Class 09- Halloween Candy Mini Project

Renee Zuhars- PID: A17329856

Table of contents

Today we will take a wee step back to some data we can taste, and explore the correlation structure and principal components of some Halloween candy.

Data Import

```
candy <- read.csv("candy-data.csv", row.names = 1)

View(candy)
```

Question 1

> How many different candy types are in this dataset?

```
nrow(candy)
```

```
[1] 85
```

There are 85 types of candy in this dataset.

Question 2

> How many fruity candy types are in the dataset?

```
sum(candy$fruity == 1)
```

```
[1] 38
```

There are 38 types of fruity candy in the dataset.

My favorite candy

Question 3

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

```r
candy["Nestle Butterfinger",]$winpercent
```

```
[1] 70.73564
```

My favorite candy is Butterfinger, and the win percent is about 71%.

Question 4

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

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

```
[1] 76.7686
```

The winpercent value for Kit Kat is about 77%.

Question 5

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

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

```
[1] 49.6535
```

The winpercent value for Tootsie Roll Snack Bars is about 50%.

Exploratory Analysis

Skimr can be useful when you want a quick overview of a dataset, for example, if you are encountering it for the first time.

```r
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	

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?

It looks like the last column `candy$winpercent` is on a different scale than the others.

Question 7

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

I think the zero represents that the candy is not chocolate, because it means the logical is false. So, the one would represent that the candy is indeed chocolate.

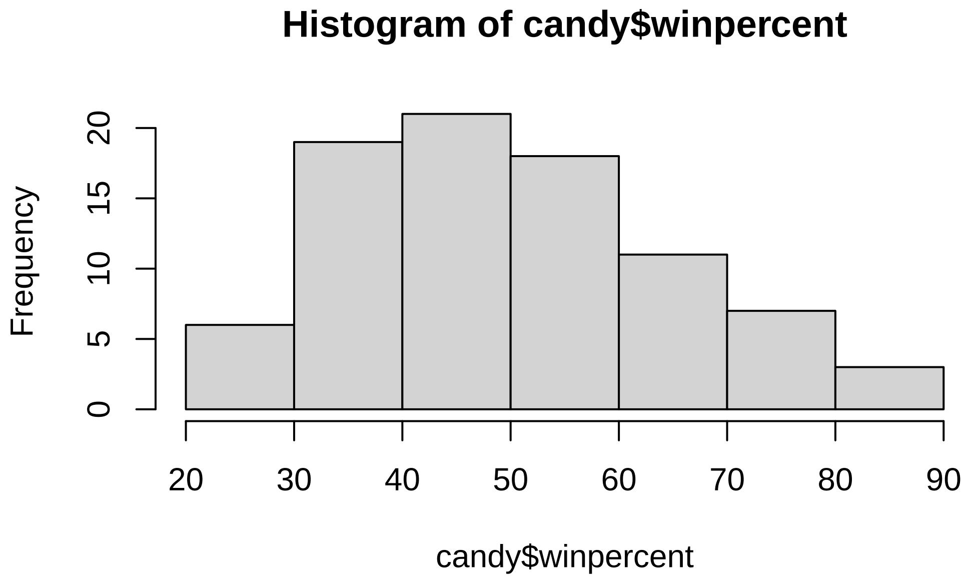
A good place to start any exploratory analysis is with a histogram. You can do this most easily with the base R function hist(). Alternatively, you can use ggplot() with geom_hist().

Question 8

> Plot a histogram of winpercent values

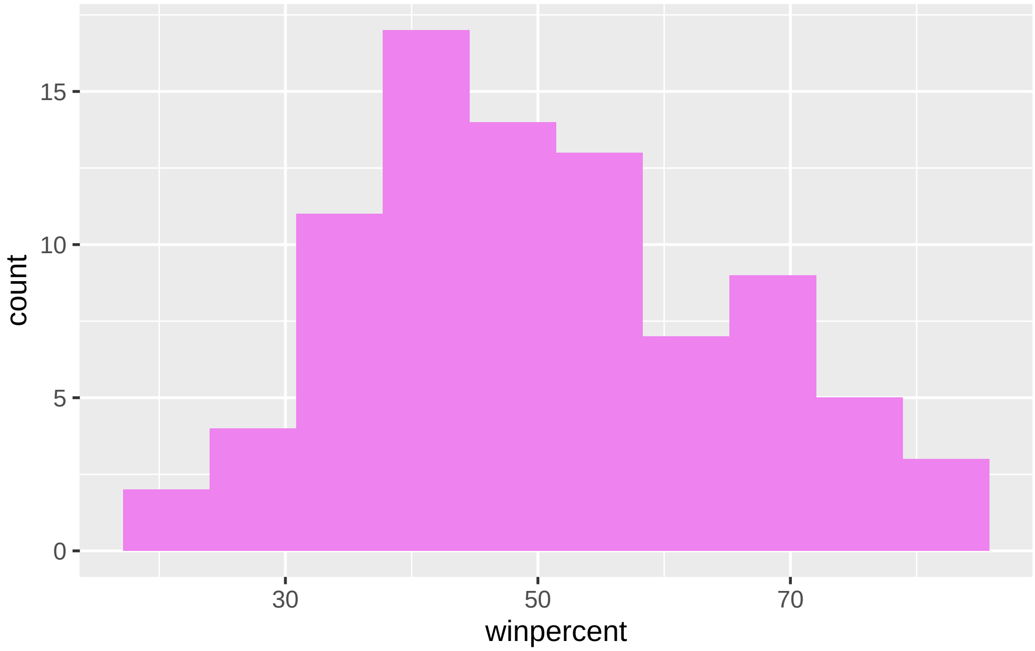
in base R:

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



in ggplot:

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



Question 9

Is the distribution of winpercent values symmetrical?

No, the distribution of winpercent values in not symmetrical.

Question 10

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

```r
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 a bit below the mean, as it is represented by the median.

Question 11

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

for chocolate candy:

```r
choc.inds <- candy$chocolate == 1
choc.candy <- candy[choc.inds,]
choc.win <- choc.candy$winpercent
mean(choc.win)
```

```
[1] 60.92153
```

for fruity candy:

```r
fruit.inds <- candy$fruity == 1
fruit.candy <- candy[fruit.inds,]
fruit.win <- fruit.candy$winpercent
mean(fruit.win)
```

```
[1] 44.11974
```

The average winpercent for chocolate candy is about 61%, while it is about 44% for fruity candy. The average winpercent for chocolate candy is higher than that of fruity candy.

Question 12

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 
```

Yes, they are statistically significant, with a P-value of 2.871e-08

Overall Candy Rankings

There are two related functions that can help here, one is the classic `sort()` and `order()`. Here's how they work:

```
x <- c(5,10,1,4)
sort(x)
```

```
[1]  1  4  5 10
```

sorts the variables directly.

```
order(x)
```

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

gives the variable position that you need to reference in order of least to greatest.

Question 13

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

```
inds <- order(candy$winpercent)
head( candy[inds,], 5) # whole candy table sorted by indeces
```

```
                   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
```

The five least liked candy types are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

Question 14

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

```r
inds <- order(candy$winpercent, decreasing = TRUE)
head( candy[inds,], 5)
```

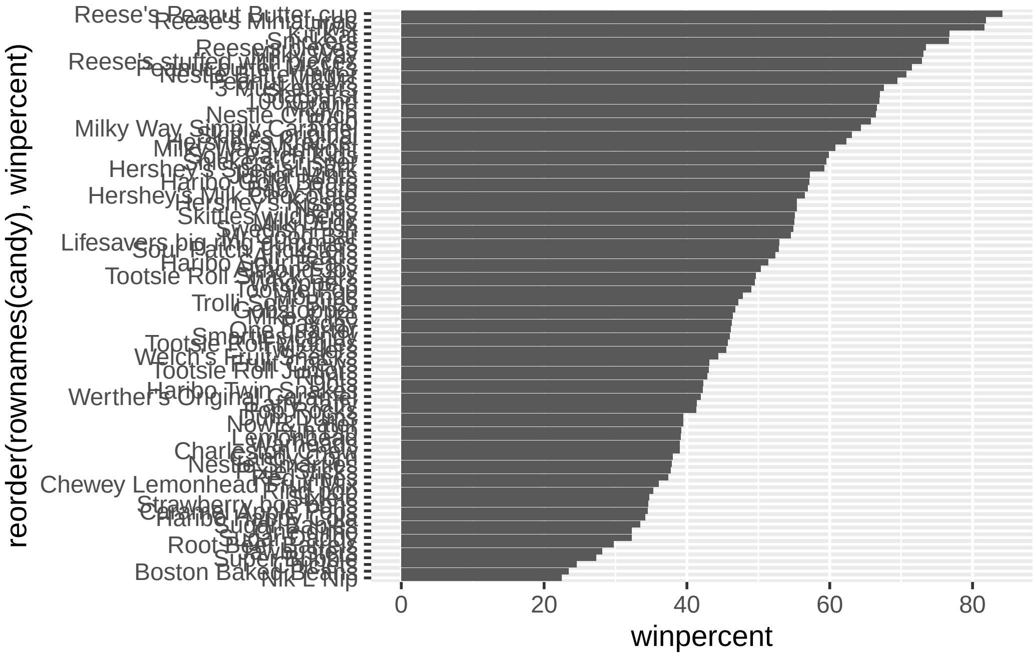
```
                          chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat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

The five most liked candies are Reese’s Peanut Butter Cup, Reese’s Miniatures, Twix, Kit Kat, Snickers.

Question 15

Make a first barplot of candy ranking based on winpercent values, with ggplot.

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

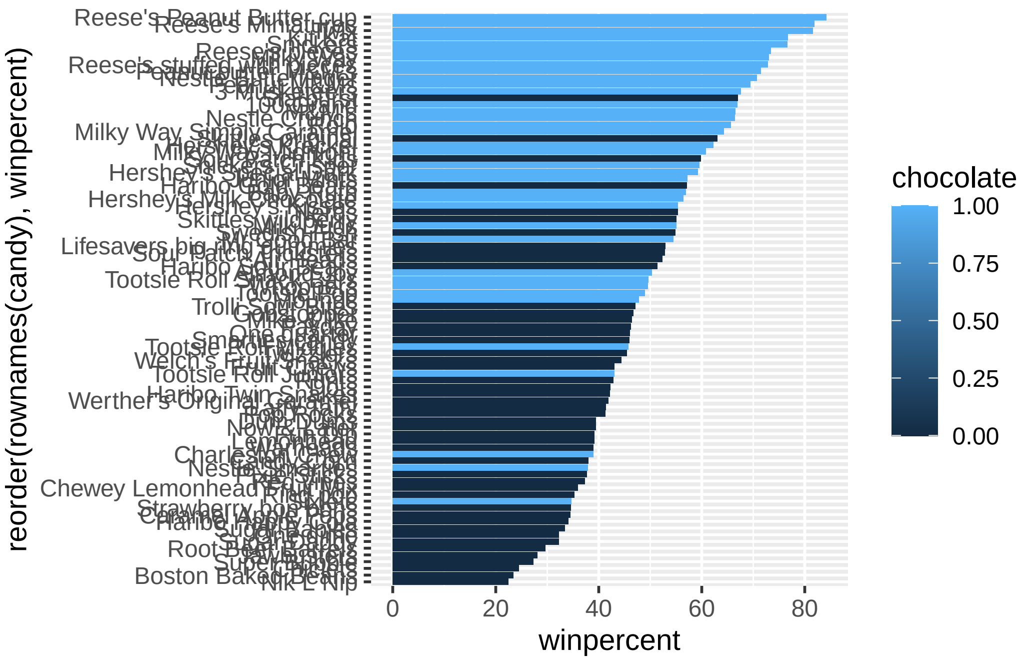


It's not the easiest plot to read.

Question 16

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

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

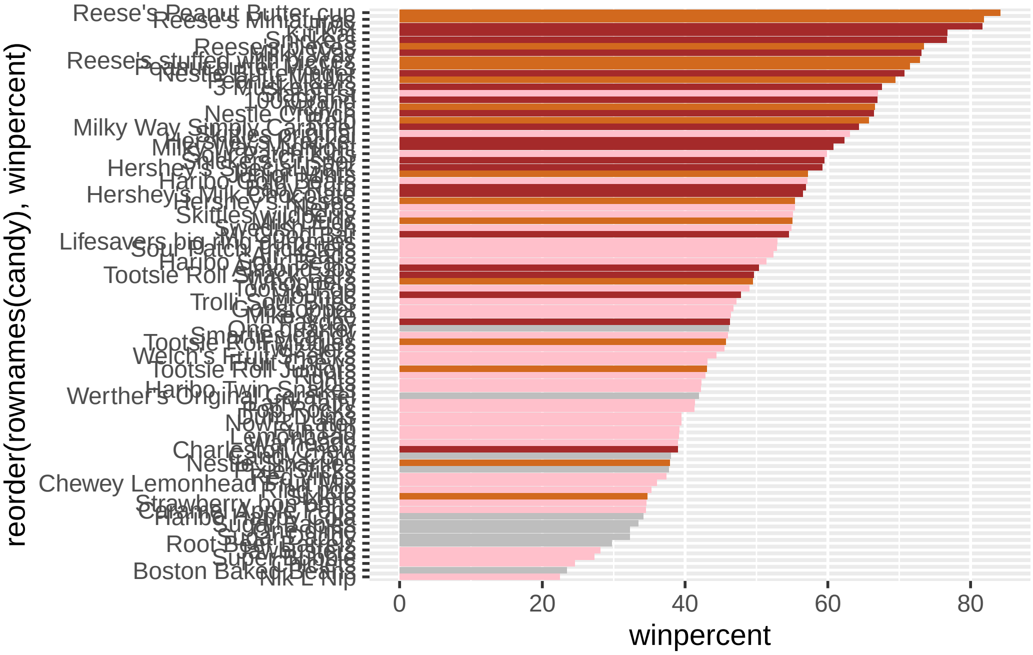


Time to add some useful color!

Here we want a custom color vector to color each bar the way we want- with `chocolate` and `fruity` candy together whether it is a `bar` or not

```
mycols <- rep("gray", nrow(candy))
mycols[as.logical(candy$chocolate)] <- "chocolate"
mycols[as.logical(candy$fruity)] <- "pink"
mycols[as.logical(candy$bar)] <- "brown"

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



```r
ggsave("mybarplot.png", width=3, height=6)
```

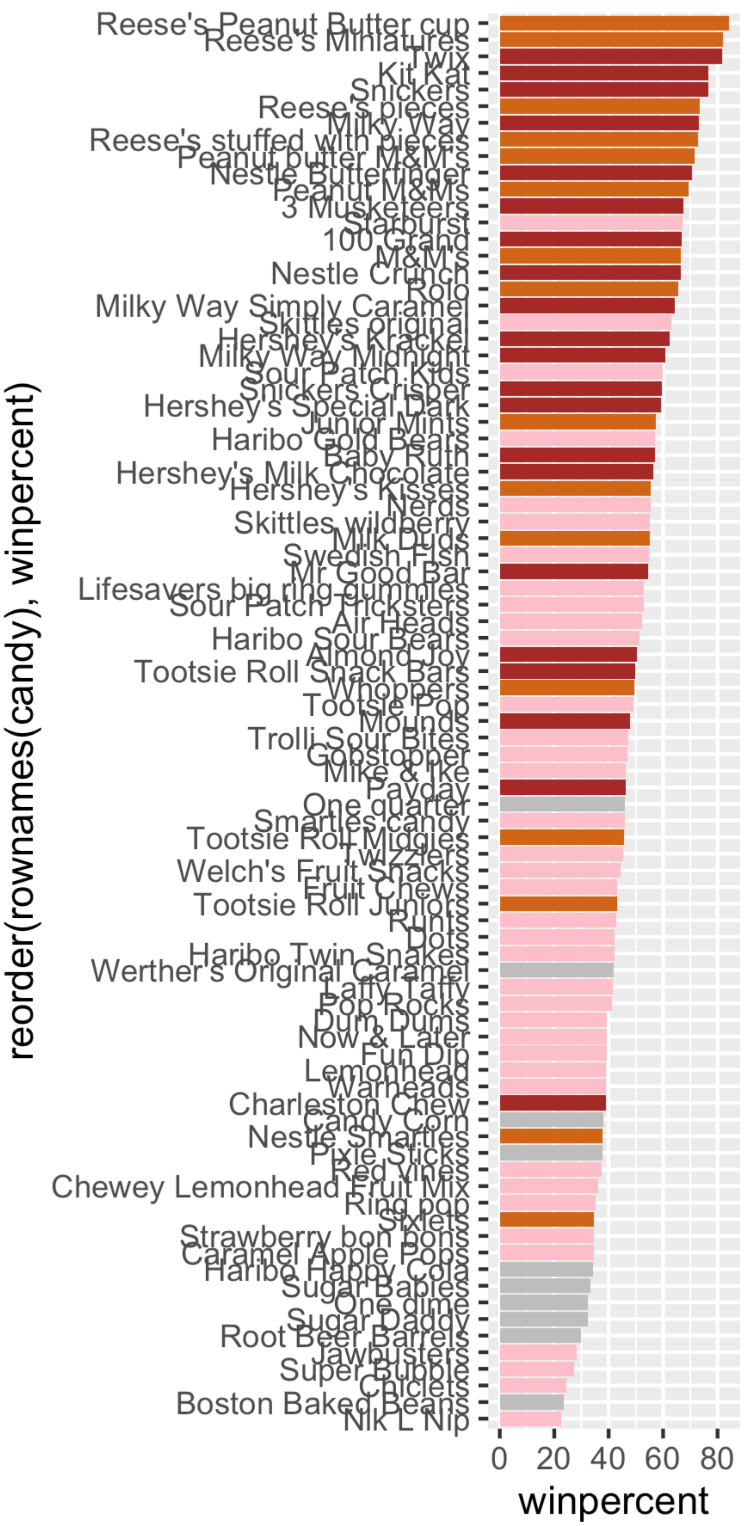


Figure 1: My silly barplot image

Question 17

> What is the worst ranked chocolate candy?

Sixlets are the worst ranked chocolate candy.

Question 18

> What is the best ranked fruity candy?

Starburst is the best ranked fruity candy.

Taking a look at pricepercent

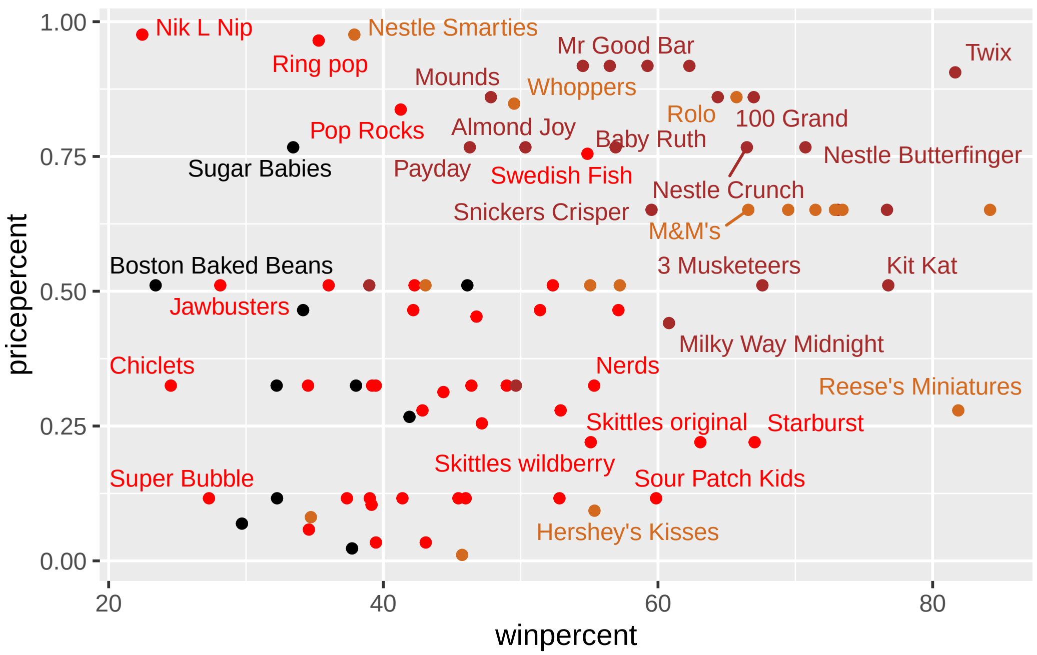
Getting the best value for your money:

```r
# Pink is too light, let's change to red
mycols <- rep("black", nrow(candy))
mycols[as.logical(candy$chocolate)] <- "chocolate"
mycols[as.logical(candy$fruity)] <- "red"
mycols[as.logical(candy$bar)] <- "brown"

library(ggrepel)

# How about a plot of price vs win
ggplot(candy) +
  aes(winpercent, pricepercent, label=rownames(candy)) +
  geom_point(col=mycols) +
  geom_text_repel(col=mycols, size=3.3, max.overlaps = 8)
```

```
Warning: ggrepel: 52 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?

Reese's Peanut Butter miniatures are the highest ranked in terms of winpercent, and have a lower price.

Question 20

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

```r
inds <- order(candy$pricepercent, decreasing = TRUE)
head( candy[inds,], 5)
```

```
                         chocolate fruity caramel peanutyalmondy nougat
Nik L Nip                        0      1       0              0      0
Nestle Smarties                  1      0       0              0      0
Ring pop                         0      1       0              0      0
Hershey's Krackel                1      0       0              0      0
```

```
Hershey's Milk Chocolate         1      0       0              0      0
                         crispedricewafer hard bar pluribus sugarpercent
Nik L Nip                               0    0   0        1        0.197
Nestle Smarties                         0    0   0        1        0.267
Ring pop                                0    1   0        0        0.732
Hershey's Krackel                       1    0   1        0        0.430
Hershey's Milk Chocolate                0    0   1        0        0.430
                         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
```

The most expensive candy types are Nik L Nip, Nestle Smarties, Ring Pop, Hershey's Krackel, and Hershey's Milk Chocolate. Of these, Nik L Nip is the least liked.

Question 21

optional, skipped

Exploring the correlation structure

```r
library(corrplot)
```

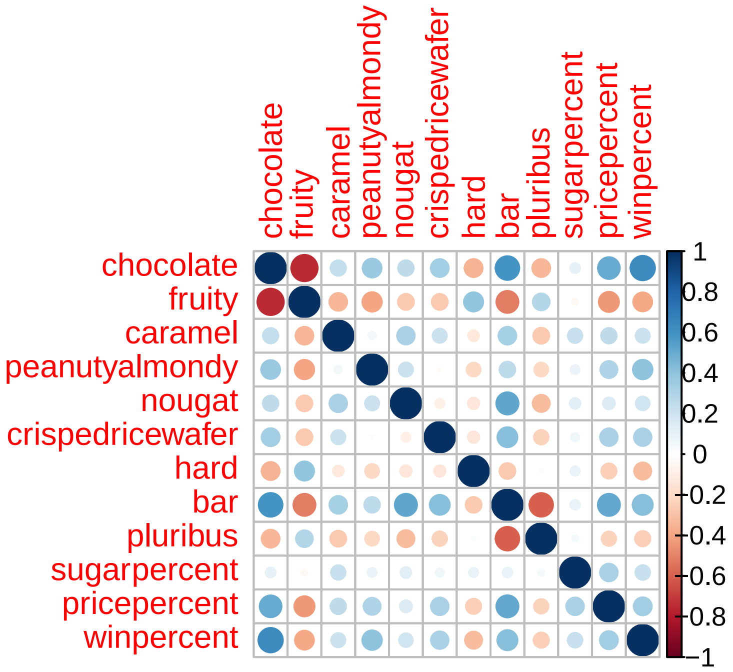
```
corrplot 0.95 loaded
```

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

```
                  chocolate      fruity     caramel peanutyalmondy      nougat
chocolate         1.0000000 -0.74172106  0.24987535     0.37782357  0.25489183
fruity           -0.7417211  1.00000000 -0.33548538    -0.39928014 -0.26936712
caramel           0.2498753 -0.33548538  1.00000000     0.05935614  0.32849280
peanutyalmondy    0.3778236 -0.39928014  0.05935614     1.00000000  0.21311310
nougat            0.2548918 -0.26936712  0.32849280     0.21311310  1.00000000
crispedricewafer  0.3412098 -0.26936712  0.21311310    -0.01764631 -0.08974359
hard             -0.3441769  0.39067750 -0.12235513    -0.20555661 -0.13867505
```

```
bar               0.5974211 -0.51506558  0.33396002     0.26041960  0.52297636
pluribus         -0.3396752  0.29972522 -0.26958501    -0.20610932 -0.31033884
sugarpercent      0.1041691 -0.03439296  0.22193335     0.08788927  0.12308135
pricepercent      0.5046754 -0.43096853  0.25432709     0.30915323  0.15319643
winpercent        0.6365167 -0.38093814  0.21341630     0.40619220  0.19937530
                 crispedricewafer        hard         bar    pluribus
chocolate              0.34120978 -0.34417691  0.59742114 -0.33967519
fruity                -0.26936712  0.39067750 -0.51506558  0.29972522
caramel                0.21311310 -0.12235513  0.33396002 -0.26958501
peanutyalmondy        -0.01764631 -0.20555661  0.26041960 -0.20610932
nougat                -0.08974359 -0.13867505  0.52297636 -0.31033884
crispedricewafer       1.00000000 -0.13867505  0.42375093 -0.22469338
hard                  -0.13867505  1.00000000 -0.26516504  0.01453172
bar                    0.42375093 -0.26516504  1.00000000 -0.59340892
pluribus              -0.22469338  0.01453172 -0.59340892  1.00000000
sugarpercent           0.06994969  0.09180975  0.09998516  0.04552282
pricepercent           0.32826539 -0.24436534  0.51840654 -0.22079363
winpercent             0.32467965 -0.31038158  0.42992933 -0.24744787
                 sugarpercent pricepercent winpercent
chocolate          0.10416906    0.5046754  0.6365167
fruity            -0.03439296   -0.4309685 -0.3809381
caramel            0.22193335    0.2543271  0.2134163
peanutyalmondy     0.08788927    0.3091532  0.4061922
nougat             0.12308135    0.1531964  0.1993753
crispedricewafer   0.06994969    0.3282654  0.3246797
hard               0.09180975   -0.2443653 -0.3103816
bar                0.09998516    0.5184065  0.4299293
pluribus           0.04552282   -0.2207936 -0.2474479
sugarpercent       1.00000000    0.3297064  0.2291507
pricepercent       0.32970639    1.0000000  0.3453254
winpercent         0.22915066    0.3453254  1.0000000
```

```r
corrplot(cij)
```



Question 22

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

The two most negatively correlated variables are chocolate and fruity.

```
round(cij["chocolate", "fruity"], 2)
```

```
[1] -0.74
```

Question 23

Similarly, what two variables are most positively correlated?

The two most positively correlated variables are either chocolate and winpercent and chocolate and bar. Let’s test:

```
round(cij["chocolate", "winpercent"], 2)
```

```
[1] 0.64
```

```
round(cij["chocolate", "bar"], 2)
```

```
[1] 0.6
```

The two most positively correlated variables are chocolate and winpercent.

Principal Component Analysis

We need to be sure to scale our input `candy` data before PCA as we have the `winpercent` column on a different scale to all others in the dataset.

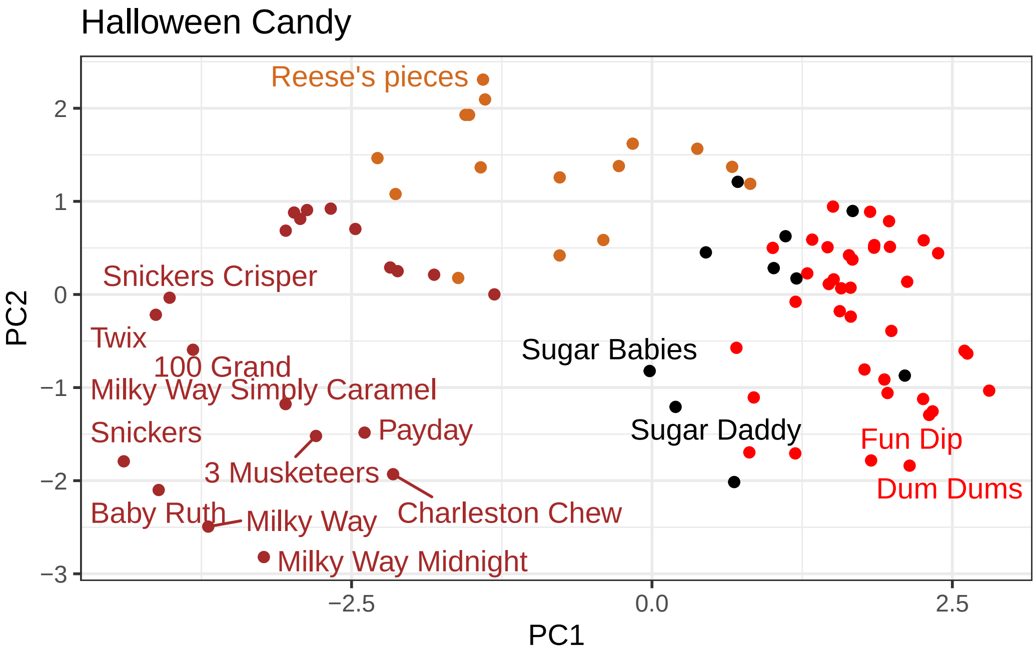
```
pca <- prcomp(candy, scale=T)
summary(pca)
```

```
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
```

First main result figure is my "PCA plot"

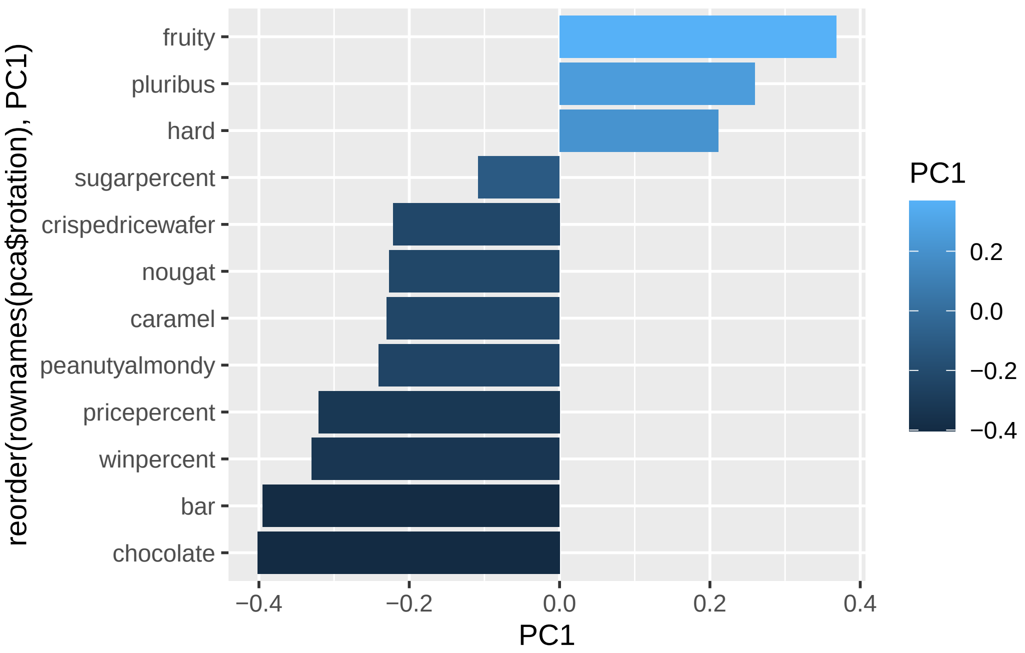
```
# pca$x
ggplot(pca$x) +
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=mycols) +
  geom_text_repel(max.overlaps = 6, col=mycols) +
  theme_bw() +
  labs(title="Halloween Candy")
```

```
Warning: ggrepel: 69 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



The second main PCA result is in the `pca$rotation`. We can plot this to generate a so-called "loadings" plot.

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



Question 24

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

The variables that are picked up strongly in the positive direction are fruity, pluribus, and hard. Yes, this makes sense to me because the variables with more positive values have been shown to exist together in the same candy more frequently. On the other hand, the variables with negative values correlate to each other.