Class 10: Halloween Mini-Project

Kira Jung

1. Importing candy data

Getting data from the FiveThirtyEight GitHub repo.

```
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
```

Question 1 How many different candy types are in the dataset?

```
nrow(candy)
```

```
[1] 85
```

Answer: there are 85 types of candy.

Question 2 How many fruity candy types are in the dataset?

```
table(candy$fruity)
```

```
 0  1
47 38
```

Answer: there are 38 fruity candy types.

2. What is your favorite candy?

Finding the percentage of people who prefer Twix over another randomly chosen candy from the dataset:

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

```
[1] 81.64291
```

Question 3 What is your favorite candy in the dataset and what is it's winpercent value?

```
candy["Haribo Gold Bears", ]$winpercent
```

```
[1] 57.11974
```

Answer: my favorite candy is Haribo Gold Bears and its winpercent value is 57.1%.

Question 4: What is the winpercent value for "Kit Kat"?

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

```
[1] 76.7686
```

Answer: the winpercent value for "Kit Kat" is 76.8%

Question 5: What is the winpercent value for "Tootsie Roll Snack Bars"?

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

```
[1] 49.6535
```

Answer: the winpercent value for Tootsie Roll Snack Bars is 49.7%.

Trying the skimr package on candy data.

First installed the skimr package. Now loading it:

```
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
winpercent	0	1	50.32	14.71	22.45	39.14	47.83	59.86	84.18	

Question 6: Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

Answer: the winpercent column appears to be on a different scale.

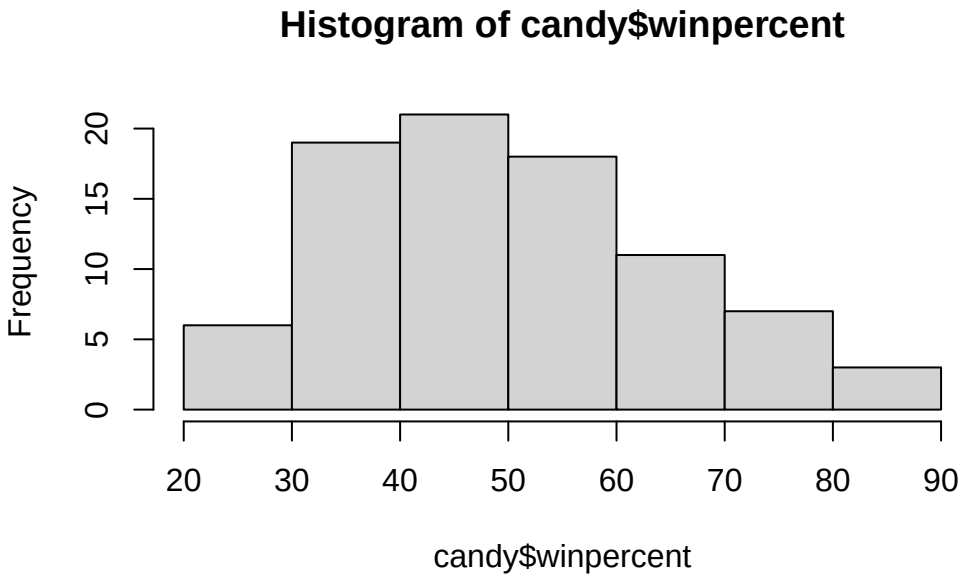
Question 7: What do you think a zero and one represent for the candy$chocolate column?

Answer: A zero value means the given candy is not a chocolate, and a value of one means the candy is a chocolate.

Making a histogram of winpercent values.

Question 8: Plot a histogram of winpercent values.

```
hist(candy$winpercent)
```



Question 9: Is the distribution of winpercent values symmetrical?

Answer: No, the distribution of winpercent values is skewed right.

Question 10: Is the center of the distribution above or below 50%?

Answer: The center of distribution is below 50% (it's between 40% to 50%).

Question 11: On average is chocolate candy higher or lower ranked than fruit candy?

```
inds <- as.logical(candy$chocolate)
chocolate <- candy[inds, ]$winpercent
inds.fruit <- as.logical(candy$fruity)
fruity <- candy[inds.fruit, ]$winpercent
```

```
mean(chocolate)
```

```
[1] 60.92153
```

```
mean(fruity)
```

```
[1] 44.11974
```

Answer: on average, chocolate candy is higher ranked than fruity candy.

Question 12: Is this difference statistically significant?

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

```
    Welch Two Sample t-test

data:  chocolate and fruity
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
```

Answer: the p-value is well below 0.05, so the difference is statistically significant.

3. Overall Candy Rankings

Question 13: 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
```

Answer: Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters are the 5 least liked candy types.

Question 14: What are the top 5 all time favorite candy types out of this set?

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

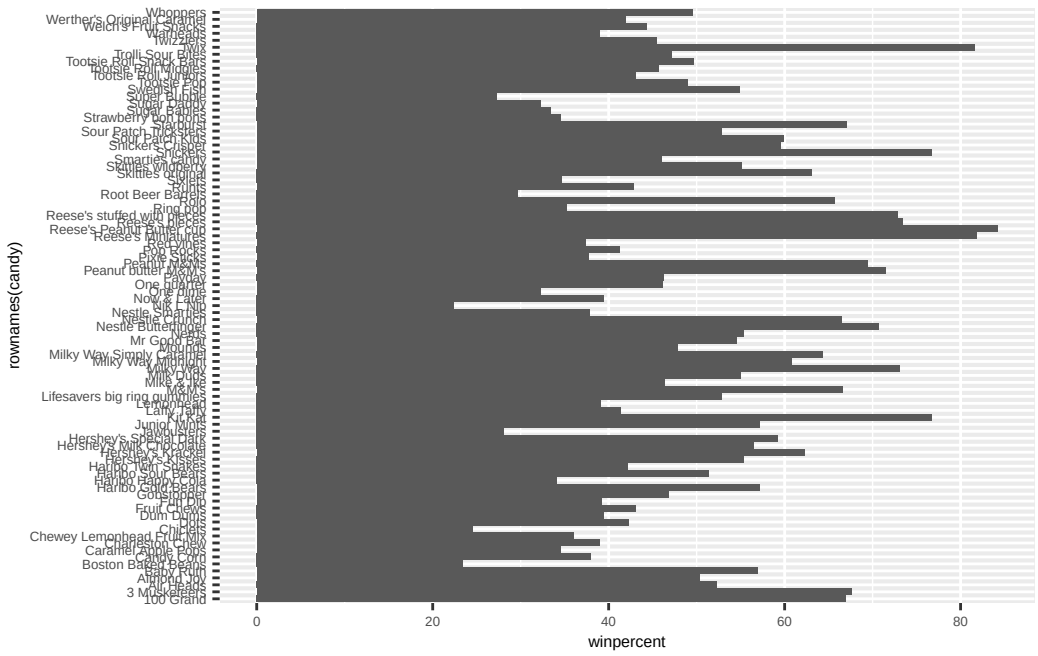
```
                        chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat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

Answer: the top 5 all time favorite candy types are Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat, and Snickers.

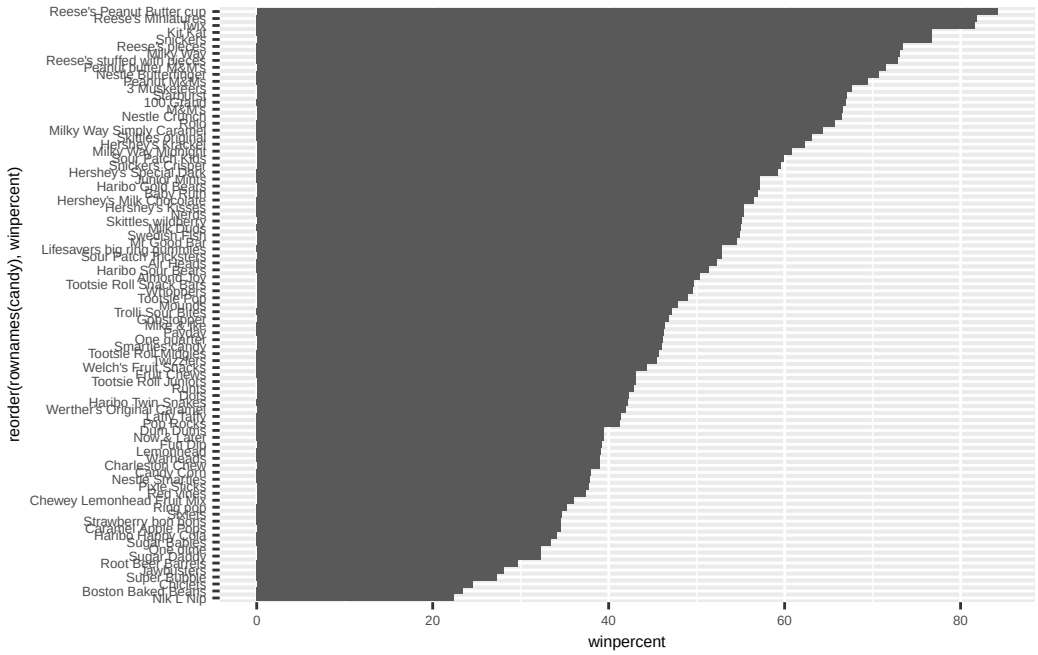
Question 15: Make a first barplot of candy ranking based on winpercent values.

```
library("ggplot2")

ggplot(candy) + aes(winpercent, rownames(candy)) + geom_col() + theme(text = element_text(
```



Question 16: 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() + theme
```



Time to add some useful color

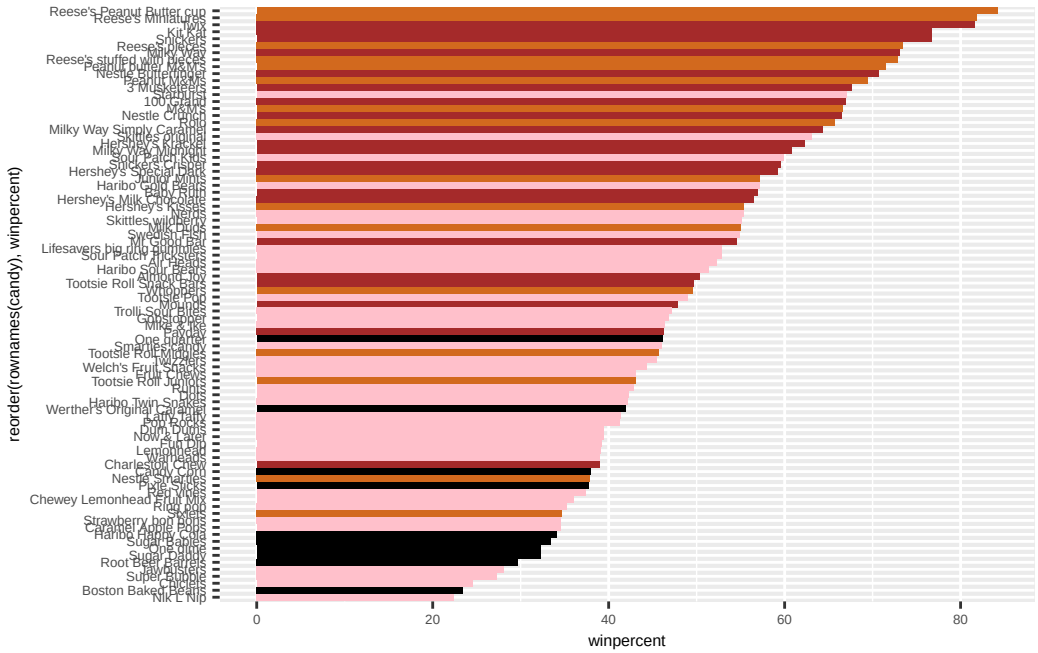
Setting up a color vector (that signifies candy type) to be used for future plots.

```
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"
```

Chocolate is for chocolate, brown is for candy bars, and pink is for fruity candy. All other candies are black.

Now making a barplot using these colors:

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



Question 17: What is the worst ranked chocolate candy?

Answer: the worst ranked chocolate candy is Sixlets.

Question 18: What is the best ranked fruity candy?

Answer: the best ranked fruity candy is Starburst.

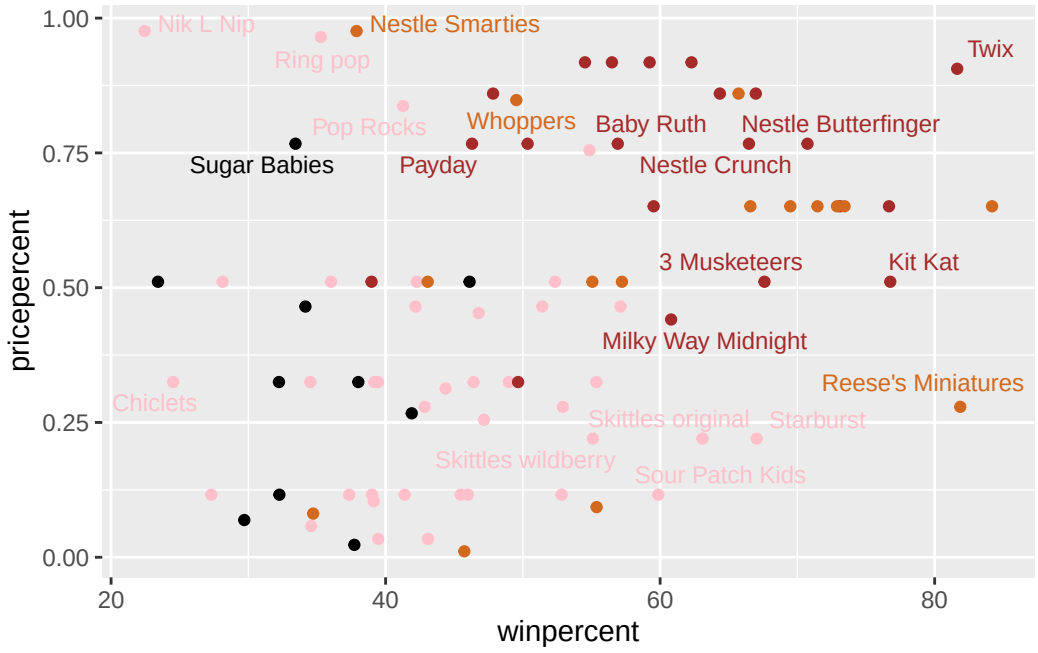
4. Taking a look at pricepercent

Pricepercent variable records the percentile rank of the candy's price against all other candies.

We can plot winpercent vs pricepercent variables to assess value for money.

```
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 = 5)
```

```
Warning: ggrepel: 65 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Question 19: Which candy type is the highest ranked in terms of winpercent for the least money - i.e. offers the most bang for your buck?

Answer: Reese's Miniatures have the highest winpercent ranking for the least money.

Question 20: What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

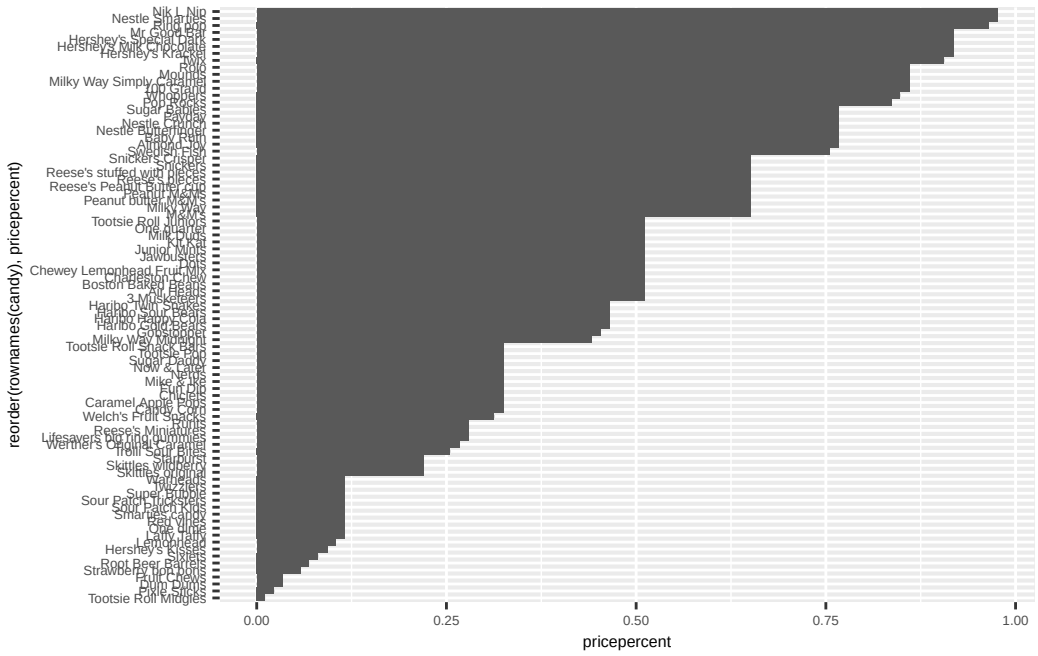
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head(candy[ord,c(11,12)], n=5 )
```

```
                      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
```

Answer: the top 5 most expensive candies are Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate. Out of these, the least popular candy is Nik L Nip (winpercent is the lowest, 22.4%).

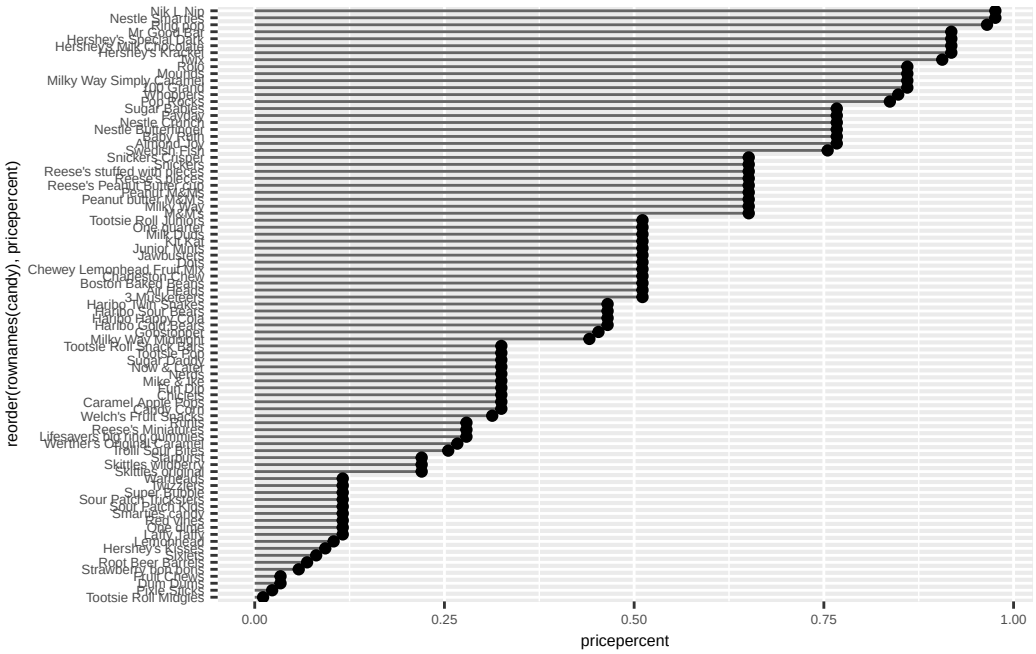
Question 21: Make a barplot again with geom_col() this time using pricepercent and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping geom_col() for geom_point() + geom_segment().

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



Now turning this into a lollipop chart:

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



5. Exploring the correlation structure

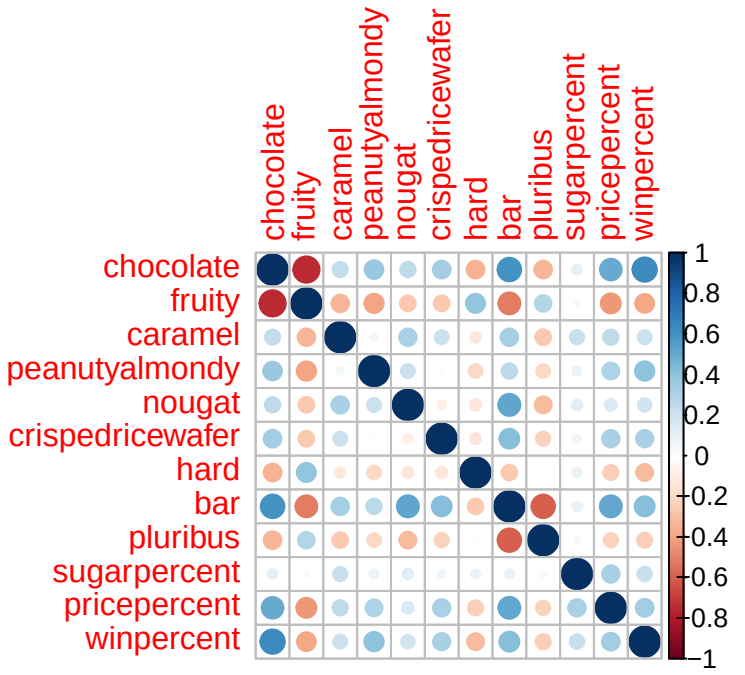
Now seeing how the variables interact with one another. Using the correlation and viewing the results with the corrplot package.

```
library(corrplot)
```

```
corrplot 0.92 loaded
```

Now plotting a correlation matrix:

```
cij <- cor(candy)
corrplot(cij)
```



Question 22: Examining this plot what two variables are anti-correlated (i.e. have minus values)?

Answer: Chocolate & fruity, pluribus & bar, fruity & bar, fruity & pricepercent, winpercent & fruity are some examples that are anti-correlated.

Question 23: Similarly, what two variables are most positively correlated?

Answer: chocolate & winpercent are the most positively correlated.

6. Principal Component Analysis

Applying PCA using the prcom() function to our candy dataset.

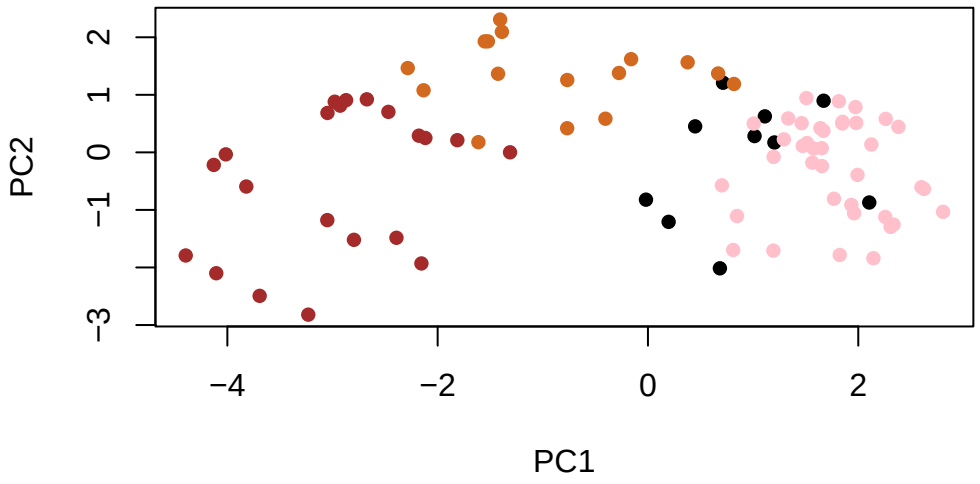
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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
```

Now plotting the main PCA score plot of PC1 vs. PC2 with color:

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

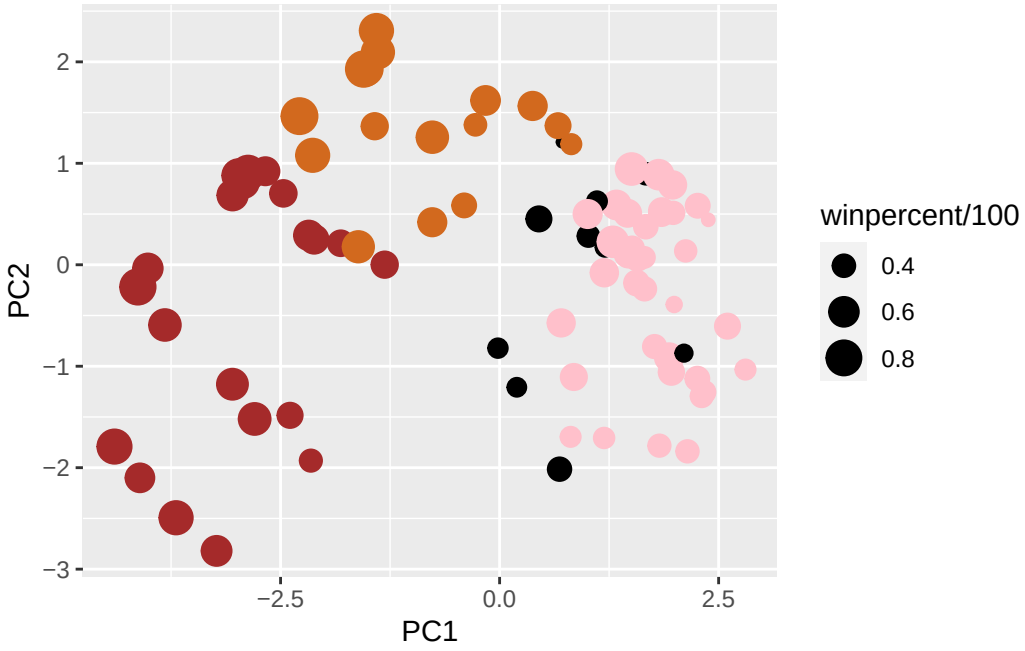


Making a new data frame with the PCA results and candy data:

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

Now using this data frame to create a scatterplot:

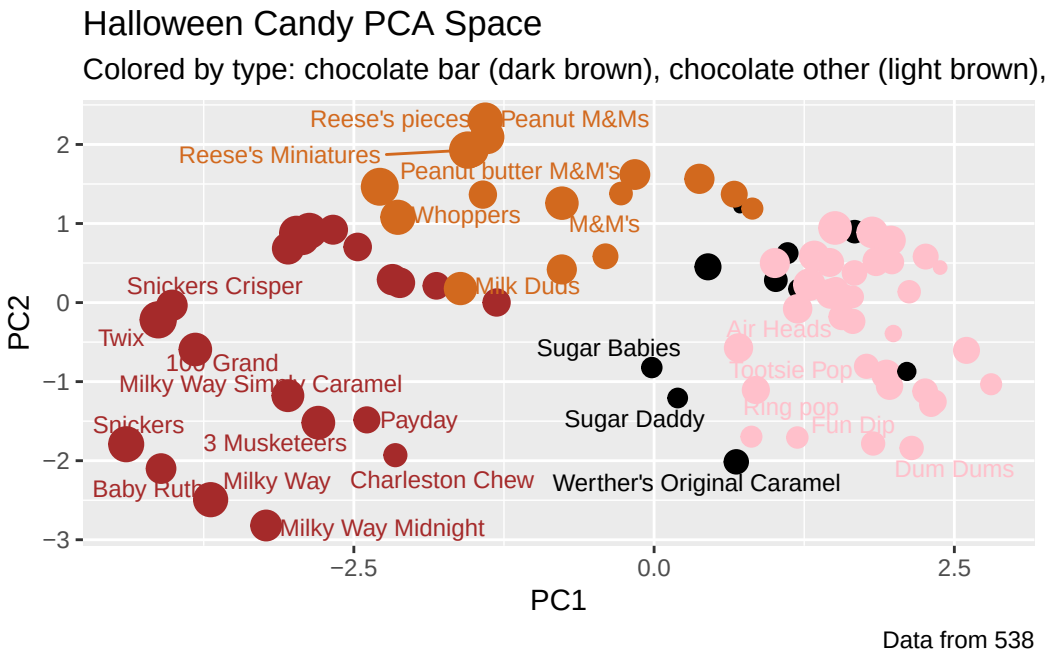
```
p <- ggplot(my_data) + aes(x=PC1, y=PC2,
size=winpercent/100, text=rownames(my_data), label=rownames(my_data)) + geom_point(col=my_

p
```



Using the ggrepel package and the function `ggrepel::geom_text_repel()` to label the plot with non overlapping candy names. Also adding a title and subtitle.

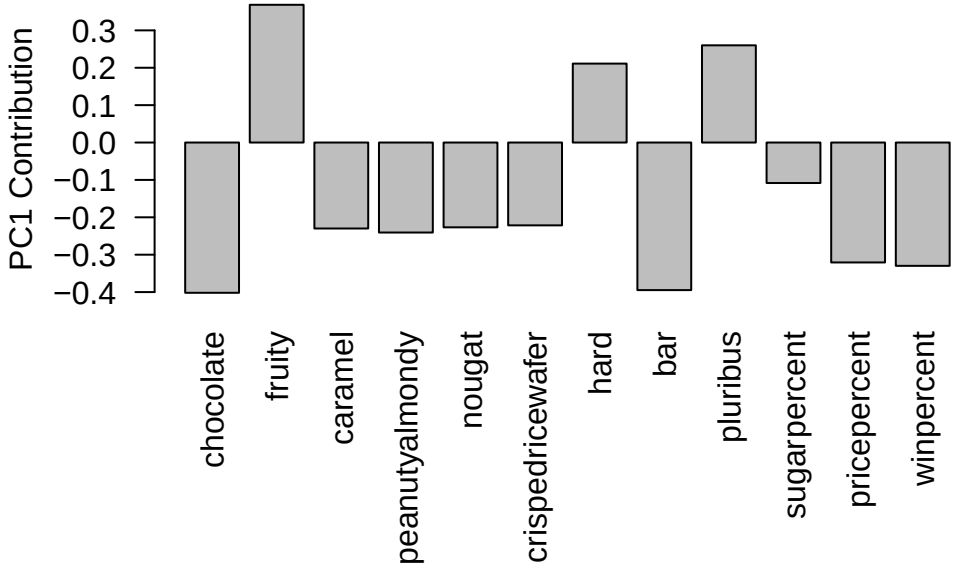
```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  + theme(legend.position = "n
```

```
Warning: ggrepel: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Finishing by looking at a barplot of PCA:

```
par(mar=c(8,4,2,2))
barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```



Question 24: What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

Answer: fruity, hard, and pluribus are picked up strongly by PC1 in the positive direction. This makes sense since these candies are up towards the positive x-axis (PC1 Contribution). This is clearly shown in the corrplot. Additionally, they are more unique compared to other candies such as chocolate or bars.

```
sessionInfo()
```

```
R version 4.2.3 (2023-03-15 ucrt)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 22621)

Matrix products: default

locale:
[1] LC_COLLATE=English_United States.utf8
[2] LC_CTYPE=English_United States.utf8
[3] LC_MONETARY=English_United States.utf8
[4] LC_NUMERIC=C
[5] LC_TIME=English_United States.utf8
```

```
attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
[1] corrplot_0.92 ggrepel_0.9.3 ggplot2_3.4.2 skimr_2.1.5

loaded via a namespace (and not attached):
 [1] Rcpp_1.0.10      pillar_1.9.0     compiler_4.2.3   base64enc_0.1-3
 [5] tools_4.2.3      digest_0.6.31    jsonlite_1.8.4   evaluate_0.21
 [9] lifecycle_1.0.3  tibble_3.2.1     gtable_0.3.3     pkgconfig_2.0.3
[13] rlang_1.1.0      cli_3.6.1        rstudioapi_0.14  yaml_2.3.7
[17] xfun_0.39        fastmap_1.1.1    repr_1.1.6       withr_2.5.0
[21] dplyr_1.1.2      stringr_1.5.0    knitr_1.43       generics_0.1.3
[25] vctrs_0.6.2      grid_4.2.3       tidyselect_1.2.0 glue_1.6.2
[29] R6_2.5.1         fansi_1.0.4      rmarkdown_2.21   farver_2.1.1
[33] purrr_1.0.1      tidyr_1.3.0      magrittr_2.0.3   scales_1.2.1
[37] htmltools_0.5.5  colorspace_2.1-0 labeling_0.4.2   utf8_1.2.3
[41] stringi_1.7.12   munsell_0.5.0
```