Class 09: Halloween Mini Project

Neha Deshpande (PID: A17567541)

Here we analyze a candy dataset from the 538 website. This is a CSV file from their GitHub repository.

Data Import

```
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
```

Exploring the dataset

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

85

```
#nrow(candy)
```

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

38

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

Q3,4. What is your favorite candy in the dataset and what is it's winpercent value?

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

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

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

The `skim()` feature in the `skimr` package gives us a quick oveview of the data.

```
#install.packages("skimr")
library("skimr")
#skim(candy)
```

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

The winpercent variable

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

A 0 represents that the candy does not have chocolate, and a 1 represents chocolate being present in the candy.

Q8. Plot a histogram of winpercent values

```
#hist(candy$winpercent, breaks=10, binwidth=5)
#library(ggplot2)
#ggplot(candy)+
  #aes(winpercent)+
  #geom_histogram(binwidth=5)
```

Q9. Is the distribution of winpercent values symmetrical?

No

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

Below

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

First find all chocolate candy and their $winpercent values, and the mean of it

```
#chocolate_winpercent<-candy$winpercent[as.logical(candy$chocolate)]
#mean(chocolate_winpercent)
```

Then look at the winpercent values for fruity candy and find their mean.

```
#fruity_winpercent<-candy$winpercent[as.logical(candy$fruity)]
#(fruity_winpercent)
```

Then, see if the mean for chocolate is greater than the mean for candy.

```
#mean(chocolate_winpercent)>mean(fruity_winpercent)
```

The chocolate candy on average is more popular than the fruity candy.

Q12. Is this difference statistically significant?

```
#t.test(chocolate_winpercent, fruity_winpercent)
```

This result is significant

Q13. What are the five least liked candy types in this set?, Q14. What are the top 5 all time favorite candy types out of this set?

`order()` shows us the indices that we need to order the data in to get them in order, and `sort()` gives you the actual value.

```
#inds<-order(candy$winpercent)
#inds
#head(candy[inds,])
#tail(candy[inds,])
```

Plotting the data

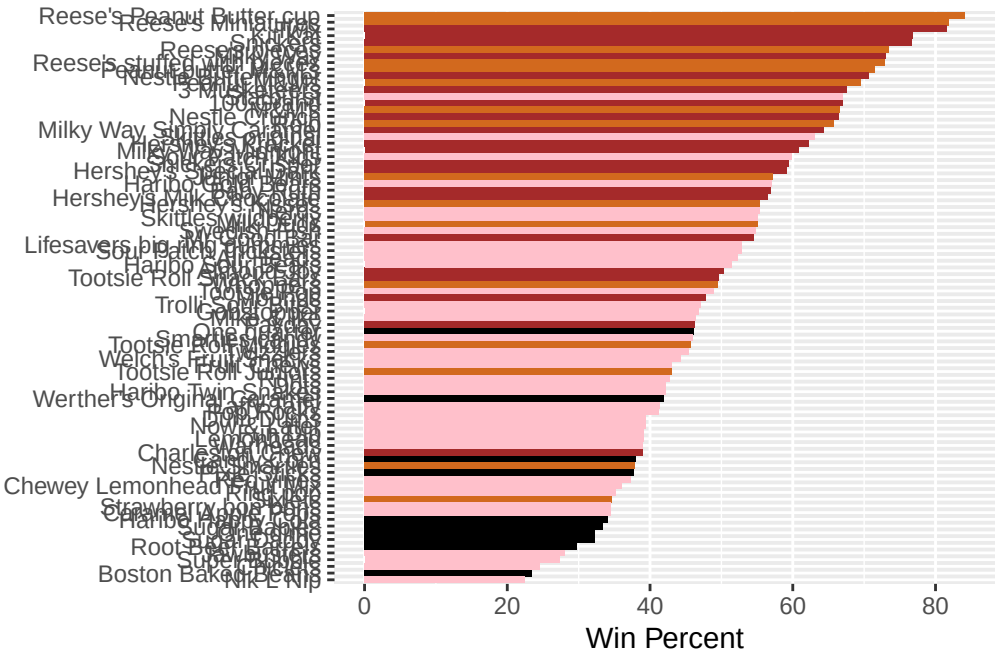
Q15. Make a first barplot of candy ranking based on winpercent values. Use the reorder() function to get the bars sorted by winpercent?

```
#library(ggplot2)
#ggplot(candy)+
  #aes(winpercent, reorder(rownames(candy), winpercent))+
  #geom_col()
```

Adding some color to increase the division. The first line of code is required to define the `my_cols()` vector.

```
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"

library(ggplot2)
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  labs(x="Win Percent", y="")+
  geom_col(fill=my_cols)
```




```
ggsave('barplot1.png',width=7, height=10)
```

For the image to show up in the quarto document, use `![]()` (Markdown syntax for inserting image)

Q17. What is the worst ranked chocolate candy?

Charleston Chews

Q18. What is the best ranked fruity candy?

Starburst

Plotting winpercent vs pricepercent. First, to insert labels onto the candy in an organized manner (without the plot looking ugly through `geom_text()`), we will use the `ggrepel` package.

```
#install.packages("ggrepel")
library(ggrepel)
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=2, max.overlaps = 5)
```

```
Warning: ggrepel: 50 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

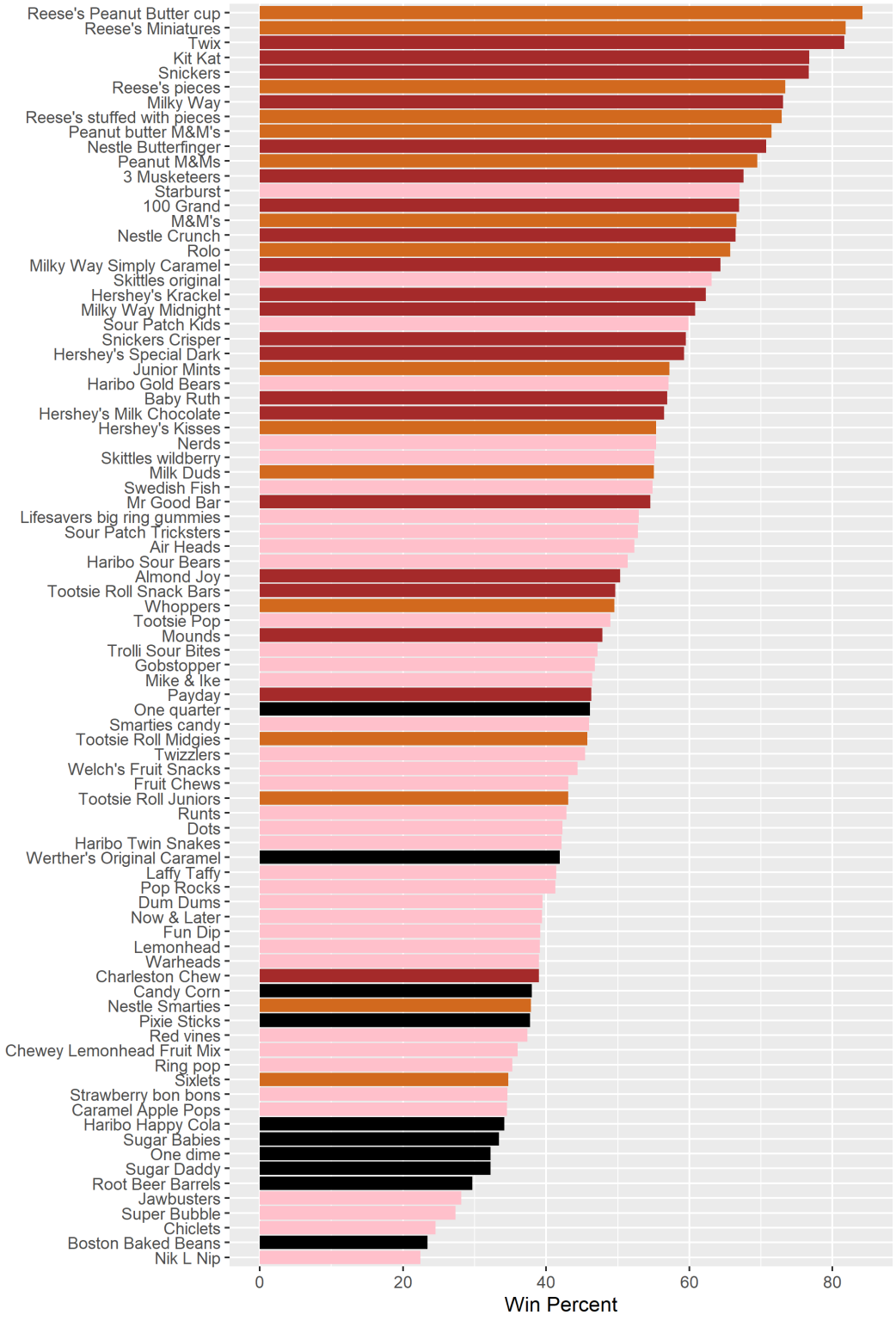


Figure 1: Caption

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?

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 )
```

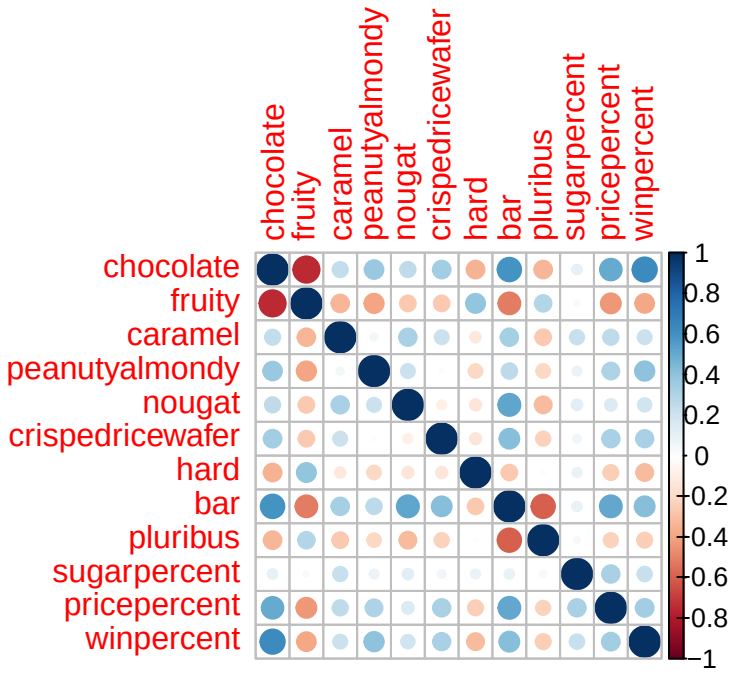
```
                        pricepercent winpercent
Nik L Nip                      0.976   22.44534
Nestle Smarties                0.976   37.88719
Ring pop                       0.965   35.29076
Hershey's Krackel              0.918   62.28448
Hershey's Milk Chocolate       0.918   56.49050
```

##Exploring the correlation structure

```
#install.packages("corrplot")
library(corrplot)
```

```
corrplot 0.92 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?

bar and chocolate

On to PCA

The main function for this is called `prcomp()` and here we need to scale our data with the `scale=TRUE` argument.

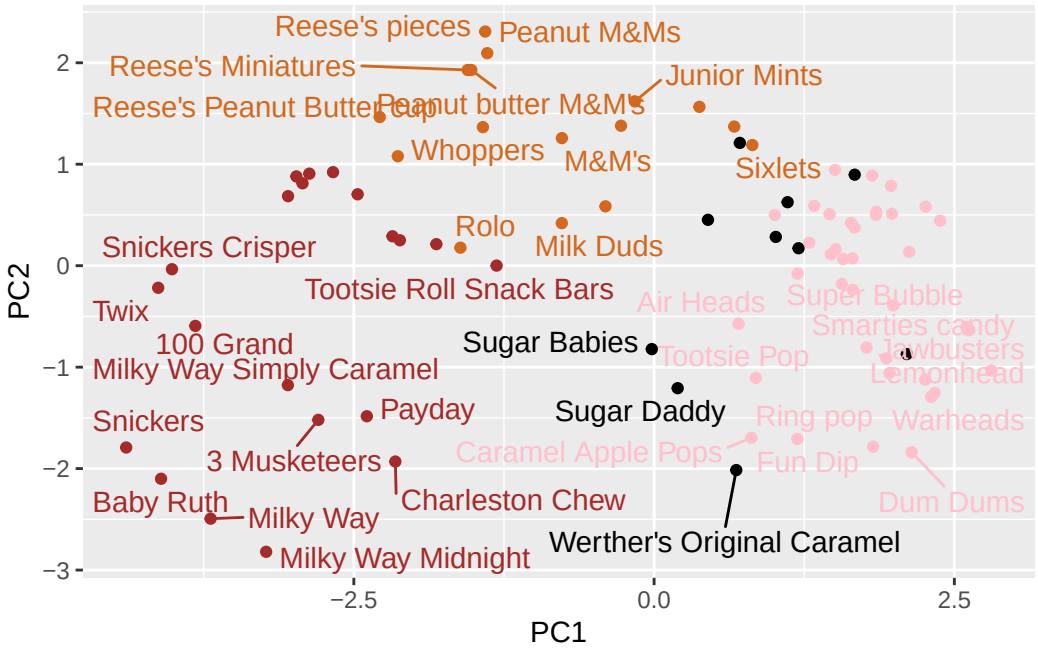
```
pca<-prcomp(candy,scale=TRUE)
#summary(pca)
```

Plot my main PCA score plot with ggplot

```
library(ggplot2)
library(ggrepel)
my_data<-cbind(candy, pca$x[,1:3])
ggplot(my_data)+
```

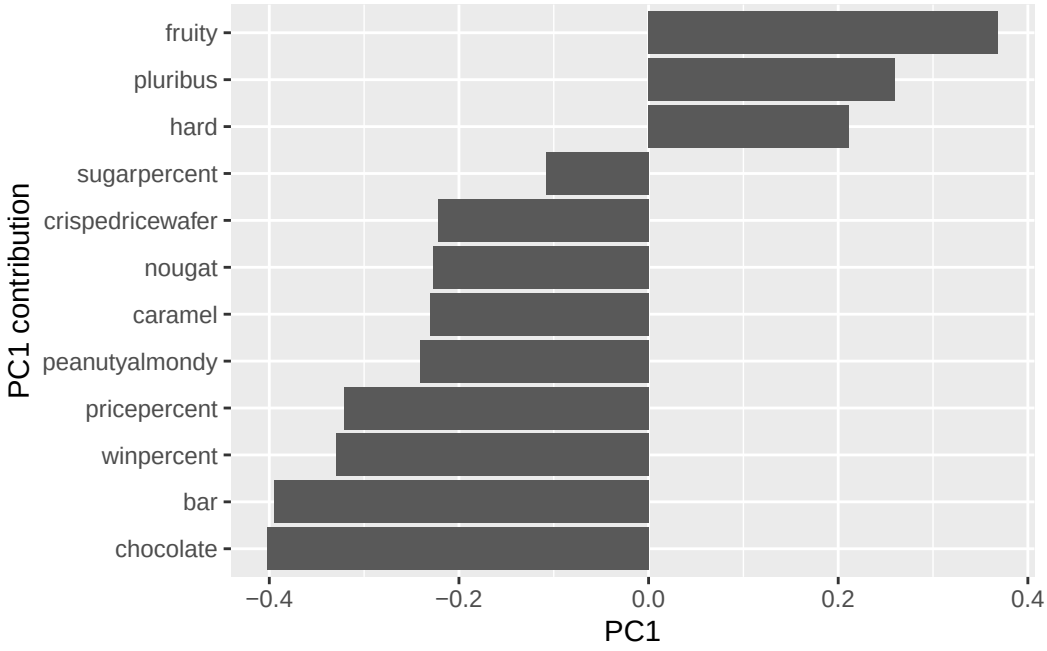
```
    aes(PC1,PC2, label=rownames(candy))+
    geom_point(col=my_cols)+
    geom_text_repel(col=my_cols,max.overlaps = 10)
```

```
Warning: ggrepel: 48 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



loadings plot:

```
loadings<-as.data.frame(pca$rotation)
library(ggplot2)
ggplot(loadings)+
  aes(PC1, reorder(rownames(loadings),PC1))+
  geom_col()+
  labs(y="PC1 contribution")
```



Q24. Does it make sense?

Yes it makes sense.