HALLOWEEN MINI PROJECT

Define the URL for the candy data

```
candy_url <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/ca

# Read the data from the URL

candy <- read.csv(url(candy_url), row.names = 1)

# Display the first few rows of the data

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?

num_candy_types <- nrow(candy)
num_candy_types
```

```
[1] 85
```

```
# Q2: How many fruity candy types are in the dataset?

fruity_candy_types <- table(candy$fruity)["yes"]

fruity_candy_types
```

```
<NA>
  NA
```

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

```
[1] 81.64291
```

```
# Q3: What is your favorite candy in the dataset and what is its

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
```

install.packages(“skimr”)

```
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?

sugrapercnet, priccepercent and winpercent look on a different scale than the others.

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

A zero most likely represents “not chocolate” and a one represents “chocolate”

```
# Q8: Plot a histogram of winpercent values

hist(candy$winpercent, main = "Histogram of Winpercent Values", xlab = "Winpercent")
```

Q9. Is the distribution of winpercent values symmetrical?

No it’s not.

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

Above %50

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

```
# Extract winpercent values for chocolate and fruit candies

chocolate_winpercent <- candy$winpercent[candy$chocolate == 1]

fruit_winpercent <- candy$winpercent[candy$chocolate == 0]

# Perform a t-test to compare the means

t_test_result <- t.test(chocolate_winpercent, fruit_winpercent)

# Print the t-test result

print(t_test_result)
```

```
    Welch Two Sample t-test

data:  chocolate_winpercent and fruit_winpercent
t = 7.3031, df = 67.539, p-value = 4.164e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 13.64744 23.91110
sample estimates:
mean of x mean of y
 60.92153  42.14226
```

chocolate candies are higher ranked than fruit candies(The mean winpercent for chocolate candies (mean of x) is 60.92153.)

Q12. Is this difference statistically significant?

Yes, the difference in winpercent values between chocolate and fruit candies is statistically significant. This is indicated by the very small p-value (4.164e-10), which is well below the typical significance level of 0.05.

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

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

least_liked
```

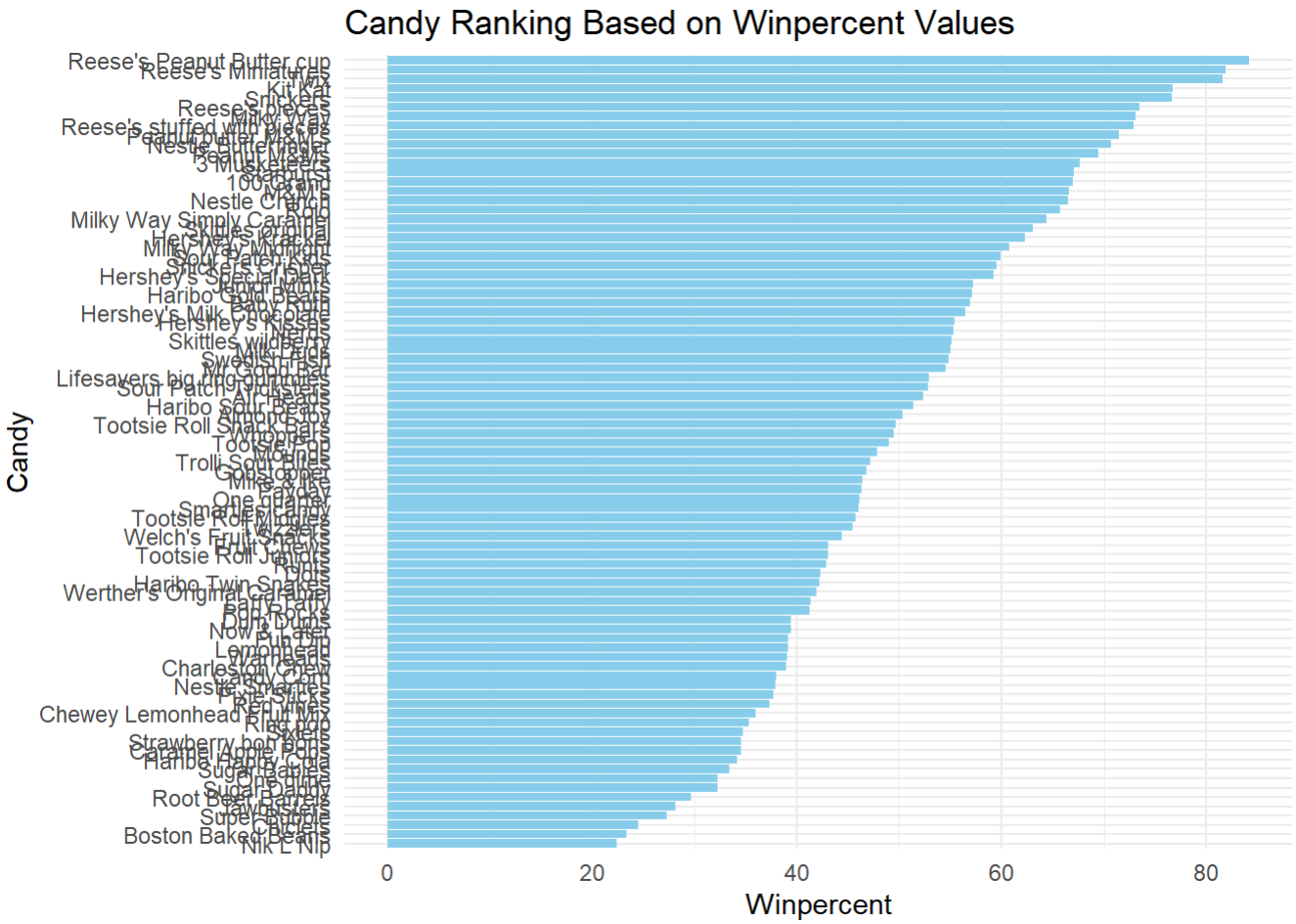
```
                  chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                 0      1       0              0      0
Boston Baked Beans        0      0       0              1      0
Chiclets                  0      1       0              0      0
Super Bubble              0      1       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
# Q14. What are the top 5 all-time favorite candy types out of this set?

favorite <- head(candy[order(candy$winpercent, decreasing = TRUE), ], n = 5)

favorite
```

```
                         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
```

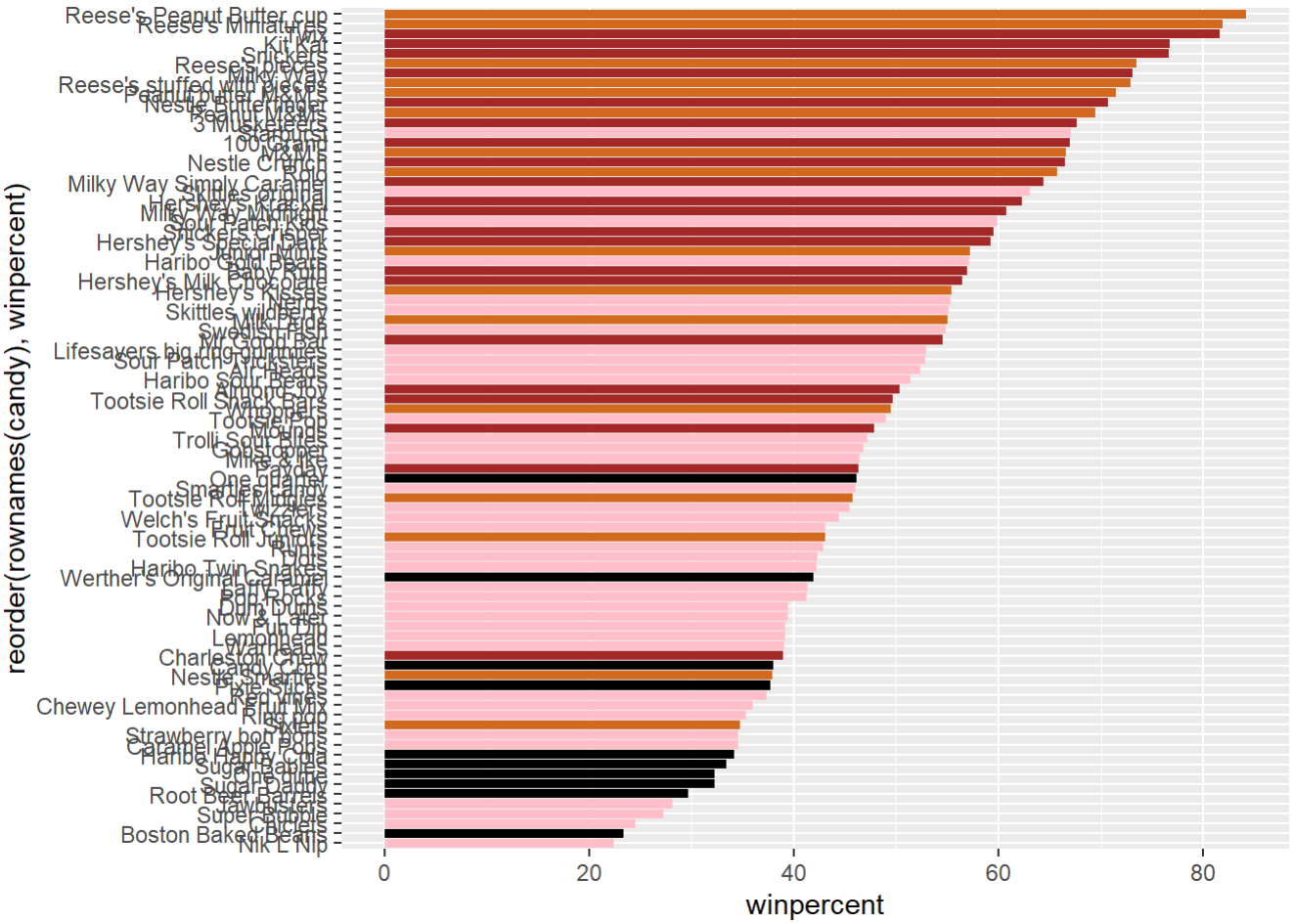
```
library(ggplot2)
```

```
ggplot(candy, aes(x = winpercent, y = reorder(row.names(candy), winpercent))) +

  geom_bar(stat = "identity", fill = "skyblue") +

  labs(title = "Candy Ranking Based on Winpercent Values",

       x = "Winpercent", y = "Candy") +

  theme_minimal()
```




```
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

- Q18. What is the best ranked fruity candy?

starburst

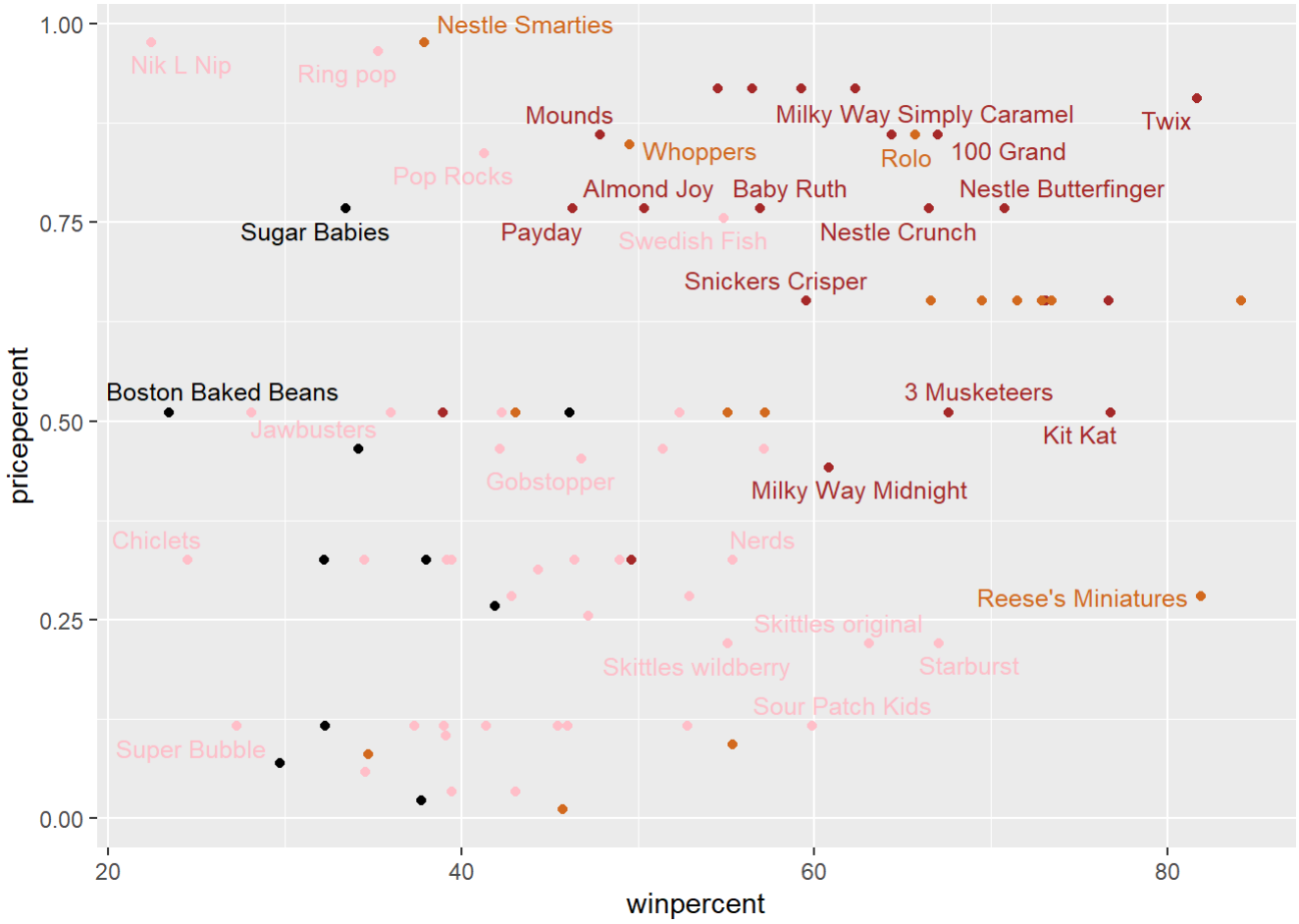
install.packages("ggrepel")

```
library(ggrepel)
```


```
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 = 5)
```

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




```
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
```

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?

Hershey's Milk Chocolate

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

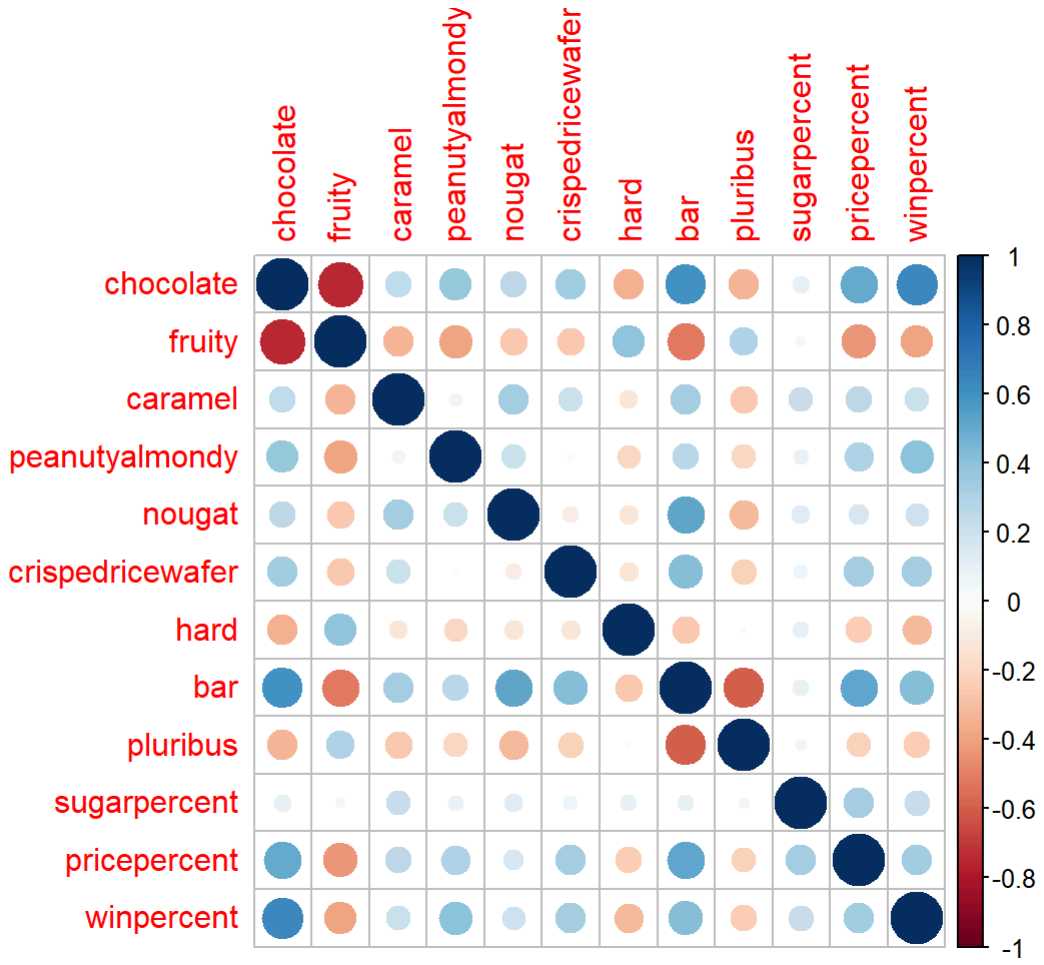
Nik L Nip

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)?

winpercent and pricepercent

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

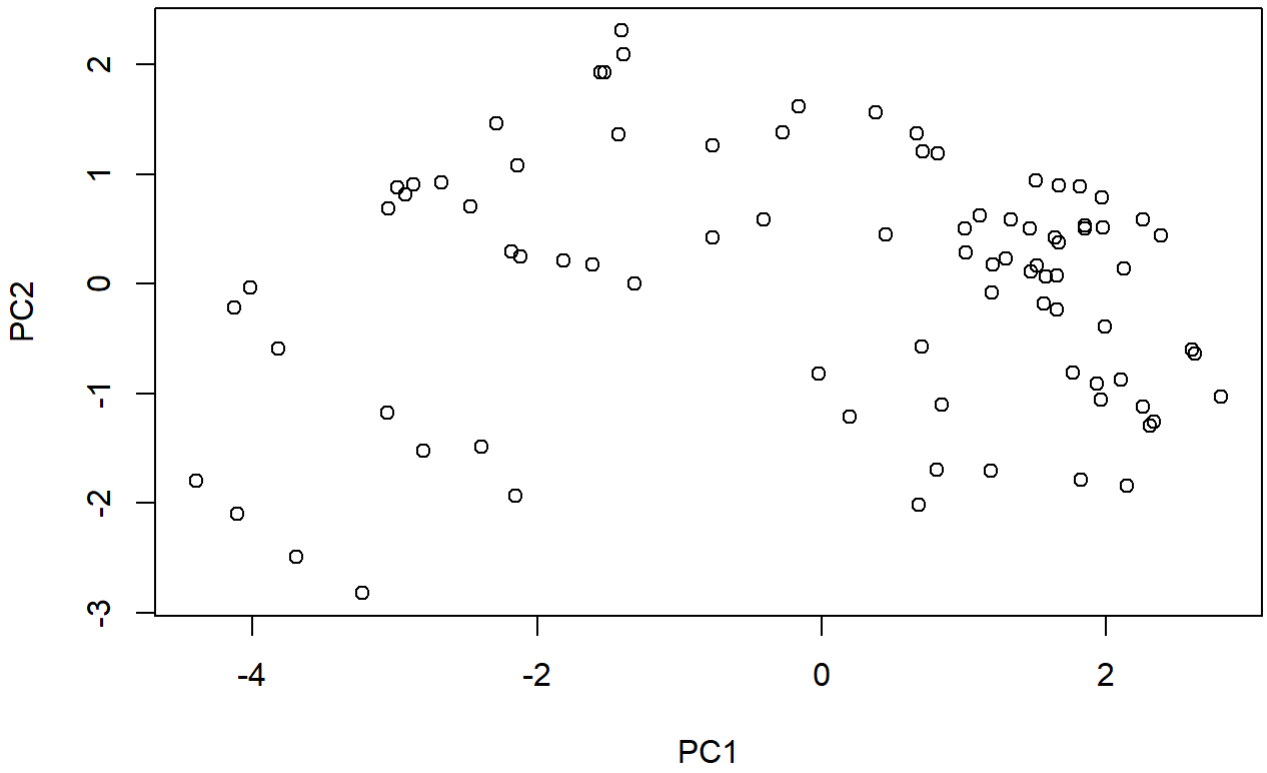
chocolate and fruity

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

summary(pca)
```

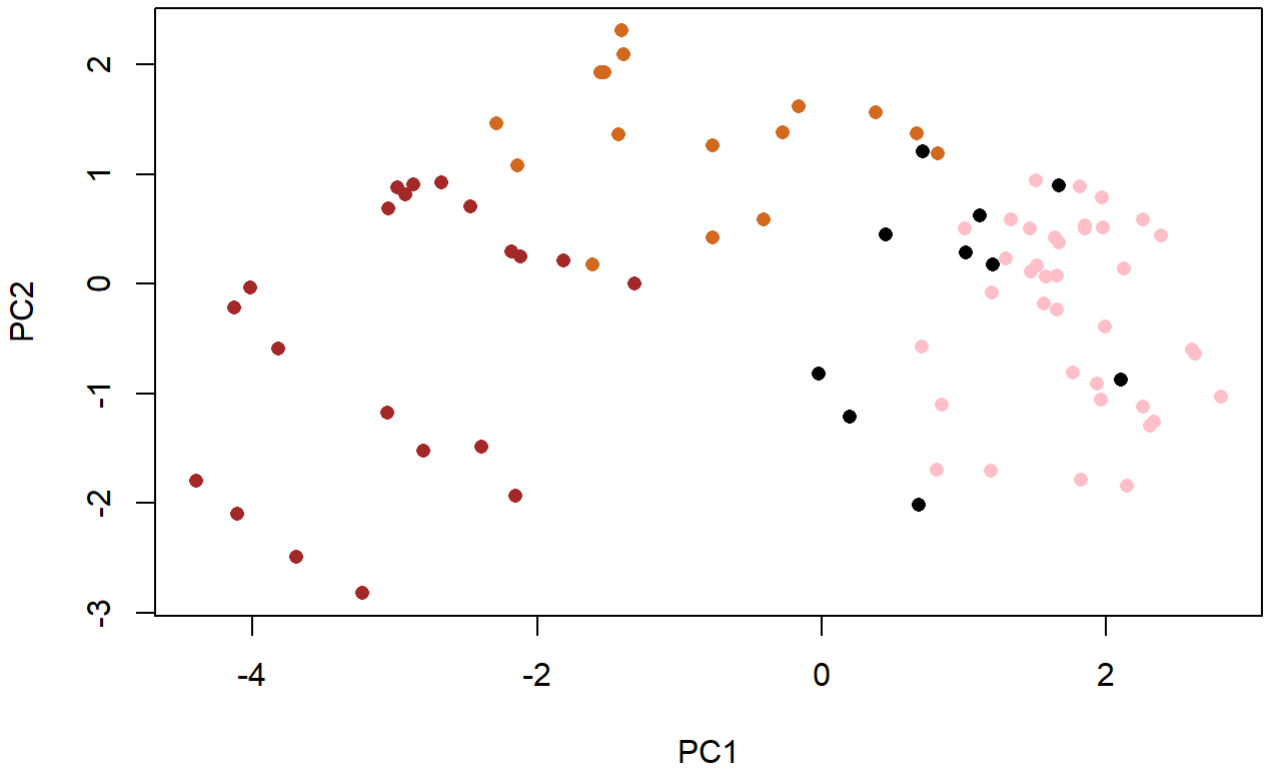
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
```

```
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
plot(pca$x[, 1], pca$x[, 2], xlab = "PC1", ylab = "PC2")
```




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

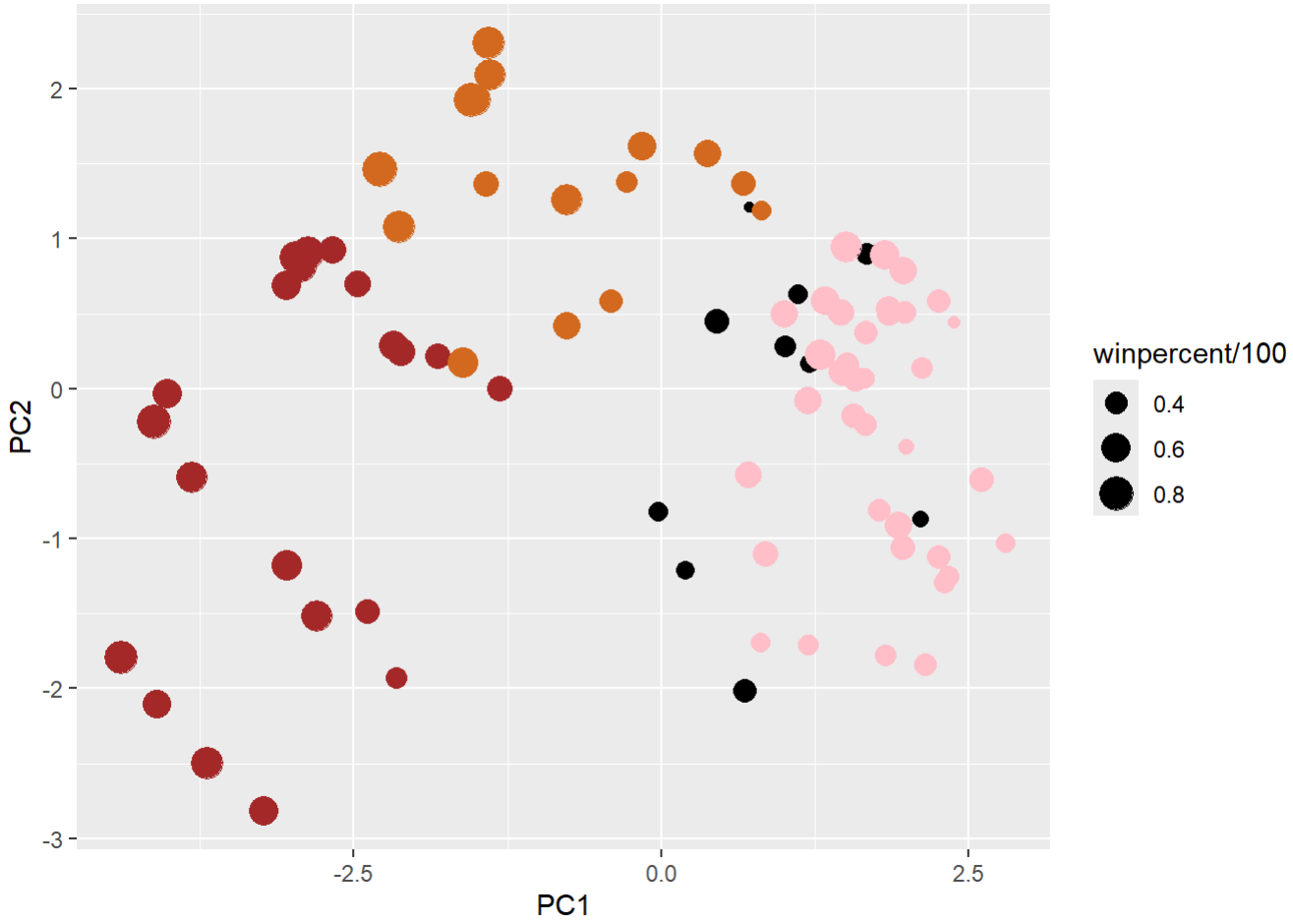



```
# Make a new data-frame with our PCA results and candy data

my_data <- cbind(candy, pca$x[,1:3])
```

install.packages("ggplot2")

```
library(ggplot2)
```

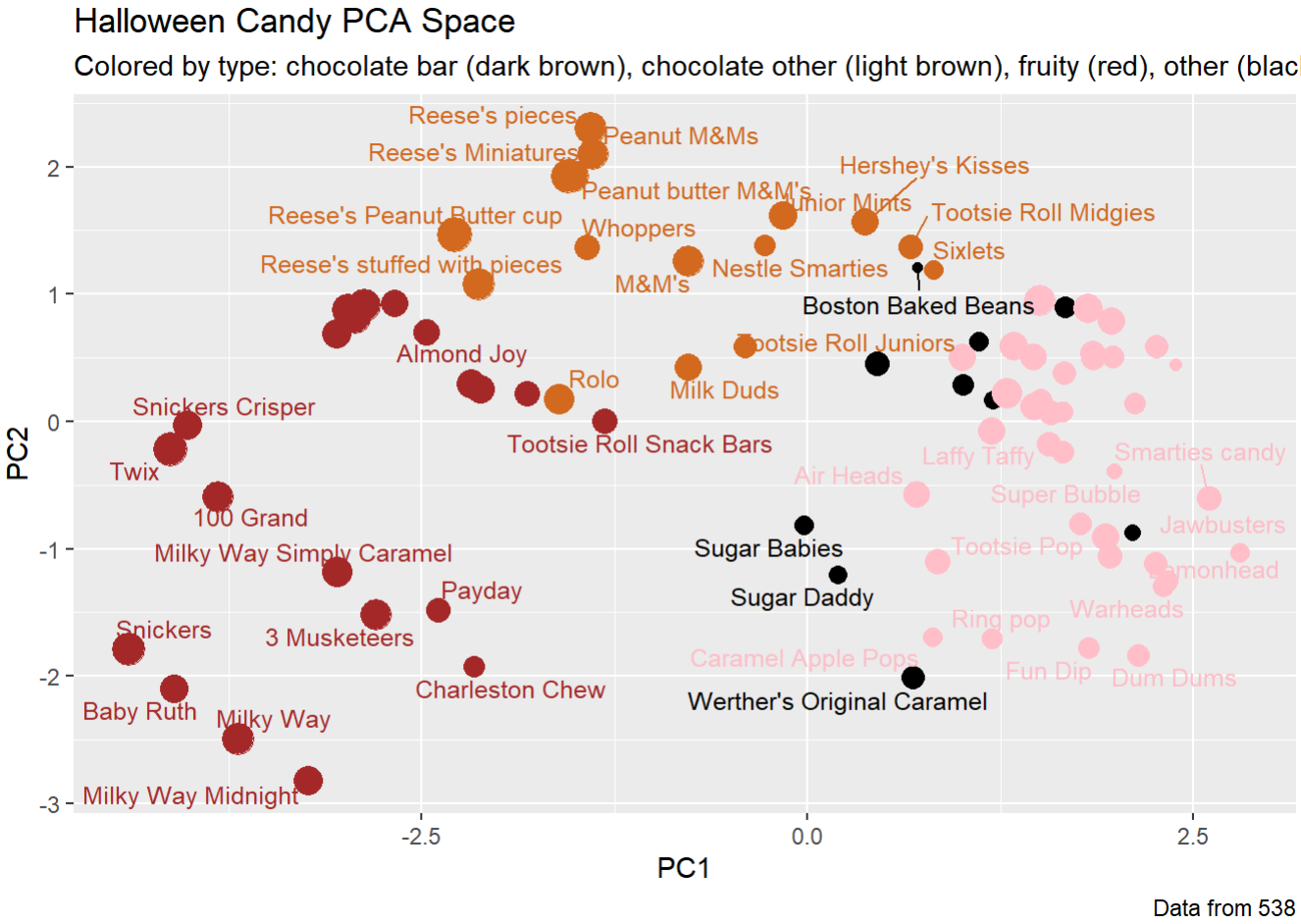
```
p <- ggplot(my_data) +

        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 = 7)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown), frui
       caption="Data from 538")
```

```
Warning: ggrepel: 40 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

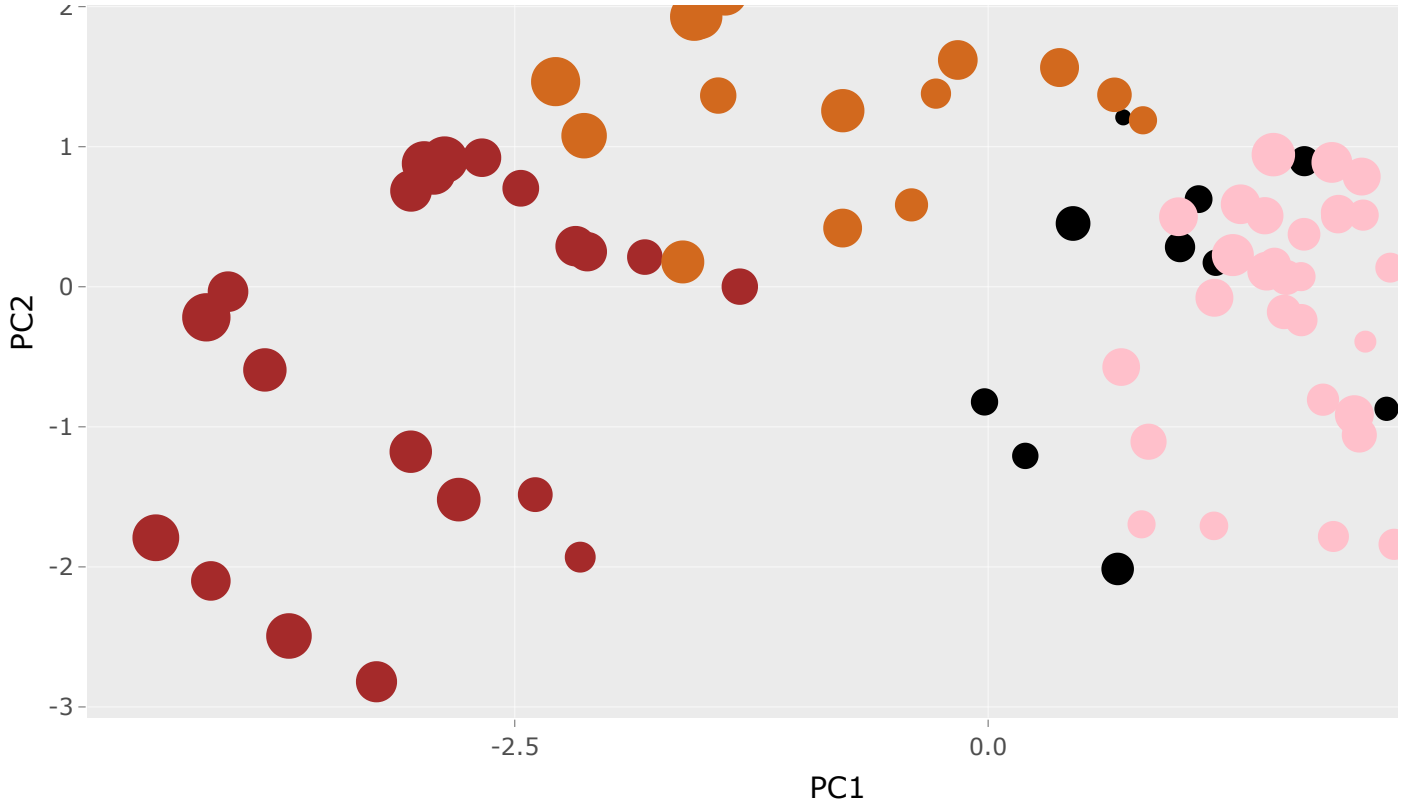


install.packages("plotly")

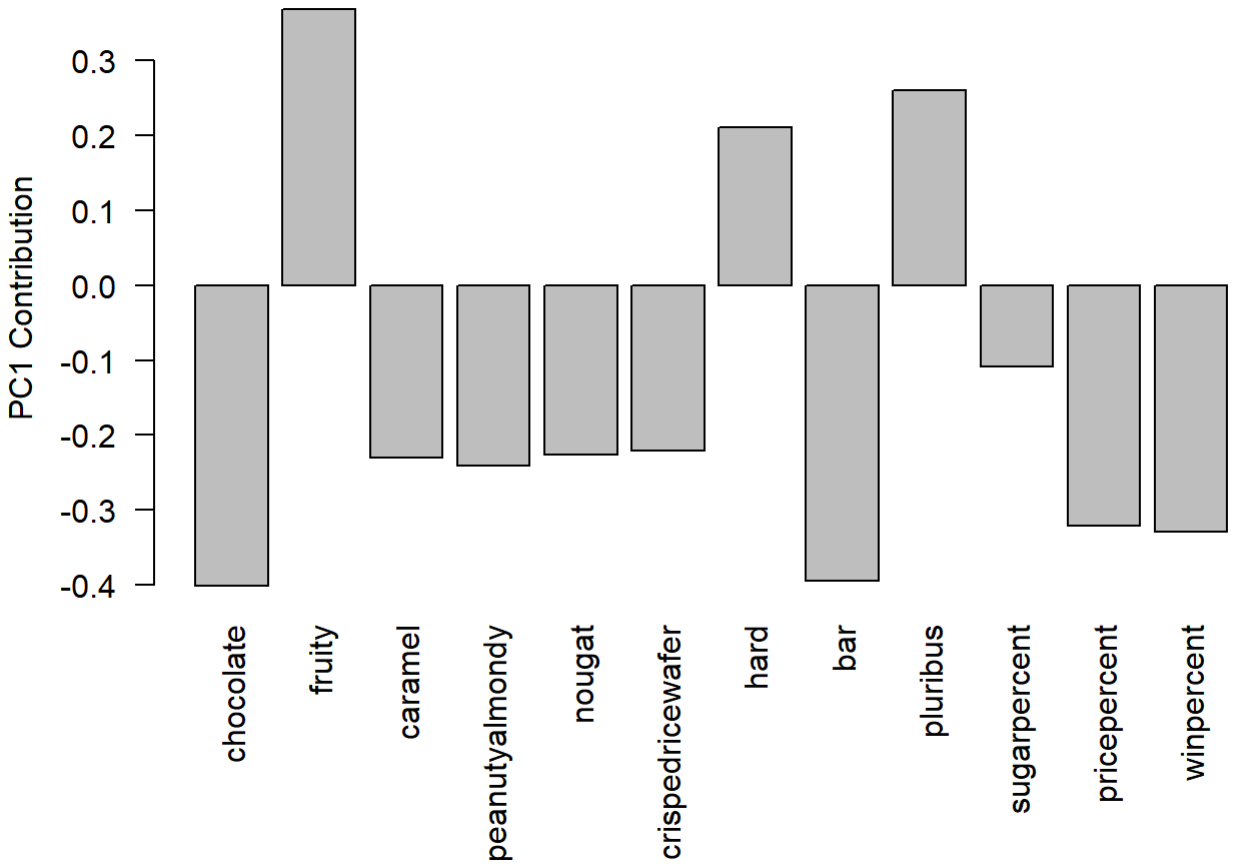
```
library(plotly)
```

```
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 candies are picked up by PC1 in the positive direction. This doesn't make sense to me. The positive association of fruity candies with PC1 suggests that candies with fruity characteristics are more likely to have higher winpercent values, indicating higher popularity, but winpercent and fruity has the opposite effects.