Class 9: Halloween Candy Mini-project

Zainab Fatima (PID: A16880407)

2025-02-04

Table of contents

Today we will examine data from 538 on common Halloween candy. In particular we will use ggplot, dplyr, and PCA to make sense of this multivariate dataset.

Importing Candy Data

```r
candy <- read.csv("https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-

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?

```r
nrow(candy)
```

```
[1] 85
```

85 different candy types

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

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

```
[1] 38
```

`Winpercent` is the value is the percentage of people who prefer this candy over another randomly chosen candy from the dataset.

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

```r
candy["Milky Way", ]$winpercent
```

```
[1] 73.09956
```

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

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

Class Question. How many chocolate candy are there in the dataset?

```r
sum(candy$chocolate)
```

```
[1] 37
```

Side note: the `skimr::skim()` function is useful for giving a summary of the dataset

```r
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?

The winpercent is different from the other variables because it is not on a 0 to 1 scale and is instead on a 0% to 100% scale. We will need to scale this dataset before analysis like PCA.

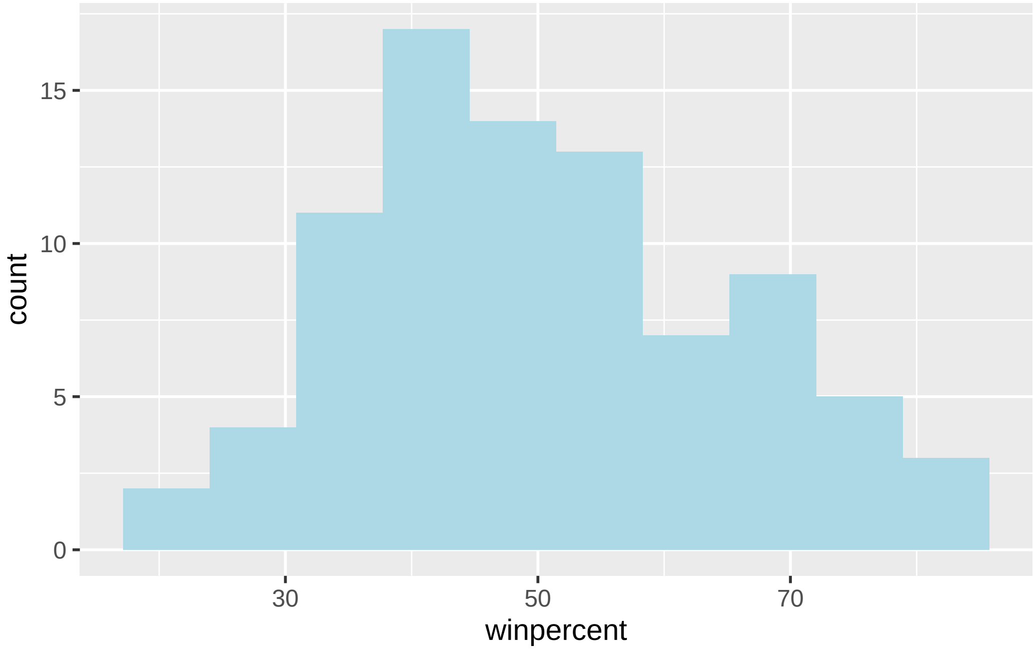
Q7. What do you think a zero and one represent for the candy$chocolate column?

The 0 shows if the candy is not chocolate, the 1 shows if the candy is chocolate.

Histogram: The function `hist()` or `ggplot() with geom_hist()`make histograms.

Q8. Plot a histogram of winpercent values

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



Note: The more bins you have, the more "spiky" the data gets and the less useful it gets

Q9. Is the distribution of winpercent values symmetrical?

No, the graph does not appear to be symmetrical.

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

```
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.14   47.83   50.32   59.86   84.18 
```

The center of distribution is below 50%.

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

Answer: Chocolate candy is highher ranked than fruit candy. Code listed below

- Step 1. Find all "chocolate" candy
- Step 2: Find their "winpercent" values
- Step 3: Summarize these values
- Step 4: Find all "fruity" candy
- Step 5: Find their "winpercent" values
- Step 6: Summarize these values
- Step 7: Compare the two summary values

1. Find all chocolate candies

```
choc.inds <- candy$chocolate == 1
#candy[choc.inds,] gives the table of chocolate vs fruit candies
```

2. Find the "winpercent" values for chocolate

```
choc.win <- candy[choc.inds,]$winpercent
```

Step 3. Summarize these winpercents for chocolate

```
choc.mean <- mean(choc.win)
#mean of winpercent is 60.9 for chocolate
choc.mean
```

```
[1] 60.92153
```

Step 4. all fruity candies

```r
fruit.inds <- candy$fruity == 1
#candy[fruit.inds,]
```

5. Find the “winpercent” values for fruity

```r
fruit.win <- candy[fruit.inds,]$winpercent
```

6. Summarize ‘winpercent’ findings for fruit

```r
fruit.mean <- mean(fruit.win)
fruit.mean
```

```
[1] 44.11974
```

7. Compare the two summary values

Clearly chocolate has a higher mean winpercent than fruit candy

```r
choc.mean
```

```
[1] 60.92153
```

```r
fruit.mean
```

```
[1] 44.11974
```

Q12. Is this difference statistically significant?

```r
t.test(choc.win, fruit.win)
```

```

    Welch Two Sample t-test

data:  choc.win and fruit.win
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 t-test above shows that the means are not equal and has a low p-value, which means that the difference is statistically significant.

This shows that people prefer chocolate candy over fruity candy.

Overall Candy Rankings

```
#sort() is not the useful, it just sorts the values
#sort(candy$winpercent)

#order() is more useful
#order() returns the rankings of each elements of the vector
#x[order(x)]
```

The `order()` function tells us how to arrange the elements of the input to make them sorted - i.e. how to order them

We can determine the order of winpercent to make them sorted and use that order to arrange the whole dateset.

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

```
ord.inds <- order(candy$winpercent)
ord.inds
```

```
 [1] 45  8 13 73 27 58 72  3 71 20 10 70 60 56 12 51 49 63  9 11 82 31 17 46 15
[26] 50 30 84 22 14 59 76 16 83 81 77 64  4 47 35 18 79 40 75 85 78  6 21  5 68
[51] 32 41 74 36 62 42 23 25  7 19 28 26 66 67 38 24 61 39 57 44 34  1 69  2 48
[76] 43 33 55 37 54 65 29 80 52 53
```

```
head(candy[ord.inds, ])
```

```
                   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
Root Beer Barrels          0      0       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
Root Beer Barrels                 0    1   0        1        0.732        0.069
                   winpercent
Nik L Nip            22.44534
Boston Baked Beans   23.41782
Chiclets             24.52499
Super Bubble         27.30386
Jawbusters           28.12744
Root Beer Barrels    29.70369
```

These are the 6 least liked candies in the dataset (top of ordedred list).

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

```r
tail(candy[ord.inds, ])
```

```
                          chocolate fruity caramel peanutyalmondy nougat
Reese's pieces                    1      0       0              1      0
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
Reese's pieces                           0    0   0        1        0.406
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
Reese's pieces                   0.651   73.43499
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
```

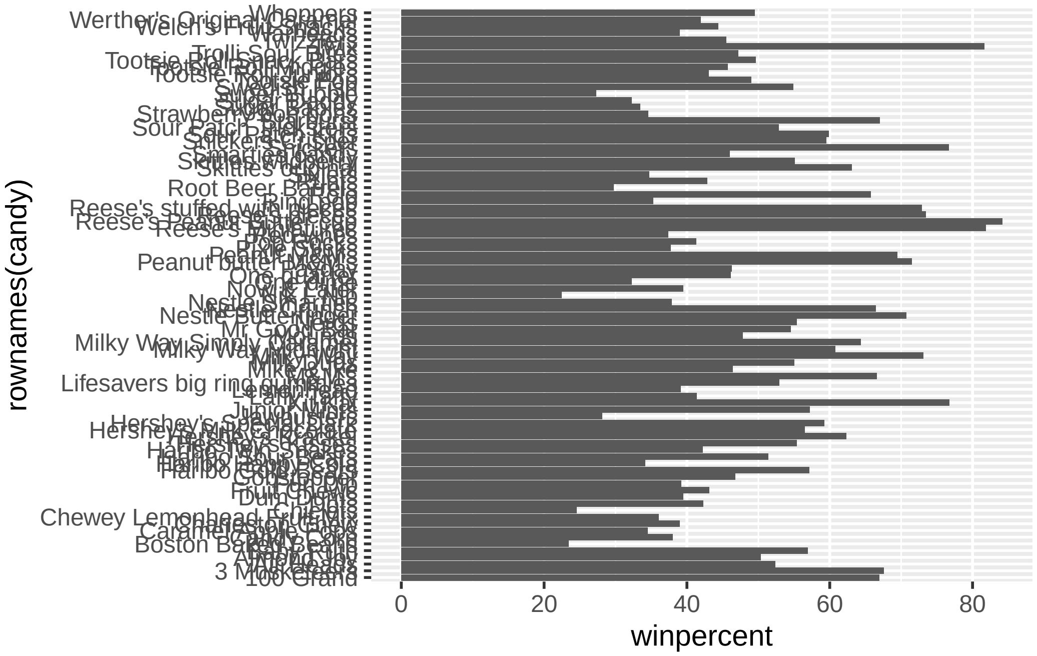
These are the 6 most liked candies (bottom of ordered list)

Note: Adding the decreasing = T argument to order can move the order so that the top are first, then to find most liked candies I could use head

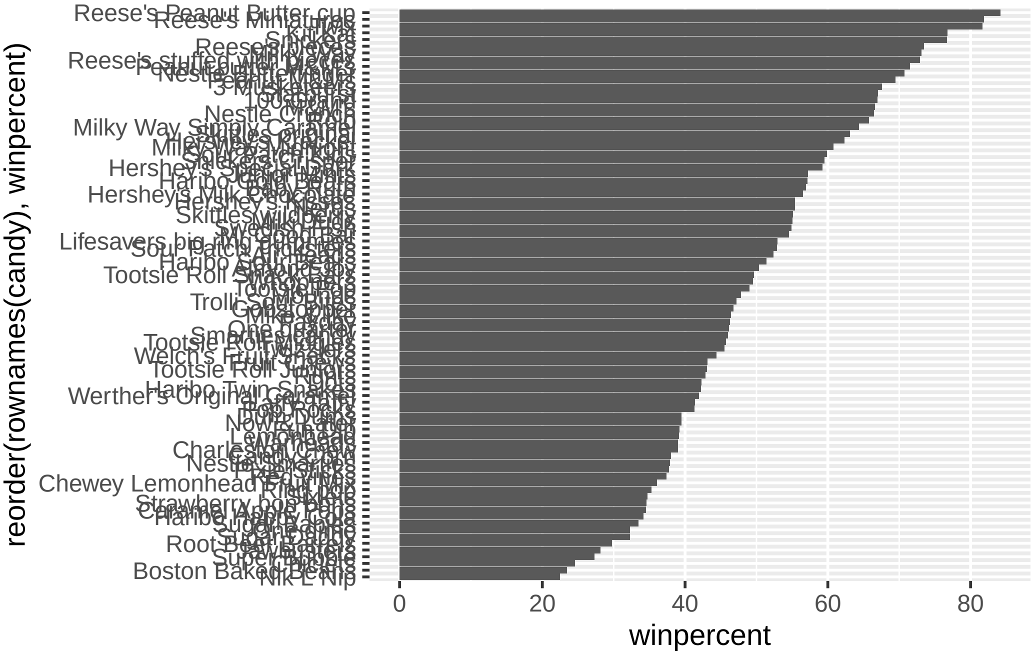
> Q15. Make a first barplot of candy ranking based on winpercent values.

Final barplot at the end.

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

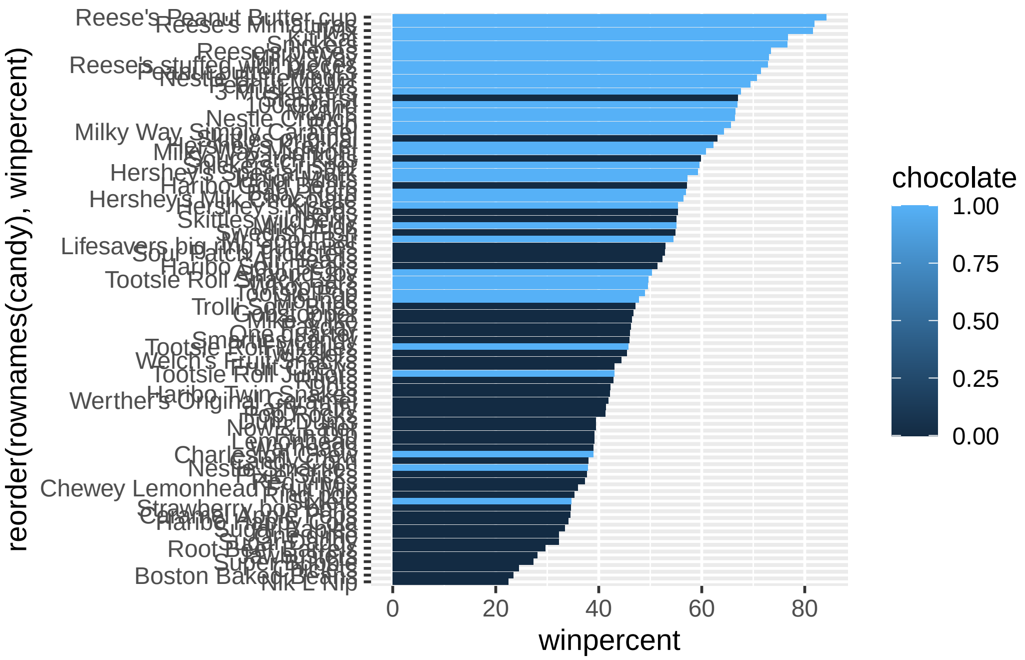


```
# Now we want to order bars by winpercent
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col()
```



Now, we can add some useful color to the plot

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



```
#not useful because it's not a color scale
```

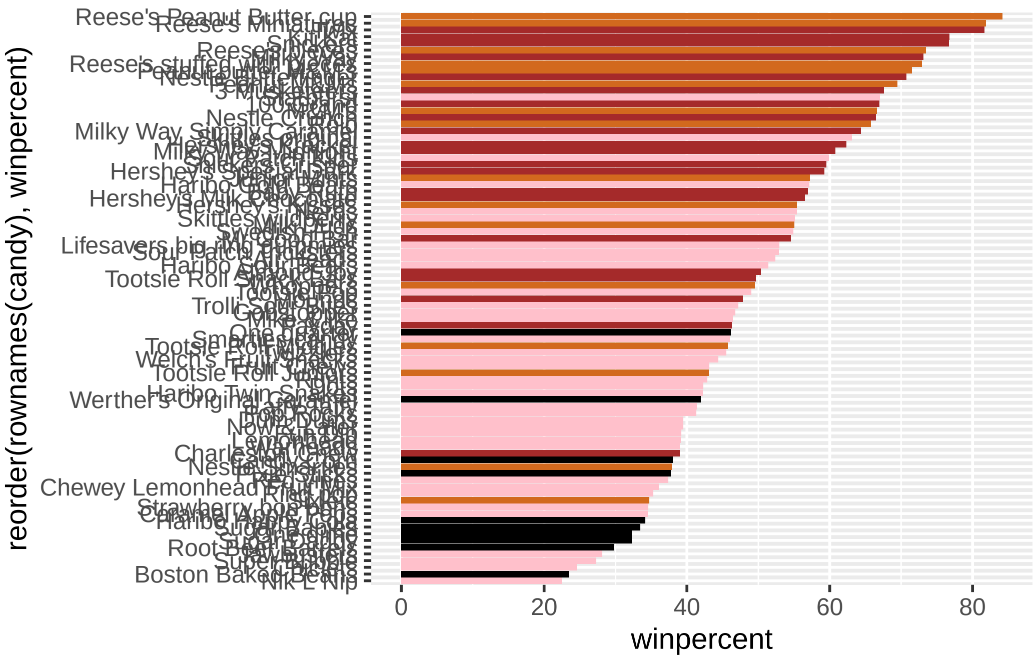
We need to make our own seperate color vector where we can spell out exactly what candy is colored a particular color

```
mycols <- rep("black", nrow(candy))
mycols[candy$chocolate == 1] <- "chocolate"
mycols[candy$bar == 1] <- "brown"
mycols[candy$fruity == 1] <- "pink"
mycols
```

```
 [1] "brown"     "brown"     "black"     "black"     "pink"      "brown"
 [7] "brown"     "black"     "black"     "pink"      "brown"     "pink"
[13] "pink"      "pink"      "pink"      "pink"      "pink"      "pink"
[19] "pink"      "black"     "pink"      "pink"      "chocolate" "brown"
[25] "brown"     "brown"     "pink"      "chocolate" "brown"     "pink"
[31] "pink"      "pink"      "chocolate" "chocolate" "pink"      "chocolate"
[37] "brown"     "brown"     "brown"     "brown"     "brown"     "pink"
[43] "brown"     "brown"     "pink"      "pink"      "brown"     "chocolate"
[49] "black"     "pink"      "pink"      "chocolate" "chocolate" "chocolate"
[55] "chocolate" "pink"      "chocolate" "black"     "pink"      "chocolate"
[61] "pink"      "pink"      "chocolate" "pink"      "brown"     "brown"
```

```
[67] "pink"      "pink"      "pink"      "pink"      "black"     "black"
[73] "pink"      "pink"      "pink"      "chocolate" "chocolate" "brown"
[79] "pink"      "brown"     "pink"      "pink"      "pink"      "black"
[85] "chocolate"
```

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



Q17. What is the worst ranked chocolate candy?

Sixlets is the worst ranked chocolate

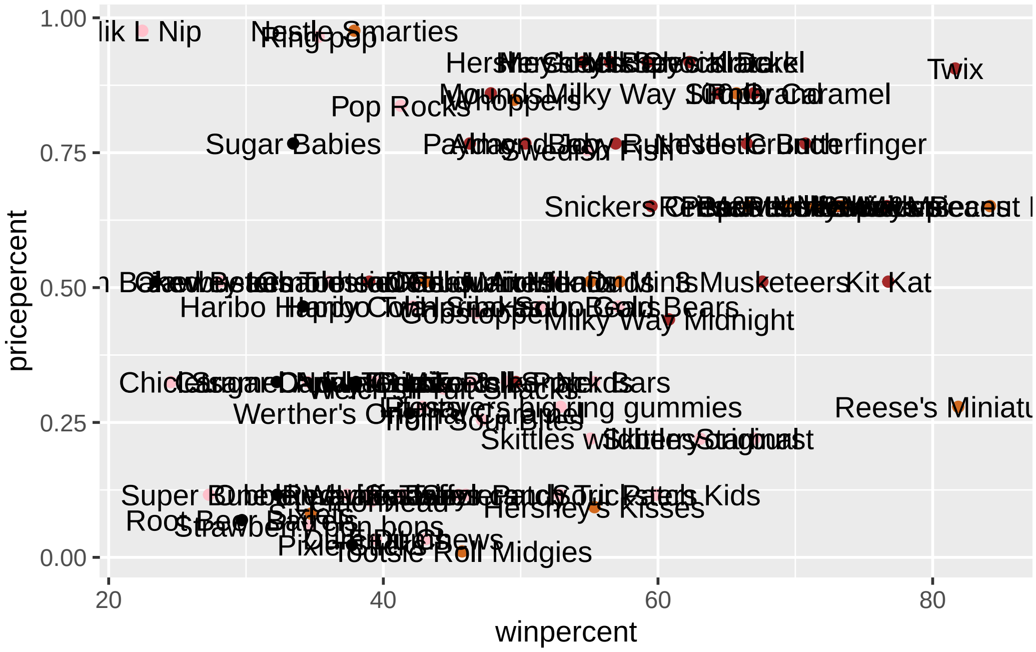
Q18. What is the best ranked fruity candy?

Starburst is the best ranked fruity candy

Taking a look at pricepercent

Make a plot of winpercent (x-axis) vs pricepercent (y-axis)

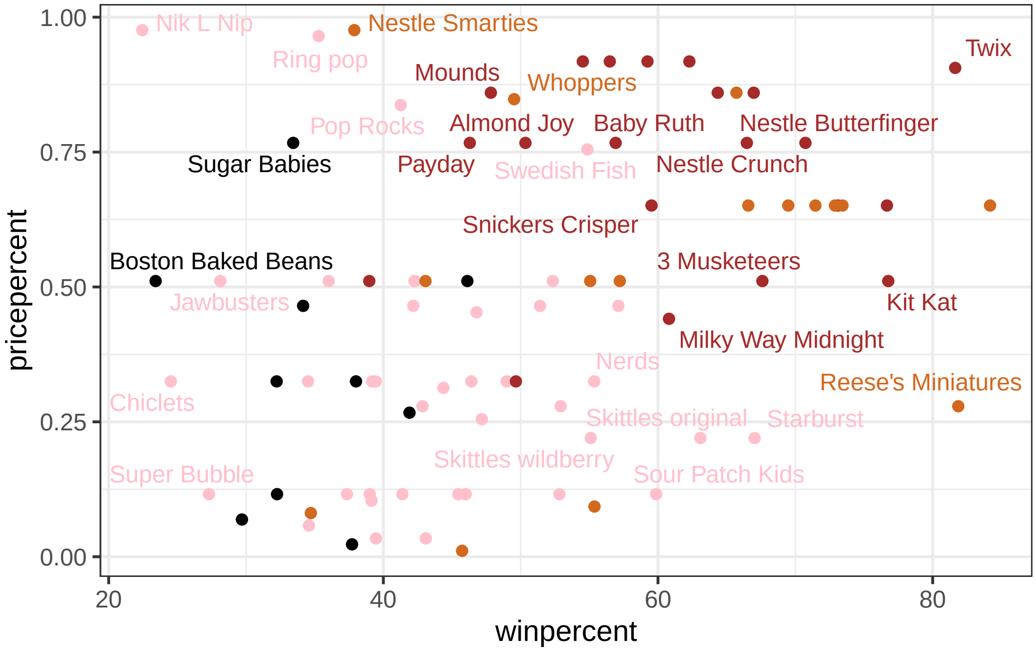
```
ggplot(candy) +
  aes(winpercent, pricepercent, label = rownames(candy)) +
  geom_point(col = mycols) +
  geom_text()
```



To avoid the overplotting of the text labels we can use the add-on package **ggrepel**

```
library(ggrepel)
ggplot(candy) +
  aes(winpercent, pricepercent, label = rownames(candy)) +
  geom_point(col = mycols) +
  geom_text_repel( col = mycols, size = 3.3, max.overlaps = 7) +
  theme_bw()
```

```
Warning: ggrepel: 57 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?

Reese's minatures

> 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 are:

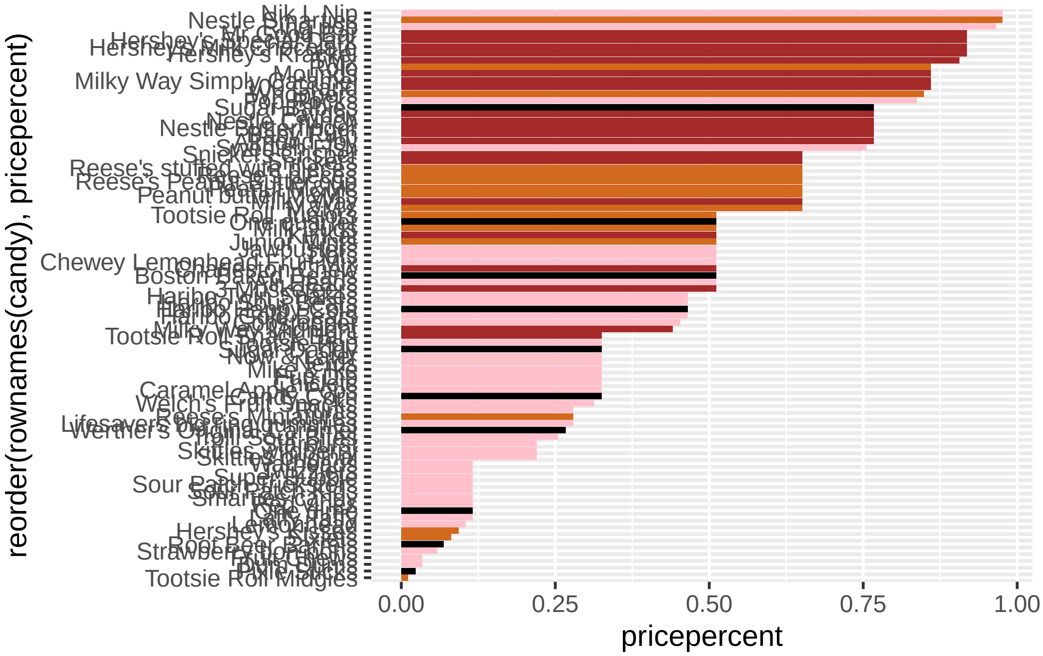
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

```
                         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
```

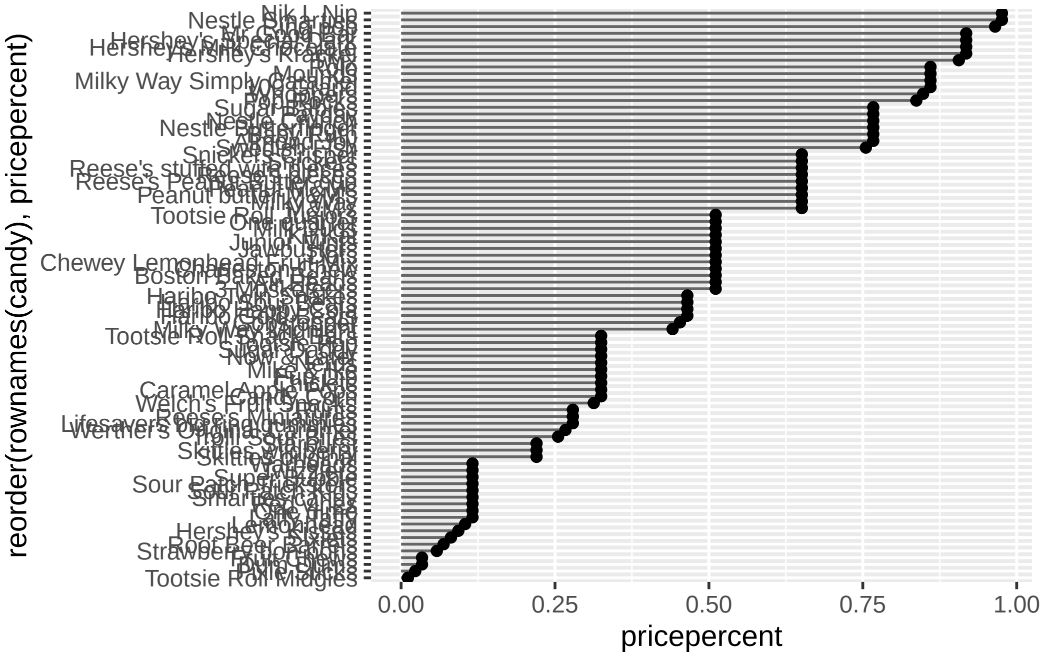
Based on the plot, the least popular is Nik N Lip.

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().

```r
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_col(fill = mycols)
```



```r
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

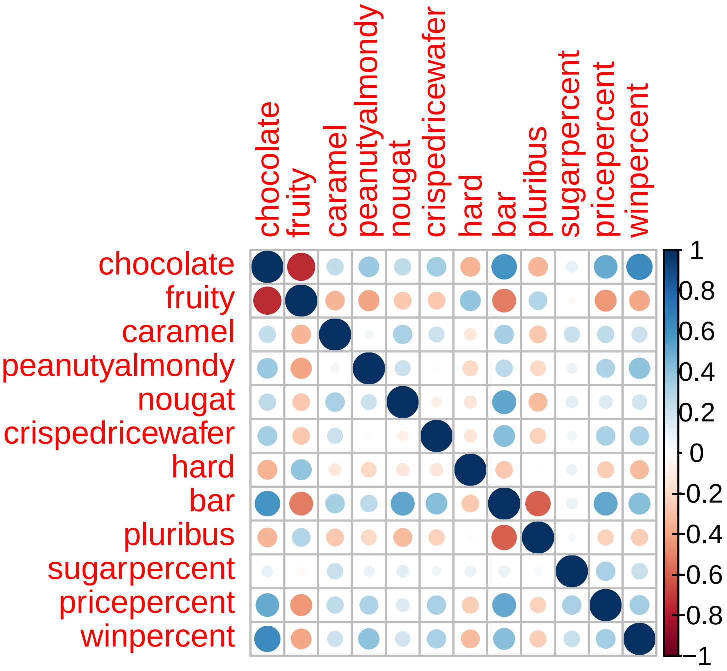
Now that we have explored the dataset a little, we will see how variables interact with one another.

First we will use correlation and view the results with the **corrplot** package to plot a correlation matrix.

```
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)?

Chocolate and fruity are anti-correlated with each other

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

Chocolate is most positively correlated with itself and fruit is most positively correlated with itself.

Chocolate is also positively correlated with caramel, peanut, nougat, bar, higher cost, and more popular.

Fruit is also positively correlated with hardness and pluribus.

Principal Component Analysis

We can apply PCA to the the `prcomp()` function to our **candy** data set.

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

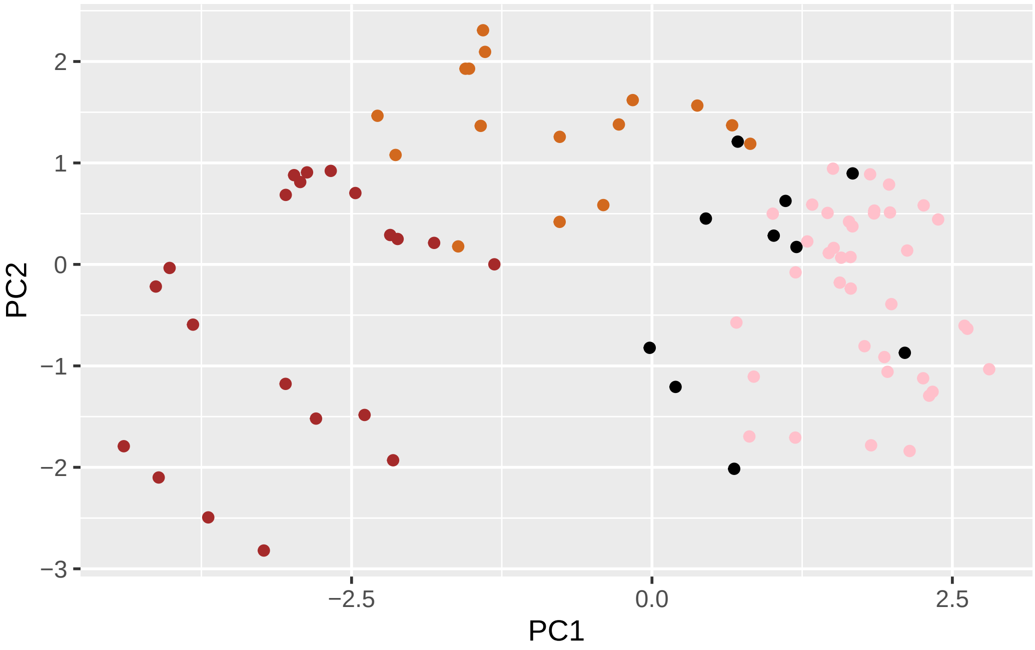
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 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
```

```
attributes(pca)
```

```
$names
[1] "sdev"     "rotation" "center"   "scale"    "x"

$class
[1] "prcomp"
```

Let’s plot our main results as a PCA “score plot”

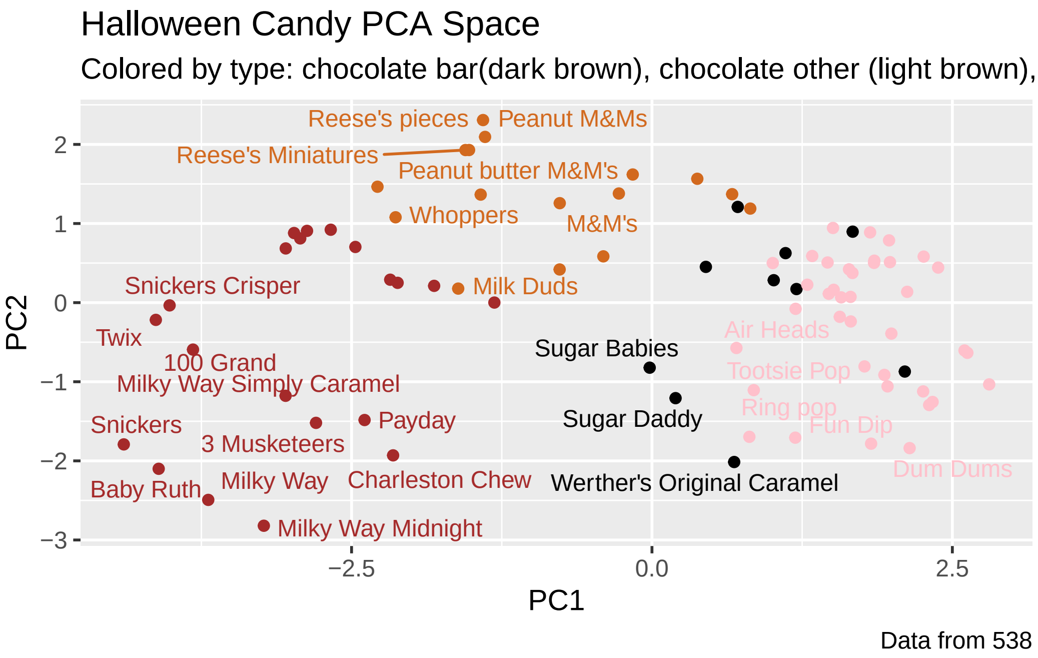
```
ggplot(pca$x) + aes(PC1, PC2, label = rownames(pca$x)) + geom_point(col = mycols)
```



```
#shows seperation of chocolate, chocolate bars, and fruity candies
```

```
ggplot(pca$x) +
  aes(PC1, PC2, label = rownames(pca$x)) +
  geom_point(col = mycols) +
  geom_text_repel(col = mycols, size = 3.3, max.overlaps = 7) +
  labs(title = "Halloween Candy PCA Space", subtitle = "Colored by type: chocolate bar(dark b
```

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

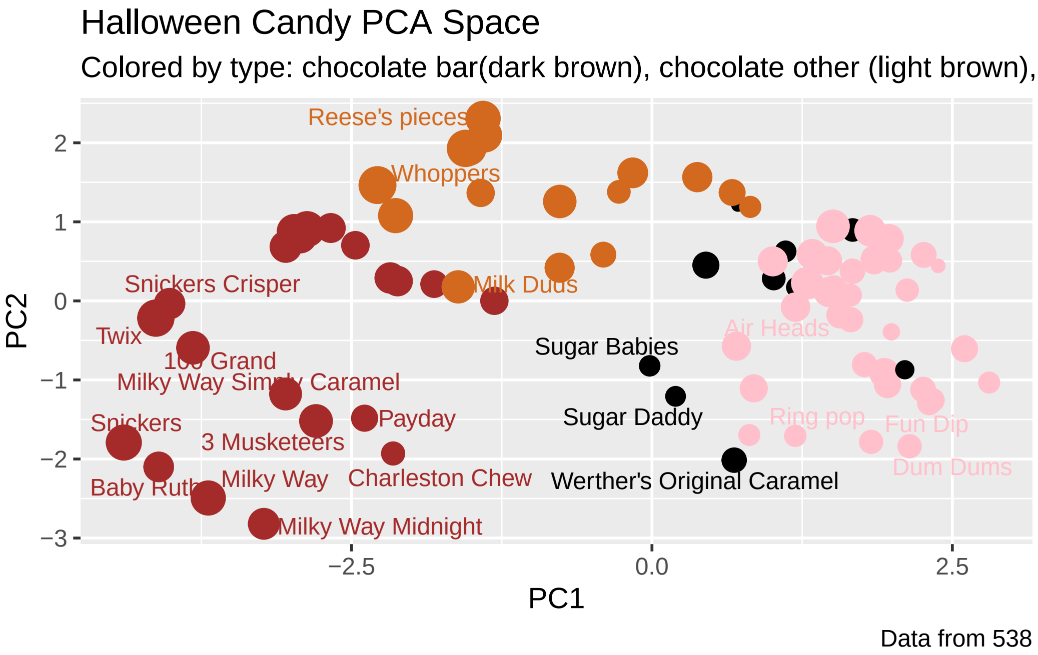


We can also make the points represent the size of `winpercent` of each point. First we will create a dataframe with our PCA data and **candy** dataset.

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

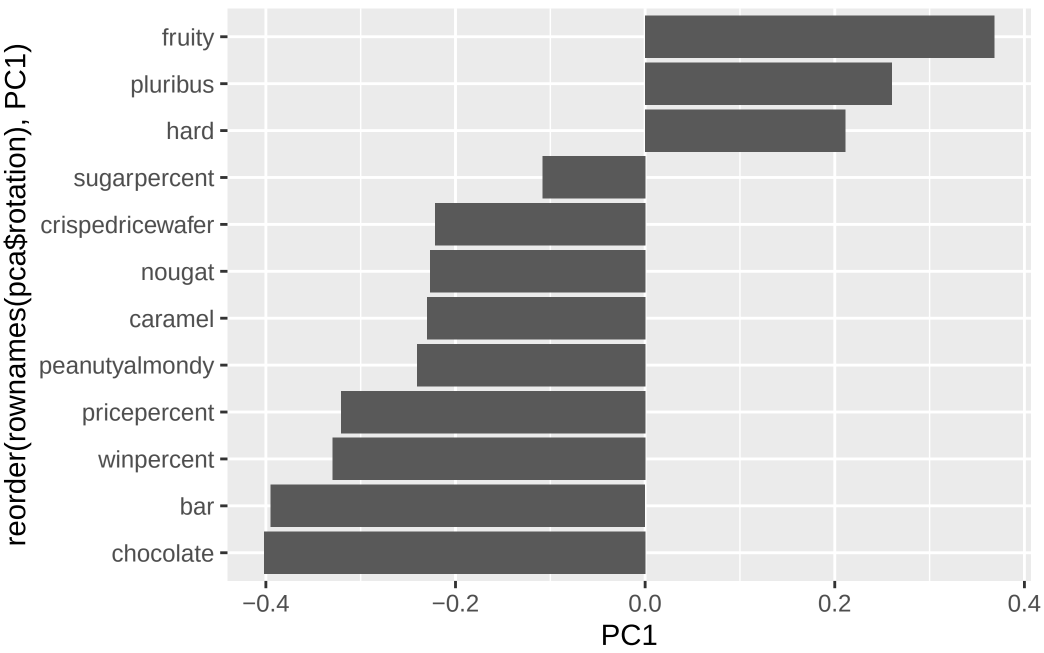
```
candy_PCA_graph <- ggplot(candy_and_PCA) +
  aes(x = PC1, y = PC2, size = winpercent/100, text = rownames(candy_and_PCA), label = rownam
  geom_point(col = mycols) +
  geom_text_repel(size = 3.3, col = mycols, max.overlaps = 6) +
  theme(legend.position = "none") +
  labs(title = "Halloween Candy PCA Space", subtitle = "Colored by type: chocolate bar(dark

candy_PCA_graph
```

```
Warning: ggrepel: 64 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Let's look at how each variable contibutes to PCs, start with PC1

```
ggplot(pca$rotation) +
  aes(PC1, reorder(rownames(pca$rotation), PC1)) +
  geom_col()
```



> 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 picked up strongly in the positive direction. This makes sense due to the earlier correlation plot where we saw that fruity candies were positively correlated with hardness and pluribus.