Haloween Candy

Tianru Zhang (PID: A15432834)

10/28/2021

```
url<-"https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-ranking/candy-data.csv"
#candy_file <- read.csv(url)

candy <- read.csv(url, row.names=1)
head(candy,n=5)
```

```
##              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
##              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
```

Q1.How many candy types: 85

There are 85 types of candies in total.

```
nrow(candy)
```

```
## [1] 85
```

Q2: How many fruity candy types?

there are 38 of them.

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

```
## [1] 38
```

#2. Favourite candy

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

```
## [1] 81.64291
```

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

```
candy["Haribo Happy Cola", ]$winpercent
```

```
## [1] 34.15896
```

My fav is Haribo Happy Cola and its winpercent value is 34%.

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

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

```
## [1] 76.7686
```

The percent is 76.77% for Kit Kat

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

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

```
## [1] 49.6535
```

The winpercent is 49.65% for Tootsie Roll Snack Bars.

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.

```
#install.packages("skimr")
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?

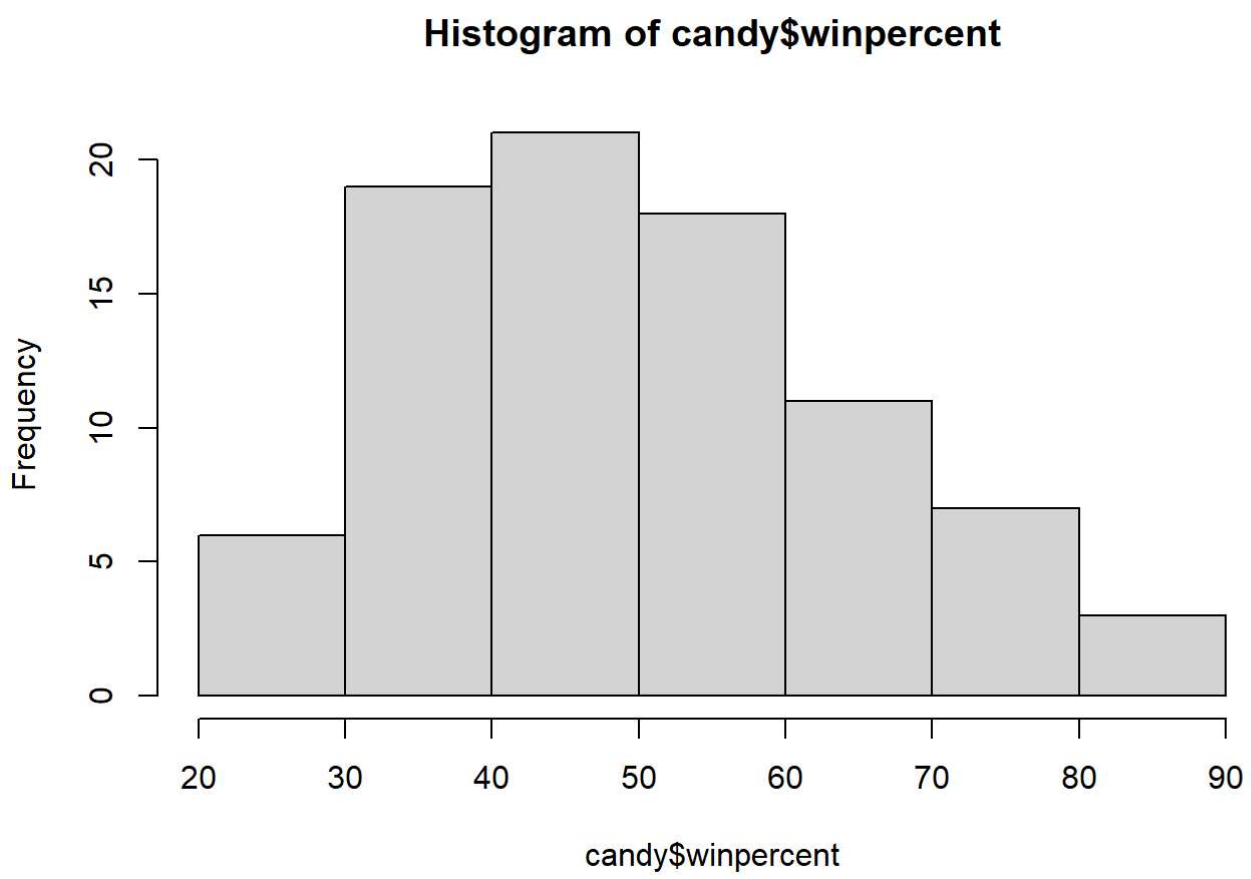
Winpercent, it looks on a very different scale.

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

I think it indicates that the candy$chocolate stands for categorical data

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical?

The distribution is not symmetrical and is slightly skewed left.

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

The center of the distribution is below 50%.

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

First, need to find all the chocolate candy rows in the 'candy' dataset

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

```
##  [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
```

```
chocolate<-candy$winpercent[as.logical(candy$chocolate)]

fruit<-candy$winpercent[as.logical(candy$fruit)]
```

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

```
##
##  Welch Two Sample t-test
##
## data:  chocolate and fruit
## 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
```

```
mean(chocolate)
```

```
## [1] 60.92153
```

```
mean(fruit)
```

```
## [1] 44.11974
```

Chocolate candies are generally higher ranked than fruity candy.

Q12. Is this difference statistically significant? Yes!

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

```
##
##  Welch Two Sample t-test
##
## data:  chocolate and fruit
## 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
```

Since the p-value is very small, this means that the difference is significant, and chcolate is more significant than fruity candy.

#3. Overall Candy Rankings Let's use the base R order() function together with head() to sort the whole dataset by winpercent. Or if you have been getting into the tidyverse and the dplyr package you can use the arrange() function together with head() to do the same thing and answer the following questions:

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

"Nik L Nip", "Boston Baked Beans", "Chiclets", "Super Bubble" and "Jawbusters"

```
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?

Kit Kat Snickers Twix ReeseÃ•s Miniatures ReeseÃ•s Peanut Butter cup

```
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
## ReeseÃ•s Miniatures                1      0       0              1      0
## ReeseÃ•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
## ReeseÃ•s Miniatures                        0    0   0        0        0.034
## ReeseÃ•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
## ReeseÃ•s Miniatures                0.279   81.86626
## ReeseÃ•s Peanut Butter cup         0.651   84.18029
```

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

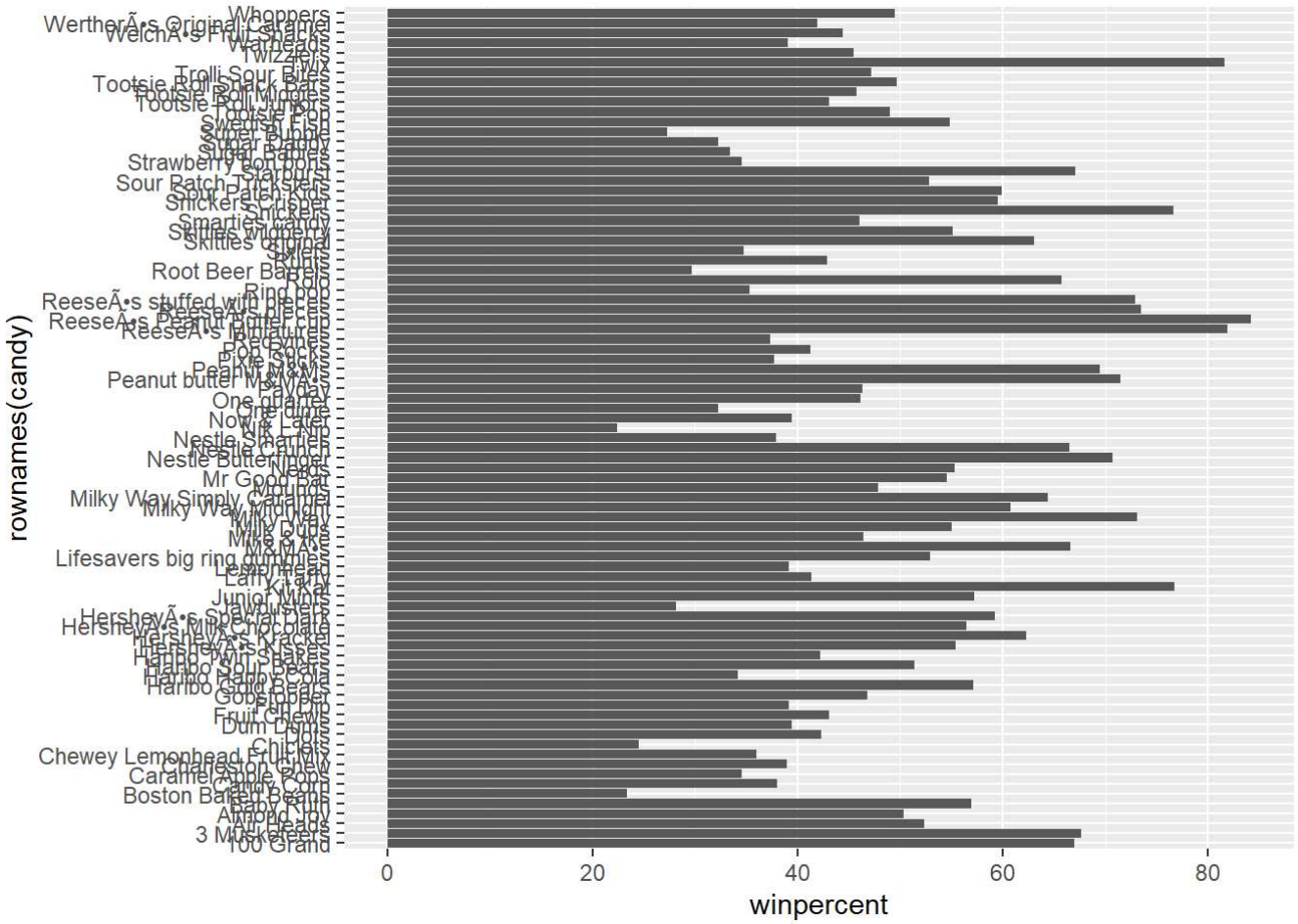
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
candy %>% arrange(winpercent) %>% head(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
```

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

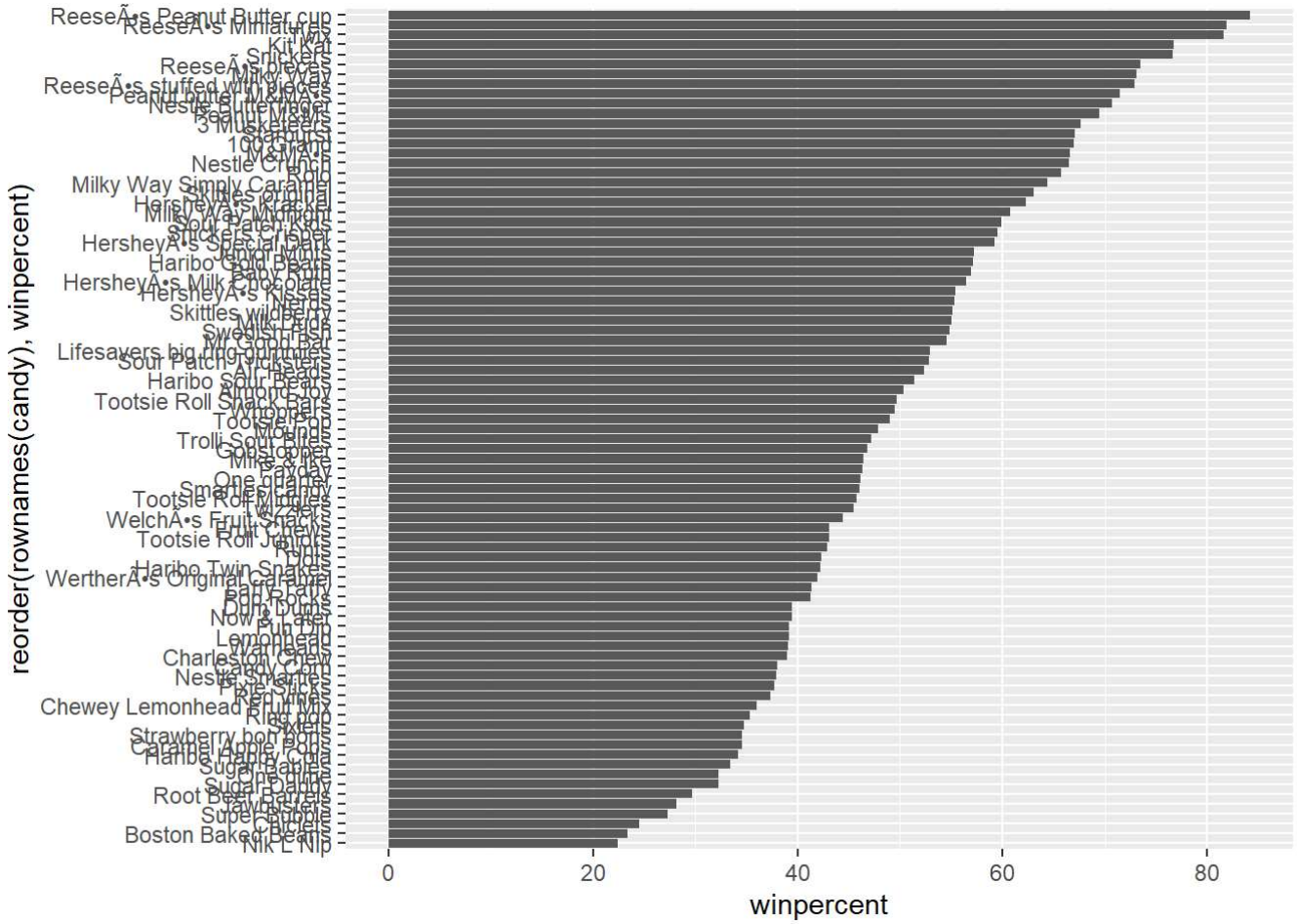
```
library(ggplot2)

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



Need to improve this plot by reordering the `winpercent` value

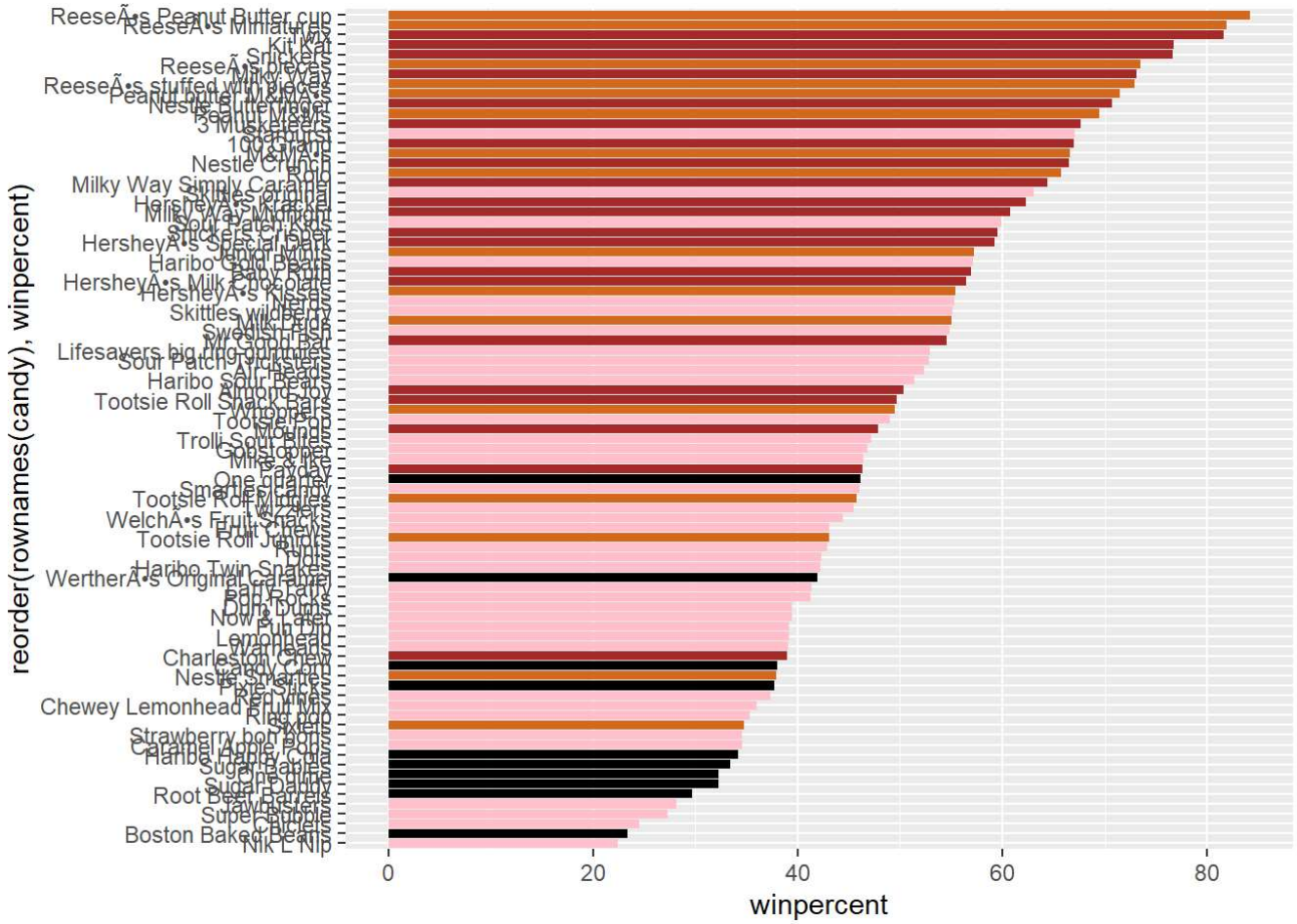
Q16. Use the reorder function to reorder the plot!

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




```
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"
```

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



Q17. What is the worst ranked chocolate candy?

The worst ranked chocolate candy is Sixlets.

Q18. What is the best ranked fruity candy?

Starbust.

#4.Take a look at the pricepercent

```
#install.packages("ggrepel")
library(ggrepel)


#change red
my_cols[as.logical(candy$fruity)]="red"
# 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=2.3, max.overlaps = 7)
```

```
## Warning: ggrepel: 14 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_A_miniature is the highest ranked.

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

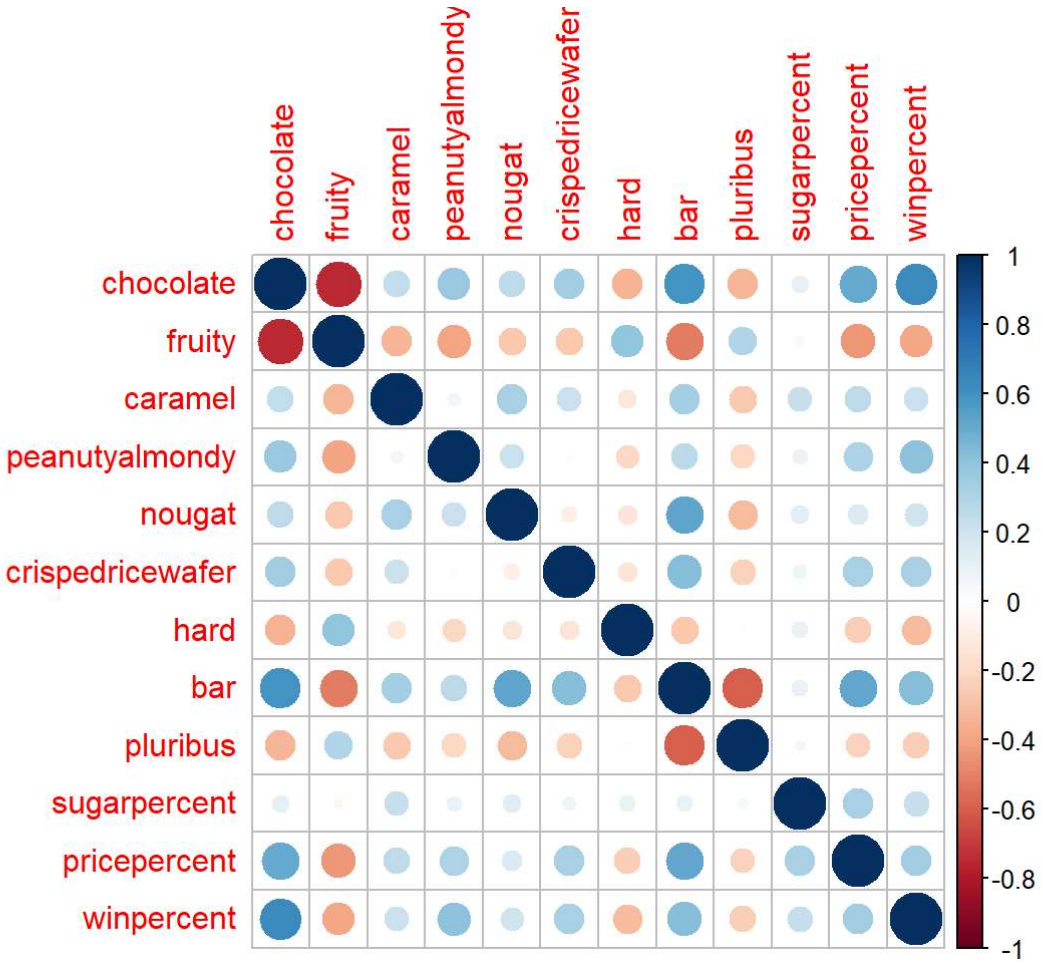
Nik l Nip, Nestie Smarties, Ring pop, Sugar babies, Pop Rocks

#5.Exploring the correlation structure

```
#install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.90 loaded
```

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



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

Chocolate and fruity variables are anti-correlated

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

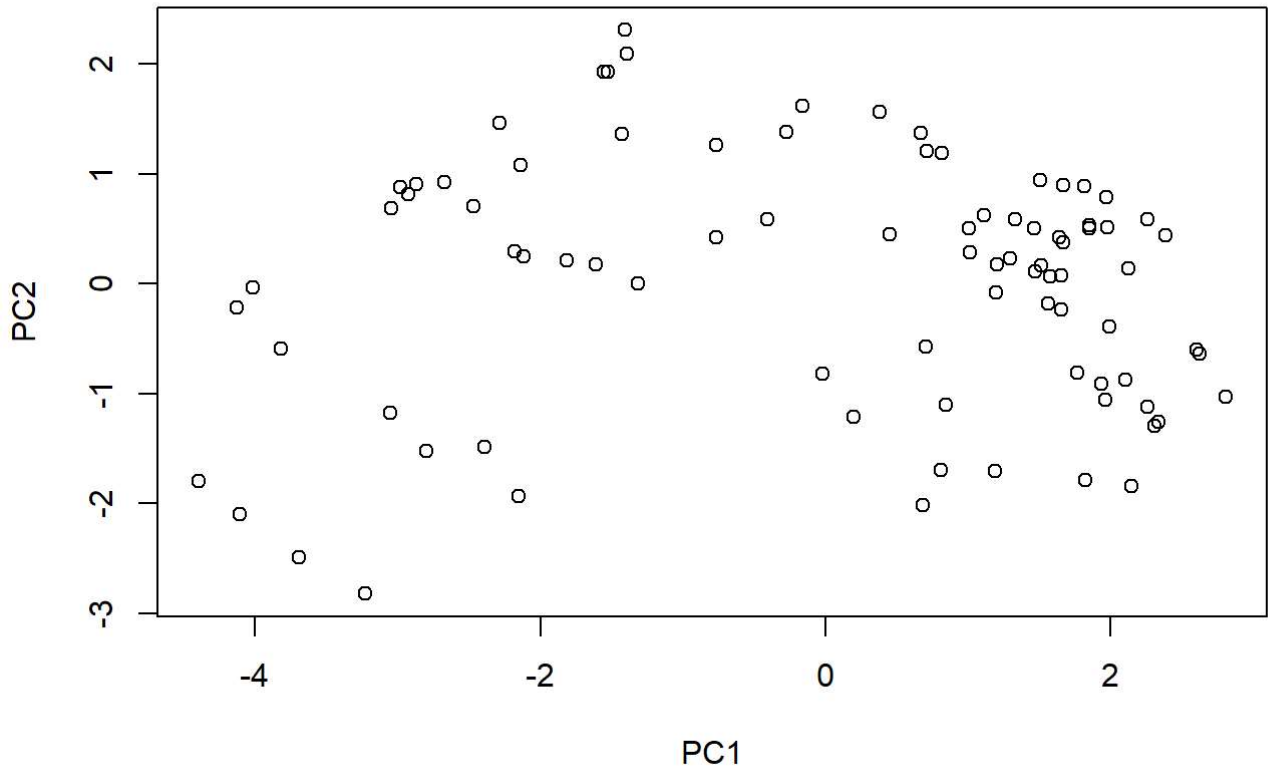
Chocolate and Bar are very positively correlated HINT: Do you like chocolaty fruity candies?

#6. PCA analysis

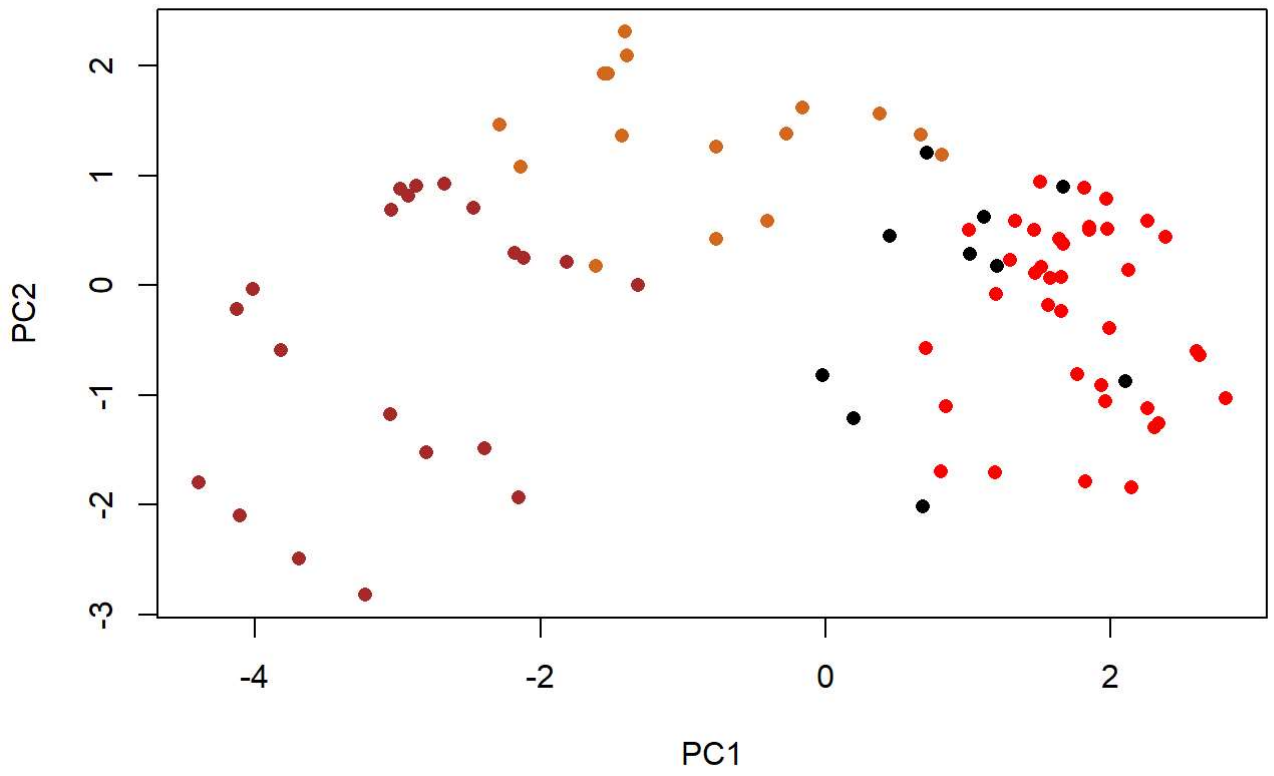
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
## 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
```

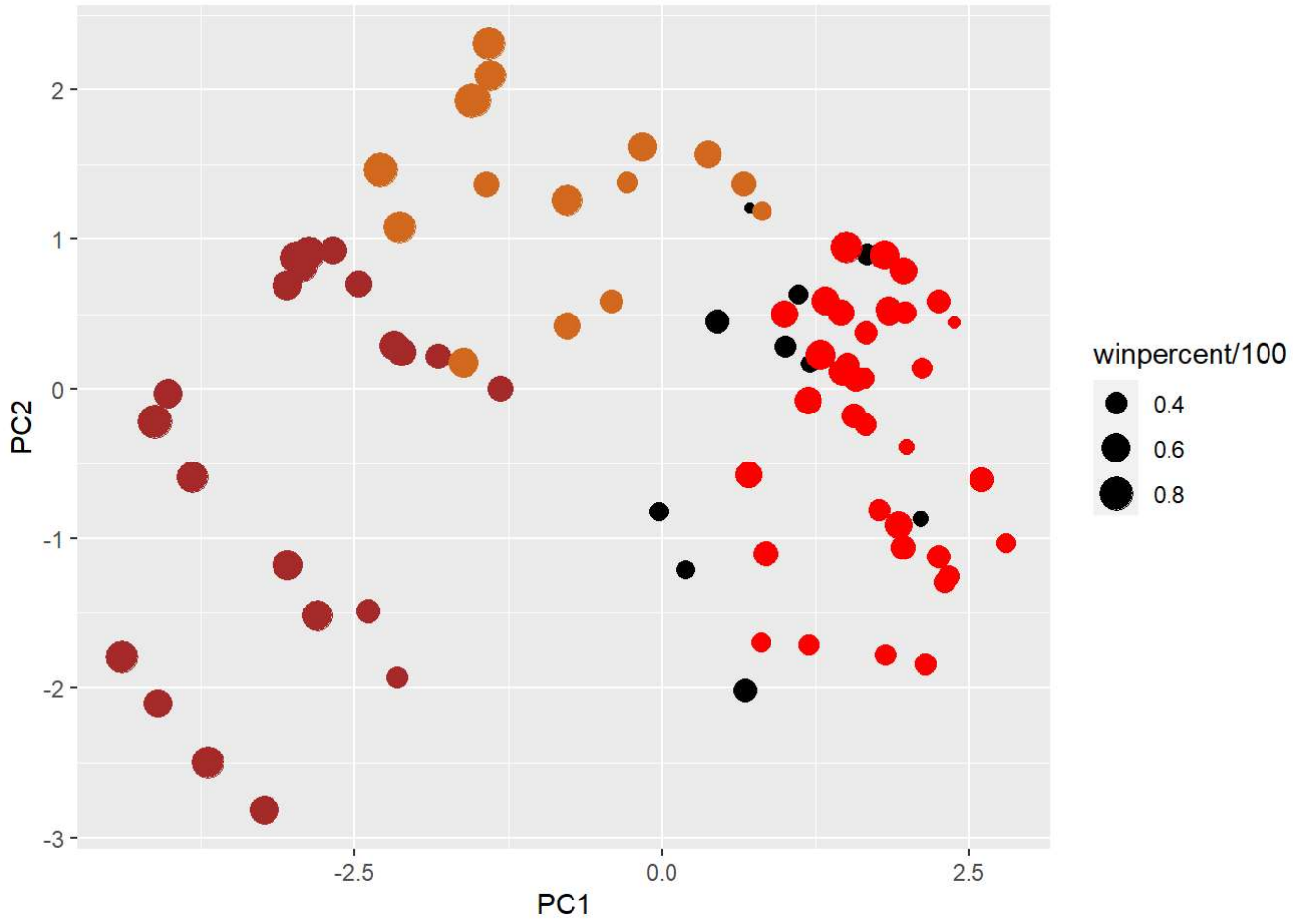
```
plot(pca$x[,1:2])
```




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



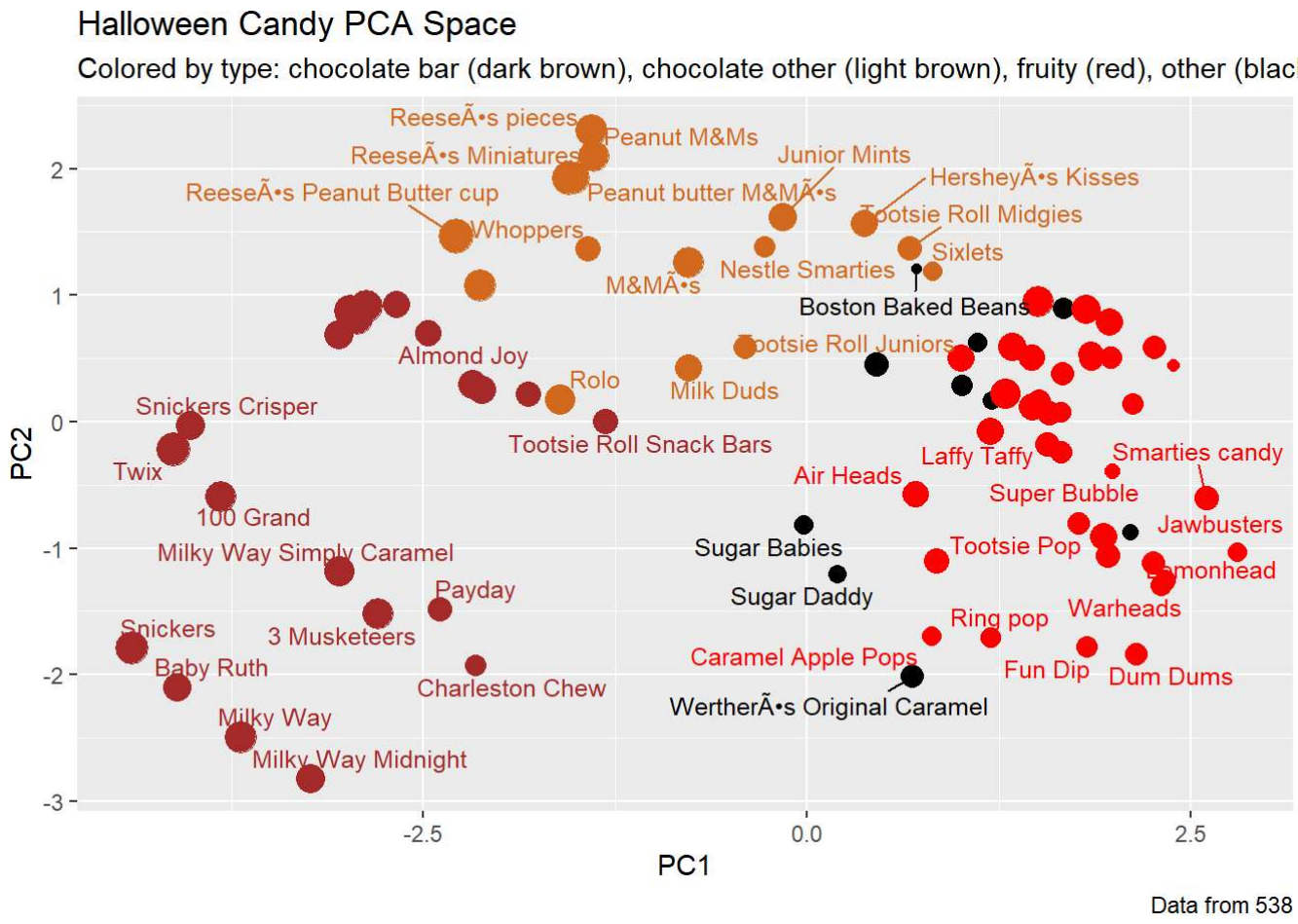

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



Again we can use the ggrepel package and the function ggrepel::geom_text_repel() to label up the plot with non overlapping candy names like. We will also add a title and subtitle like so:

```
library(ggrepel)

p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown), fru
ity (red), other (black)",
       caption="Data from 538")
```

```
## Warning: ggrepel: 41 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
#install.packages("plotly")
library(plotly)
```

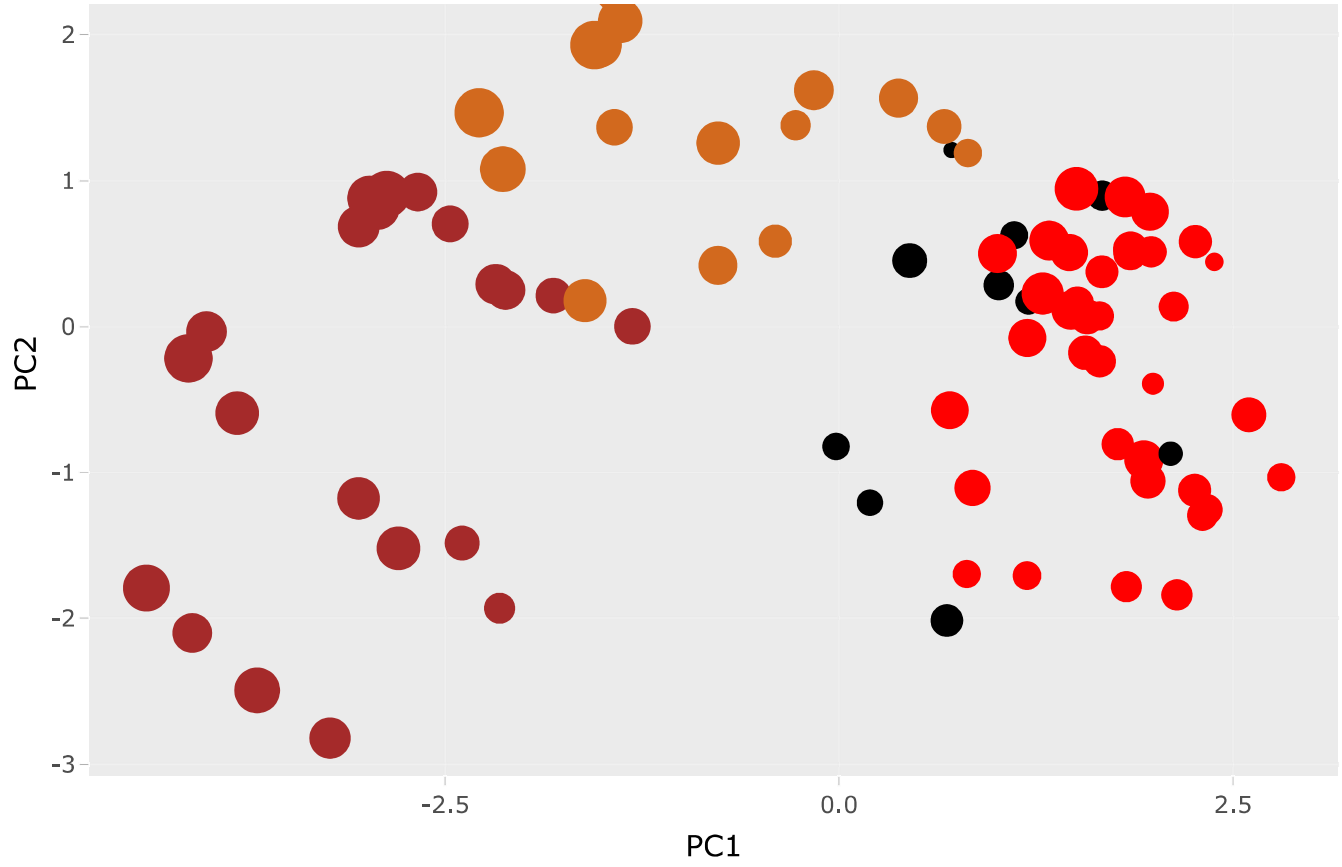
```
##
## Attaching package: 'plotly'
```

```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

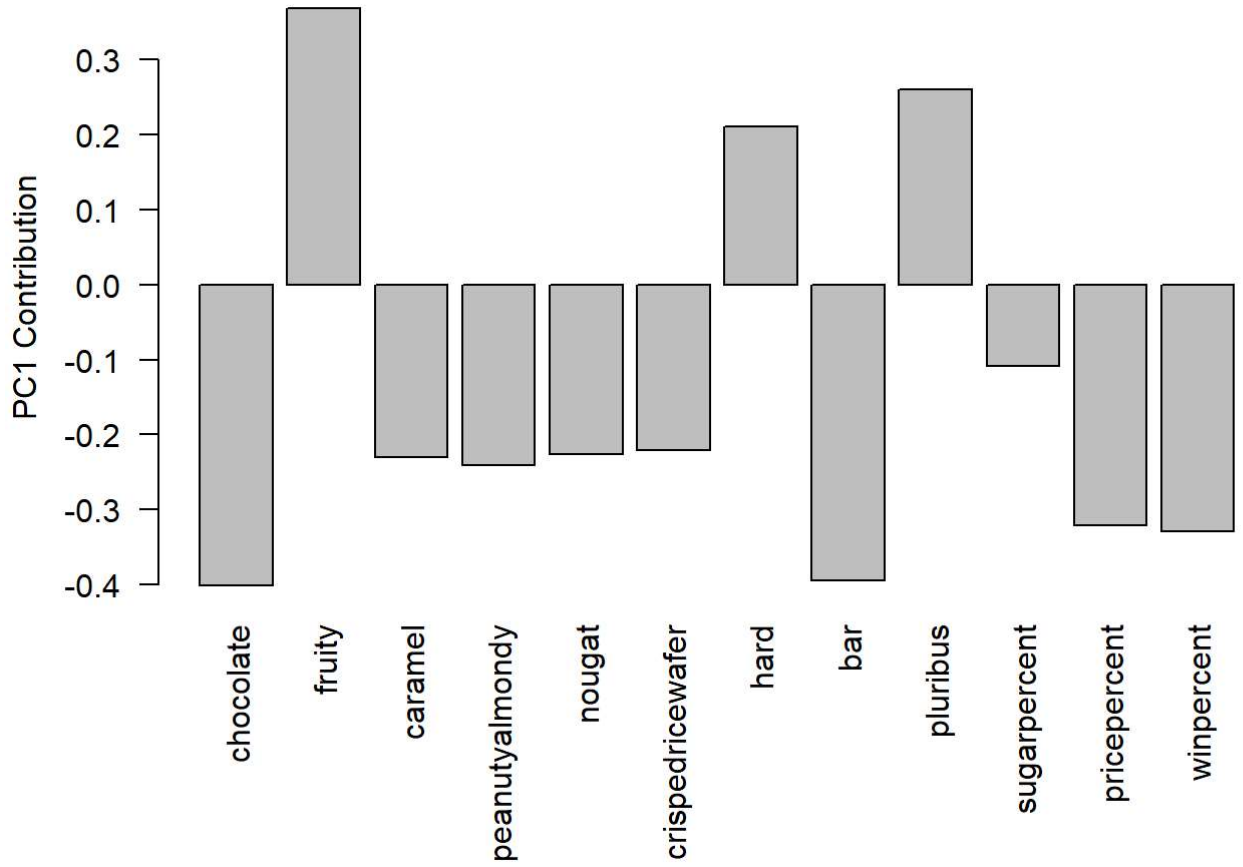
```
## The following object is masked from 'package:stats':
##
##     filter
```

```
## 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")
```



>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 in the positive direction. This makes prefect sense.