Class 10: Halloween Mini-Project

Selma Cifric (PID A69042976)

Table of contents

Importing candy data 1
What is your favorate candy? 3
Taking a look at pricepercent 13
Exploring the correlation structure 15
Principal Component Analsysis 16

Importing candy data

```r
candy_file <- "candy-data.csv"

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
```

```r
library(flextable)
flextable::flextable(head(candy))
```

chocolate	fruity	caramel	peanutyalmondy	nougat	crispedricewafer	hard	bar	pluribus	s
1	0	1	0	0	1	0	1	0	
1	0	0	0	1	0	0	1	0	
0	0	0	0	0	0	0	0	0	
0	0	0	0	0	0	0	0	0	
0	1	0	0	0	0	0	0	0	
1	0	0	1	0	0	0	1	0	

```r
library(dplyr)
```

The functions dim(), nrow(), table() and sum() may be useful for answering the first 2 questions.

> Q1. How many different candy types are in this dataset?

```r
nrow(candy)
```

```
[1] 85
```

There are 85 different candies in this dataset.

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

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

```
[1] 38
```

There are 38 fruity candy types in this dataset.

What is your favorate candy?

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

My favorite candy is Sour Patch Kids.

```
candy["Sour Patch Kids", ]$winpercent
```

```
[1] 59.864
```

Their win percent value is 59.864.

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

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

```
[1] 76.7686
```

Winpercent value for Kit Kat is 76.76.

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

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

```
[1] 49.6535
```

Winpercent value for Tootsie Roll Snack Bars is 49.65.

Side-note: the skimr::skim() function. There is a useful skim() function in the skimr package that can help give you a quick overview of a given dataset. Let's install this package and try it on our candy data.

```
library("skimr")
skim(candy)
```

Table 2: 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?

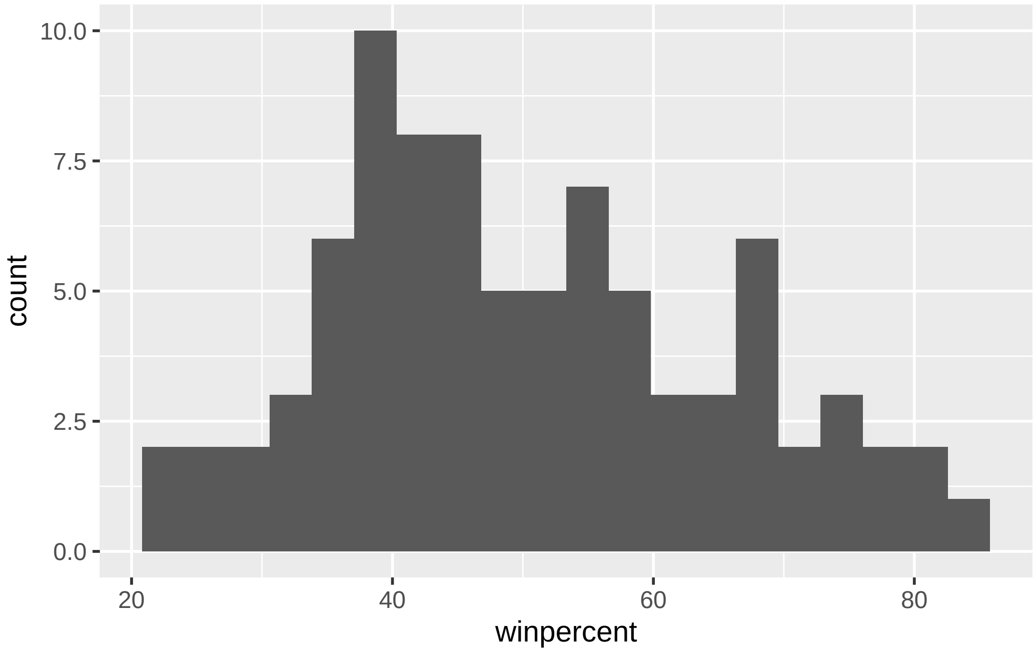
winpercent mean is much higher than the rest of the variables' means, by 50%, and it doesn't use 0-1 scale like others.

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

0 or 1 mean yes or no, where 1 means that candy has the chocolate and 0 means no chocolate present in that candy.

> Q8. Plot a histogram of winpercent values

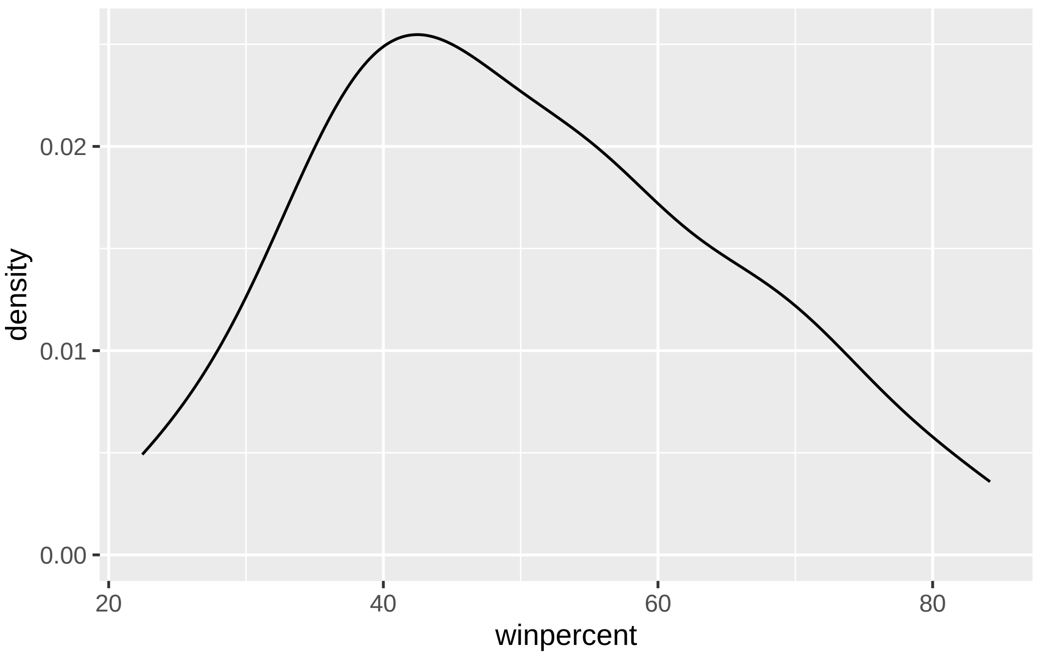
```
library(ggplot2)
ggplot(candy) +
  aes(winpercent) +
  geom_histogram(bins=20)
```



Q9. Is the distribution of winpercent values symmetrical?

No, the distribution is slightly skewed to the left.

```
ggplot(candy) +
  aes(winpercent) +
  geom_density()
```



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

```r
mean(candy$winpercent)
```

```
[1] 50.31676
```

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

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

From median, we can tell the distribution is under 50% which could be a better measure of centrality than mean. The mean suggests that the center of distribution is right at 50%.

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

```r
# 1. Find all chocolate candy in the dataset.
choc.inds <- as.logical(candy$chocolate)
choc.candy <- candy[choc.inds, ]
head(choc.candy)
```

```
                 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
Hershey's 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
Hershey's Kisses                0    0   0        1        0.127        0.093
                 winpercent
100 Grand          66.97173
3 Musketeers       67.60294
Almond Joy         50.34755
Baby Ruth          56.91455
Charleston Chew    38.97504
Hershey's Kisses   55.37545
```

```r
# 2. Extract their `winpercent` values.

choc.win <- choc.candy$winpercent
choc.win
```

```
 [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
# 3. Find the mean of these values.

mean(choc.win)
```

```
[1] 60.92153
```

```
# 4-6. Do the same for the fruity candy.

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

```
                           chocolate fruity caramel peanutyalmondy nougat
Air Heads                          0      1       0              0      0
Caramel Apple Pops                 0      1       1              0      0
Chewey Lemonhead Fruit Mix         0      1       0              0      0
Chiclets                           0      1       0              0      0
Dots                               0      1       0              0      0
Dum Dums                           0      1       0              0      0
                           crispedricewafer hard bar pluribus sugarpercent
Air Heads                                 0    0   0        0        0.906
Caramel Apple Pops                        0    0   0        0        0.604
Chewey Lemonhead Fruit Mix                0    0   0        1        0.732
Chiclets                                  0    0   0        1        0.046
Dots                                      0    0   0        1        0.732
Dum Dums                                  0    1   0        0        0.732
                           pricepercent winpercent
Air Heads                         0.511   52.34146
Caramel Apple Pops                0.325   34.51768
Chewey Lemonhead Fruit Mix        0.511   36.01763
Chiclets                          0.325   24.52499
Dots                              0.511   42.27208
Dum Dums                          0.034   39.46056
```

```
fruit.win <- fruit.candy$winpercent
fruit.win
```

```
 [1] 52.34146 34.51768 36.01763 24.52499 42.27208 39.46056 43.08892 39.18550
 [9] 46.78335 57.11974 51.41243 42.17877 28.12744 41.38956 39.14106 52.91139
[17] 46.41172 55.35405 22.44534 39.44680 41.26551 37.34852 35.29076 42.84914
[25] 63.08514 55.10370 45.99583 59.86400 52.82595 67.03763 34.57899 27.30386
[33] 54.86111 48.98265 47.17323 45.46628 39.01190 44.37552
```

```
mean(fruit.win)
```

```
[1] 44.11974
```

```r
# 7. Which mean value is higher?
mean(choc.win)>mean(fruit.win)
```

```
[1] TRUE
```

Yes, chocolate candy is higher ranked than fruity candy.

> Q12. Is this difference statistically significant?

Hint: The chocolate, fruity, nougat etc. columns indicate if a given candy has this feature (i.e. one if it has nougart, zero if it does not etc.). We can turn these into logical (a.k.a. TRUE/FALSE) values with the as.logical() function. We can then use this logical vector to access the coresponding candy rows (those with TRUE values). For example to get the winpercent values for all nougat contaning candy we can use the code: candy*winpercent*[*as.logical*(*candy*nougat)]. In addation the functions mean() and t.test() should help you answer the last two questions here.

```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 
```

Yes, the difference is statistically significant indicated by the p-value of 2.871e-08.

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

Example of using `order()` function.

```r
x <- c(10, 2, 5, 1)
order(x)
```

```
[1] 4 2 3 1
```

Using `head()` and `order()` function we can see the least favorite candy.

```
ord.ind <- order(candy$winpercent)
head(candy[ord.ind, ], 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
```

These are the least 5 liked candy: Nik L Nip, Boston Baked Bean, Chiclets, Super bubble, and Jawbusters.

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

```
tail(candy[ord.ind, ], 5)
```

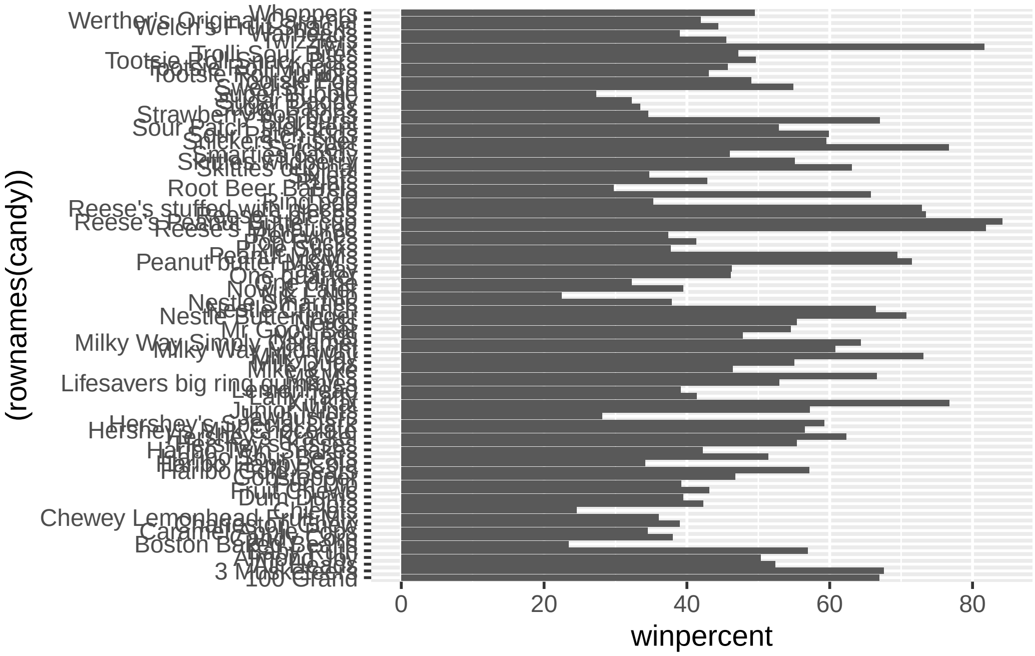
```
                          chocolate fruity caramel peanutyalmondy nougat
Snickers                          1      0       1              1      1
Kit Kat                           1      0       0              0      0
Twix                              1      0       1              0      0
Reese's Miniatures                1      0       0              1      0
Reese's 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
```

```
Reese's Miniatures                        0    0   0        0        0.034
Reese's 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
Reese's Miniatures               0.279   81.86626
Reese's Peanut Butter cup        0.651   84.18029
```

Top 5 most liked candy are Snickers, Kit Kat, Twin, Reese's Miniatures, and Reese's Peanut Butter Cups.

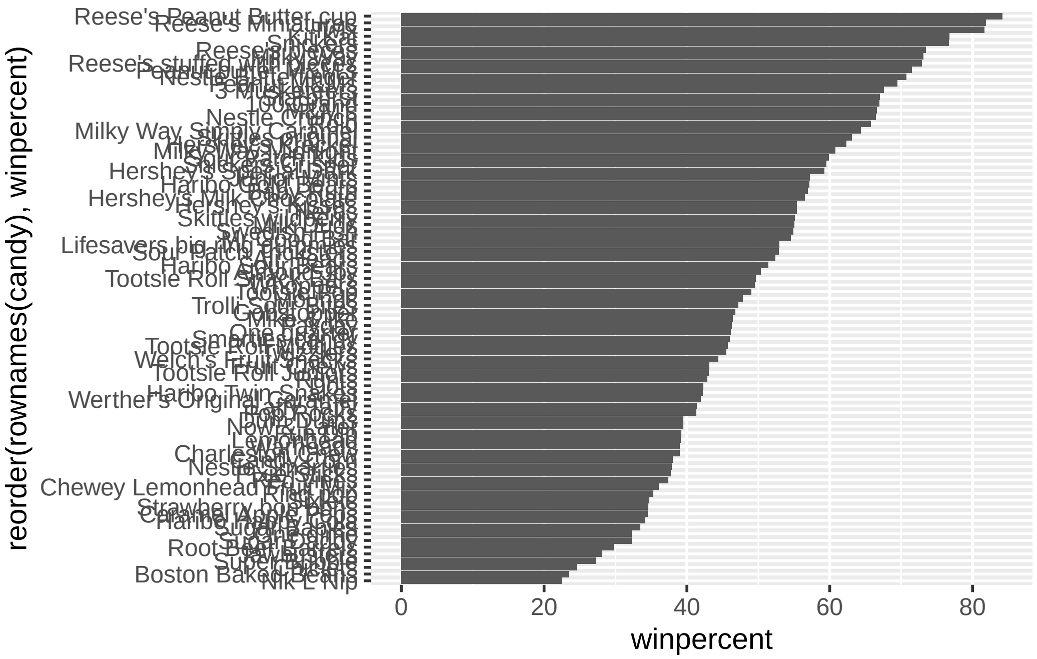
> Q15. Make a first barplot of candy ranking based on winpercent values.

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



> 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()
```



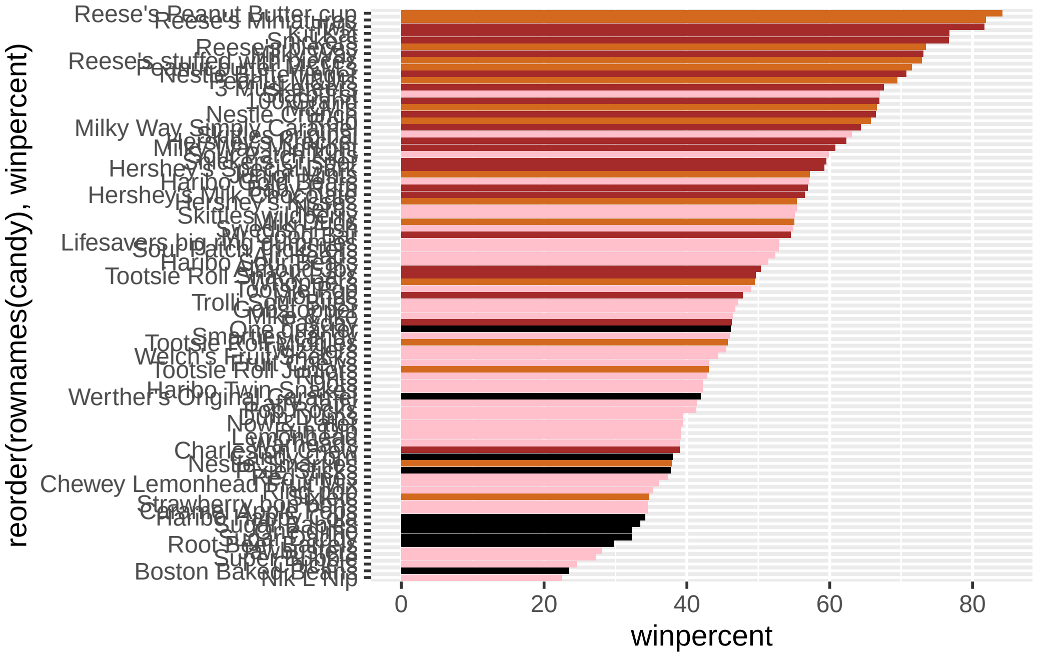
Let's add some color to this plot:

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

```
 [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(x=winpercent,
      y=reorder( rownames(candy), winpercent)) +
  geom_col(fill=my_cols)
```



Q17. What is the worst ranked chocolate candy?

Sixlets are the least favorite chocolate candy.

Q18. What is the best ranked fruity candy?

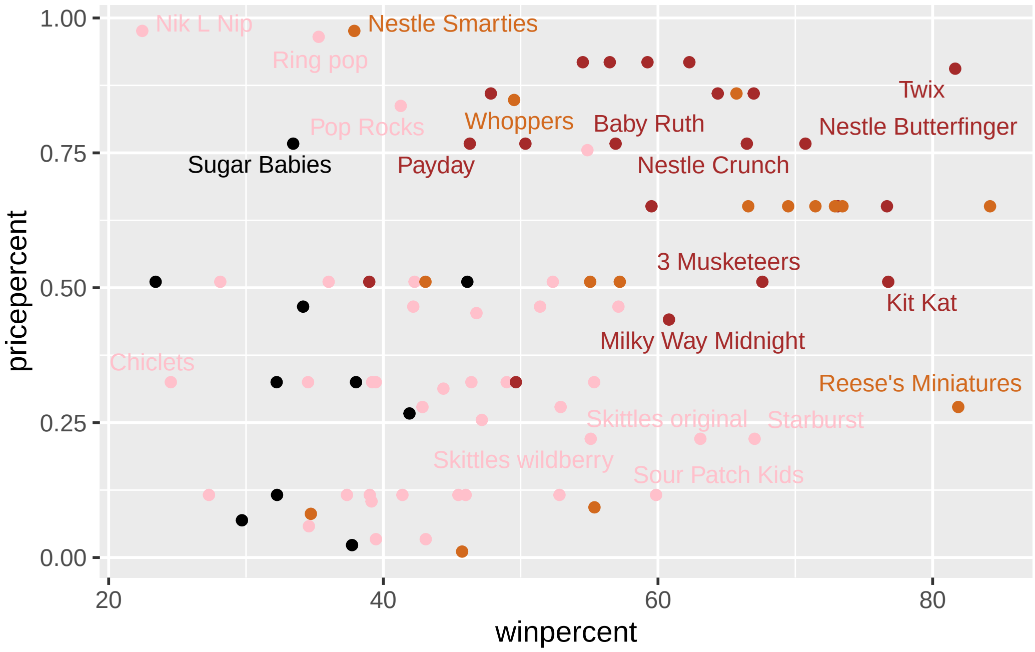
Starburts are the best ranked fruity candy (top most pink bar).

Taking a look at pricepercent

```r
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
```



> 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?

It's chocolate candy Reese's Miniatures.

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

```r
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
```

Top 5 most expensive but least popular candy are: Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate.

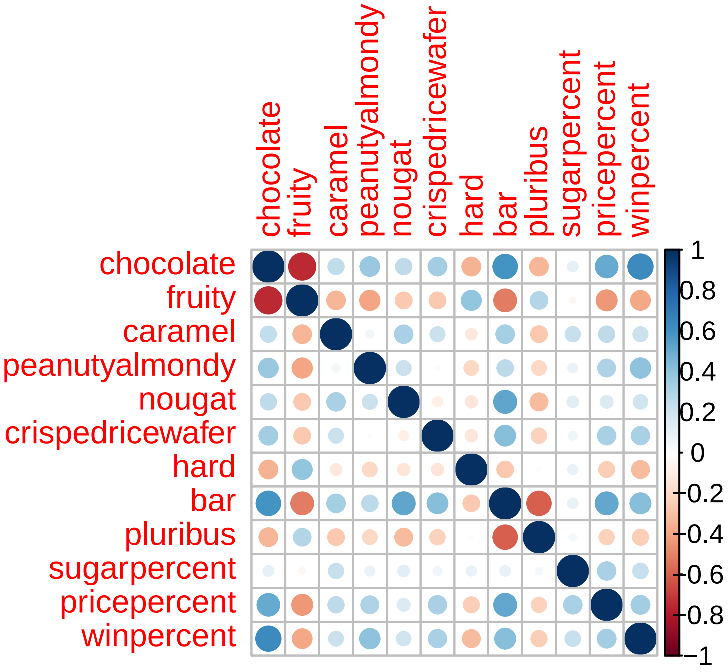
Exploring the correlation structure

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

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

```
corrplot(cij)
```



> Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

Chocolate and fruity candy are anti-correlated.

> Q23. Similarly, what two variables are most positively correlated? HINT: Do you like chocolaty fruity candies?

Chocolate and winpercent as well as chocolate and bar. I actually like a nice dark chocolate raspberry mix, which seems to be an unpopular opinion.

Principal Component Analsysis

The main function in base R for this is `prcomp()` and we want to set the `scale=T` here:

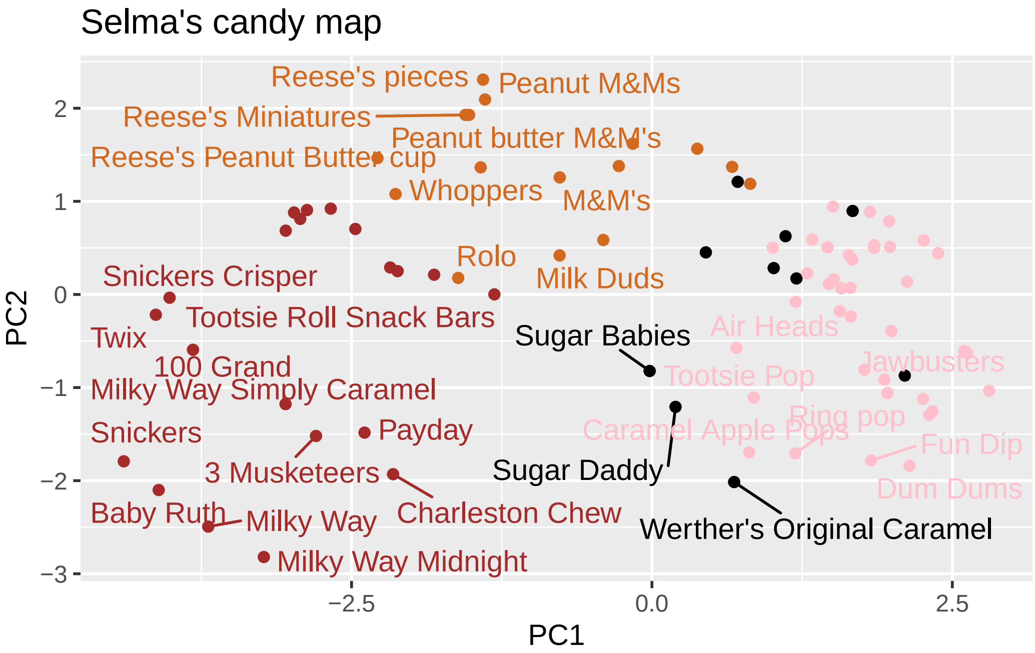
```r
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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
```

Let’s look at our first main result figure - the “PCA plot” or PC1 vs PC2

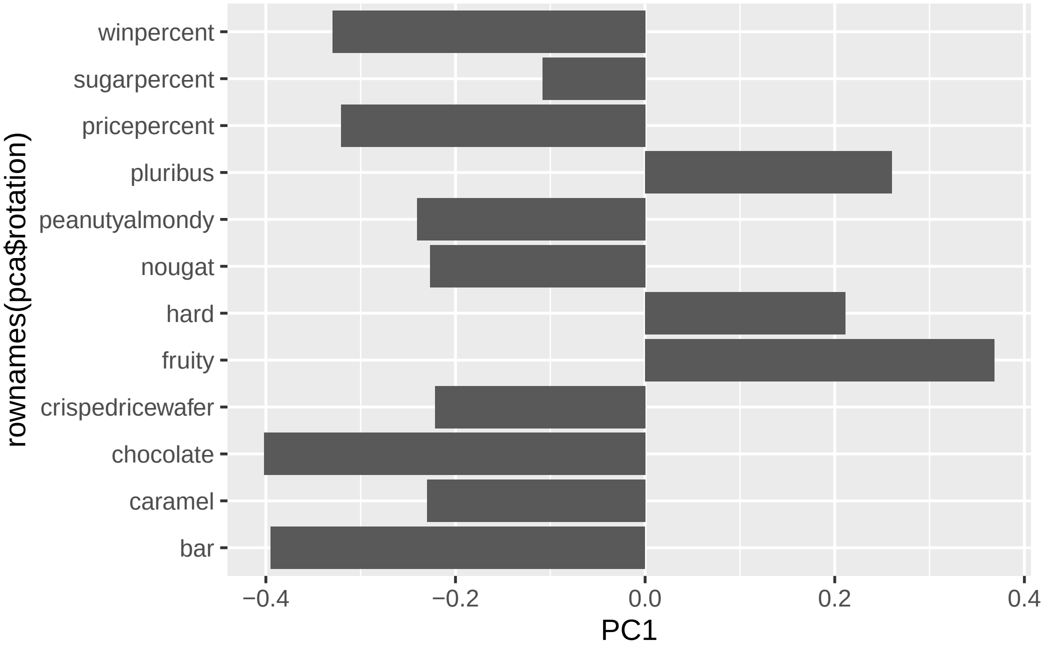
```r
ggplot(pca$x) +
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols) +
  labs(title="Selma's candy map")
```

```
Warning: ggrepel: 54 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Don't forget about your variable "loadings" - how the original vairable contribute to your new PCs…

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



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

Fruity, hard and pluribus are picked up strongly by PC1. It makes sense, most fruity candy are hard and come in a pack of multiples.