Halloween project

Author
Cameron Jones

1. Importing candy data

```
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
```

```
ncol(candy)
```

```
[1] 12
```

There are 85 different individual candies in this data, with them being broken down into 12 different categories

Q2. How many fruity candy types are in the dataset

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

```
[1] 38
```

There are 38 fruity candies in this data set.

##2. What is your favorate candy? One of the most interesting variables in the dataset is winpercent. For a given candy this value is the percentage of people who prefer this candy over another randomly chosen candy from the dataset (what 538 term a matchup). Higher values indicate a more popular candy.

We can find the winpercent value for Twix by using its name to access the corresponding row of the dataset. This is because the dataset has each candy name as rownames (recall that we set this when we imported the original CSV file). For example the code for Twix is:

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

```
[1] 81.64291
```

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

```
candy["Junior Mints", ]$winpercent
```

```
[1] 57.21925
```

My favorite candy, Junior Mints, has a min percentage of 57.22%

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

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

```
[1] 76.7686
```

The winpercent value for Kit Kats is 76.77%

> 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 rolls is 49.65%

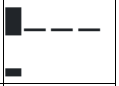
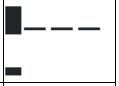
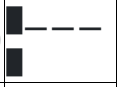
The skim function

```
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	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, looking above, it's clear that the winpercent variable is from a 0-100 scale, whereas the rest have values between 0-1.

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

```
candy$chocolate
```

```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

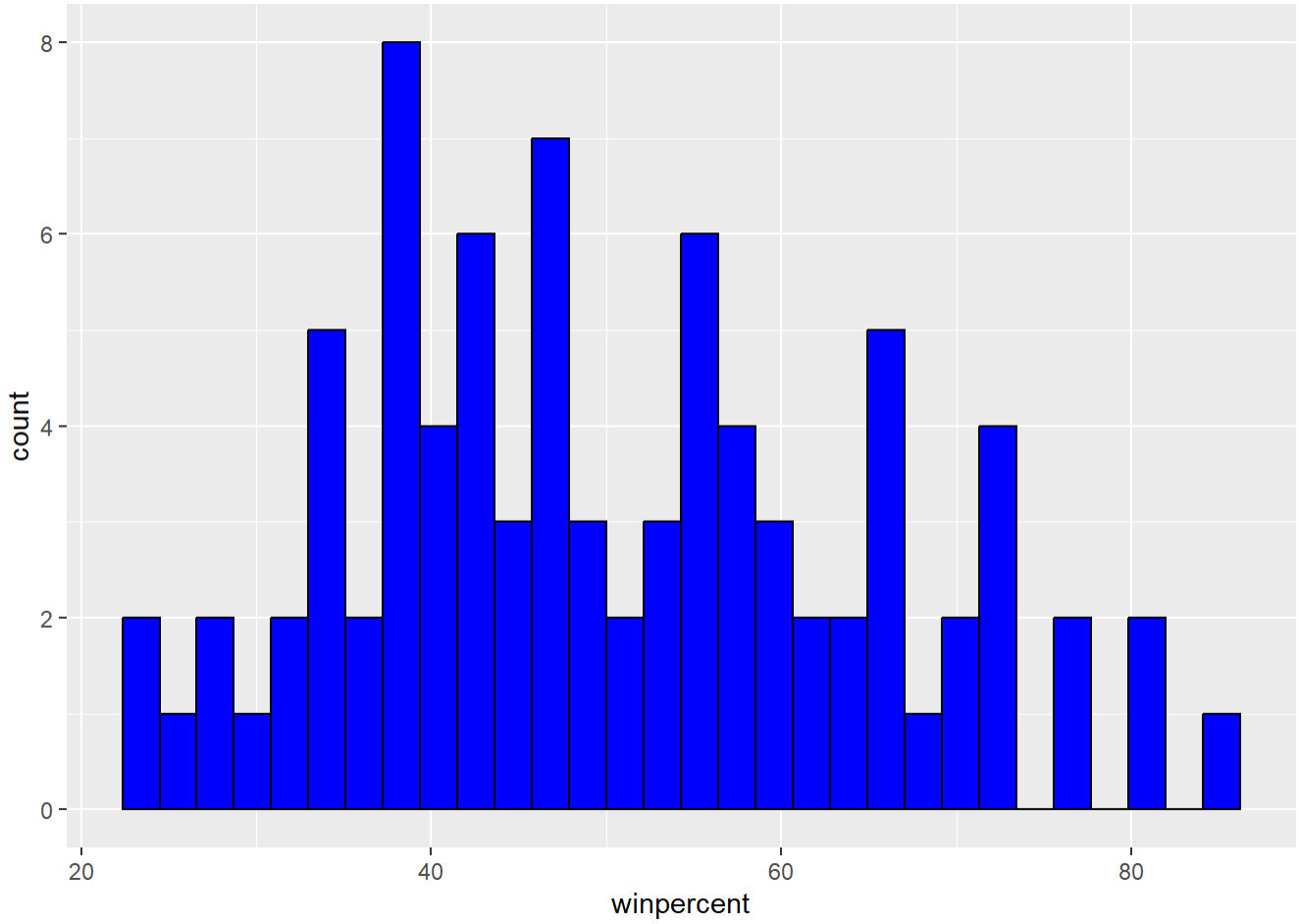
A 1 represents that this candy is a chocolate-type candy, and a 0 means it's not at all chocolate.

A good place to start any exploratory analysis is with a histogram:

> Q8. Plot a histogram of winpercent values

```
library(ggplot2)
p<-ggplot(candy, aes(x=winpercent)) +
 geom_histogram(color="black", fill="blue")
p
```

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

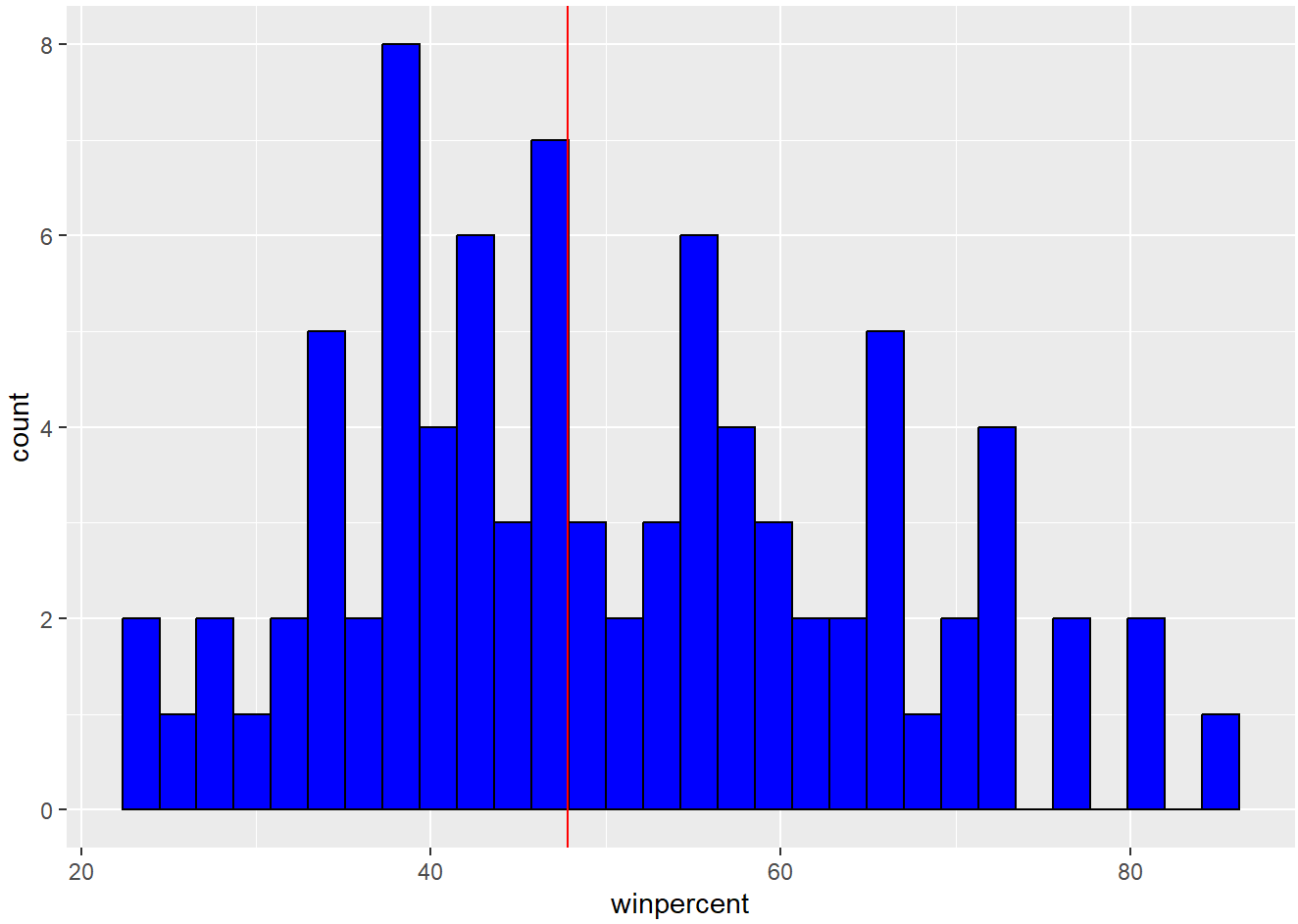


Q9. Is the distribution of winpercent values symmetrical?

No, we see the peak to be approximately in the middle of the 2nd quartile.

```
library(ggplot2)
p<-ggplot(candy, aes(x=winpercent)) +
 geom_histogram(color="black", fill="blue") +
 geom_vline(xintercept = median(candy$winpercent), color = "red") +
 scale_fill_gradient("red")
p
```

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



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

Below, the median is at 47.82

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

```
[1] 47.82975
```

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

```
fruitywin<- candy$winpercent[as.logical(candy$fruity)]
chocolatewin <- candy$winpercent[as.logical(candy$chocolate)]
```

```
mean(fruitywin)
```

```
[1] 44.11974
```

```
mean(chocolatewin)
```

```
[1] 60.92153
```

We can conclude from this that chocolate candy is ranked higher than fruity candy.

Q12. Is this difference statistically significant?

```
t.test(fruitywin, chocolatewin)
```

```
  Welch Two Sample t-test

data:  fruitywin and chocolatewin
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:
 -22.15795 -11.44563
sample estimates:
mean of x mean of y
 44.11974  60.92153
```

Based on this result, we can say that these results are pretty significant. We know this from the very low p value (below 0.05 is usually enough), but the 95% confidence interval and t value also tell us we have signicant results.

##3. Overall Candy Rankings

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
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
```

Nik L Nip, Boston Baked Bean, Chiclets, Super Bubbler and Jawbusters are all the least liked candies.

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

```
head(candy[order(-candy$winpercent),], n=5)
```

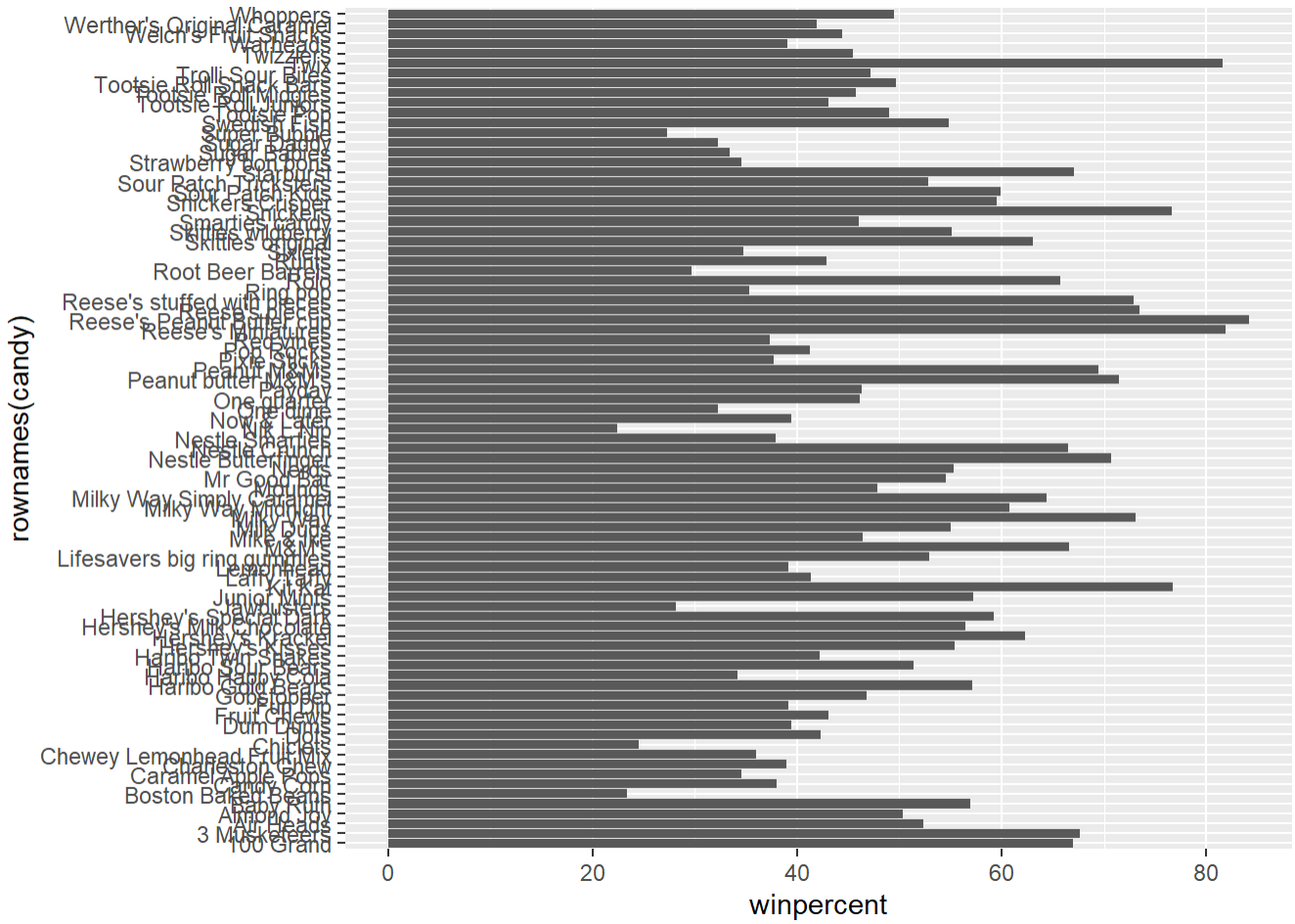
```
                    chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0            1      0
Reese's Miniatures                1      0       0            1      0
Twix                              1      0       1            0      0
Kit Kat                           1      0       0            0      0
```

```
Snickers                    1     0      1           1     1
                  crispedricewafer hard bar pluribus sugarpercent
Reese's Peanut Butter cup             0    0  0       0       0.720
Reese's Miniatures                  0    0  0       0       0.034
Twix                              1    0  1       0       0.546
Kit Kat                           1    0  1       0       0.313
Snickers                           0    0  1       0       0.546
                  pricepercent winpercent
Reese's Peanut Butter cup       0.651   84.18029
Reese's Miniatures            0.279   81.86626
Twix                        0.906   81.64291
Kit Kat                     0.511   76.76860
Snickers                    0.651   76.67378
```

The top 5 candies are Reese's Peanut Butter Cup, Reese's Miniatures, Twix, Kit Kat, and Snickers.

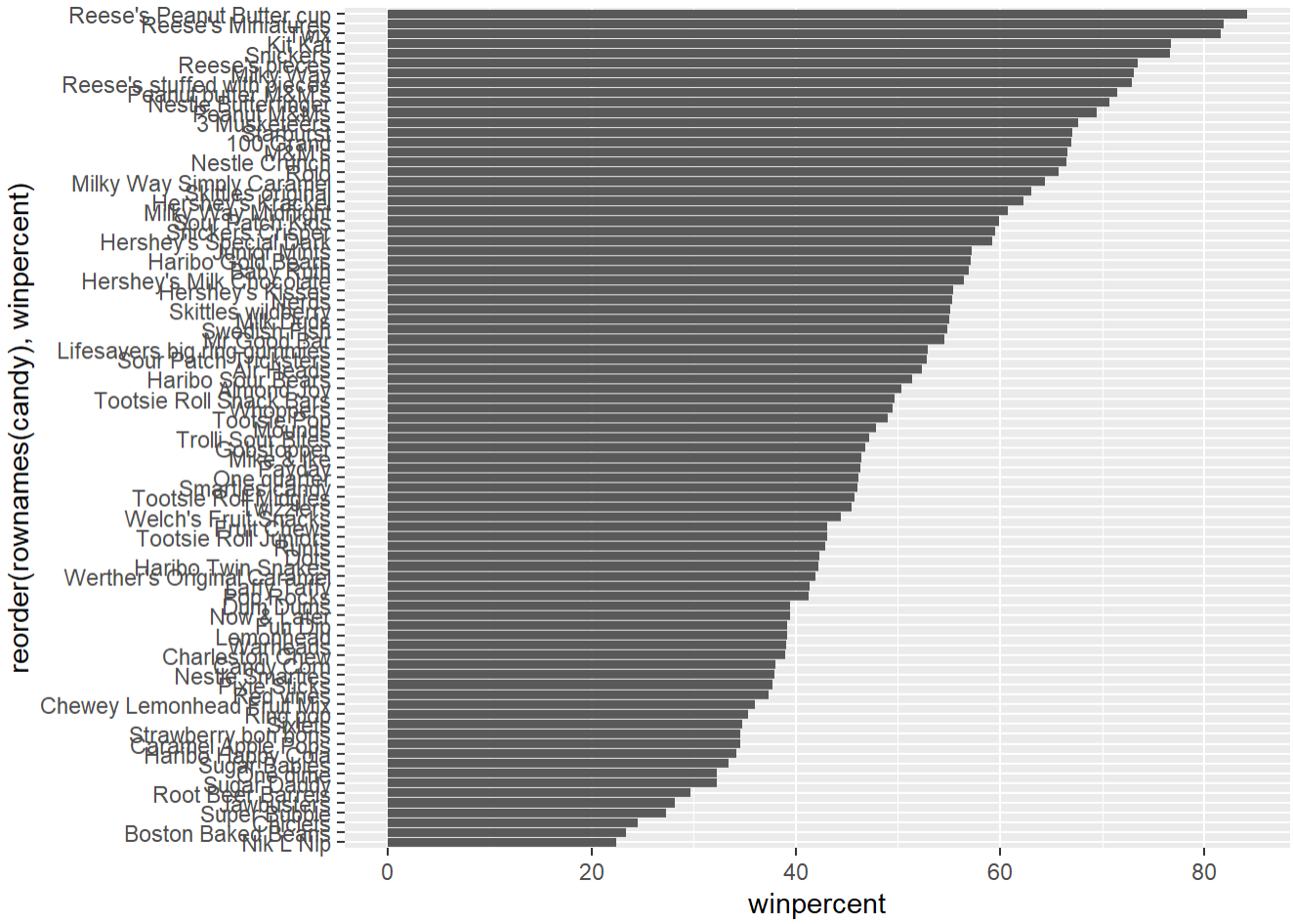
Q15. Make a first barplot of candy ranking based on winpercent values.

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

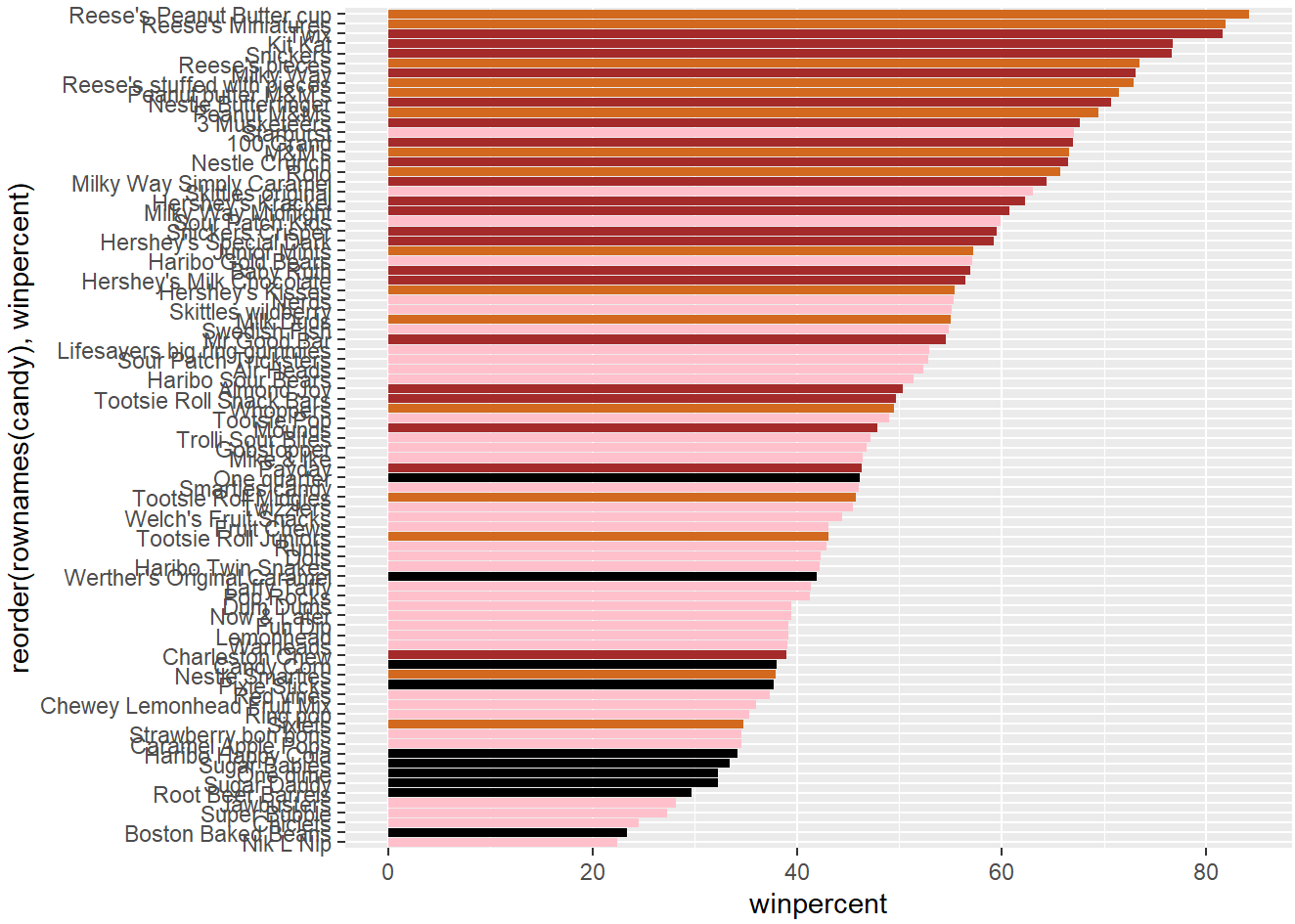


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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




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



Q17. What is the worst ranked chocolate candy?

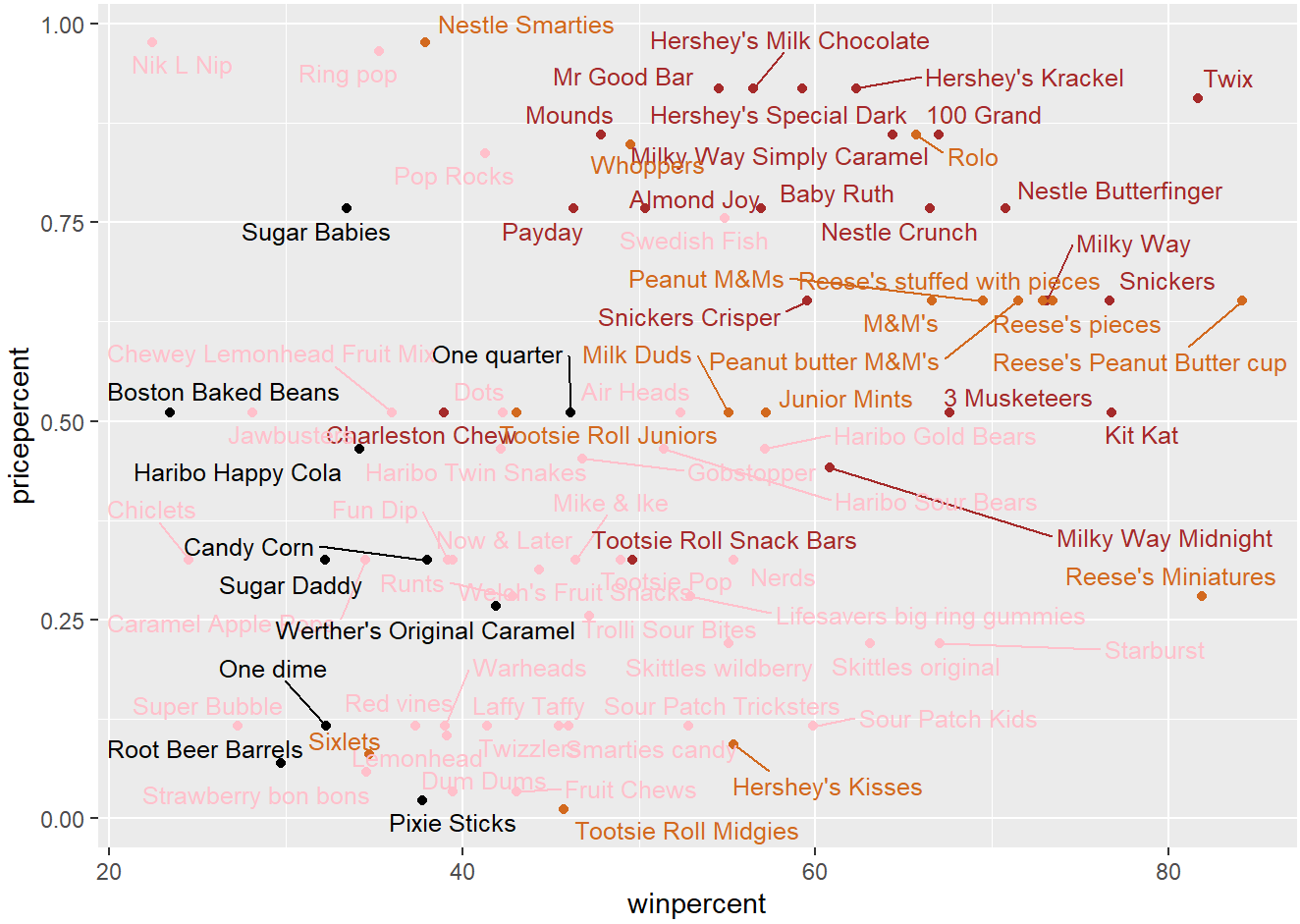
Sixlets. And this is factually accurate

Q18. What is the best ranked fruity candy?

Starburst are the best rated fruit candy.

##4. Taking a look at pricepercent

```
library(ggrepel)
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 100)
```



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 are among the highest in winpercent while being relatively low on pricepercent, meaning they are the best bang for one's buck.

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

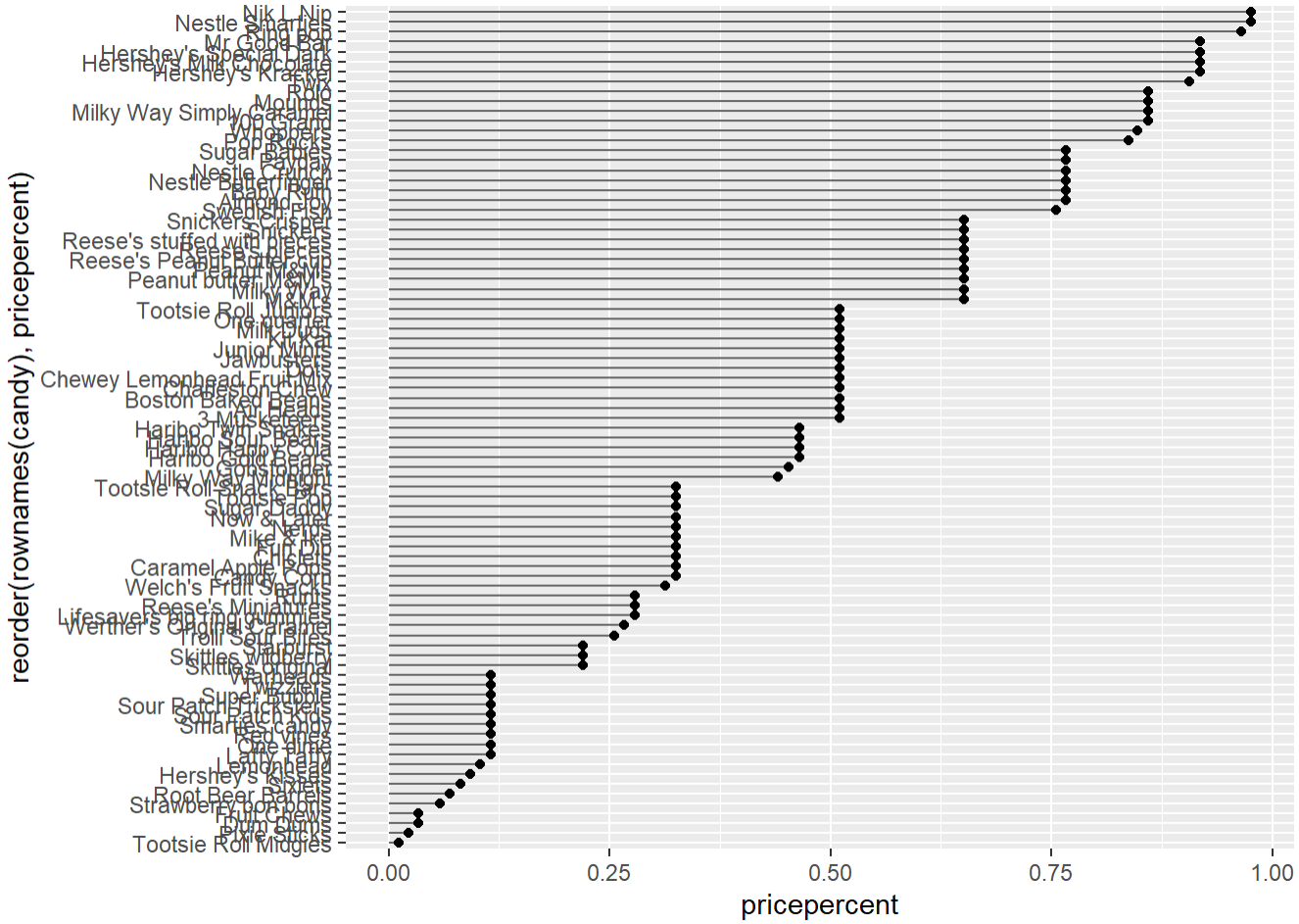
Nik L Nip, Ring pop, Smarties, Hershey's Krackel and Hershey's Milk Chocolate are the most expensive candies, and Nik L Nip is the least popular.

```
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
```

Q21. Make a barplot again

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

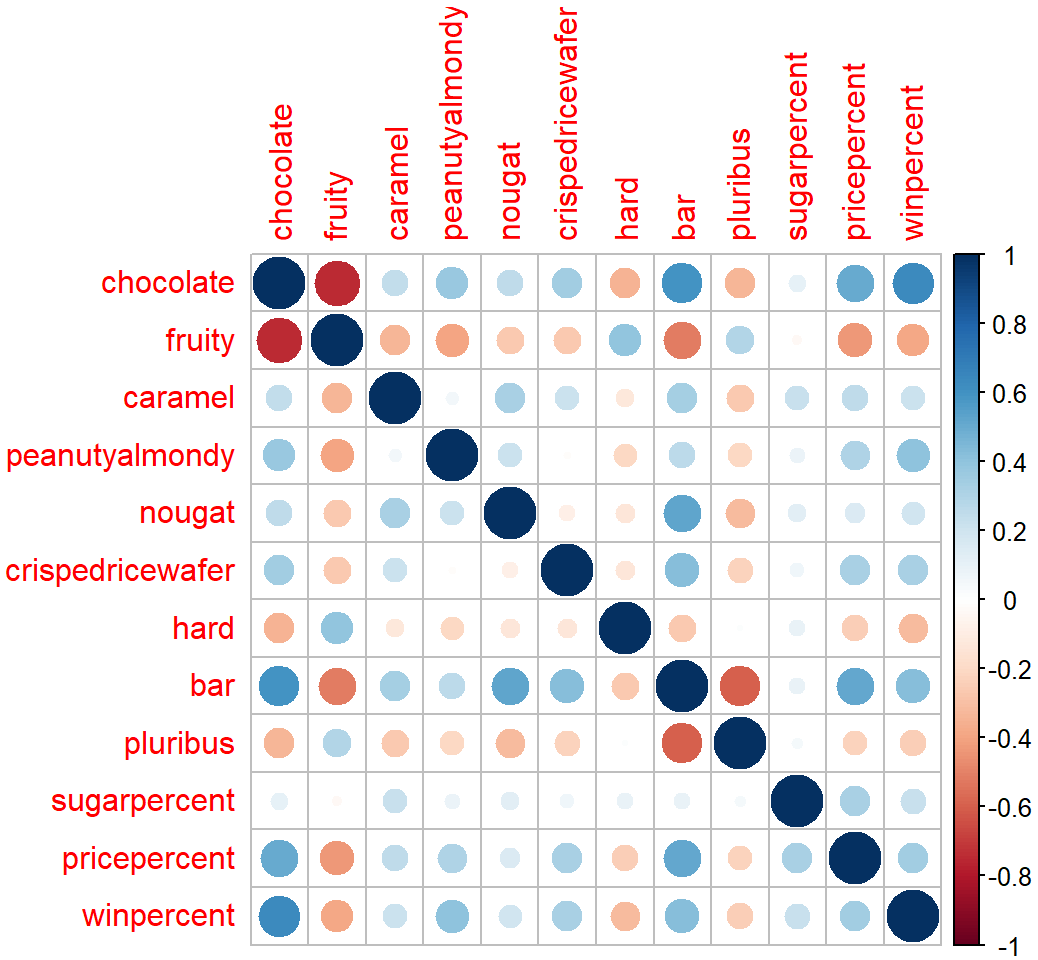


##5 Exploring the correlation structure

```
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 are the most anti-corrlated, so people don't typically like chocolate and fruity candy. Personally, chocolate with real fruit is a great combination.

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

winpercent and chocolate, as well as chocolate and bar, are the best correlations. This means people are most likely to choose chocolate over another option AND like chocolate in bar form.

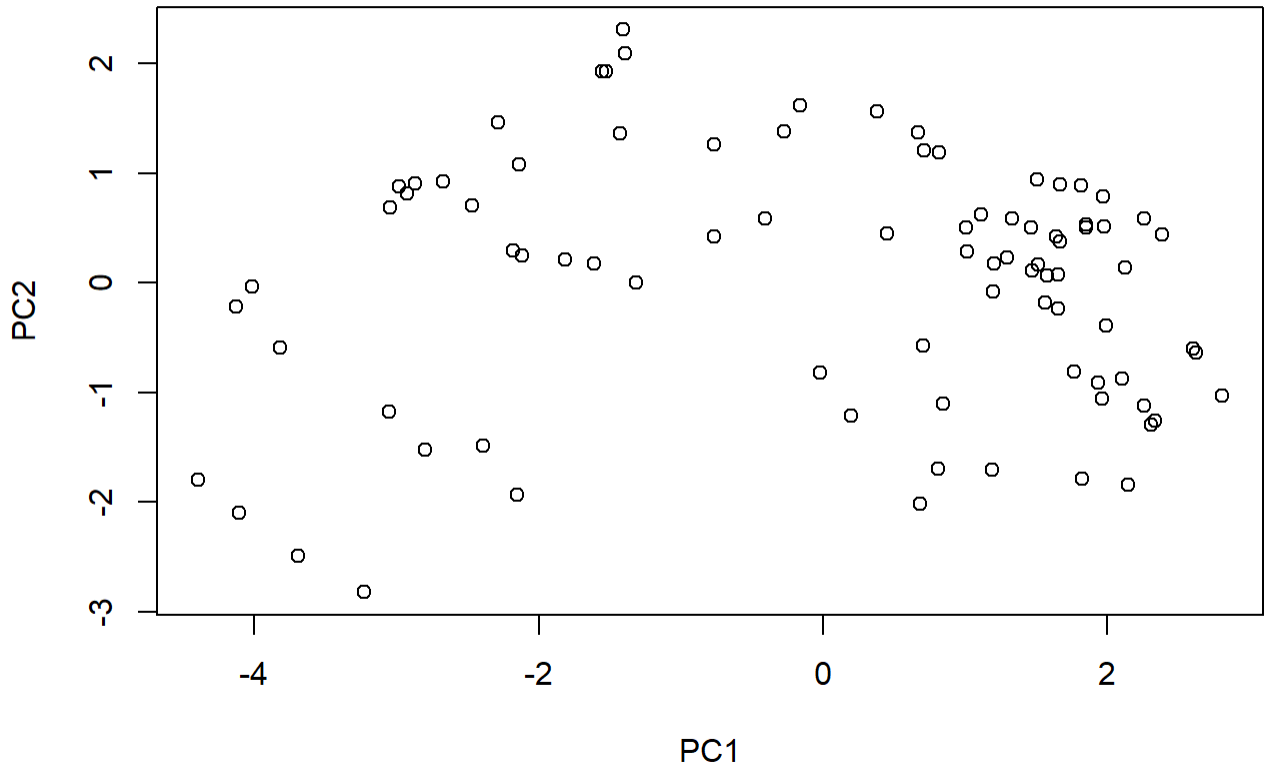
##6. Principal Component Analysis

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

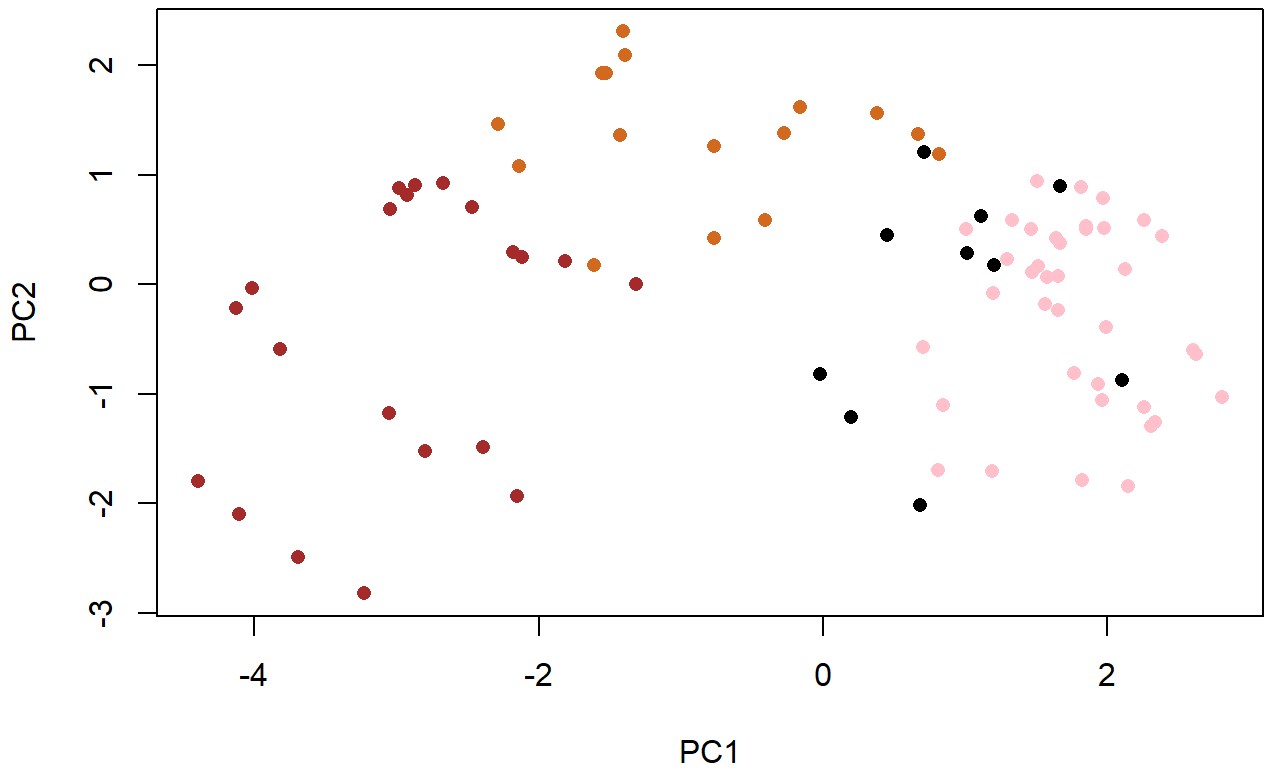
```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
```

```
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
```

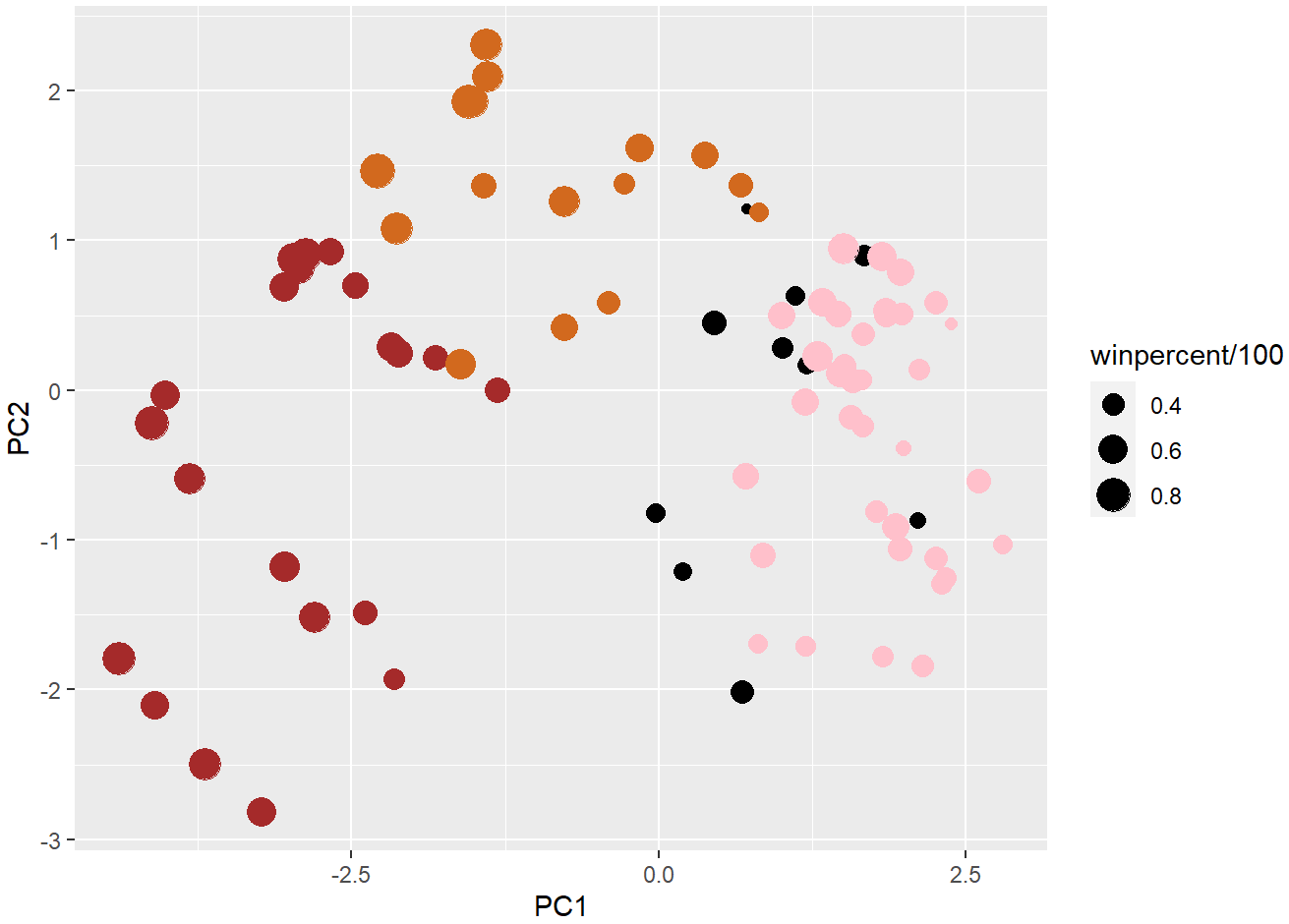
```
plot(pca$x[,1:2])
```



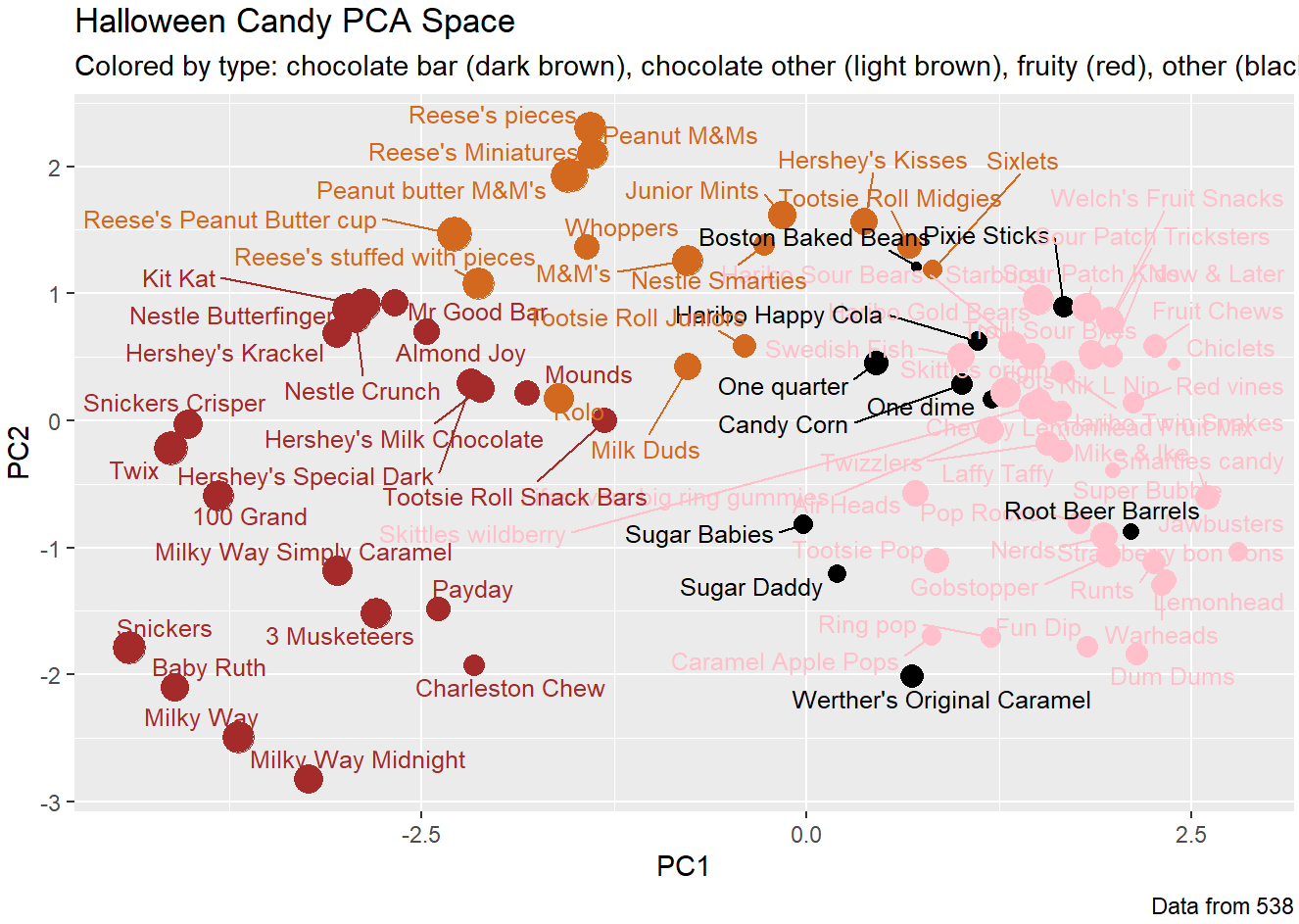

```
plot(pca$x[,1:2], col=my_cols, pch=16)
```




```
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 = 100)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown), fruity (red), other (black)",
       caption="Data from 538")
```

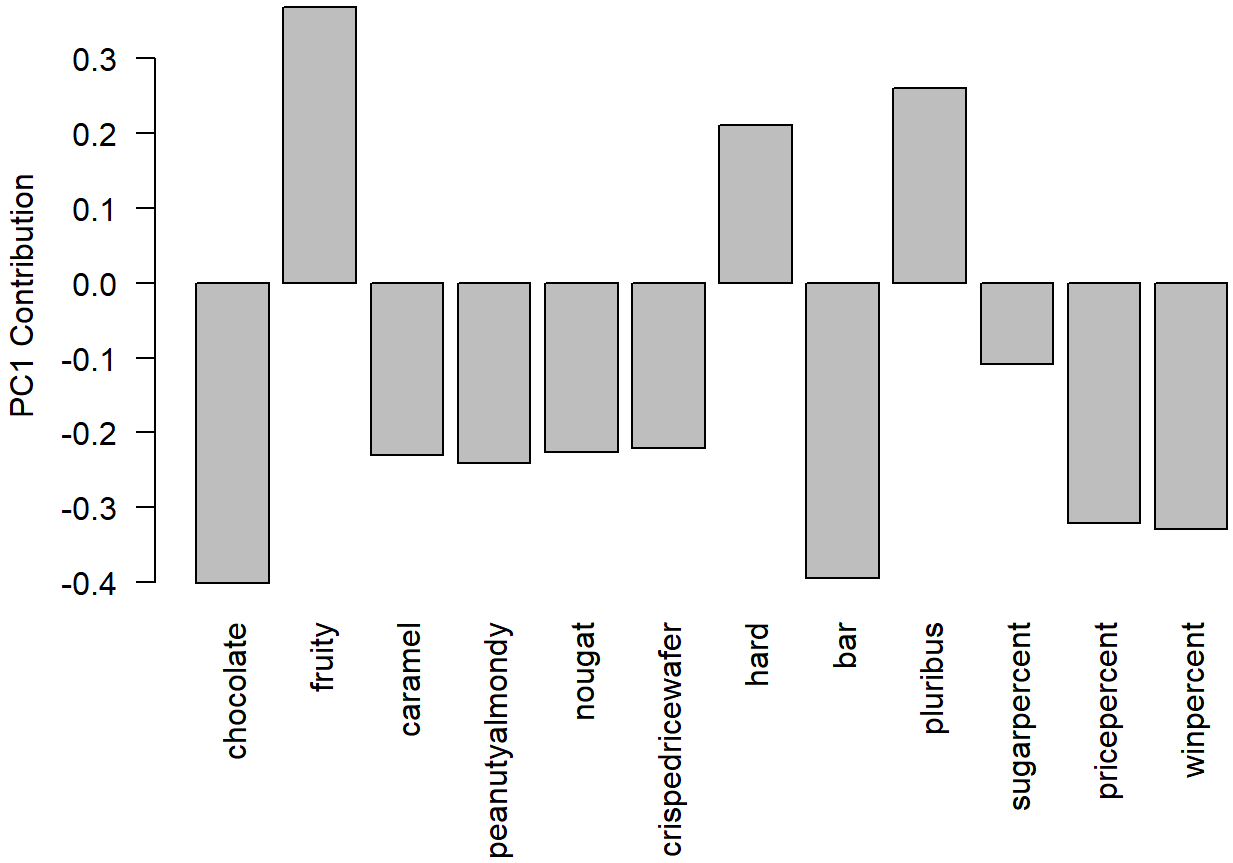



```
library(plotly)
```

```
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)
```

```
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, pluribus and hard are all picked up positively by PC1. This makes sense considering some of the popular candies like skittles and mike and ikes, which are fruity, hard, and come in a package of many (pluribus).