sugar Daddy
3 Musketeers
Baby Ruth
Milky Way
Charleston Chew
Werther's Original Caramel
Milky Way Midnight
PC2
PC1
2
1
0
−1
−2
−3
−2.5
0.0
2.5
Data from 538