Class 09

Kaliyah Adei-Manu (A18125684)

Background

In this mini project, you will explore FiveThirtyEight's Halloween candy dataset.

We will use lots of **ggplot** some basic stats, correlation analysis and PCA to make sense of the landscape of US candy, something hopefully more relatable than protenomics and transcriptomics.

Data Import

Our dataset is a CSV file so we use `read.csv()`

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

There are 85 rows in this dataset

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

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

```
[1] 38
```

38 out of 85 types of candies are fruity > Q3. What is your favorite candy (other than Twix) in the dataset and what is it's winpercent value?

My favorite candy is 3 Musketeers

```
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(row.names(candy)== "3 Musketeers") |>
  select(winpercent)
```

```
             winpercent
3 Musketeers   67.60294
```

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

```
library(dplyr)
candy |>
  filter(row.names(candy)== "Kit Kat") |>
  select(winpercent)
```

```
         winpercent
Kit Kat    76.7686
```

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

```
library(dplyr)
candy |>
  filter(row.names(candy)== "Kit Kat") |>
  select(winpercent)
```

```
         winpercent
Kit Kat    76.7686
```

```
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	

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

Yes!

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

```
skim(candy$chocolate)
```

Table 3: Data summary

Name	candy$chocolate
Number of rows	85
Number of columns	1
_____	
Column type frequency:	
numeric	1
_____	
Group variables	None

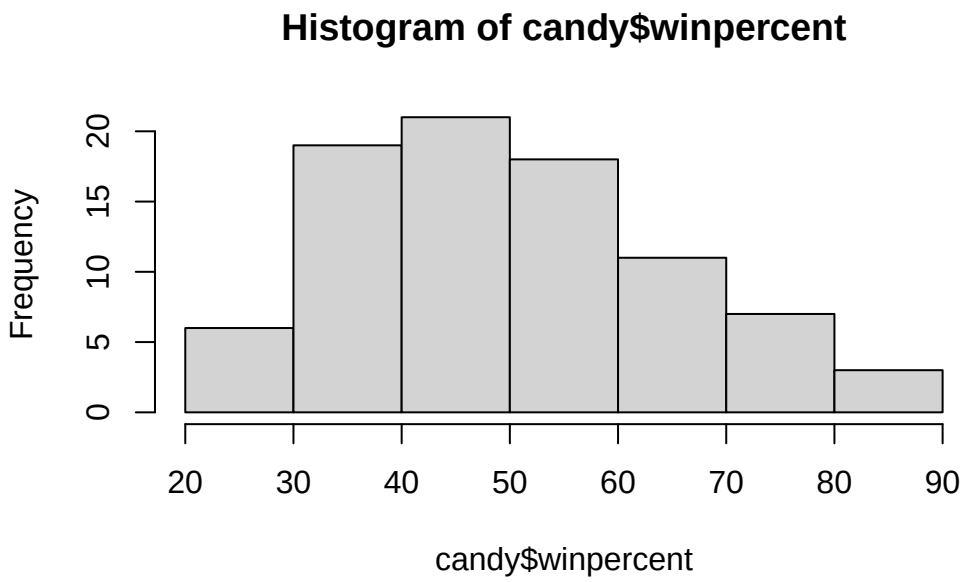
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	0	1	0.44	0.5	0	0	0	1	1	

Exploratory analysis

Q8. Plot a histogram of winpercent values

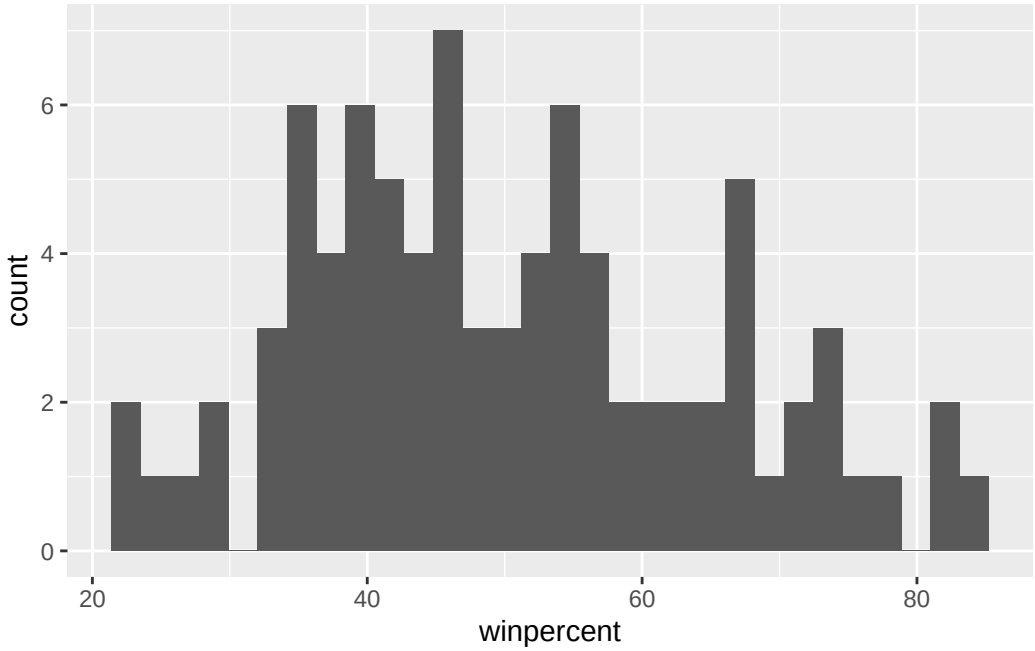
```
hist(candy$winpercent)
```




```
library(ggplot2)

ggplot(candy)+
  aes(winpercent)+
  geom_histogram()
```

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



Q9. Is the distribution of winpercent values symmetrical?

No

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

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

```
summary(candy$winpercent)
```

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

The eman is above but the median is below 50.

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

1. Find all chocolate candy
2. Get their winpercent values 3.Find mean 4.Find all fruit candy 5.Get their winpercent values 6.Find the mean

3. Comapre the two means

```
choc.candy <- candy[candy$chocolate ==1,]
choc.win <- choc.candy$winpercent
mean(choc.win)
```

```
[1] 60.92153
```

```
fruit.win <- candy[candy$fruity ==1,]$winpercent
mean(fruit.win)
```

```
[1] 44.11974
```

```
mean(candy$winpercent[candy$chocolate==1])
```

```
[1] 60.92153
```

```
mean(candy$winpercent[candy$fruity==1])
```

```
[1] 44.11974
```

On average, chocolate candy is ranked higher than chocolate candy

Q12. Is this difference statistically significant?

```
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
```

Yes this difference is significant, the p-value is 2.871e-08 which is below 0.05 making the results significant.

Overall Candy Ranking

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

```
y <- c("y", "a", "z")
sort(y)
```

```
[1] "a" "y" "z"
```

```
y
```

```
[1] "y" "a" "z"
```

```
order(y)
```

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

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

```
                  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
```

```
head(candy[ord.ind,],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
```

bottom 5 are...

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

```
tail(candy[ord.ind,],5)
```

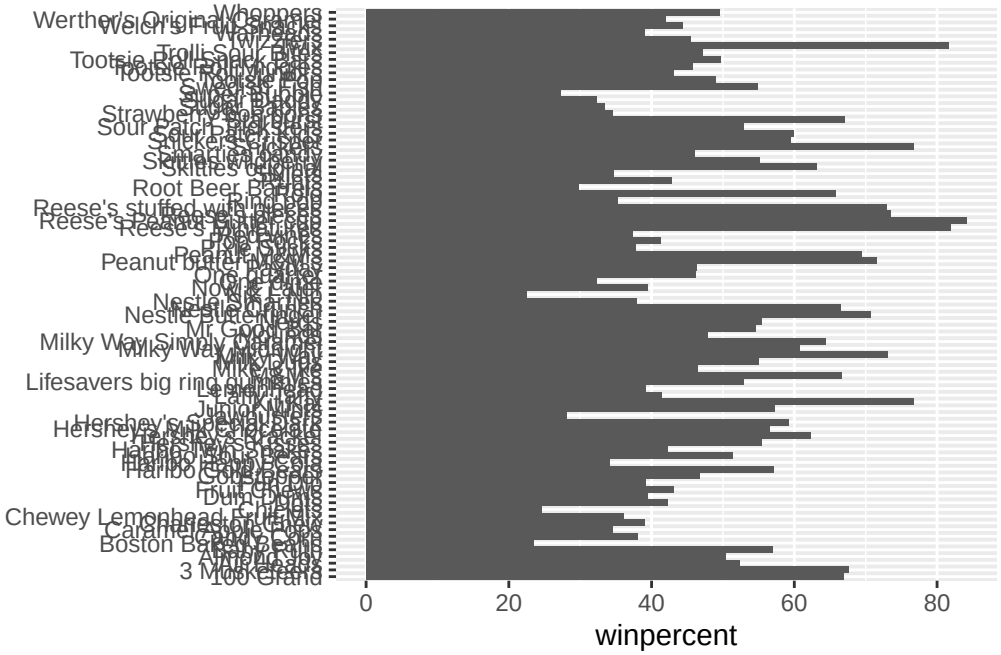
```
                        chocolate fruity caramel peanutyalmondy nougat
Snickers                        1      0       1              1      1
Kit Kat                         1      0       0              0      0
Twix                            1      0       1              0      0
Reese's Miniatures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

top 5 are...

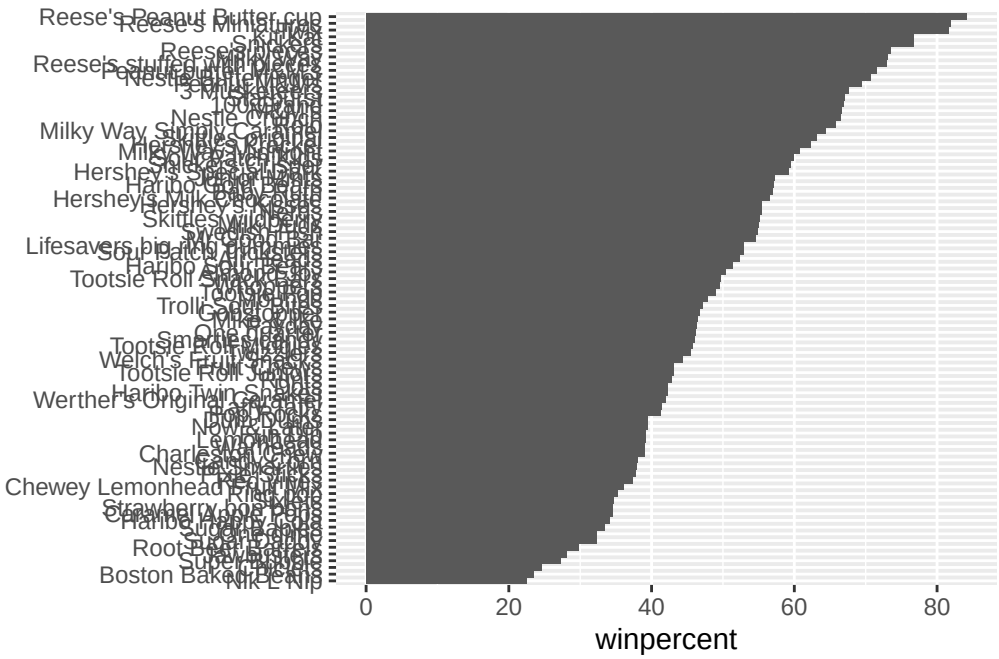
Q15. Make a first barplot of candy ranking based on winpercent values.

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



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

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

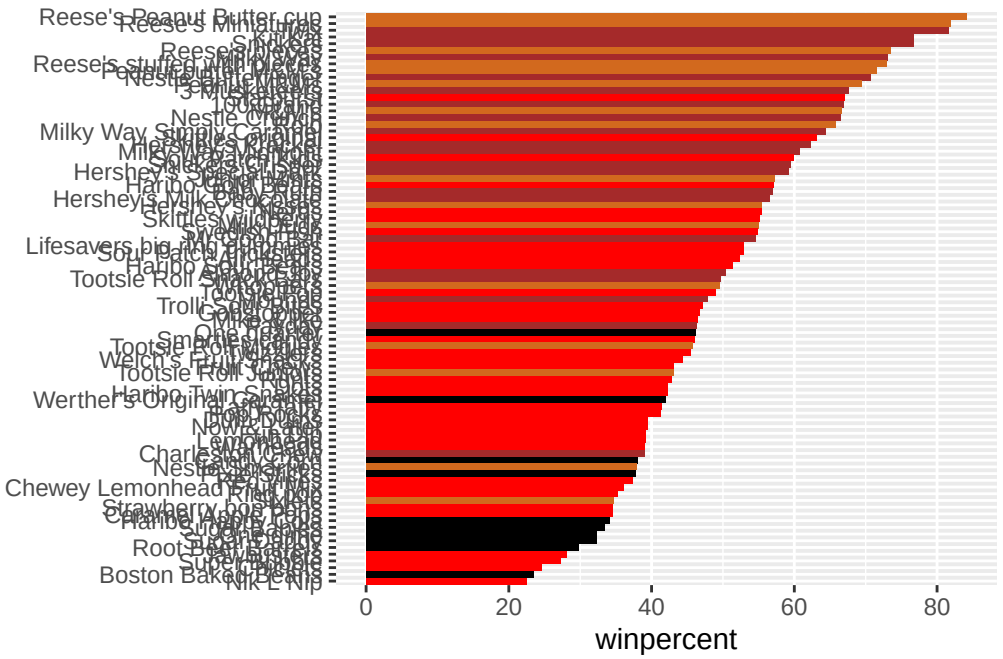


we need custom color vector

```
my_cols <- rep("black", nrow(candy))
my_cols[candy$chocolate==1] <- "chocolate"
my_cols[candy$bar==1] <- "brown"
my_cols[candy$fruity==1] <- "red"
my_cols
```

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

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



Q17. What is the worst ranked chocolate candy?

The worst ranked chocolate candy is sixlets.

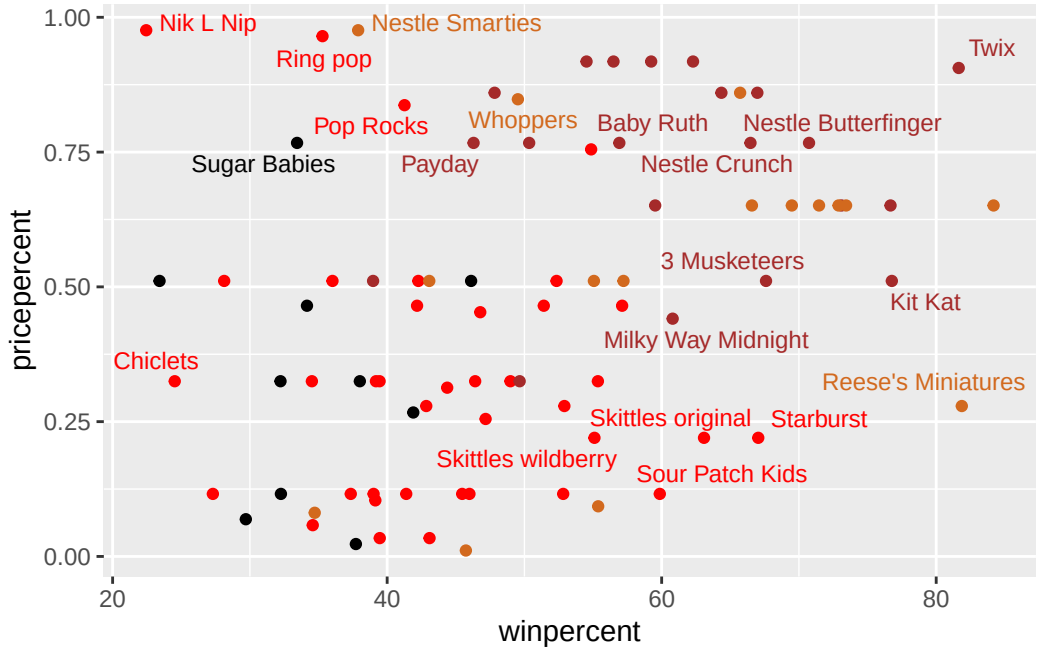
Q18. What is the best ranked fruity candy?

The best ranked fruity candy is starbursts

```
library(ggrepel)

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

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

```
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
```

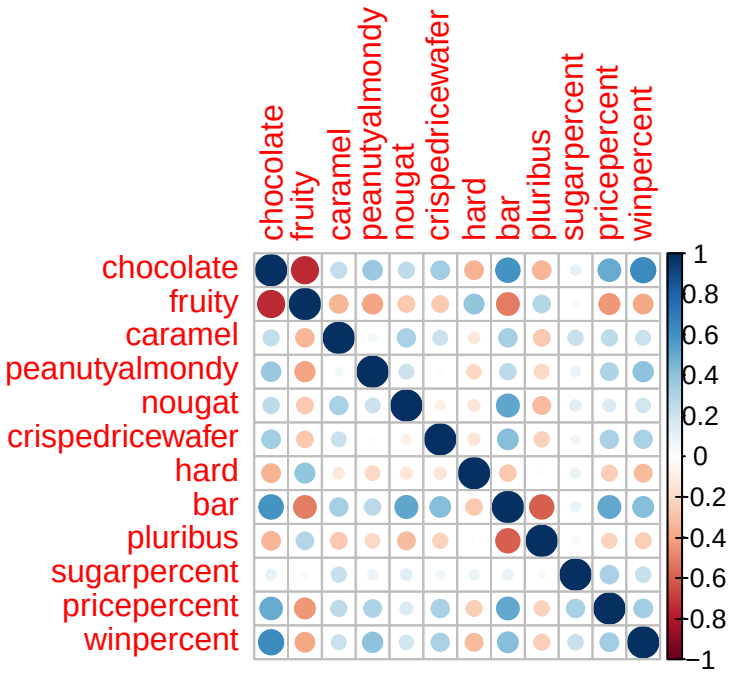
The least popular is Nik L Nip as shown on the graph.

Exploring the correlation structure

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



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

Two variables that are anti-correlated are chocolate and fruity, and also bar and pluribus.

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

Chocolate is strongly positively correlated with winpercent meaning it is more popular. Chocolate is also strongly positively correlated with bar.

Principal Component Analysis

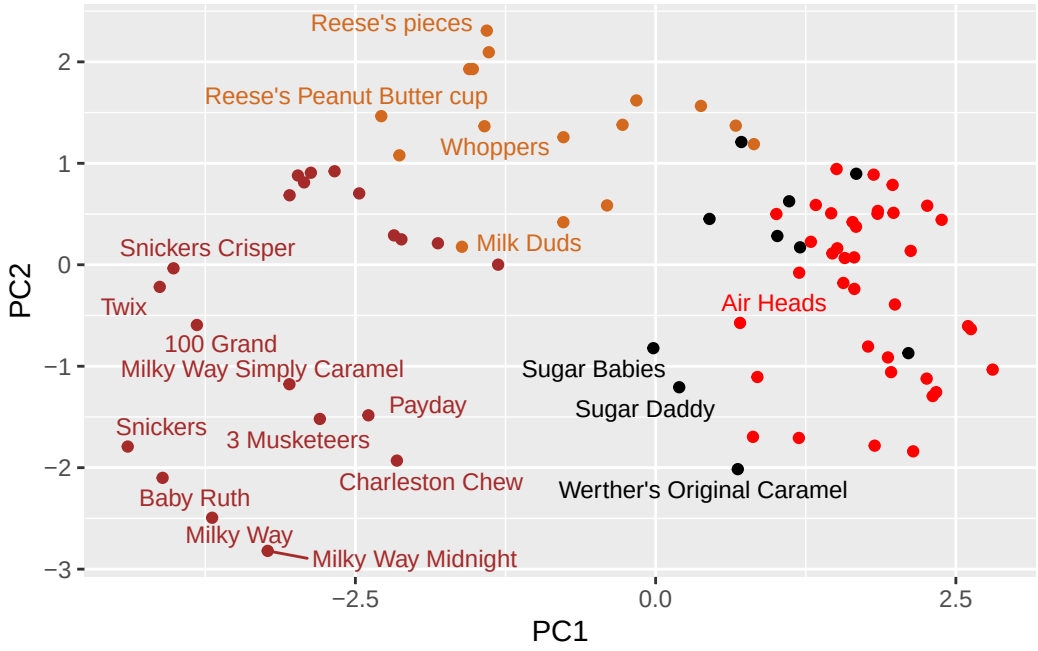
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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
```

Score Plot...

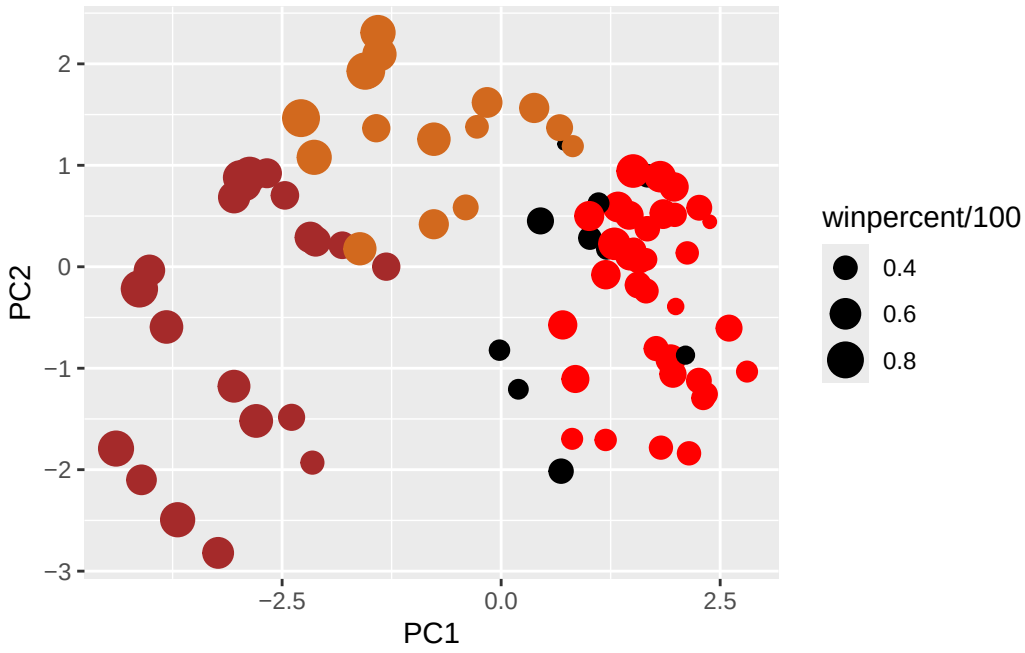
```
ggplot(pca$x)+
  aes(PC1, PC2, label=row.names(pca$x))+
  geom_point(col= my_cols)+
  geom_text_repel(max.overlaps = 5, size=3.3,col=my_cols)
```

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




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

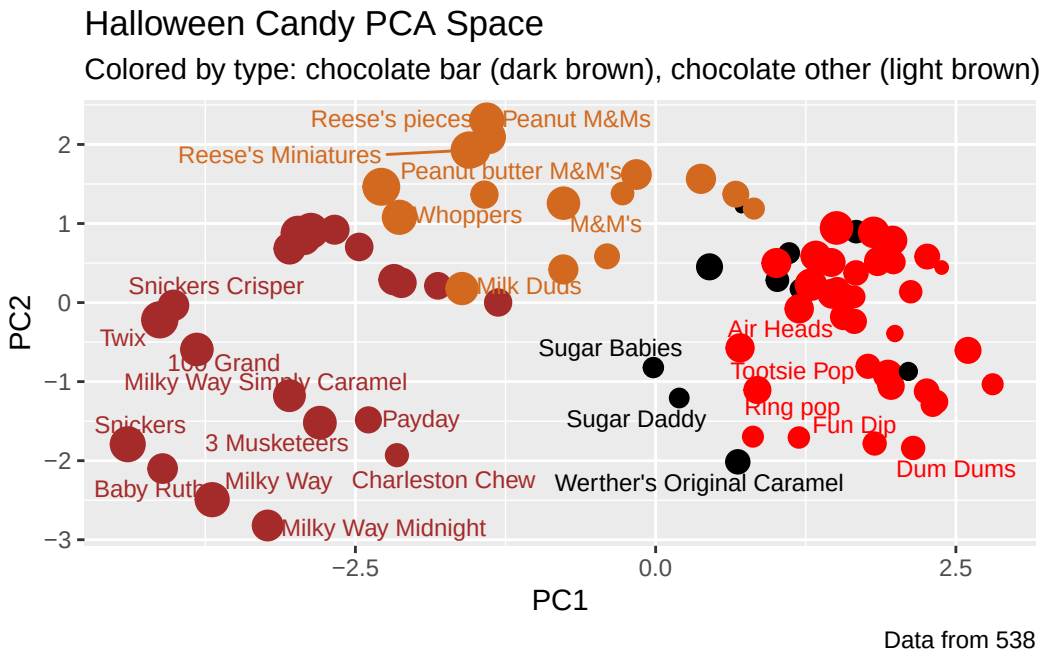
```
 p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col= my_cols)
p
```




```
library(ggrepel)

p + geom_text_repel(size=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")
```

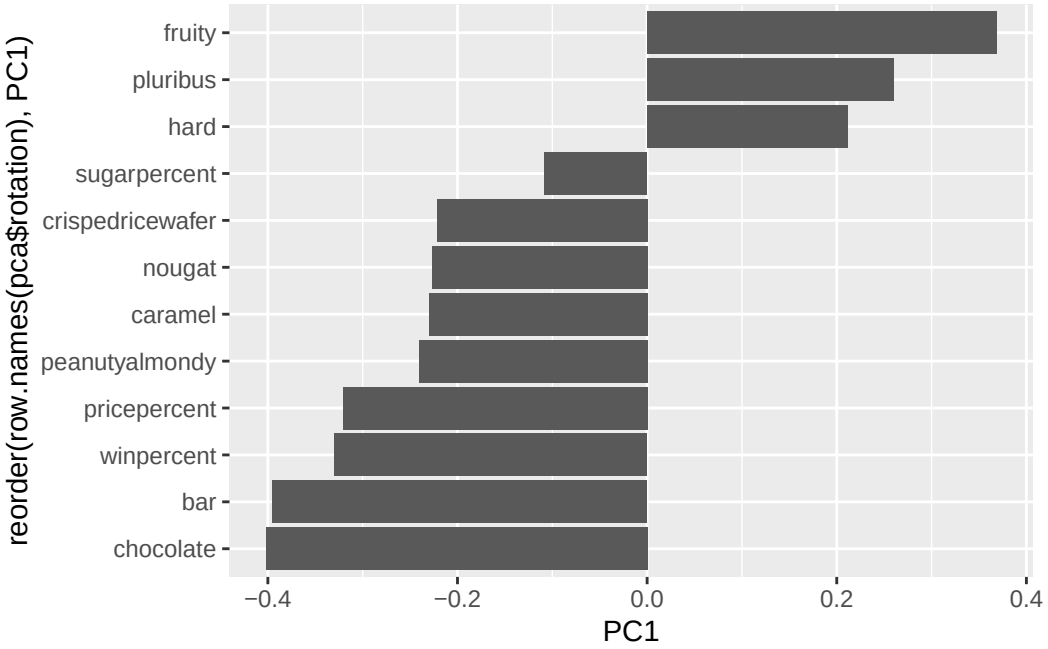
```
Warning: ggrepel: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




```
##library(plotly)
##ggplotly(p)
```

Q24. Complete the code to generate the loadings plot above. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you? Where did you see this relationship highlighted previously?

```
ggplot(pca$rotation)+
  aes(PC1,reorder(row.names(pca$rotation),PC1))+
  geom_col()
```



Pluribus, fruity, and hard are correlated, and so are bar, chocolate, pricpercent,a nd winpercent. These results make sense and go along with what the previous corrplot showed.

> Q25. Based on your exploratory analysis, correlation findings, and PCA results, what combination of characteristics appears to make a "winning" candy? How do these different analyses (visualization, correlation, PCA) support or complement each other in reaching this conclusion?

A "winning" candy tends to be chocolate, a bar,have peanuts,or caramel. This findings are shown in both the correlation plots and PCA.