Class10

AUTHOR

Ryan Fong

We will first download and import the 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?

```
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

The winpercent variable can be used to see a popularity of a candy relative to the other candies.

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

```
[1] 81.64291
```

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

My favorite candy are air heads.

```
candy["Air Heads", ]$winpercent
```

```
[1] 52.34146
```

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

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

The skimr package will be downloaded so the `skimr()` function can be used to find the overview of the dataset.

the overview of the dataset.

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

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	p7
chocolate	0	1	0.44	0.50	0.00	0.00	0.00	1.0
fruity	0	1	0.45	0.50	0.00	0.00	0.00	1.0
caramel	0	1	0.16	0.37	0.00	0.00	0.00	0.0
peanutyalmondy	0	1	0.16	0.37	0.00	0.00	0.00	0.0
nougat	0	1	0.08	0.28	0.00	0.00	0.00	0.0
crispedricewafer	0	1	0.08	0.28	0.00	0.00	0.00	0.0
hard	0	1	0.18	0.38	0.00	0.00	0.00	0.0
bar	0	1	0.25	0.43	0.00	0.00	0.00	0.0
pluribus	0	1	0.52	0.50	0.00	0.00	1.00	1.0
sugarpercent	0	1	0.48	0.28	0.01	0.22	0.47	0.7
pricepercent	0	1	0.47	0.29	0.01	0.26	0.47	0.6
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.8

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

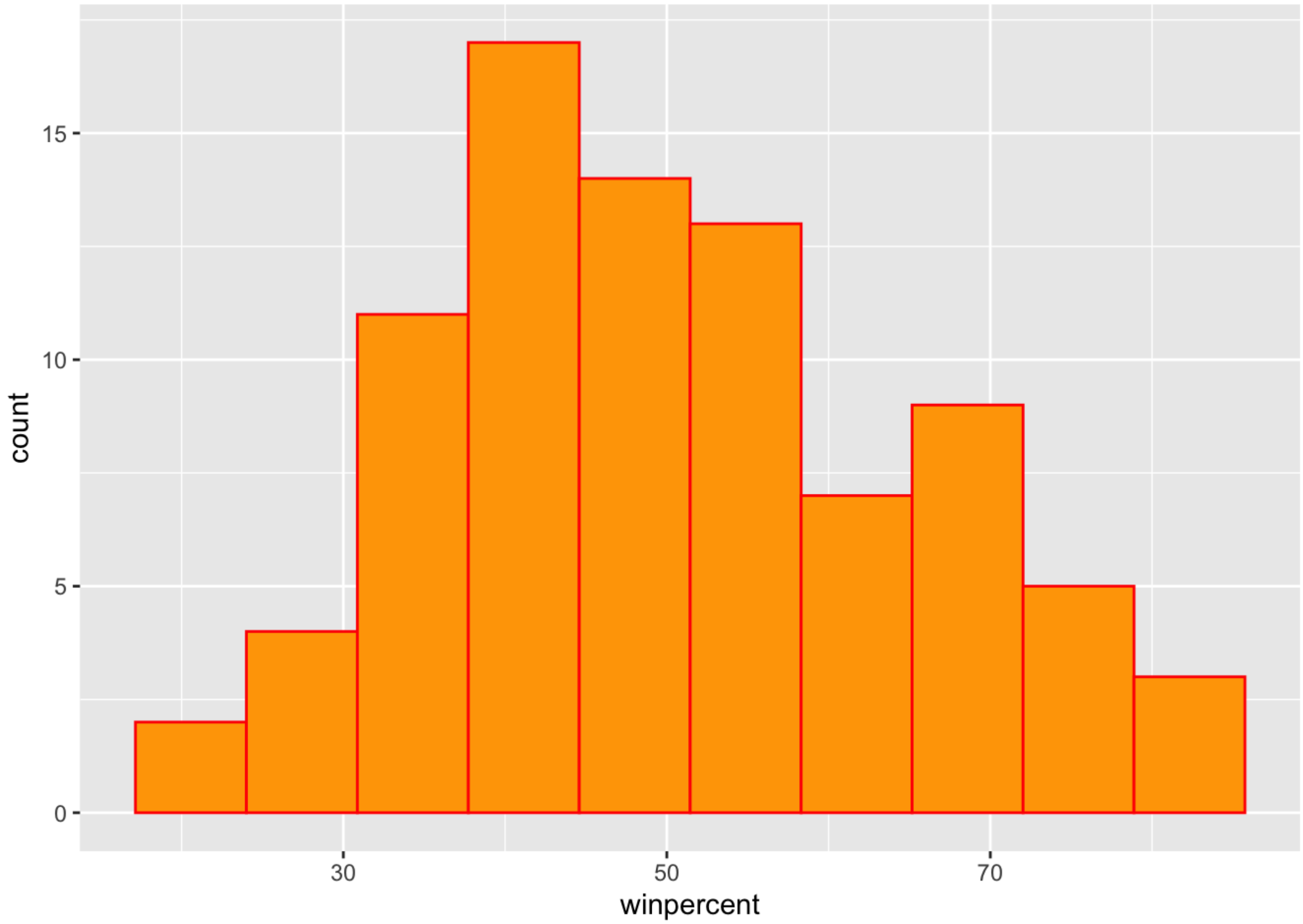
Yes, it would be winpercent.

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

0 represents that there are no chocolate in the candy and 1 means that there are chocolate in the candy.

> Q8. Plot a histogram of winpercent values

```r
library(ggplot2)

ggplot(candy) +
  aes(winpercent) +
  geom_histogram(bins=10, col="red",fill="orange")
```



> Q9. Is the distribution of winpercent values symmetrical?

Not really because the distribution is slightly skewed to the right.

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

The center of distribution is below 50 percent

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

```r
chocolate.inds <- as.logical(candy$chocolate)
chocolate.wins <- candy[chocolate.inds,]$winpercent
mean(chocolate.wins)
```

```
[1] 60.92153
```

```
fruity.inds <- as.logical(candy$fruity)
fruity.wins <- candy[fruity.inds,]$winpercent
mean(fruity.wins)
```

```
[1] 44.11974
```

Chocolate candy is higher ranked than fruity candy

> Q12. Is this difference statistically significant?

```
t.test(chocolate.wins,fruity.wins)
```

```

	Welch Two Sample t-test

data:  chocolate.wins and fruity.wins
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, because the p-value is small.

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

```
head(candy[order(candy$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
```

```
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
```

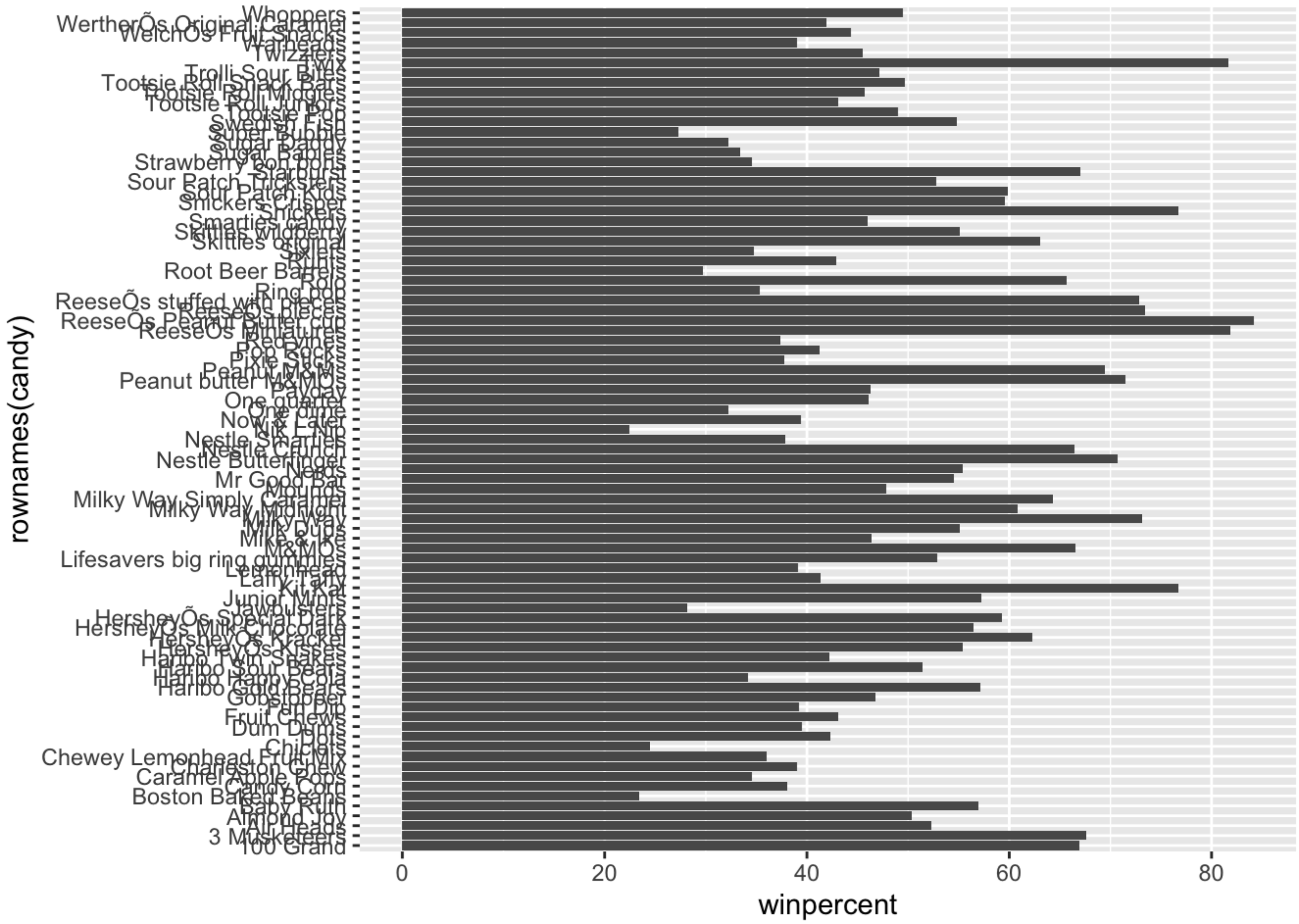
> Q14. What are the top 5 all time favorite candy types out of this set?

```
tail(candy[order(candy$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
ReeseÕs Miniatures                1      0       0              1
0
ReeseÕs 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
Twix                                      1    0   1        0
0.546
ReeseÕs Miniatures                        0    0   0        0
0.034
ReeseÕs 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
ReeseÕs Miniatures               0.279   81.86626
ReeseÕs Peanut Butter cup        0.651   84.18029
```

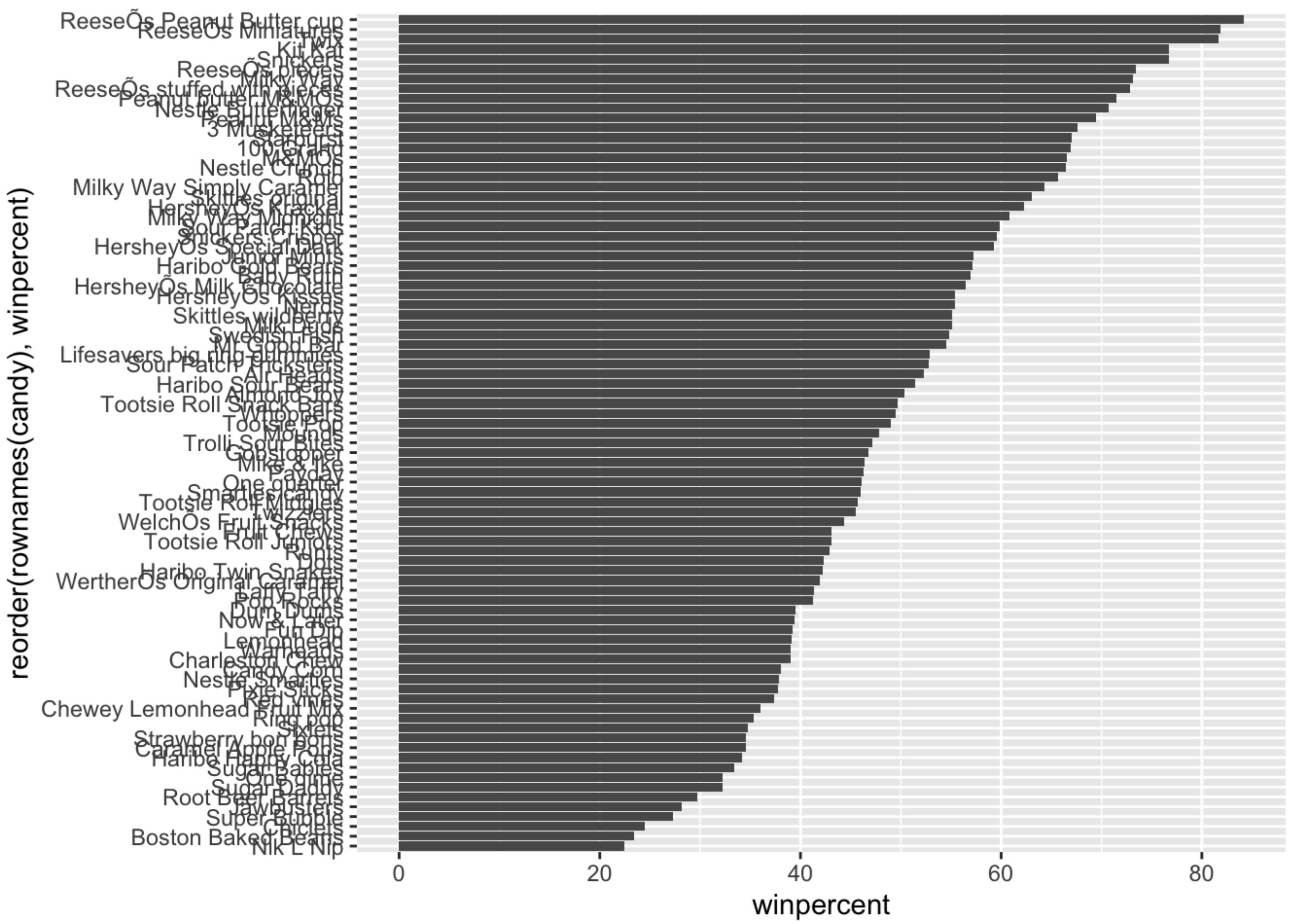
> Q15. Make a first barplot of candy ranking based on winpercent values.

```
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?

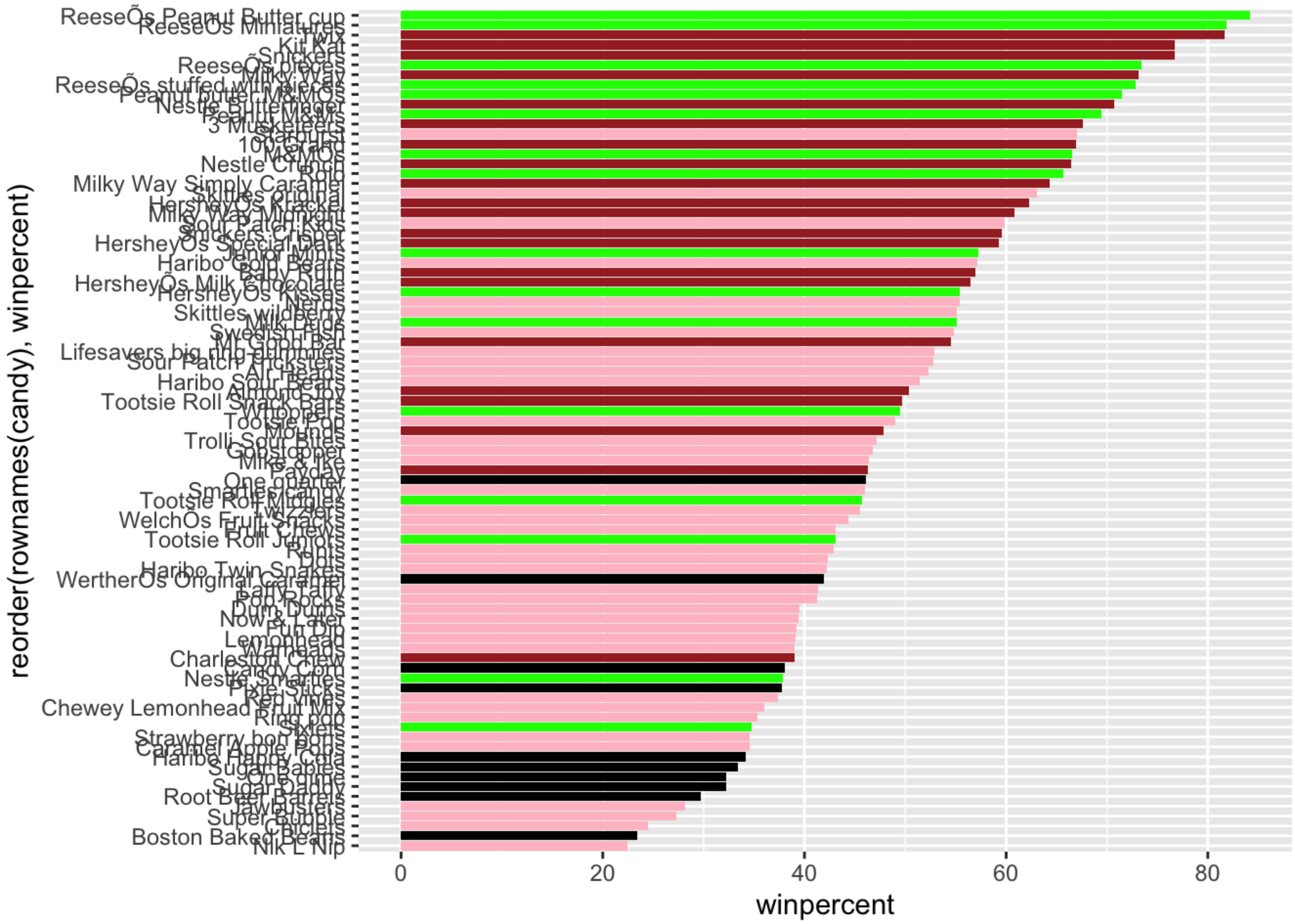
```
library(ggplot2)

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



Now colors will be added

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

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



> Q17. What is the worst ranked chocolate candy?

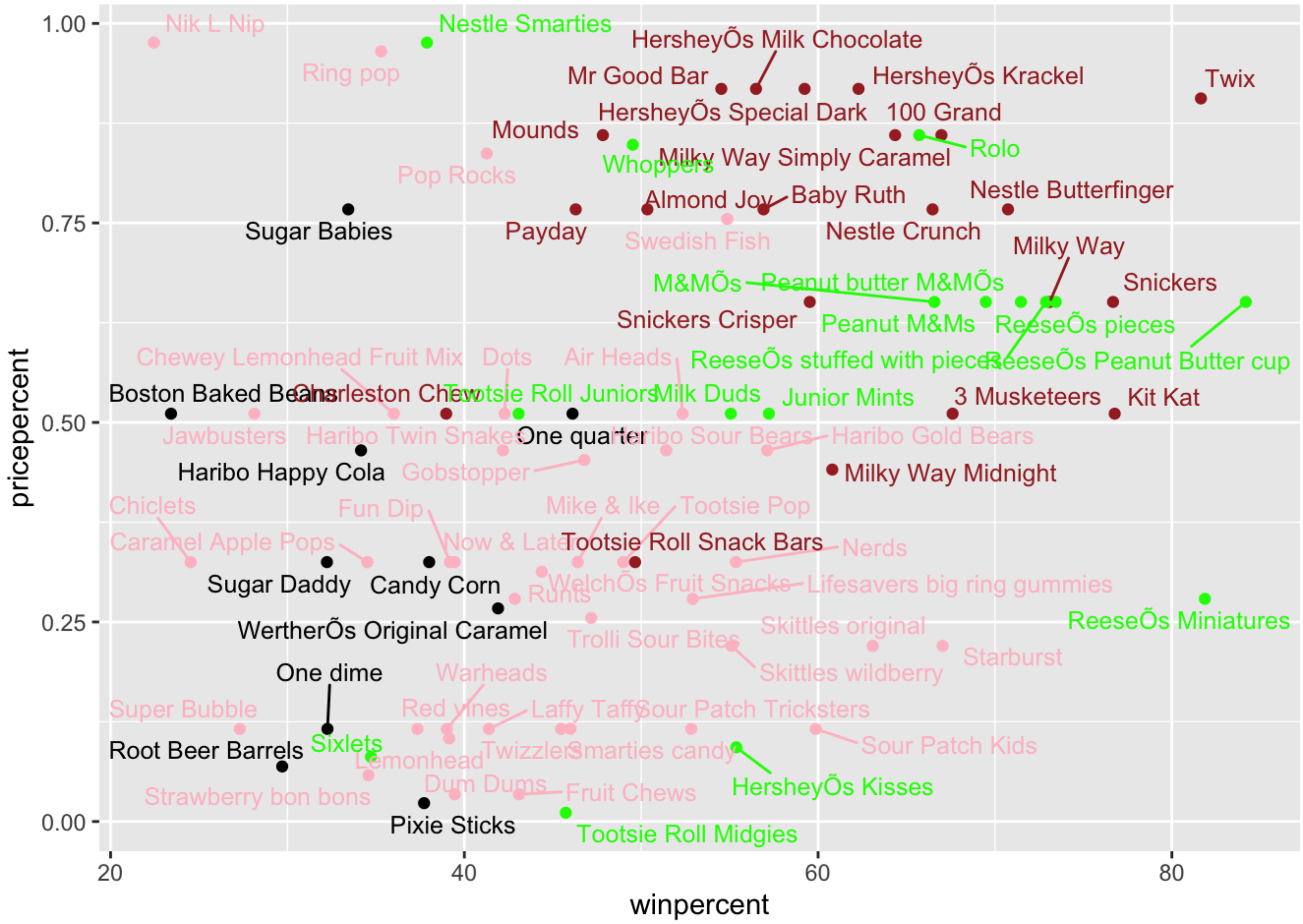
sixlets

> Q18. What is the best ranked fruity candy?

starbursts

A plot is made to compare the winpercent and pricepercent to compare the best value candy. First, ggrepel must be downloaded

```
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=3.3, max.overlaps = 25)
```



> 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?

Resses Miniatures

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

```r
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
HersheyÕs Krackel               0.918   62.28448
HersheyÕs Milk Chocolate        0.918   56.49050
```

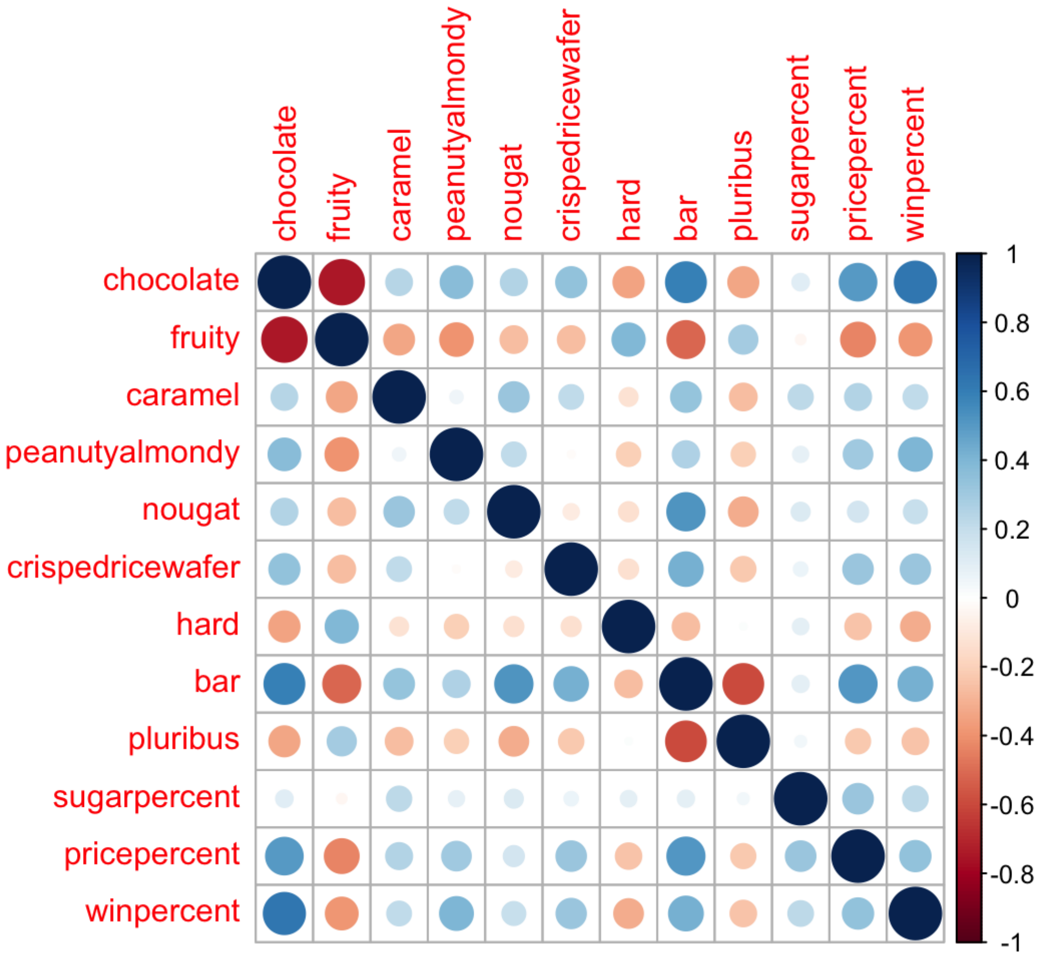
In the top 5 most expensive candies listed above, Nik L Nip is the least popular

corrplot will be use to see how the variables are related to one another. corrplot will first be downloaded

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

```
## corrplot 0.90 loaded
cij <- cor(candy)
corrplot(cij)
```



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

chocolate and fruity

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

chocolate and winpercent

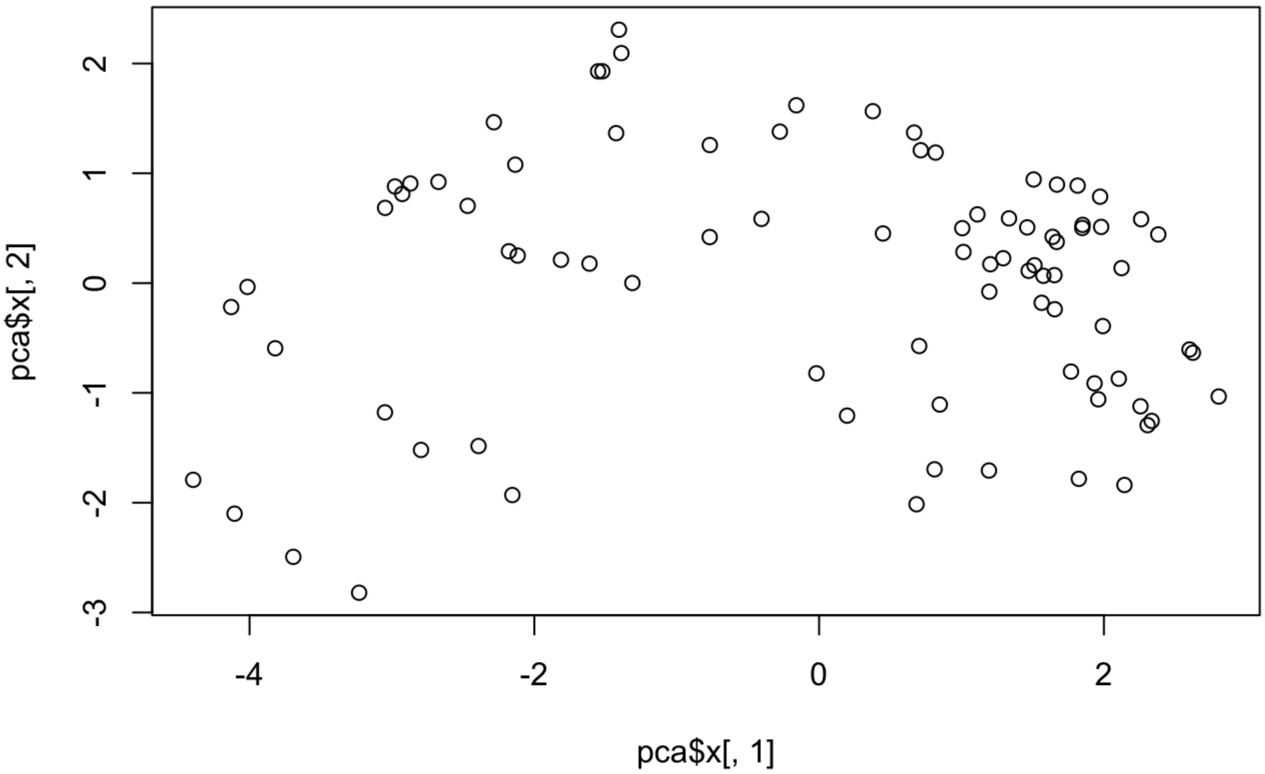
PCA will be applied using `prcomp()`

```
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
```

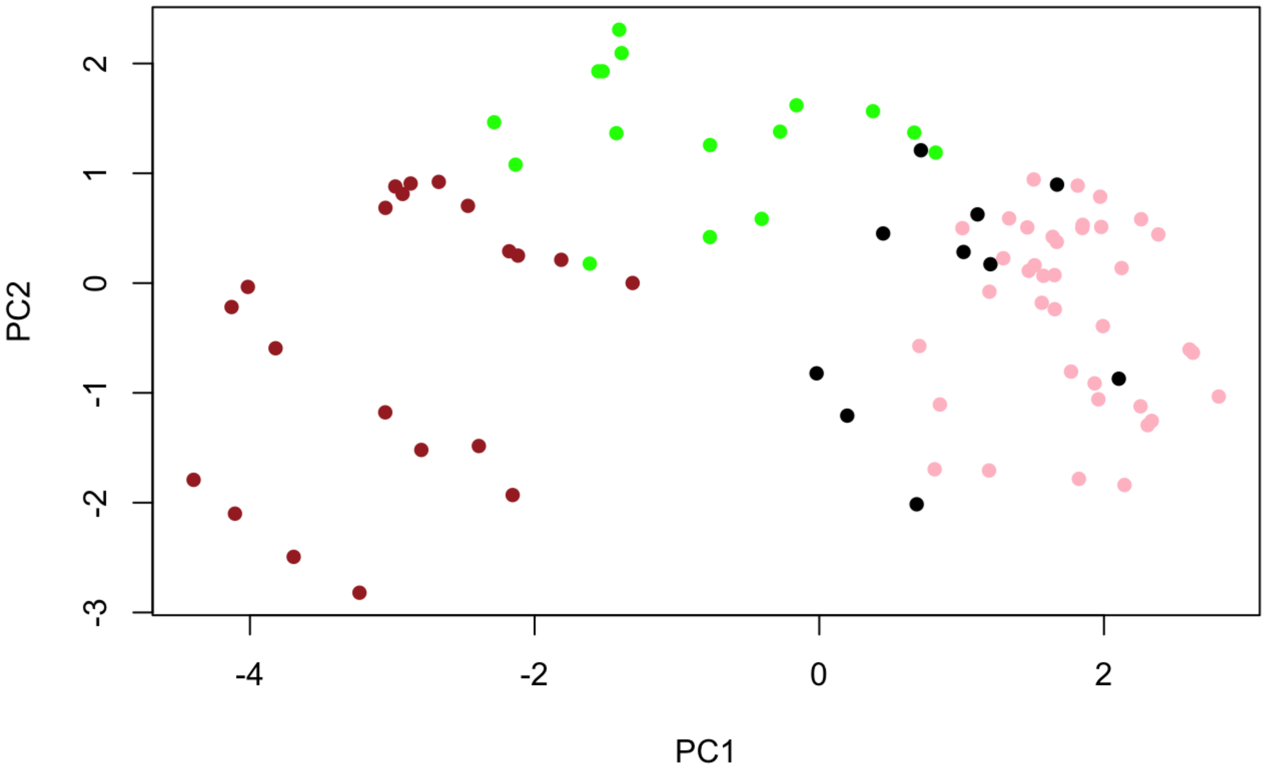
A plot will be made

```
plot(pca$x[,1],pca$x[,2])
```



Character changed and colors added

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

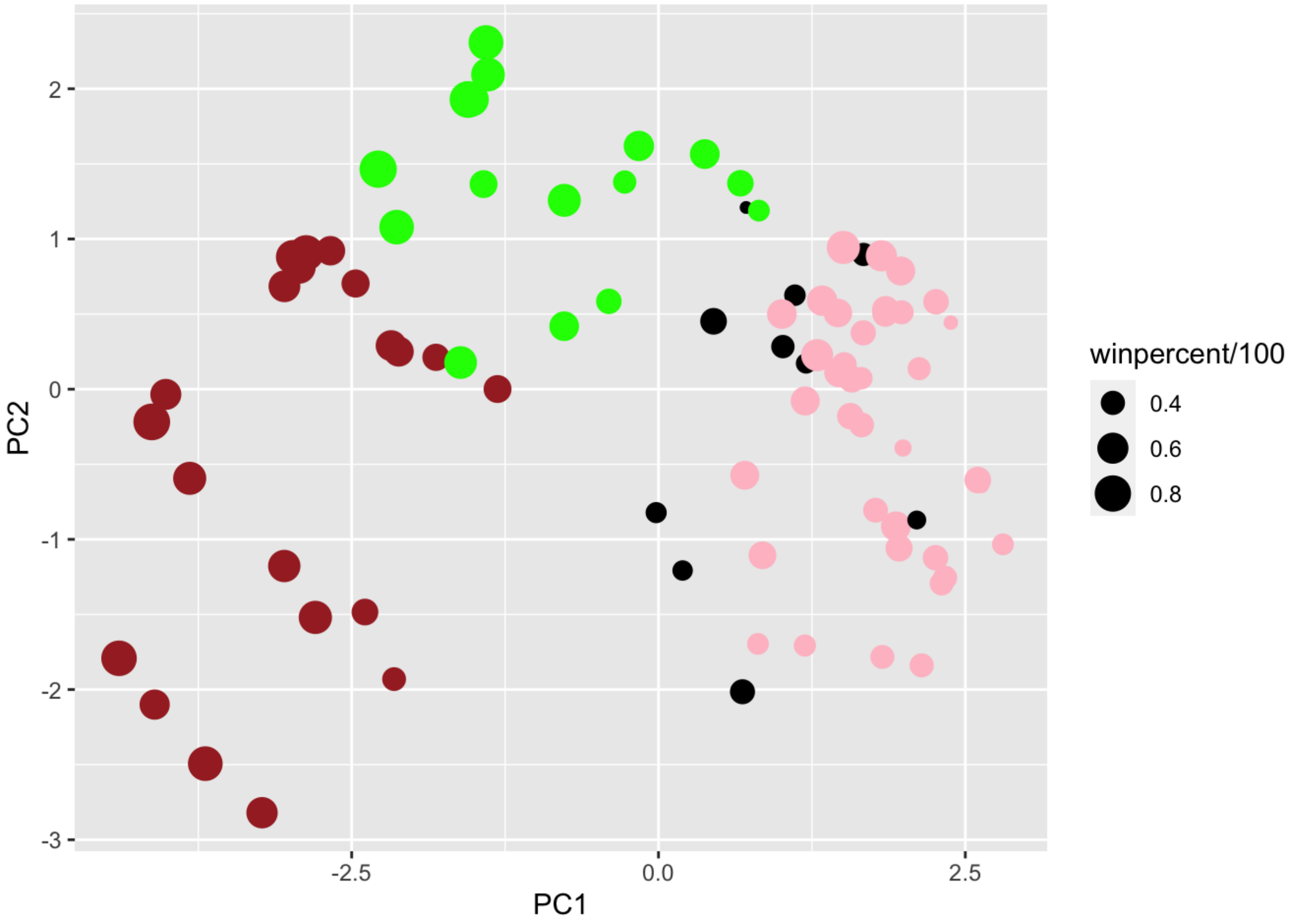


A new data frame is made so seperate columns can be included to make ggplot look nicer

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

Now this can be ploted setting ggplot equal to p

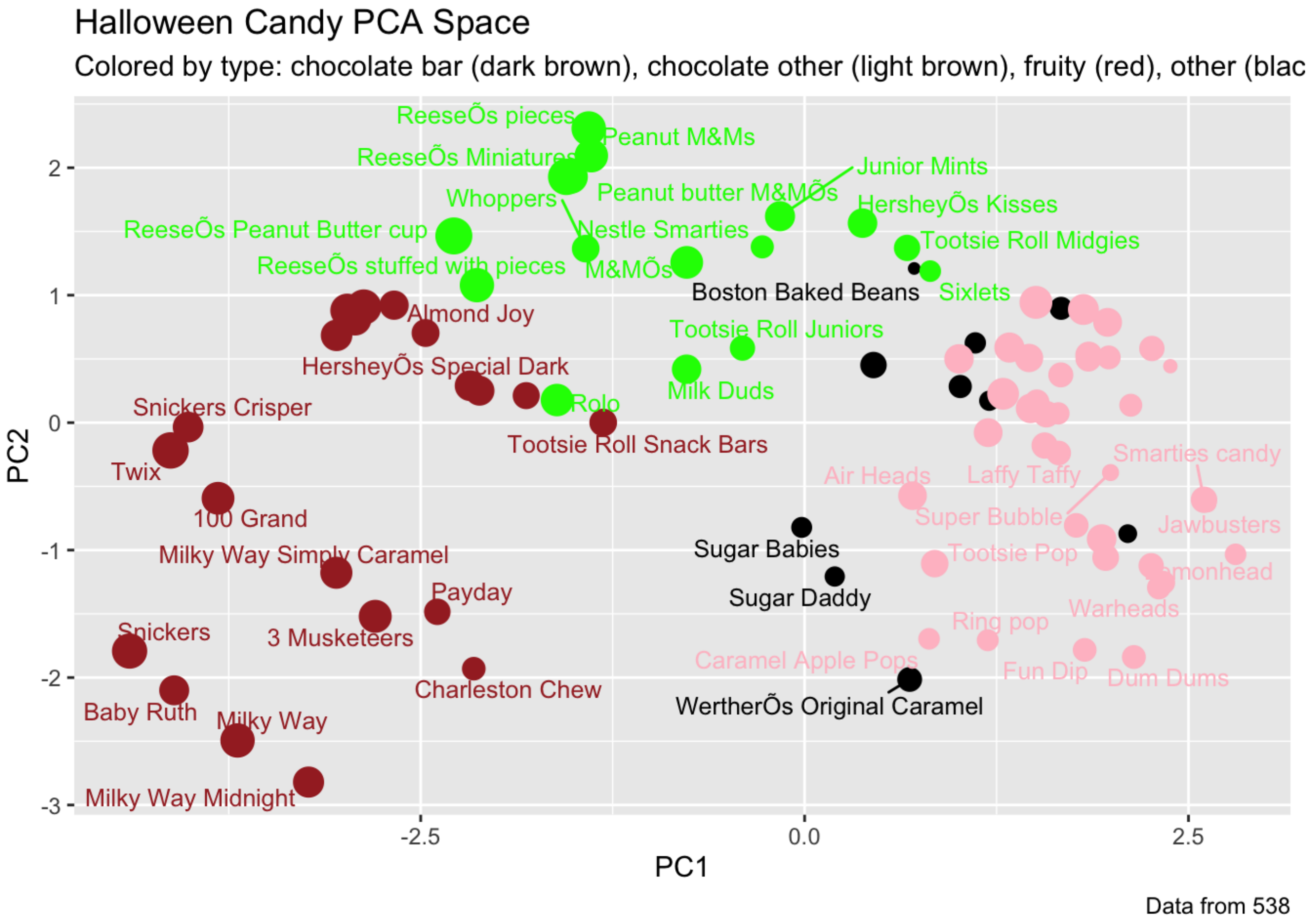
```
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)


p
```



ggrepel in implemented to add the labels of the candy on the plot

```
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
       caption="Data from 538")
```

```
Warning: ggrepel: 39 unlabeled data points (too many overlaps).
Consider
increasing max.overlaps
```

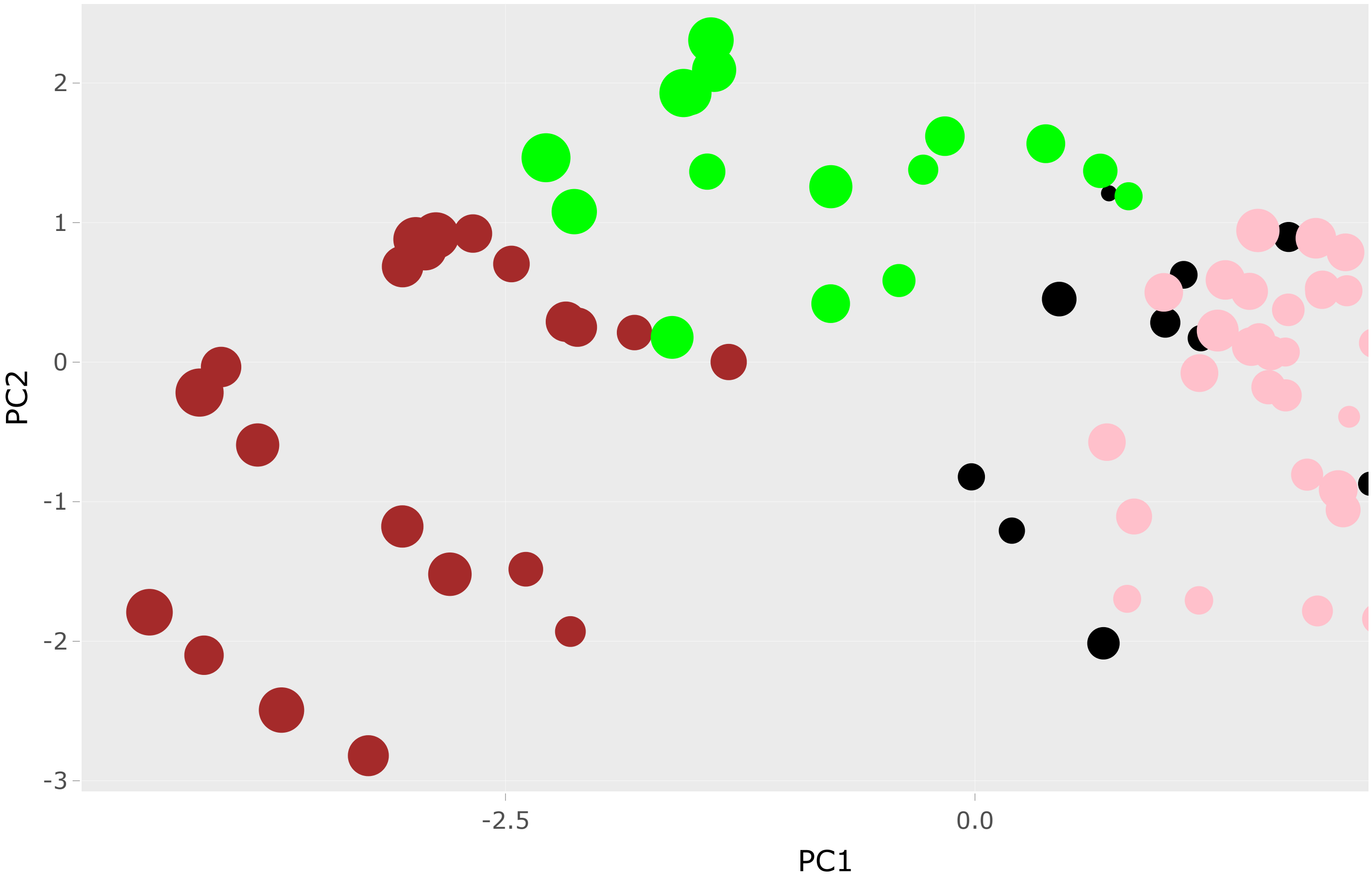


plotly is used to generate an interactive which will be downloaded first

```
library(plotly)
```

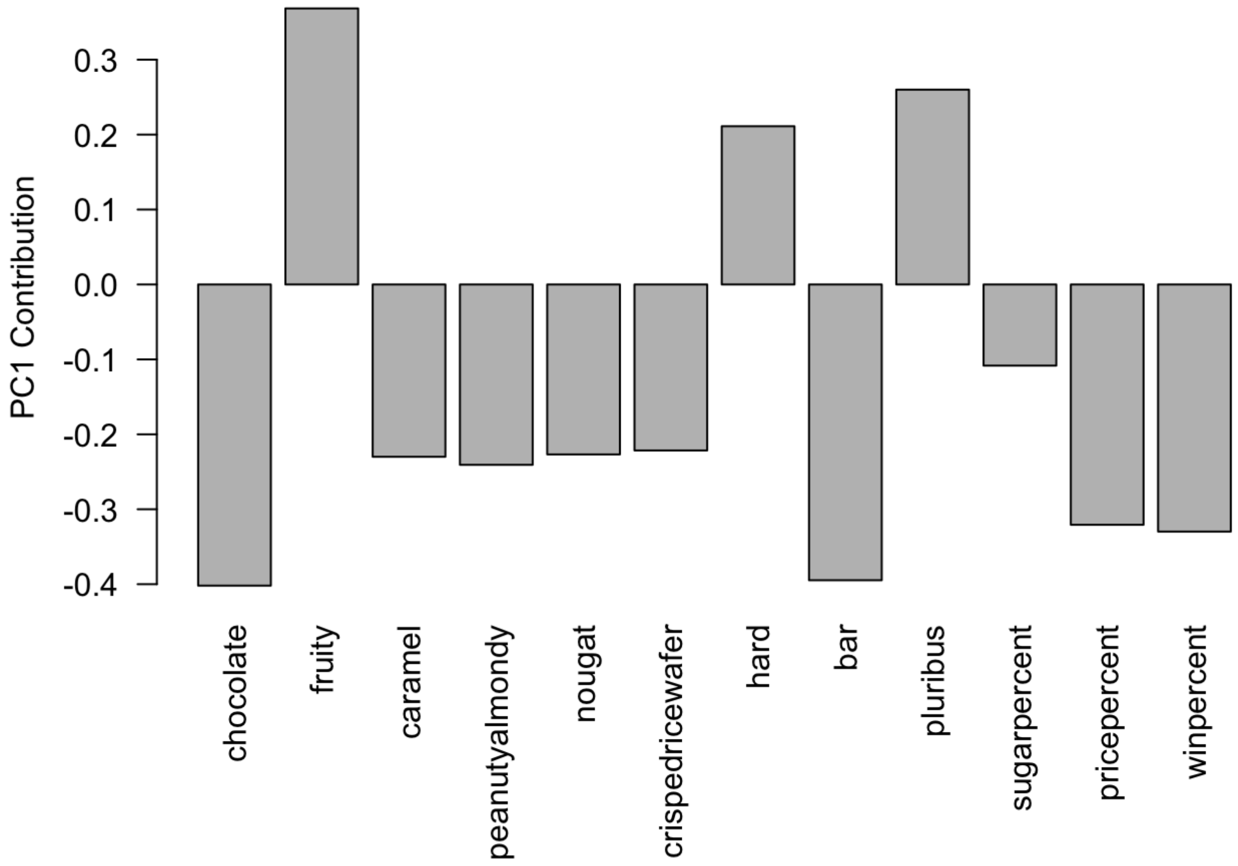
```
Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

    last_plot

The following object is masked from 'package:stats':

    filter

The following object is masked from 'package:graphics':

    layout
```

```
ggplotly(p)
```



A barplot is made to compared the PCA combination of each category

```
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?

Fruity, hard, and pluribus. This makes sernse because these 3 categories have little to no correlation to the other categories in PC1 by the negative direction.