Class 10: Exploratory Analysis of Halloween Candy

AUTHOR
Duc Nguyen

1. 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
```

> **Ans1: 85 different candy types**

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

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

```
[1] 38
```

> **Ans2: 38 fruity candy types**

2. What is your favorate candy?

```
row.names(candy)
```

```
 [1] "100 Grand"                   "3 Musketeers"
 [3] "One dime"                    "One quarter"
 [5] "Air Heads"                   "Almond Joy"
 [7] "Baby Ruth"                   "Boston Baked Beans"
 [9] "Candy Corn"                  "Caramel Apple Pops"
 [9] "Candy Corn"                  "Caramel Apple Pops"
[11] "Charleston Chew"             "Chewey Lemonhead Fruit Mix"
[13] "Chiclets"                    "Dots"
[15] "Dum Dums"                    "Fruit Chews"
[17] "Fun Dip"                     "Gobstopper"
[19] "Haribo Gold Bears"           "Haribo Happy Cola"
[21] "Haribo Sour Bears"           "Haribo Twin Snakes"
[23] "HersheyÕs Kisses"            "HersheyÕs Krackel"
[25] "HersheyÕs Milk Chocolate"    "HersheyÕs Special Dark"
[27] "Jawbusters"                  "Junior Mints"
[29] "Kit Kat"                     "Laffy Taffy"
[31] "Lemonhead"                   "Lifesavers big ring gummies"
[33] "Peanut butter M&MÕs"         "M&MÕs"
[35] "Mike & Ike"                  "Milk Duds"
[37] "Milky Way"                   "Milky Way Midnight"
[39] "Milky Way Simply Caramel"    "Mounds"
[41] "Mr Good Bar"                 "Nerds"
[43] "Nestle Butterfinger"         "Nestle Crunch"
[45] "Nik L Nip"                   "Now & Later"
[47] "Payday"                      "Peanut M&Ms"
[49] "Pixie Sticks"                "Pop Rocks"
[51] "Red vines"                   "ReeseÕs Miniatures"
[53] "ReeseÕs Peanut Butter cup"   "ReeseÕs pieces"
[55] "ReeseÕs stuffed with pieces" "Ring pop"
[57] "Rolo"                        "Root Beer Barrels"
[59] "Runts"                       "Sixlets"
[61] "Skittles original"           "Skittles wildberry"
[63] "Nestle Smarties"             "Smarties candy"
[65] "Snickers"                    "Snickers Crisper"
[67] "Sour Patch Kids"             "Sour Patch Tricksters"
[69] "Starburst"                   "Strawberry bon bons"
[71] "Sugar Babies"                "Sugar Daddy"
[73] "Super Bubble"                "Swedish Fish"
[75] "Tootsie Pop"                 "Tootsie Roll Juniors"
[77] "Tootsie Roll Midgies"        "Tootsie Roll Snack Bars"
[79] "Trolli Sour Bites"           "Twix"
[81] "Twizzlers"                   "Warheads"
[83] "WelchÕs Fruit Snacks"        "WertherÕs Original Caramel"
[85] "Whoppers"
```

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

```
candy["HersheyÕs Kisses", ]$winpercent
```

```
[1] 55.37545
```

> **Ans3: My favorite candy is not in the dataset. However, my daughter’s favorite candy is Hershey’s Kisses, and it’s winpercent value is 55.37545.**

> **Q4. What is the winpercent value for “Kit Kat”?**

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

```
[1] 76.7686
```

> **Ans4: The winpercent value for “Kit Kat” is 76.7686.**

> **Q5. What is the winpercent value for “Tootsie Roll Snack Bars”?**

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

```
[1] 49.6535
```

> **Ans5: The winpercent value for “Tootsie Roll Snack Bars” is 49.6535.**

Note: Install the “skimr” package first using ‘install.packages(“skimr”)’ 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?**

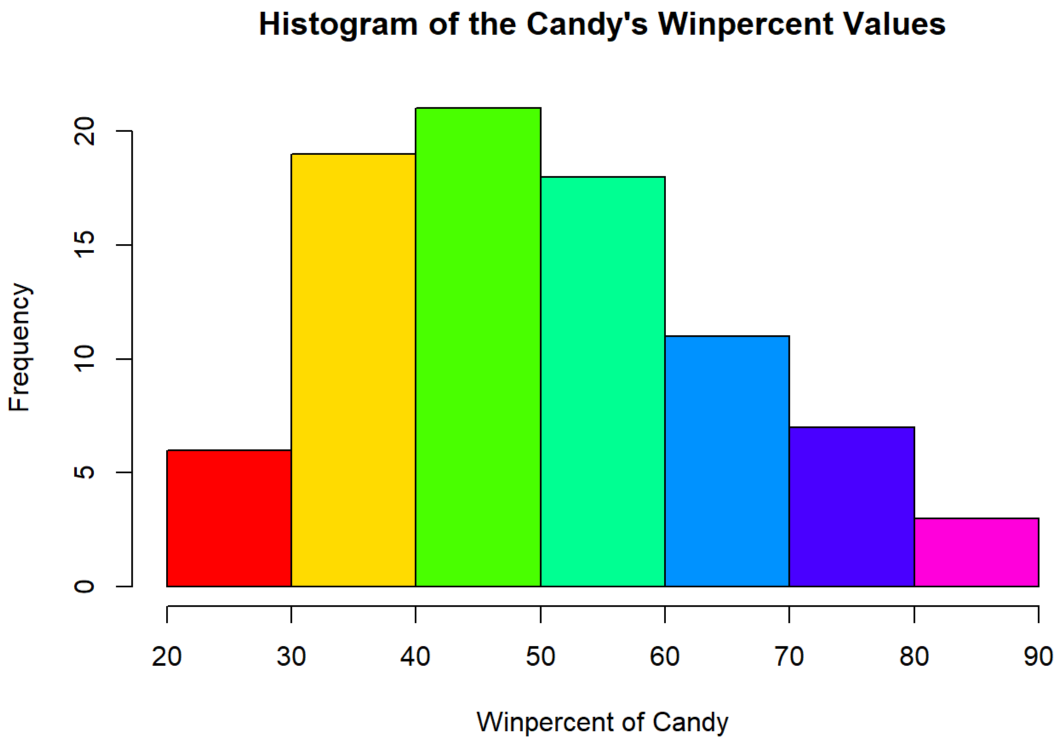
> **Ans6: Yes, the “winpercent” variable looks to be on a different scale to the majority of the others.**

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

> **Ans7: One (1) represents for the candy which contains chocolate in its ingredients, and zero (0) represents for the candy which does not contain chocolate in its ingredients.**

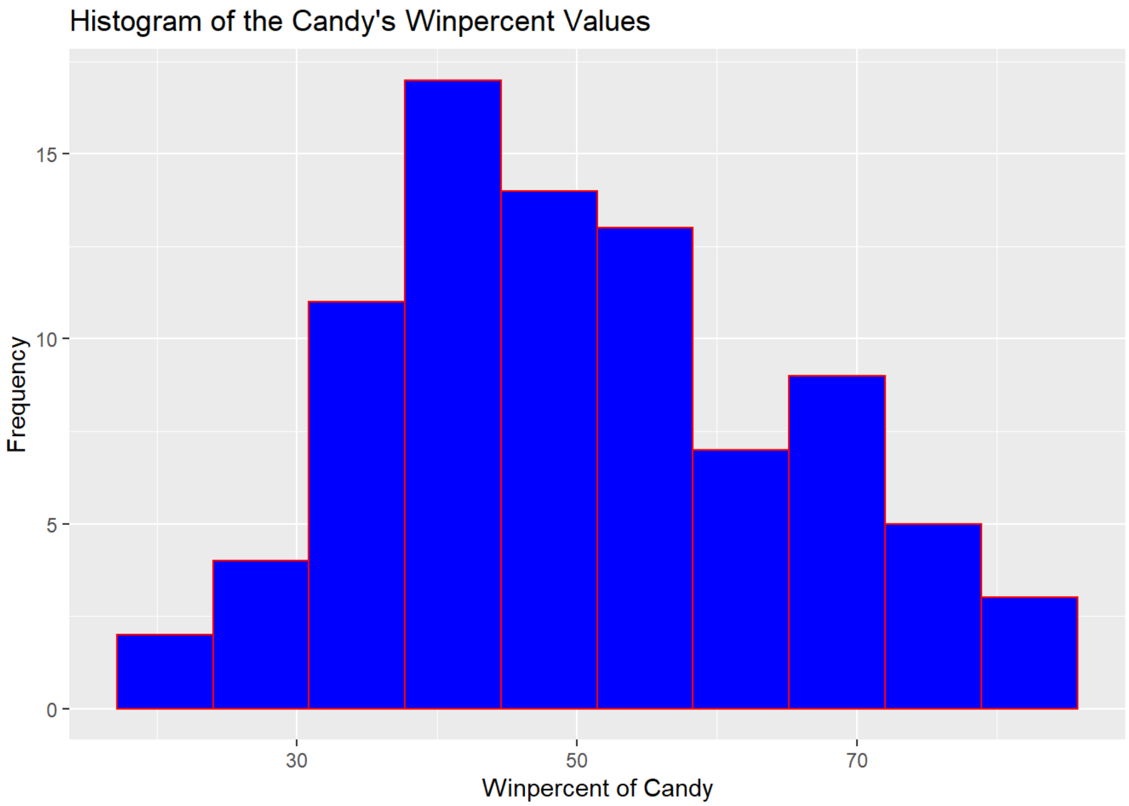
> **Q8. Plot a histogram of winpercent values**

```
hist(candy$winpercent, col = rainbow(7),
     main = "Histogram of the Candy's Winpercent Values",
     xlab = "Winpercent of Candy", ylab = "Frequency")
```



Or using ggplot2 packages

```
library(ggplot2)
ggplot(candy) +
  aes(winpercent) +
  geom_histogram(bins = 10, col = "red", fill = "blue") +
  labs(title = "Histogram of the Candy's Winpercent Values",
       x = "Winpercent of Candy", y = "Frequency")
```



> **Q9. Is the distribution of winpercent values symmetrical?**

> **Ans9: No, the distribution of winpercent values is not symmetrical.**

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

> **Ans10: The center of the distribution is above 50%**

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

```r
chocolate.ind <- as.logical(candy$chocolate)
head(candy[chocolate.ind,])
```

```
                 chocolate fruity caramel peanutyalmondy nougat
100 Grand                1      0       1              0      0
3 Musketeers             1      0       0              0      1
Almond Joy               1      0       0              1      0
Baby Ruth                1      0       1              1      1
Charleston Chew          1      0       0              0      1
HersheyÕs Kisses         1      0       0              0      0
                 crispedricewafer hard bar pluribus sugarpercent pricepercent
100 Grand                       1    0   1        0        0.732        0.860
3 Musketeers                    0    0   1        0        0.604        0.511
Almond Joy                      0    0   1        0        0.465        0.767
Baby Ruth                       0    0   1        0        0.604        0.767
Charleston Chew                 0    0   1        0        0.604        0.511
HersheyÕs Kisses                0    0   0        1        0.127        0.093
                 winpercent
100 Grand          66.97173
```

```
100 Grand          66.97173
3 Musketeers       67.60294
Almond Joy         50.34755
Baby Ruth          56.91455
Charleston Chew    38.97504
HersheyÕs Kisses   55.37545
```

```r
chocolate.wins <- candy[chocolate.ind,]$winpercent
chocolate.wins
```

```
 [1] 66.97173 67.60294 50.34755 56.91455 38.97504 55.37545 62.28448 56.49050
 [9] 59.23612 57.21925 76.76860 71.46505 66.57458 55.06407 73.09956 60.80070
[17] 64.35334 47.82975 54.52645 70.73564 66.47068 69.48379 81.86626 84.18029
[25] 73.43499 72.88790 65.71629 34.72200 37.88719 76.67378 59.52925 48.98265
[33] 43.06890 45.73675 49.65350 81.64291 49.52411
```

```r
round(mean(chocolate.wins), 2) # Average winpercent of chocolate candy
```

```
[1] 60.92
```

```r
fruity.ind <- as.logical(candy$fruity)
fruity.wins <- candy[fruity.ind,]$winpercent
round(mean(fruity.wins), 2) # Average winpercent of fruity candy
```

```
[1] 44.12
```

> **Ans11: On average, the chocolate candy (60.92%) is HIGHER ranked than the fruit candy (44.12%).**

> **Q12. Is this difference statistically significant?**

```r
t.test(chocolate.wins, fruity.wins)
```

```

    Welch Two Sample t-test

data:  chocolate.wins and fruity.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 
```

> **Ans12: Yes, this is difference statistically significant because the p-value = 2.871e-08, which is less than 0.05.**

3. Overall Candy Ranking

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

```r
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
```

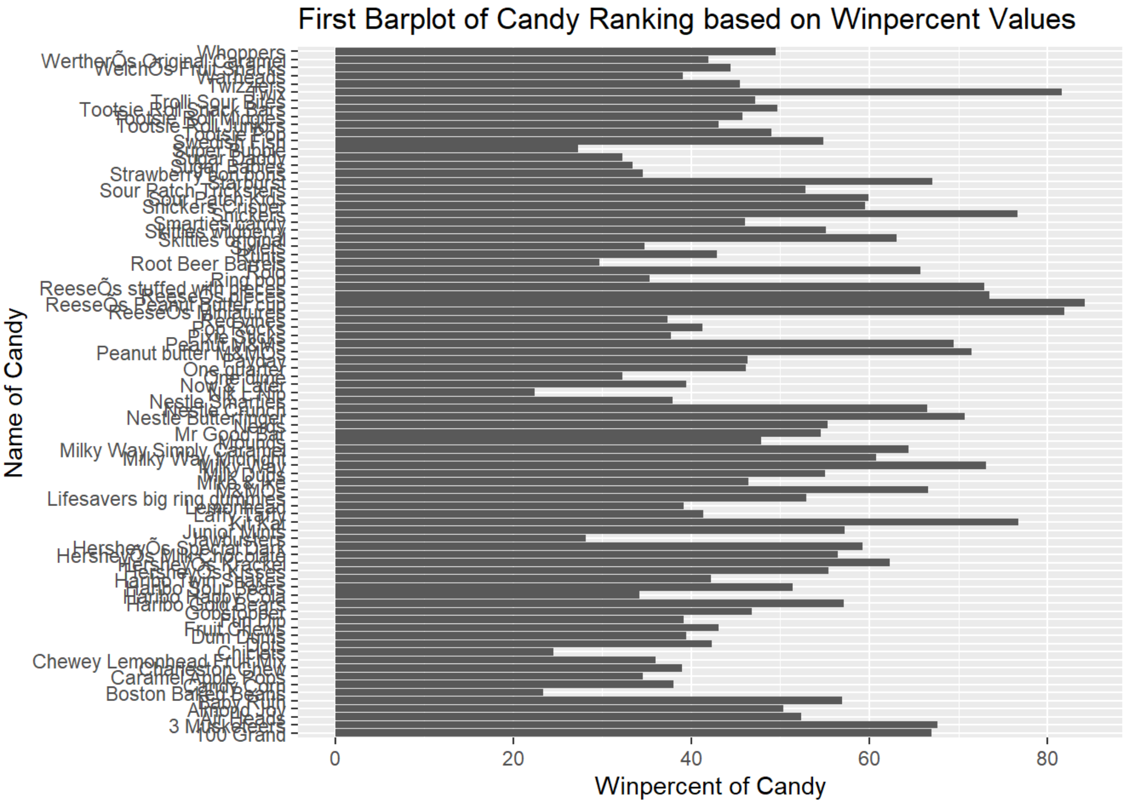
> **Q14. What are the top 5 all time favorite candy types out of this set?**

```
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
ReeseÕs Miniatures                1      0       0              1      0
ReeseÕs 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
ReeseÕs Miniatures                       0    0   0        0        0.034
ReeseÕs 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
ReeseÕs Miniatures               0.279   81.86626
ReeseÕs Peanut Butter cup        0.651   84.18029
```

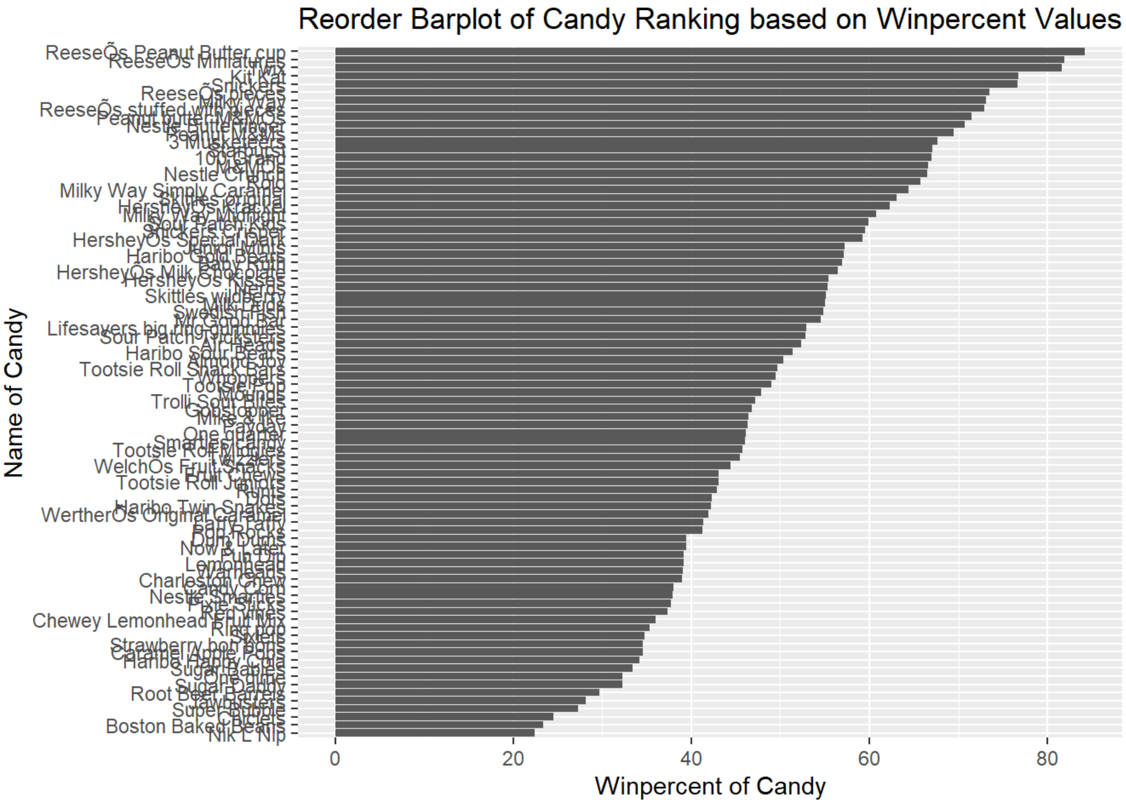
> **Q15. Make a first barplot of candy ranking based on winpercent values**

```
library(ggplot2)
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col() +
  labs(title = "First Barplot of Candy Ranking based on Winpercent Values",
       x = "Winpercent of Candy", y = "Name of Candy")
```



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

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col() +
  labs(title = "Reorder Barplot of Candy Ranking based on Winpercent Values",
       x = "Winpercent of Candy", y = "Name of Candy")
```



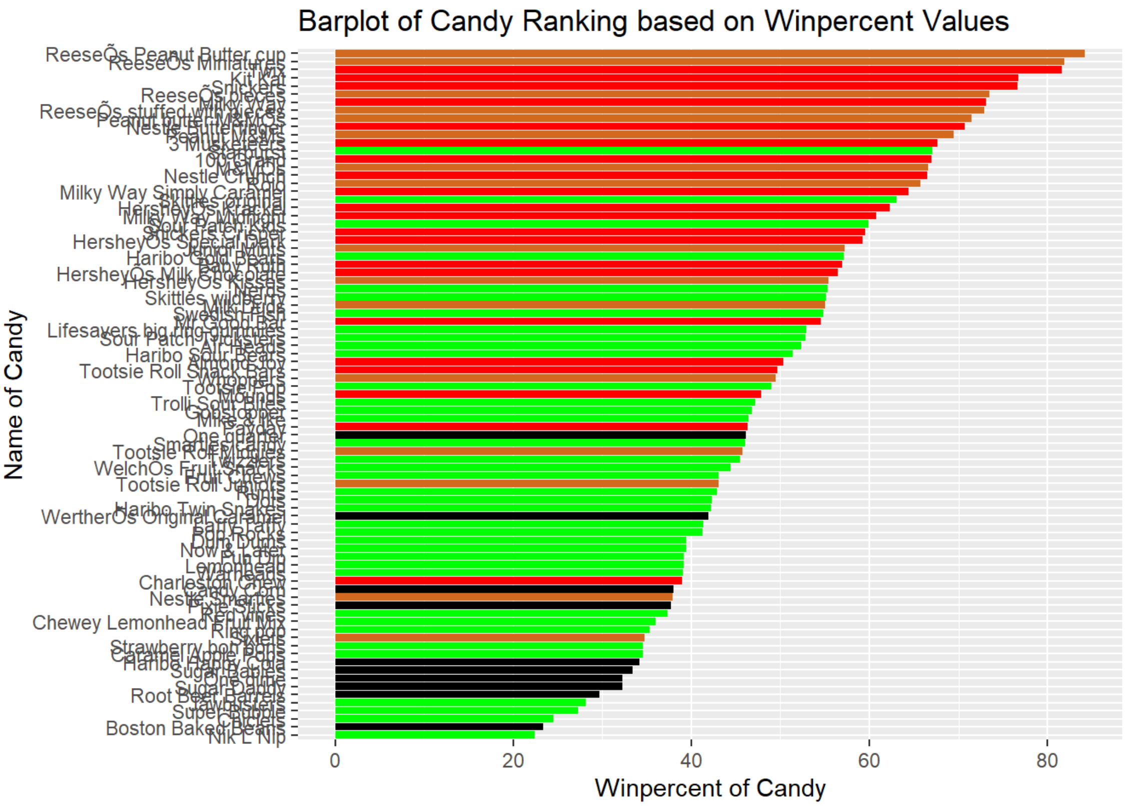
Time to add some useful color

Setup a color vector

```
my_cols=rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "red"
my_cols[as.logical(candy$fruity)] = "green"
```

Try improve barplot with these colors

```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill = my_cols) +
  labs(title = "Barplot of Candy Ranking based on Winpercent Values",
       x = "Winpercent of Candy", y = "Name of Candy")
```



```
ggsave("tmp.png") # To take a picture of the graph above
```

```
Saving 7 x 5 in image
```

> **Q17. What is the worst ranked chocolate candy?**

> **Ans17: Sixlets**

> **Q18. What is the best ranked fruity candy?**

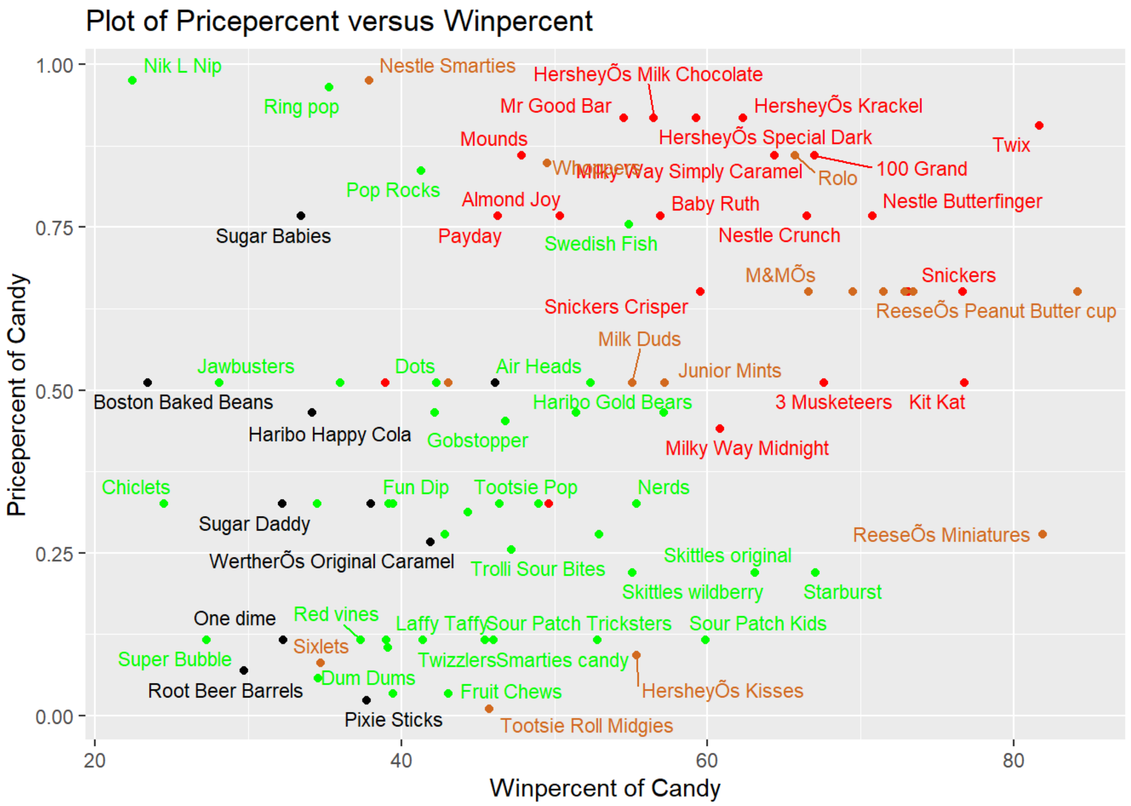
> **Ans18: Starburst**

4. Taking a look at pricepercent

Note: Install the "ggrepel" package first by using 'install.packages("ggrepel")' function

```
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, max.overlaps = 9) +
  labs(title = "Plot of Pricepercent versus Winpercent",
       x = "Winpercent of Candy", y = "Pricepercent of Candy")
```

```
Warning: ggrepel: 22 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?**

> **Ans19: Fruity candy type**

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

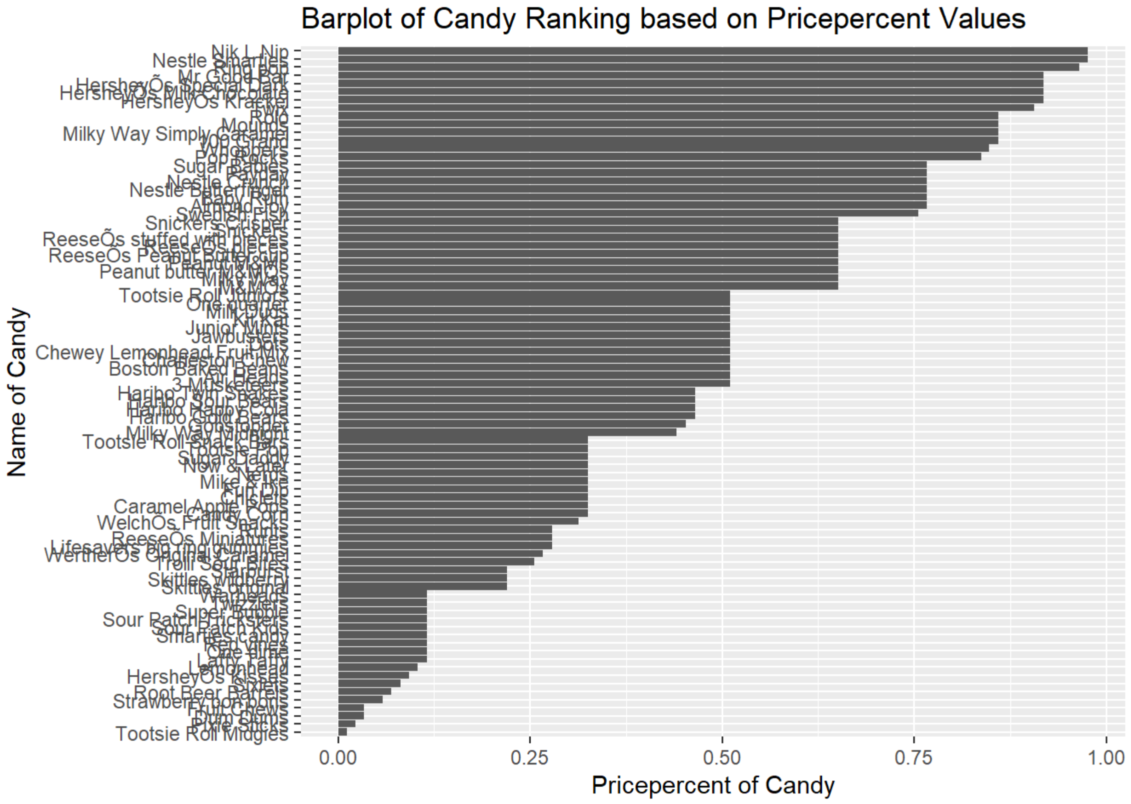
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

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

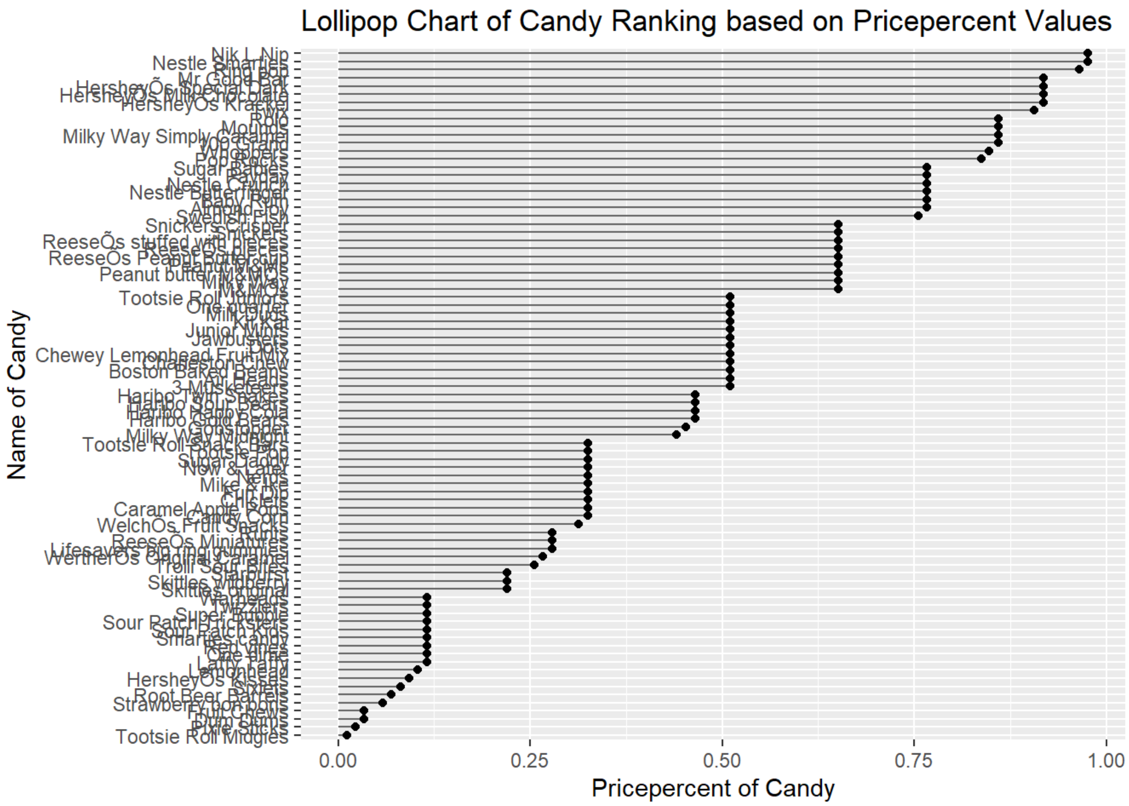
Optional

> **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 finally making 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_col() +
  labs(title = "Barplot of Candy Ranking based on Pricepercent Values",
       x = "Pricepercent of Candy", y = "Name of Candy")
```



```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
  geom_point() +
  labs(title = "Lollipop Chart of Candy Ranking based on Pricepercent Values",
       x = "Pricepercent of Candy", y = "Name of Candy")
```



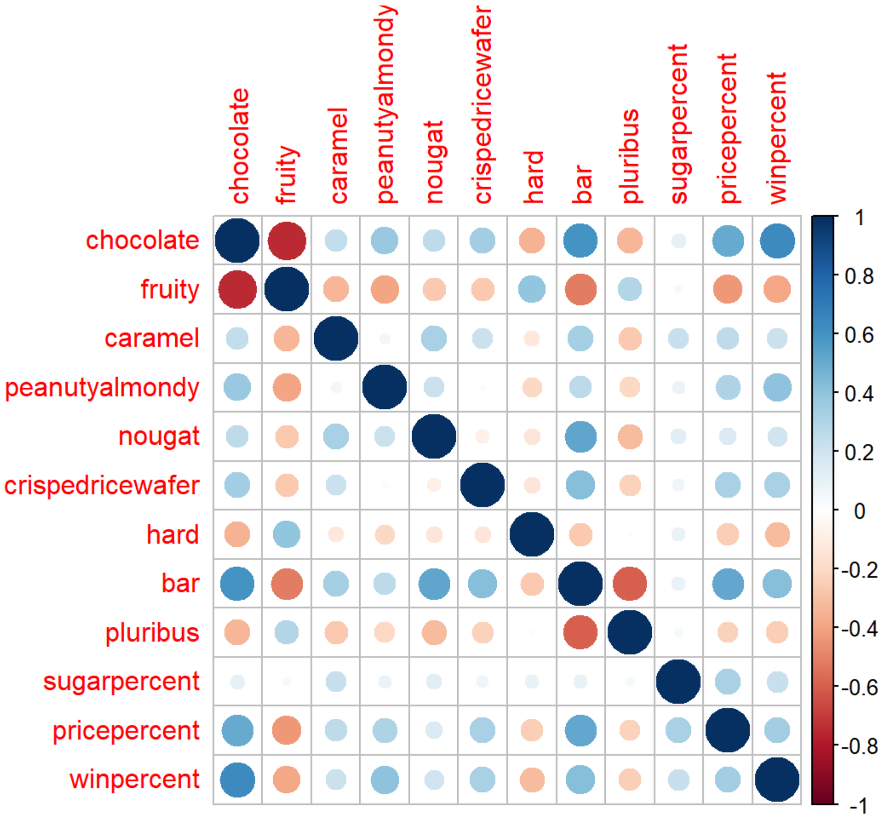
5. Exploring the correlation structure

Note: Install the "corrplot" package first by using 'install.packages("corrplot")' function

```
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)?**

> **Ans22: Chocolate and Fruity are anti-correlated.**

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

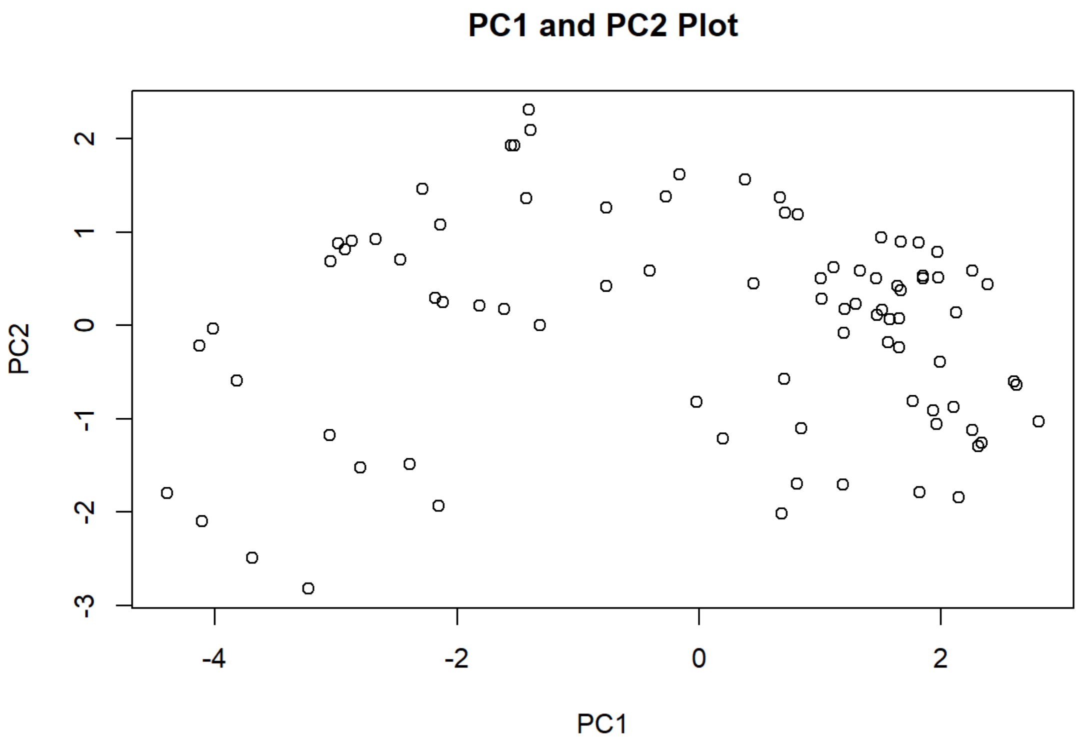
> **Ans23: Chocolate and Bar (or Chocolate and Winpercent) are most positively correlated**

6. Principal Component Analysis

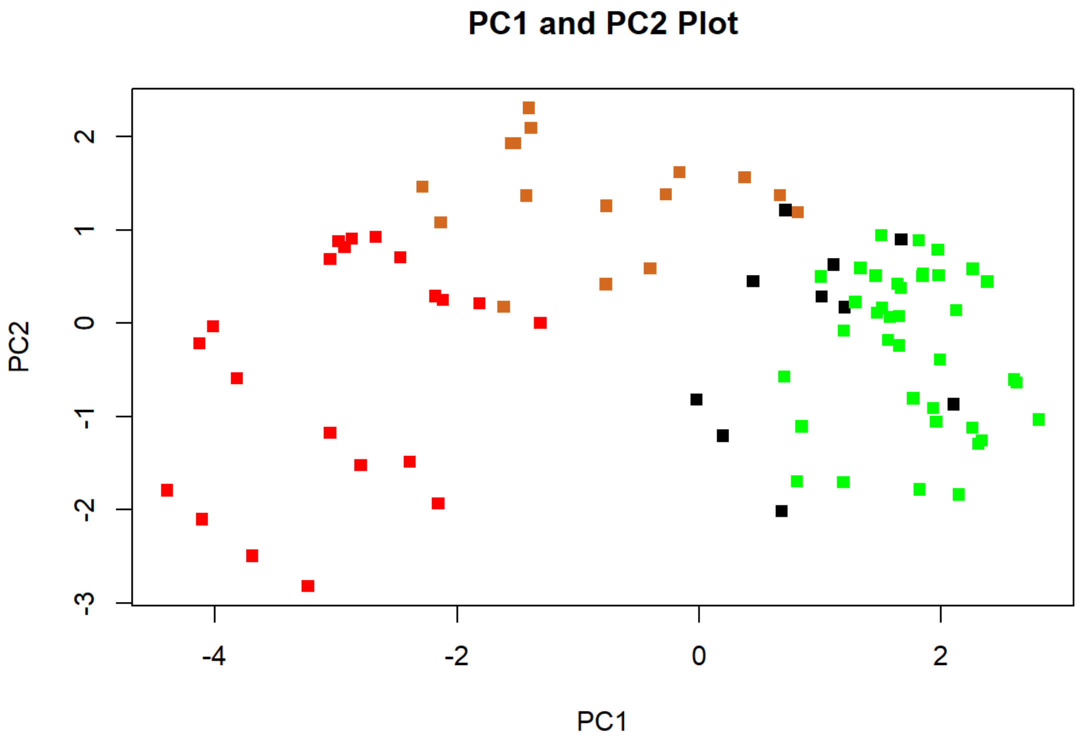
```
pca <- prcomp(candy, scale = TRUE)
summary(pca)
```

```
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$x[,1], pca$x[,2],
     main = "PC1 and PC2 Plot", xlab = "PC1", ylab = "PC2")
```

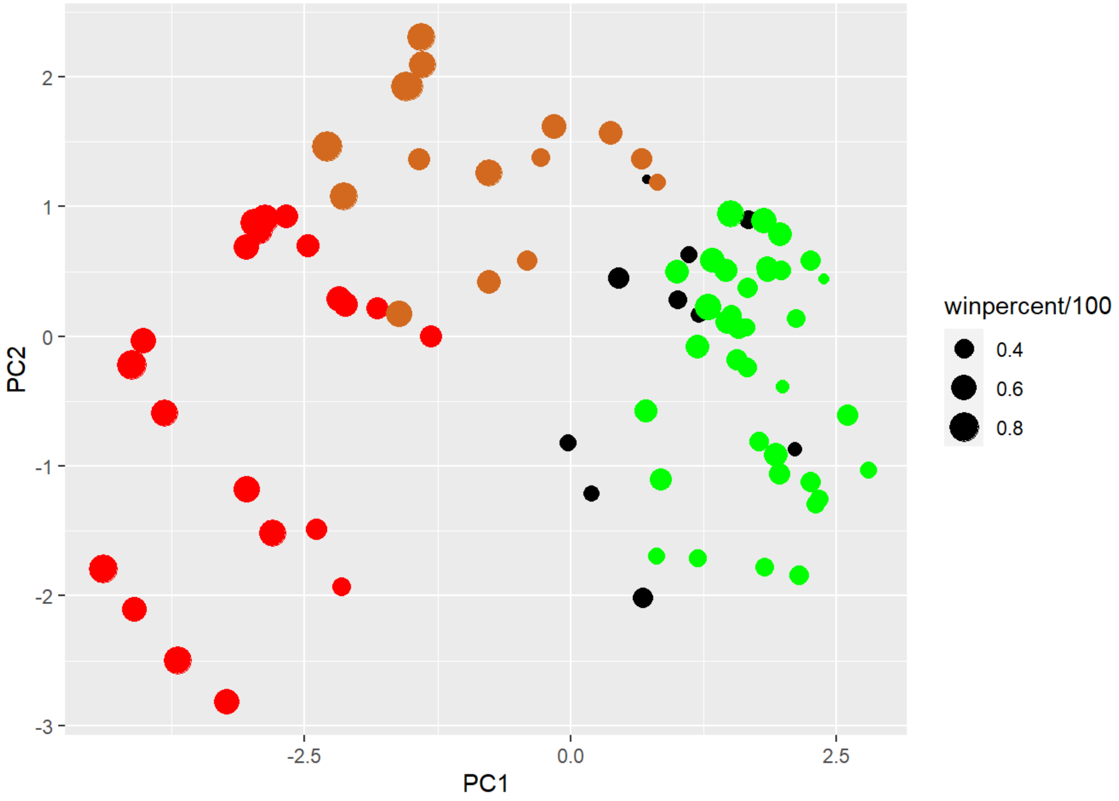


```
plot(pca$x[,1:2], col = my_cols, pch = 15,
     main = "PC1 and PC2 Plot", xlab = "PC1", ylab = "PC2")
```



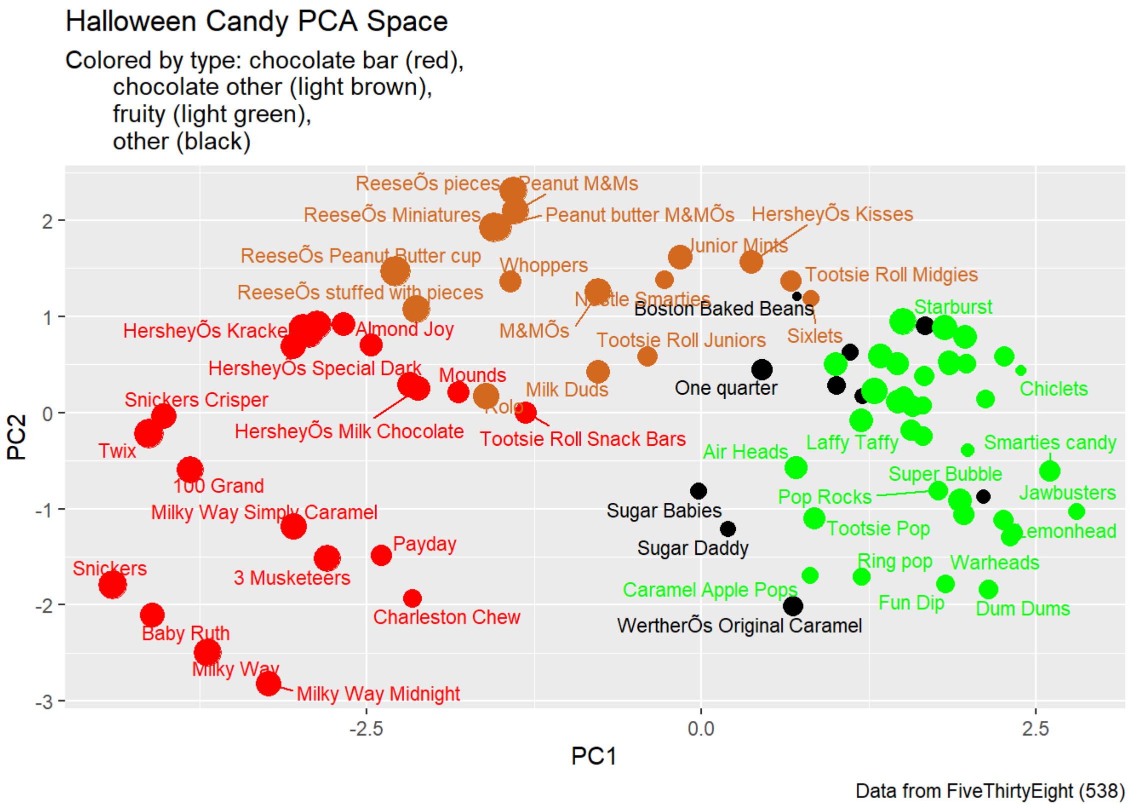
```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

```
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)

p
```



```
library(ggrepel)

p + geom_text_repel(size = 3, col = my_cols, max.overlaps = 9)  +
  theme(legend.position = "none") +
  labs(title = "Halloween Candy PCA Space",
       subtitle = "Colored by type: chocolate bar (red),
       chocolate other (light brown),
       fruity (light green),
       other (black)",
       caption = "Data from FiveThirtyEight (538)")
```

```
Warning: ggrepel: 32 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```


Halloween Candy PCA Space

Colored by type: chocolate bar (red),
chocolate other (light brown),
fruity (light green),
other (black)

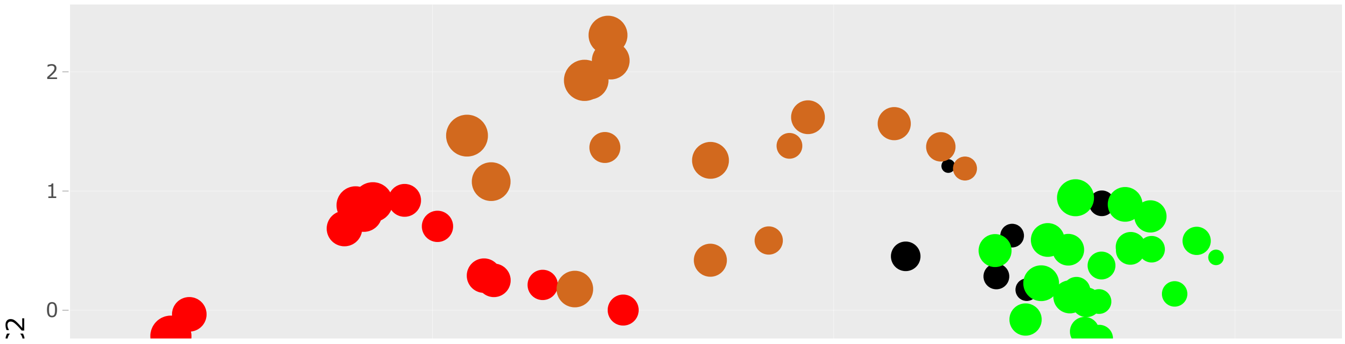
Data from FiveThirtyEight (538)

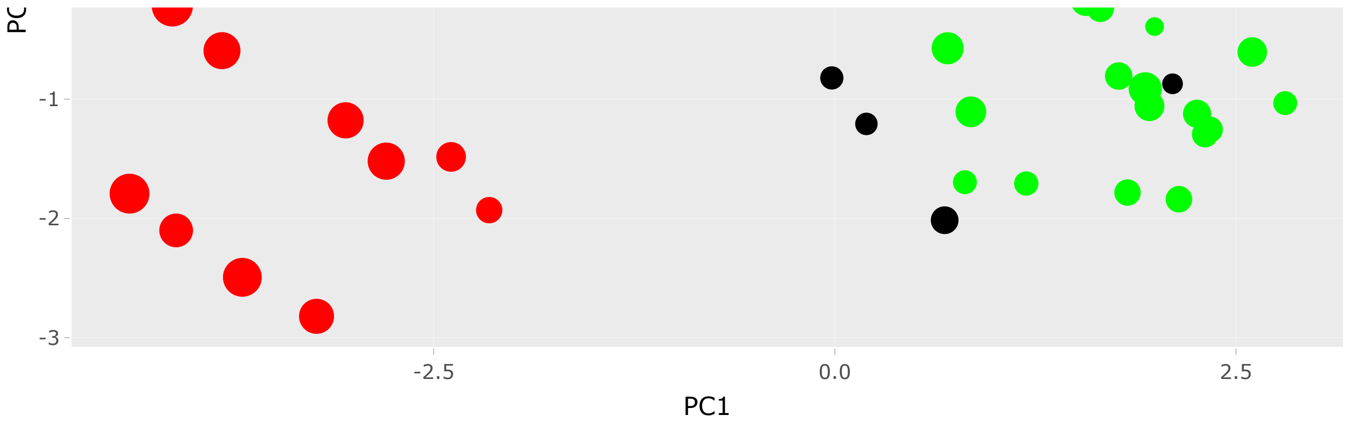
Note: Install "plotly" package first by using 'install.packages("plotly")' function

```
library(plotly)
```

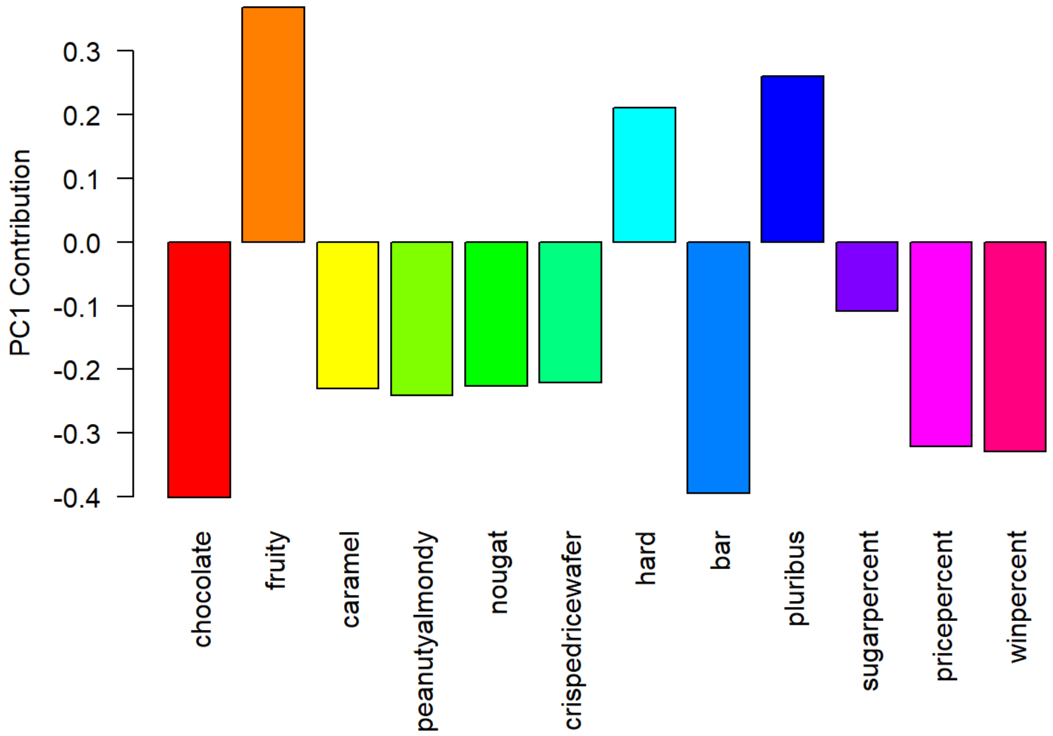
```

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", col = rainbow(12))
```



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

> **Ans24: Fruity, Hard, and Pluribus are picked up strongly by PC1 in the positive direction. These make sense since they are positive correlations, fruity candies are usually hard, and they are usually set in a bag or a box of multiple fruity candy flavors.**

Comment: Since I used the "plotly" package, which only works in HTML format, I could not render in PDF format. Thus, I rendered it in HTML format and then printed it in PDF format.