Class09 Halloween Mini Project

Hugh (Trevor) Redford PID:A17067426

Importing 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
```

There are 85 different candy types in the dataset.

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

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

```
[1] 38
```

There are 38 fruit candy types in the dataset.

What is your favorite candy?

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

My favorite candy is Reese's Peanut Butter cup and its win percent is 84.18029

```
candy["Reese's Peanut Butter cup", ] $winpercent
```

```
[1] 84.18029
```

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

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

```
[1] 76.7686
```

The winpercent value for Kit Kat is 78.7686

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

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

```
[1] 49.6535
```

The winpercent value for Tootsie Roll Snack Bars is 49.6535

Skim Function

```
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	
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?

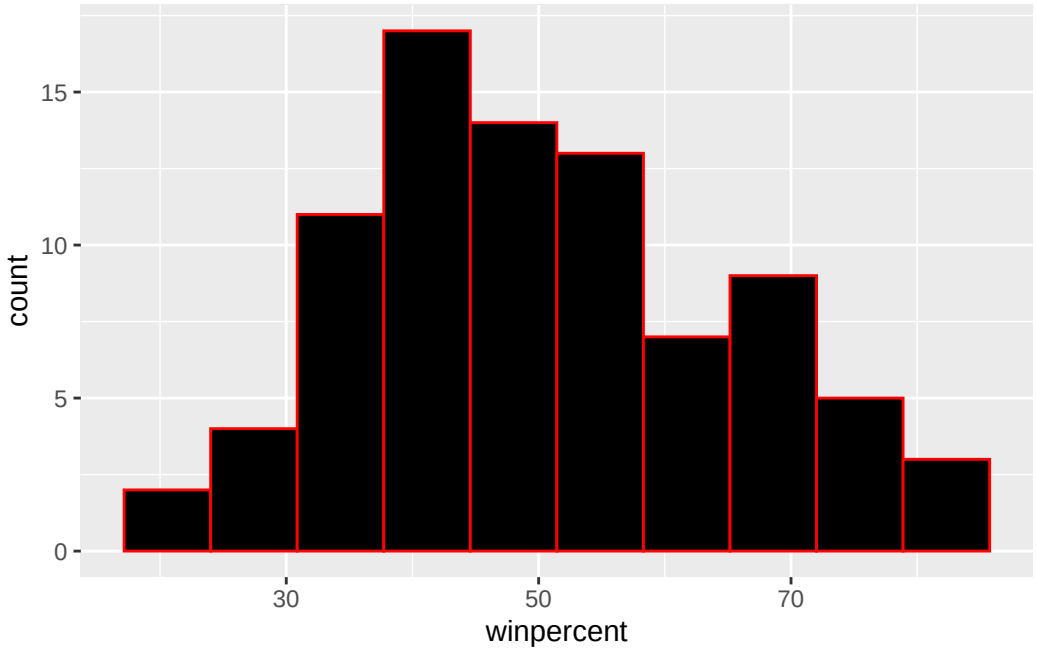
Yes, the winpercent variable stands out as being on a significantly different scale than the other columns. While most variables have values ranging between 0 and 1, winpercent has a much higher mean (50.32), a larger standard deviation (14.71), and considerably greater percentile values. This suggests that winpercent is measured on a different numerical scale, making it an outlier in terms of magnitude compared to the other features in the dataset.

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

It is likely that a value of 1 indicates that a candy contains chocolate, while a value of 0 signifies that the candy does not contain chocolate. This binary representation helps distinguish chocolate-based candies from non-chocolate ones in the dataset.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)
ggplot(candy) +
aes(winpercent) +
geom_histogram(bins = 10, col="red", fill="black")
```



Q9. Is the distribution of winpercent values symmetrical?

No it is not symmetrical.

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

It is below 50%.

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

```
# Chocolate Average Rank
chocolate.inds <- as.logical(candy$chocolate)
chocolate.wins <- candy[chocolate.inds,]$winpercent
mean(chocolate.wins)
```

```
[1] 60.92153
```

```
#Fruit Average Rank
fruit.inds <- as.logical(candy$fruity)
fruit.wins <- candy[fruit.inds,]$winpercent
mean(fruit.wins)
```

```
[1] 44.11974
```

On average, chocolate candy (60.92153) is higher ranked than fruit candy (44.11974)

Q12. Is this difference statistically significant?

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

```
    Welch Two Sample t-test

data:  chocolate.wins and fruit.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
```

The difference is statistically significant because the p-value from the Welch Two Sample t-test is 2.871e-08, which is far smaller than the commonly used significance threshold of 0.05. This indicates strong evidence against the null hypothesis, which assumes no difference between the mean wins of chocolate and fruit candies.

Overall Candy Rankings

```
# 5 least liked
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
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
```

```r
# 5 most liked
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
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
```

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

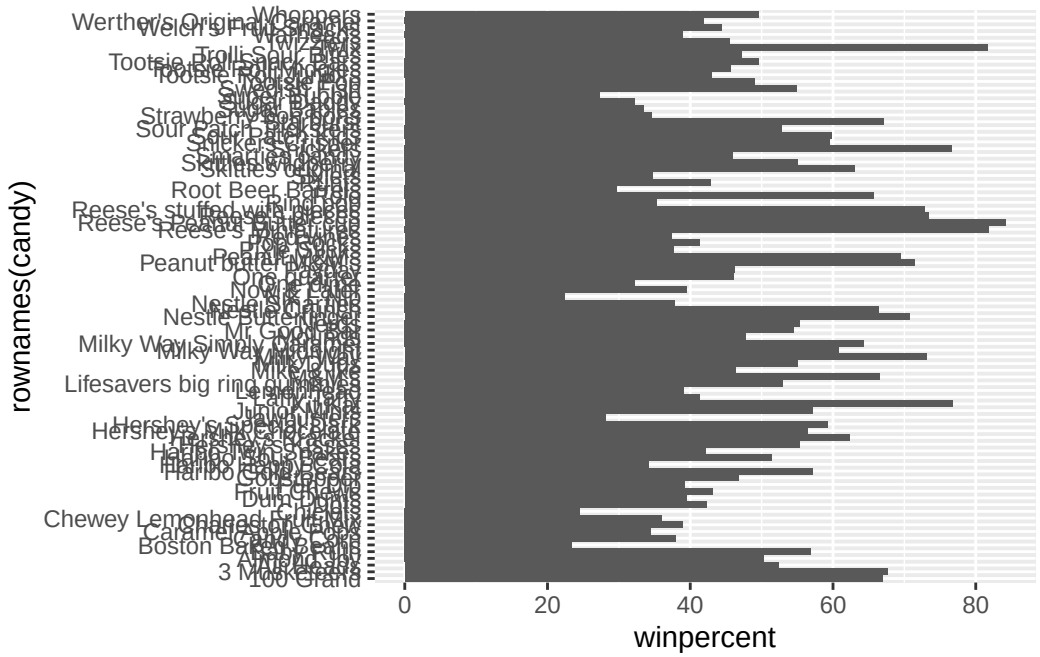
Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters

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

Snickers, Kit Kat, Twix, Reese's Miniatures, Reese's Peanut Butter Cup

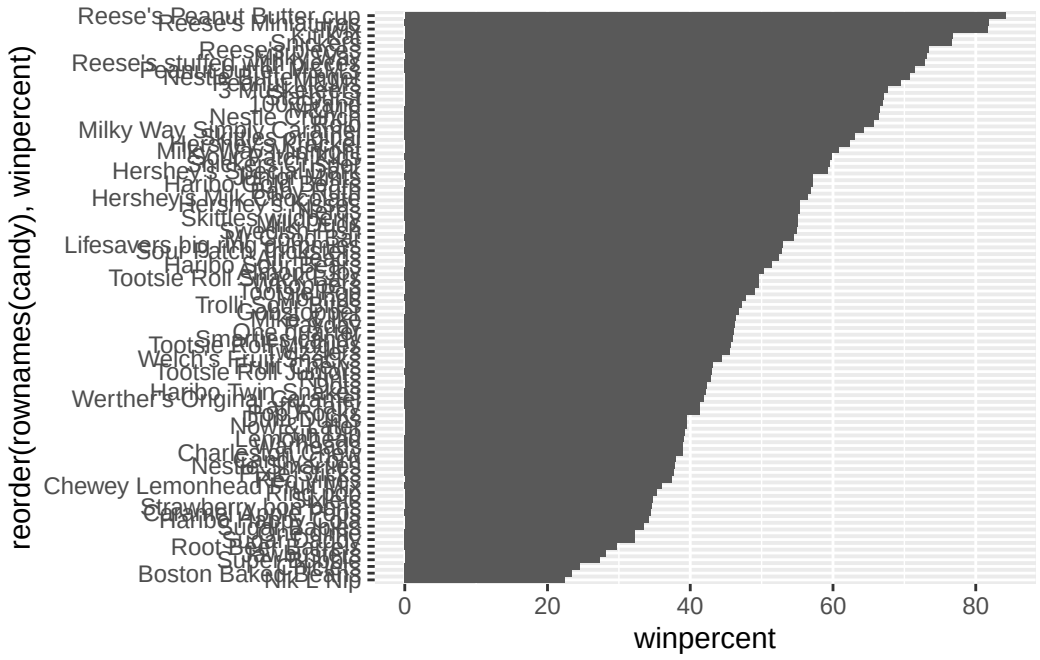
Q15. Make a first barplot of candy ranking based on winpercent values.

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



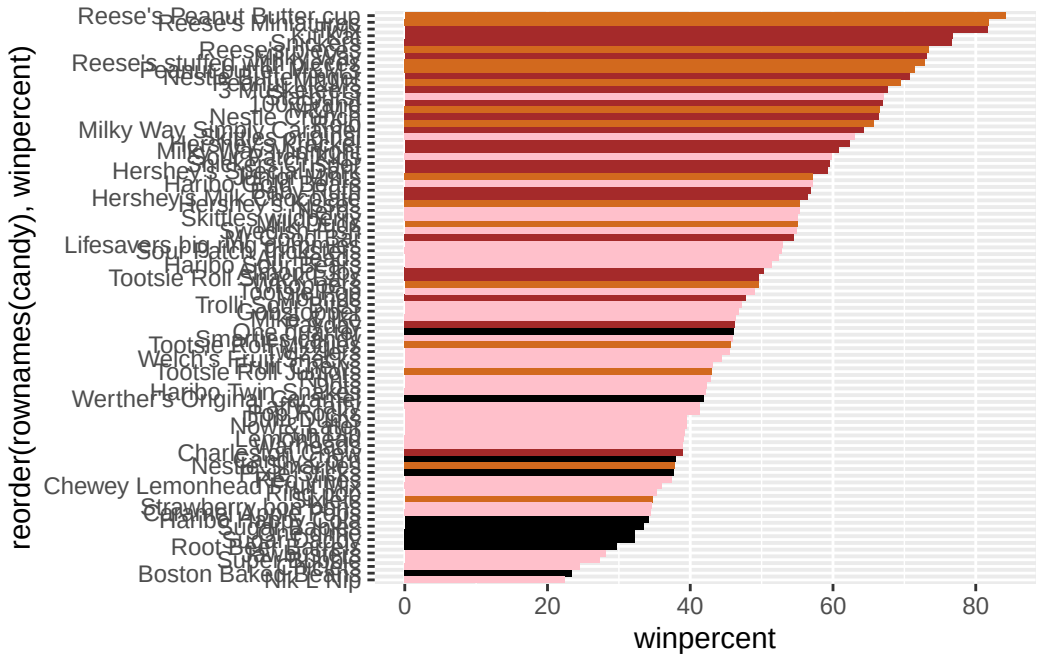
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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




```
#Set color for each candy type
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)) +
geom_col(fill=my_cols)
```



Q17. What is the worst ranked chocolate candy?

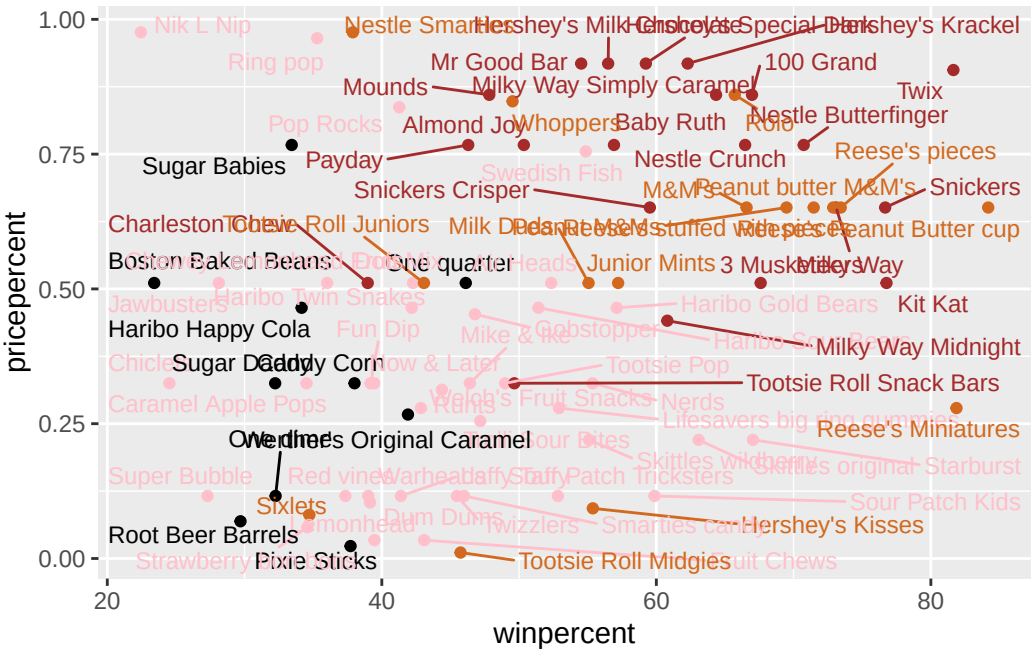
Nik L Nip

Q18. What is the best ranked candy?

Reese's Peanut Butter Cup

Taking a Look at Pricepercent

```
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 = 50)
```



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?

Reeses Miniatures

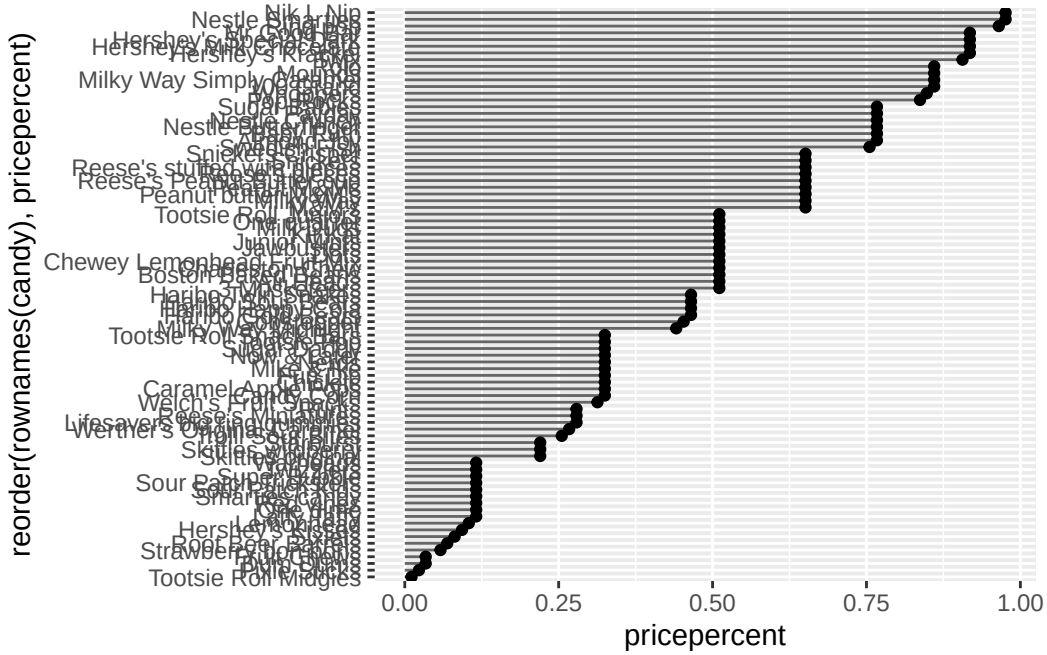
Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

The top 5 most expensive candy types are Nik L Nip, Nestle Smarties, Hershey's Krackel, Hershey's Milk Chocolate, and Mr Good Bar. The least popular is Nik L Nip

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 finallymaking a so called "dot chat" or "lollipop" chart by swapping geom_col() for geom_point() + geom_segment().

```
ggplot(candy) +
aes(pricepercent, reorder(rownames(candy), pricepercent)) +
geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
  xend = 0), col="gray40") +

geom_point()
```

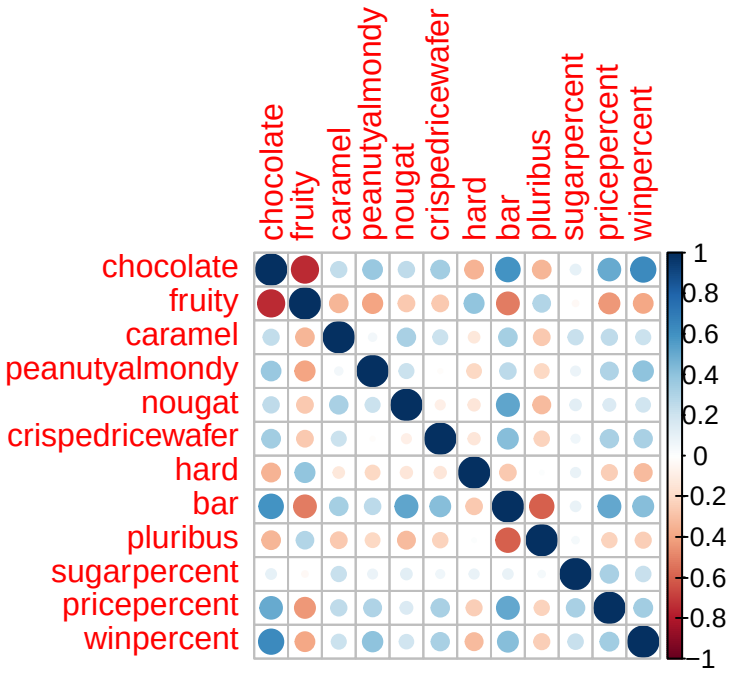


Exploring the Correlation Structure

```
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)?

Fruity and Chocolate

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

Win Percent and Chocolate

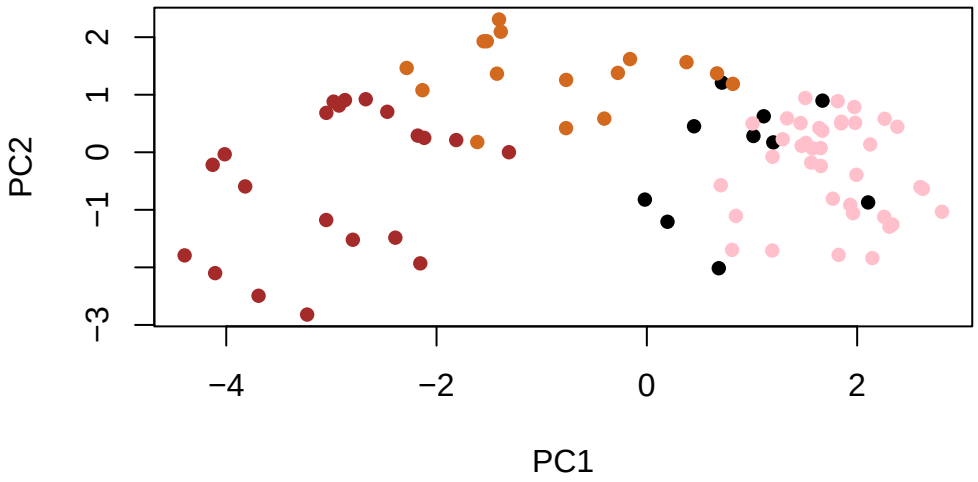
Principal Component Analysis (PCA)

```
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
```

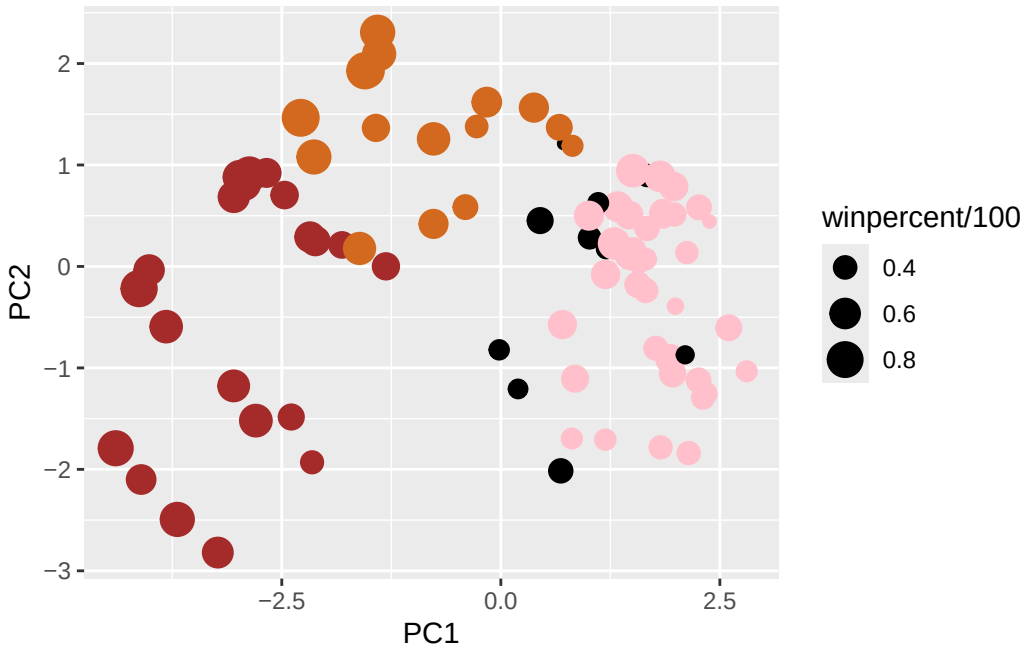
```
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

```r
#Main PCA score plot of PC1 vs. PC2
plot(pca$x[,1:2], col=my_cols, pch = 16)
```



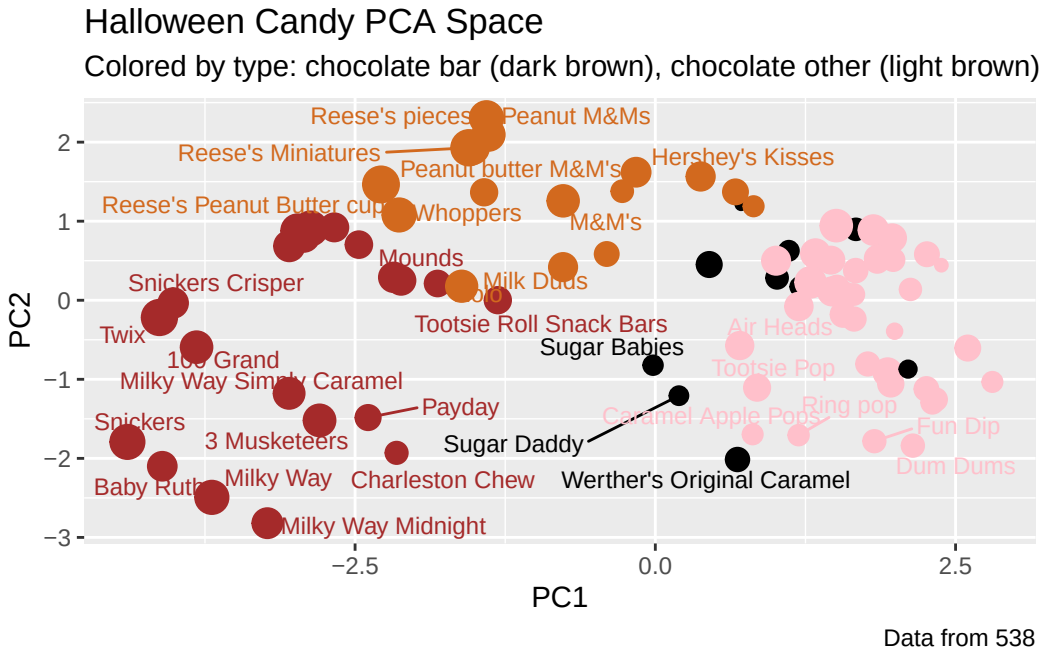

```r
#Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
p <- ggplot(my_data) +
aes(x=PC1, y=PC2,
size=winpercent/100,
text=rownames(my_data),
label=rownames(my_data)) +
geom_point(col=my_cols)

print(p)
```




```
library(ggrepel)
p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 10) +
theme(legend.position = "none") +
labs(title="Halloween Candy PCA Space",
subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown), fruity
caption="Data from 538")
```

```
Warning: ggrepel: 53 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

The primary variable strongly associated with PC1 in the positive direction is fruity. This makes sense because chocolate and fruity candies appear to be inversely correlated. Since chocolate-heavy candies tend to have negative PC1 values, it follows that fruity candies would be positioned toward the positive end of PC1. This suggests that PC1 effectively captures the contrast between chocolate-based and fruity candies in the dataset.