Lab 9: Candy Lab

Rachel Lamm (PID: A18518313)

Table of contents

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 the proteomics and transcriptomics work that we will use these methods on throughout the rest of the course.

Data Import

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

```
candy_file <- "candy-data.csv"
candy = read.csv("candy-data.csv", 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
```

Q3. What is your favorite candy (other than Twix) in the dataset and what is it's winpercent value?

```
#My favorite candy is sour patch kids.
candy["Sour Patch Kids", "winpercent"]
```

```
[1] 59.864
```

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
```

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!

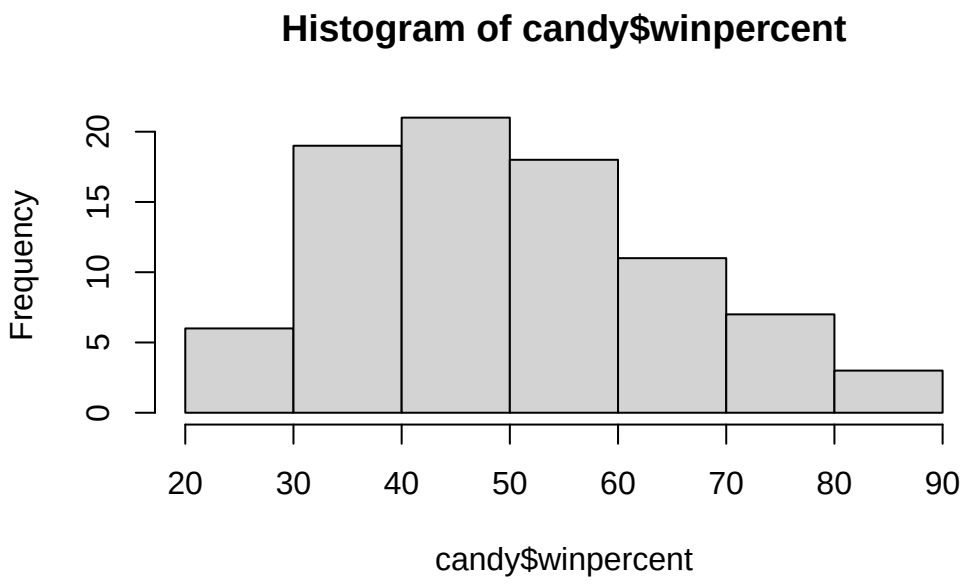
Q7. What do you think a zero and one represent for the `candy$chocolate` column?

The zero and ones represent whether or not the candy contains what's being specified in the column. So, 1 for yes if it does contain chocolate and 0 for no it doesn't.

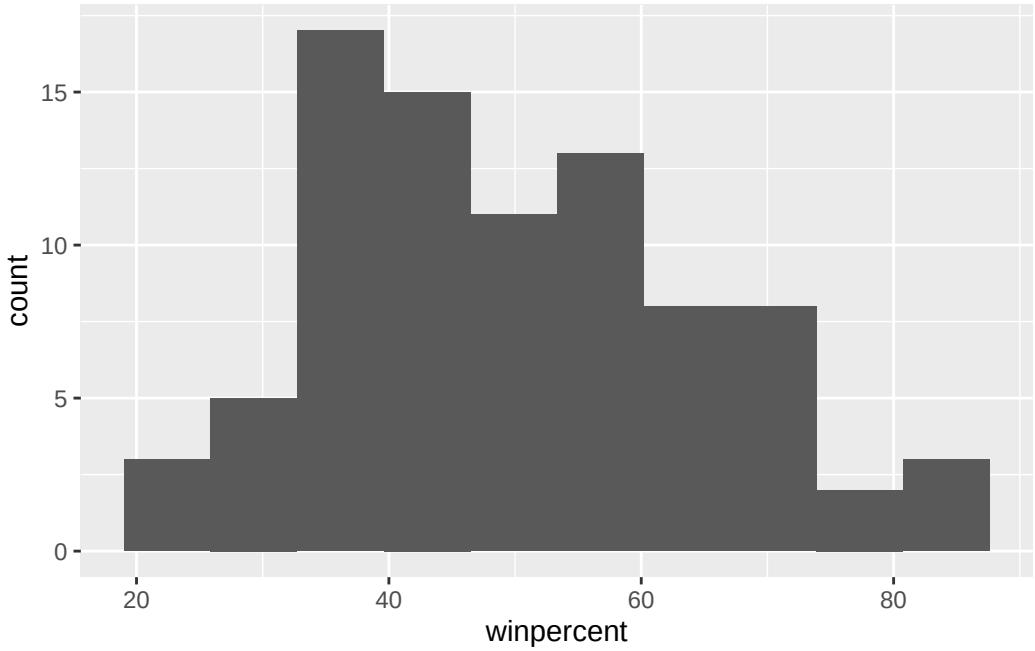
Exploratory Analysis

Q8. Plot a histogram of winpercent values

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



```
library(ggplot2)

ggplot(candy)+
  aes(x=winpercent)+
    geom_histogram(bins=10)
```



Q9. Is the distribution of winpercent values symmetrical?

Not really, it appears to be slightly right skewed.

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

It is below 50%.

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

```
[1] 47.82975
```

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

On average, chocolate candy is higher ranked than fruit candy.

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

Q12. Is this difference statistically significant?

Yes, the difference is statistically significant.

```
choc.win <- candy$winpercent[candy$chocolate == 1]
fruit.win <- candy$winpercent[candy$fruity == 1]

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
```

Overall Candy Rankings

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

```
ord.ind <-order(candy$winpercent)
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
```

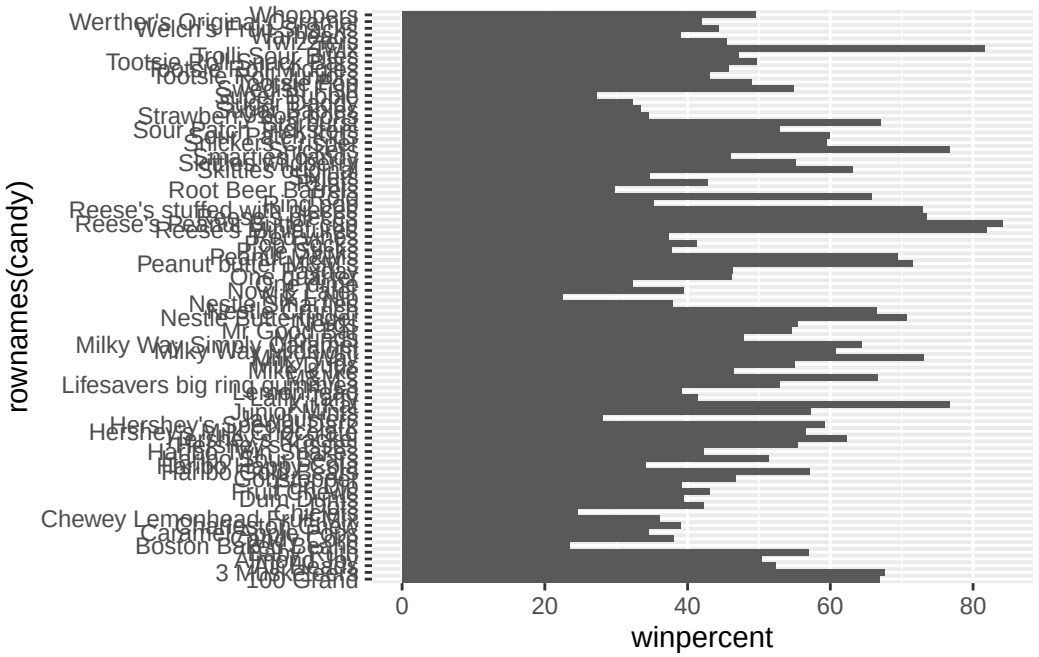
Q14. What are the top 5 all time favorite candy types out of this set?

```
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
```

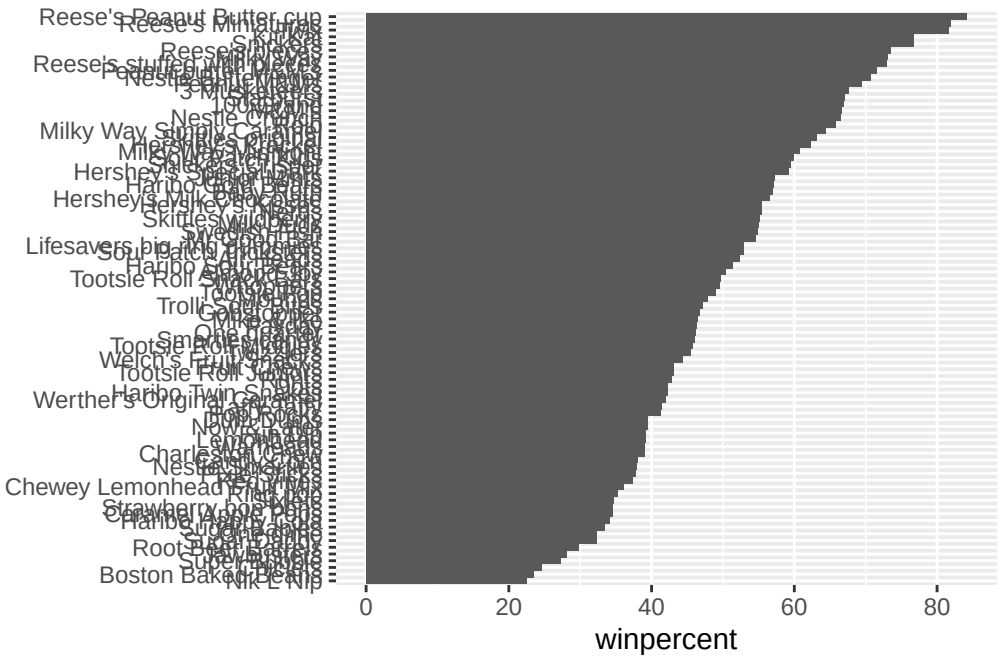
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

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



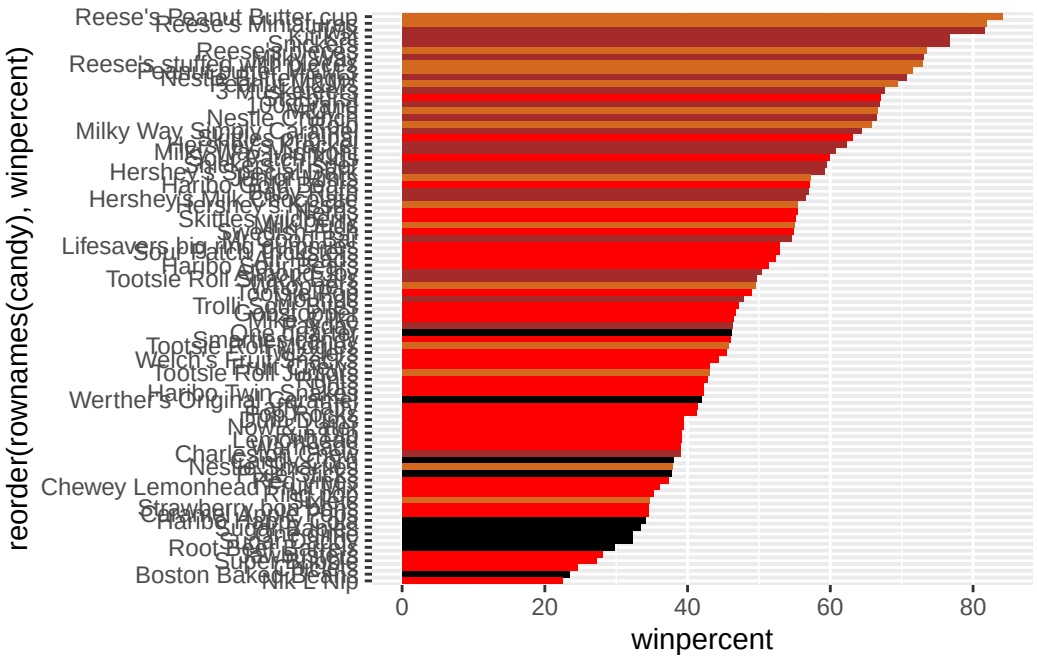
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
library(ggplot2)

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




```
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)] = "red"
```

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



Q17. What is the worst ranked chocolate candy?

The worst ranked chocolate candy is Sixlets.

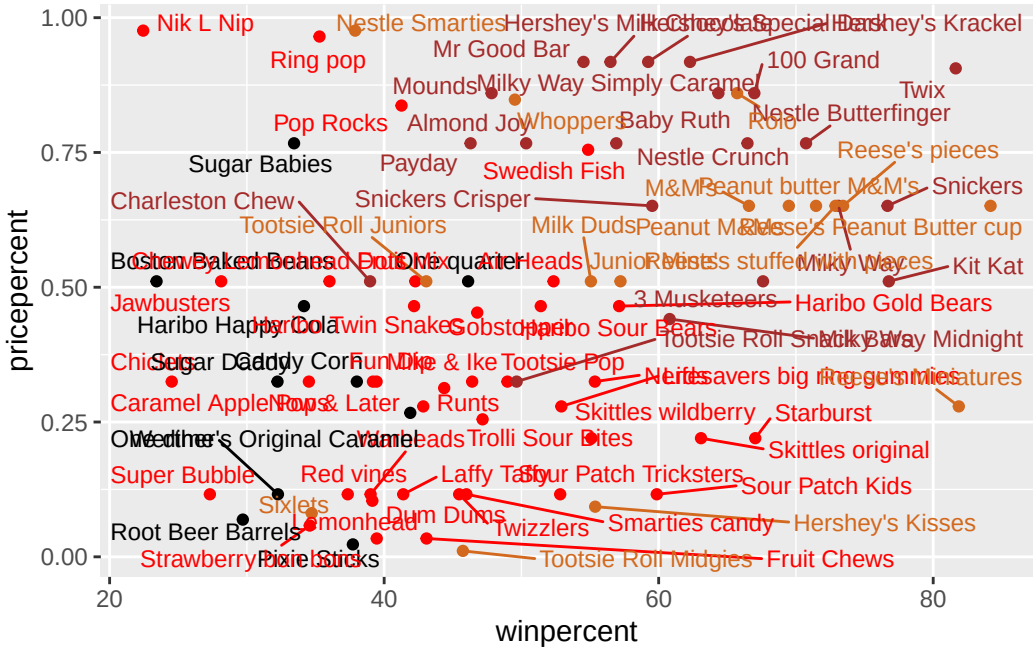
Q18. What is the best ranked fruity candy?

The best ranked fruity candy is Starburst.

Taking a look at pricepercent

```
library(ggrepel)
library(ggplot2)


ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 20)
```

```
Warning: ggrepel: 1 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?

The candy that is the highest ranked in terms of win percent and maintaining a relatively lower price percent would be the Reese's miniatures.

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

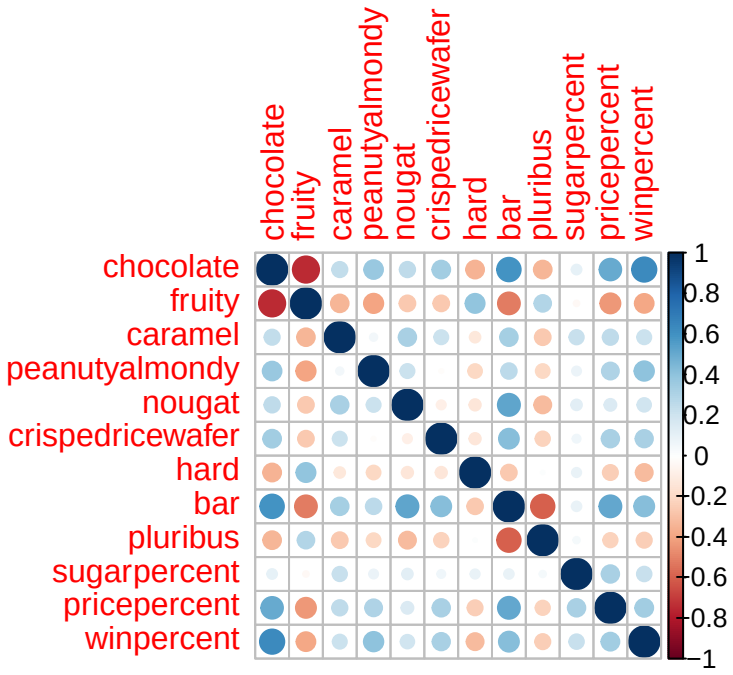
The top 5 most expensive candy types are Nik L Nips, Nestle Smarties, Ring Pops, Hershey Krackels, and Hershey's Milk Chocolate. The least popular candy is Nik L Nips.

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)?

The two variables that are anti-correlated would be chocolate and fruity.

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

The two variables that are most positively correlated would be chocolate and bar.

Principal Component Analysis (PCA)

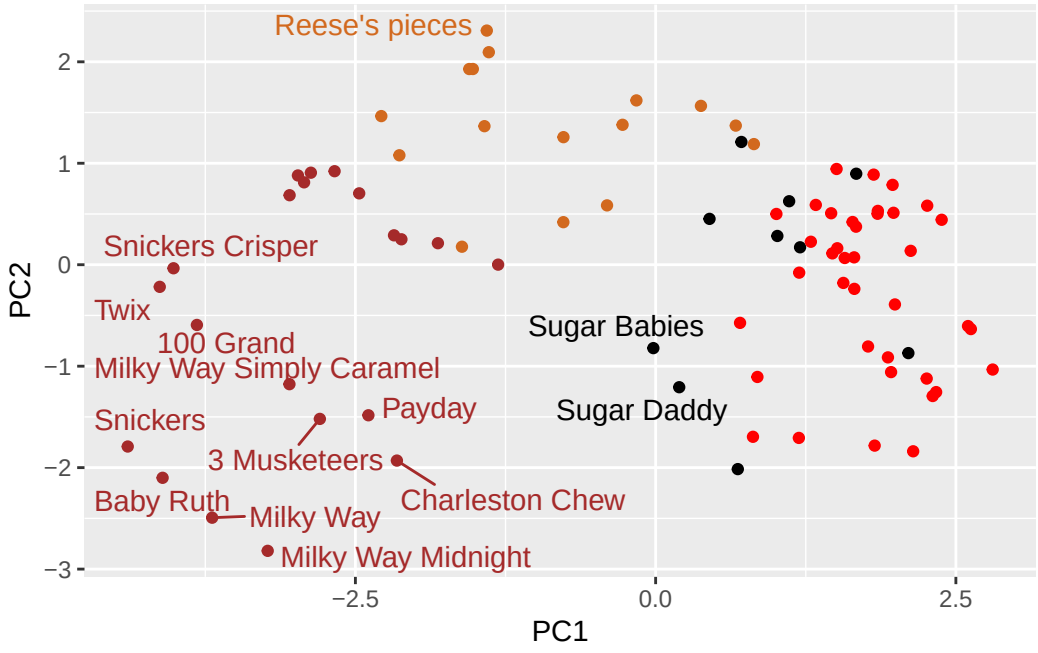
```
pca <- prcomp(candy, scale= TRUE)
summary(pca)
```

```
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
```

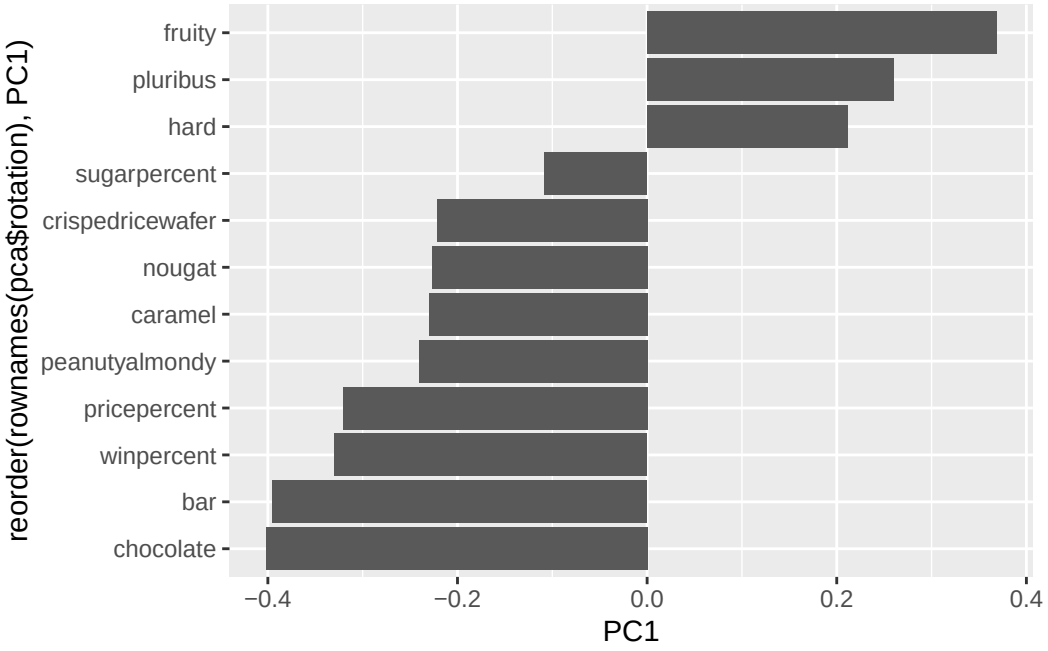
```
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

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

```
Warning: ggrepel: 71 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```




```
ggplot(pca$rotation, aes(x = PC1,
  y = reorder(rownames(pca$rotation), PC1))) +
  geom_col()
```




```
#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?

In the positive direction, the variables fruity, pluribus, and hard are strongly picked up by PC1. This does make sense to me because we saw this in the correlation questions. We identified that fruity and chocolate are negatively correlated which matches the plot because it separates these variables from the chocolate candies.

> 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?

What seems to make a winning candy is one that is chocolate based and is in the shape of a bar. Additionally, ones that contain peanuts or caramel seem to be more popular. The different analyses such as the plots, correlation analysis, and PCA all complement each other by showing chocolate candies are more popular while fruity candies are less popular. It proves this by showing that chocolate is positively associated with win percent. Also, the PCA shows further evidence by separating the chocolate candies from the fruity ones.