

Class 10 :Halloween Mini Project

Peyton Chiu (PID:A18145937)

Table of contents

Background	1
Data Import	1
A quick overview of the dataset	4
Winpercent and Price Percent	11
Exploring the correlation structure	13
Principal Component Analysis	14

Background

As it is nearly Halloween

Data Import

```
candy_file <- read.csv("candy-data.csv")
candy = data.frame(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

```
flectable::flectable(head(candy,10))
```

chocolate	fruity	caramel	peanut	almond	nougat	crisp	rice	wafer	hard	bar	pluribus s
1	0	1	0	0	0	1	0	0	1	0	
1	0	0	0	0	1	0	0	0	1	0	
0	0	0	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	0	0	
1	0	0	1	0	0	0	0	0	1	0	
1	0	1	1	1	1	0	0	0	1	0	
0	0	0	1	0	0	0	0	0	0	1	
0	0	0	0	0	0	0	0	0	0	1	
0	1	1	0	0	0	0	0	0	0	0	

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

```
nrow(candy)
```

```
[1] 85
```

There are 85 candy types in this data set

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

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

```
[1] 38
```

There are 38 fruity candy types in the dataset

```
library(dplyr)
```

Warning: package 'dplyr' was built under R version 4.4.3

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 %>%  
  filter(rownames(candy)=="Twix") %>%  
  select(winpercent)
```

```
      winpercent  
Twix      81.64291
```

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

```
candy["Swedish Fish",]$winpercent
```

```
[1] 54.86111
```

My favorite candy in the data set is Swedish fish and its winpercent value is 54.86%

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

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

```
[1] 76.7686
```

The winpercent of 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
```

The winpercent for Tootsie Roll Snack Bars is 49.65%

A quick overview of the dataset

```
library("skimr")
```

Warning: package 'skimr' was built under R version 4.4.3

```
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	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
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?

Yes, the variable for winpercent is on much higher scale than the other columns shown.

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

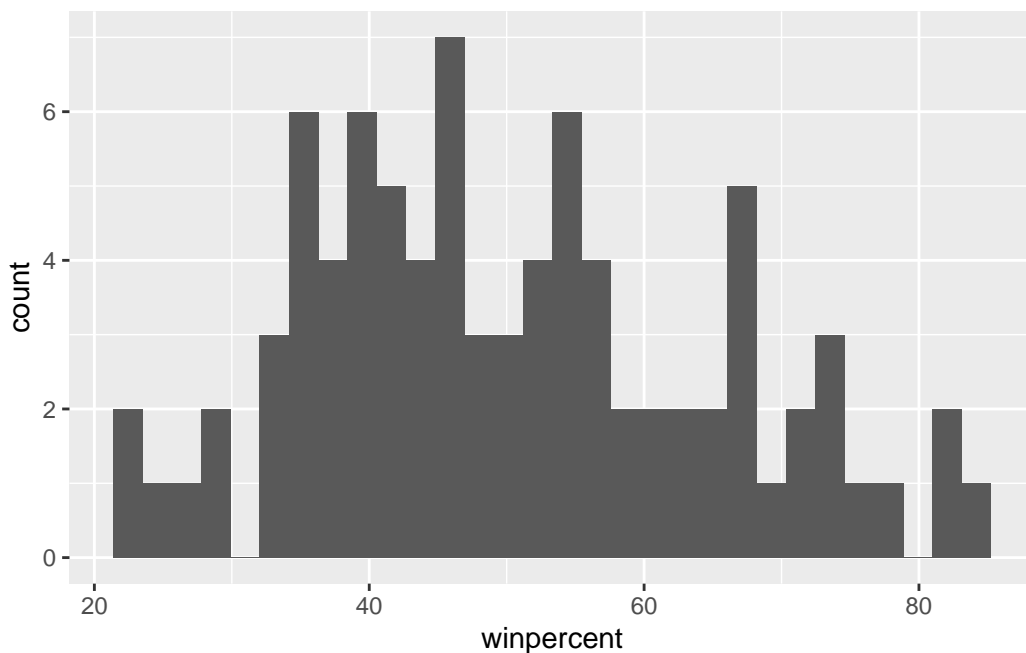
That the candy does not contain chocolate

Q8. Plot a histogram of winpercent values

```
library(ggplot2)
```

Warning: package 'ggplot2' was built under R version 4.4.3

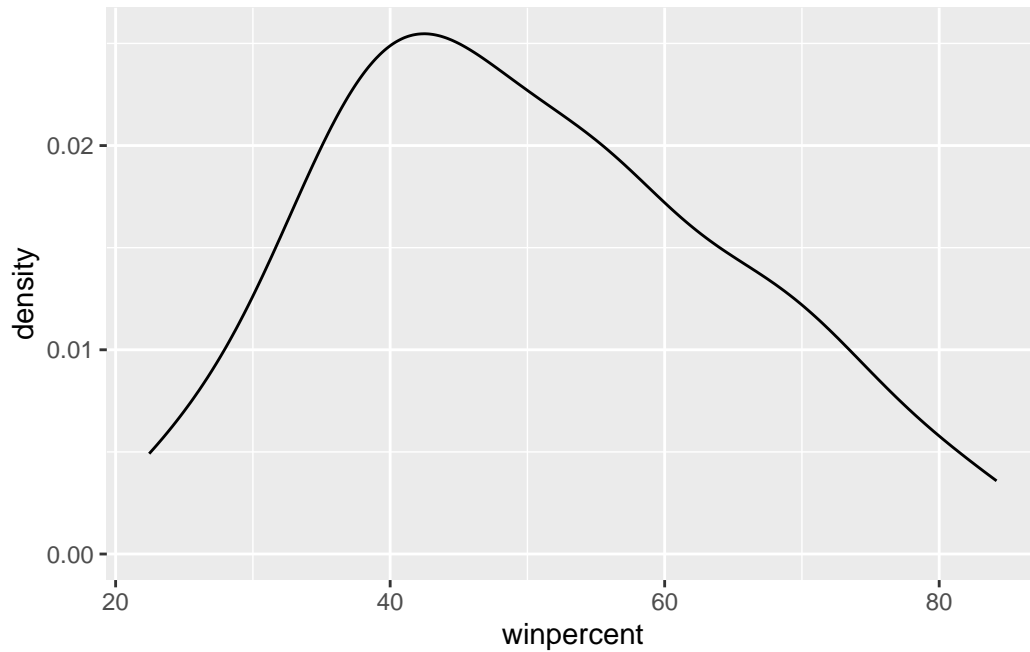
```
ggplot(candy, aes(x = winpercent)) +  
  geom_histogram(bins = 30)
```



Q9. Is the distribution of winpercent values symmetrical?

```
library(ggplot2)

ggplot(candy, aes(x = winpercent)) +
  geom_density()
```



No the distribution is not 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 the distribution is lower than 50 as the median is 47.83

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

```
#1 find all chocolate candy
#2 find their winpercent values
#3 calculate the mean of these values
```

```
#4-6 do the same for fruit candy
#7 compare the means
#8 pick the higher one
```

```
library(dplyr)
choco_candy <- candy %>%
  filter(chocolate==1)

fruity_candy <- candy %>%
  filter(fruity==1)

mean(choco_candy$winpercent)
```

```
[1] 60.92153
```

```
mean(fruity_candy$winpercent)
```

```
[1] 44.11974
```

On average chocolate candy is higher ranked than than fruit candy

Q12. Is this difference statistically significant

```
t.test(choco_candy$winpercent,fruity_candy$winpercent)
```

Welch Two Sample t-test

```
data: choco_candy$winpercent and fruity_candy$winpercent
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
```

These differences are statically significant

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

```
candy %>% arrange(winpercent) %>% head(5)
```

	chocolate	fruity	caramel	peanut	almond	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		
	crisped	rice	wafer	hard	bar	pluribus	sugar	percent
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 lip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters

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

```
candy %>% arrange(desc(winpercent)) %>% head(5)
```

	chocolate	fruity	caramel	peanut	almond	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		
	crisped	rice	wafer	hard	bar	pluribus	sugar	percent
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

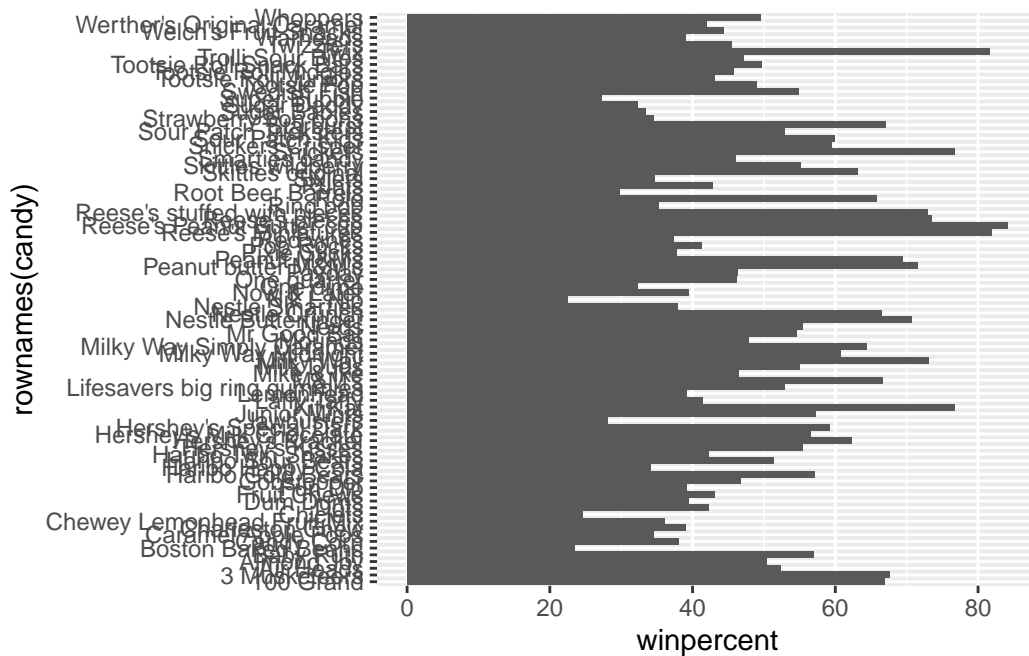
The top 5 all time favorite candy are Reese's Peanut Butter cup, Reese's miniatures, Twix, Kit Kat, Snickers

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

```
library(ggplot2)

library(ggplot2)

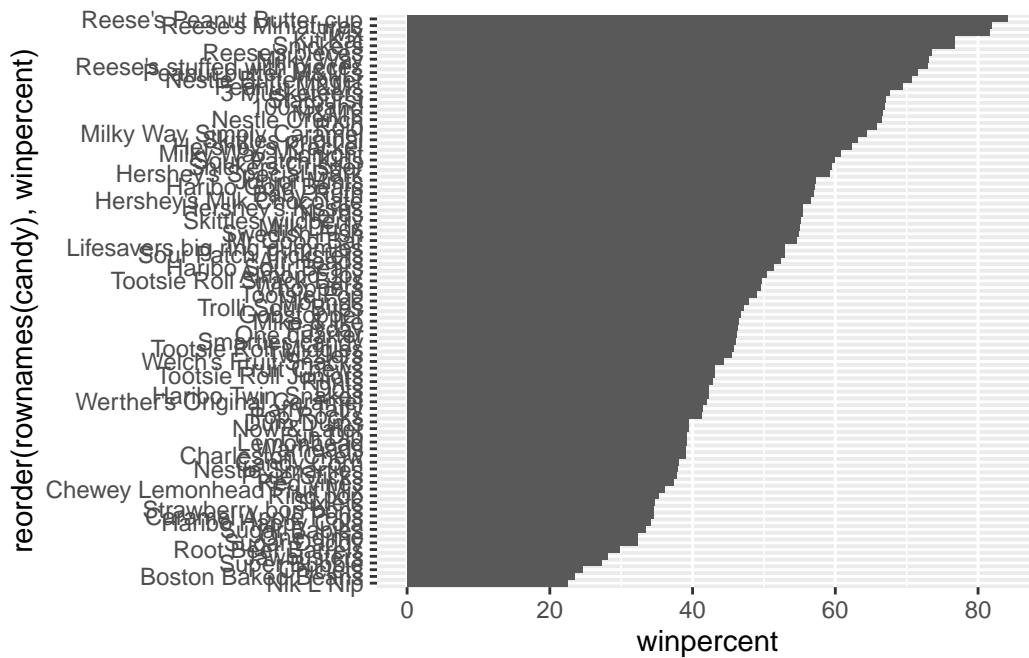
ggplot(candy, aes(winpercent,rownames(candy))) +
  geom_col()
```



Q16. This is quite ugly, use the `reorder()` function to get the bars sorted by `winpercent`?

```
library(ggplot2)

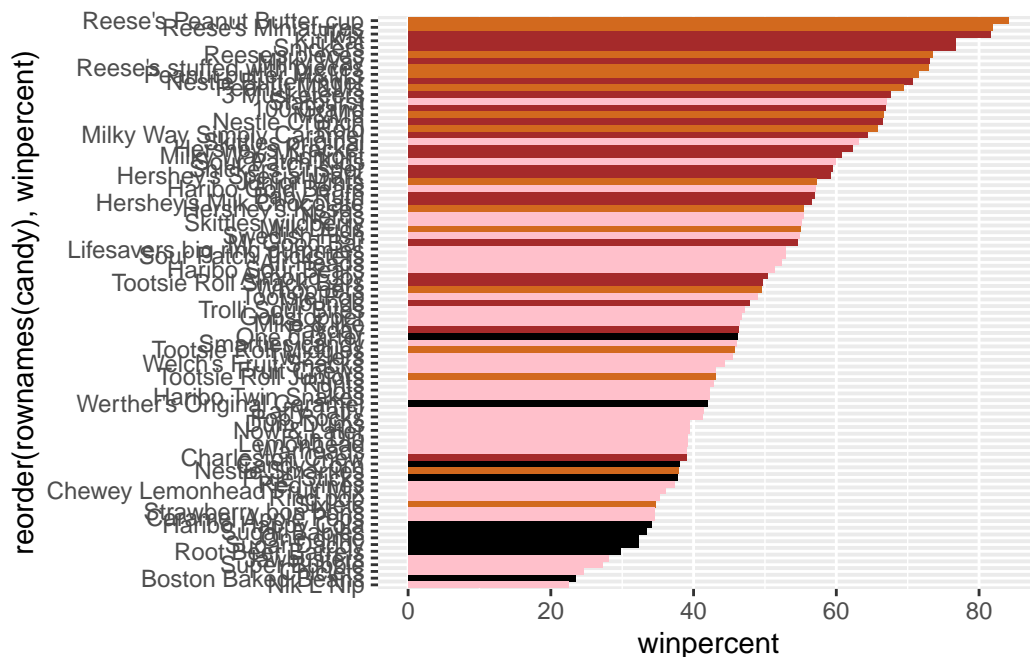
ggplot(candy, aes(winpercent, reorder(rownames(candy), winpercent))) +
  geom_col()
```



Q17. What is the worst ranked chocolate candy?

```
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)
```



The worst ranked chocolate candy was Sixlets

Q18. What is the best ranked fruity candy

The best ranked fruit candy is star bursts

Winpercent and Price Percent

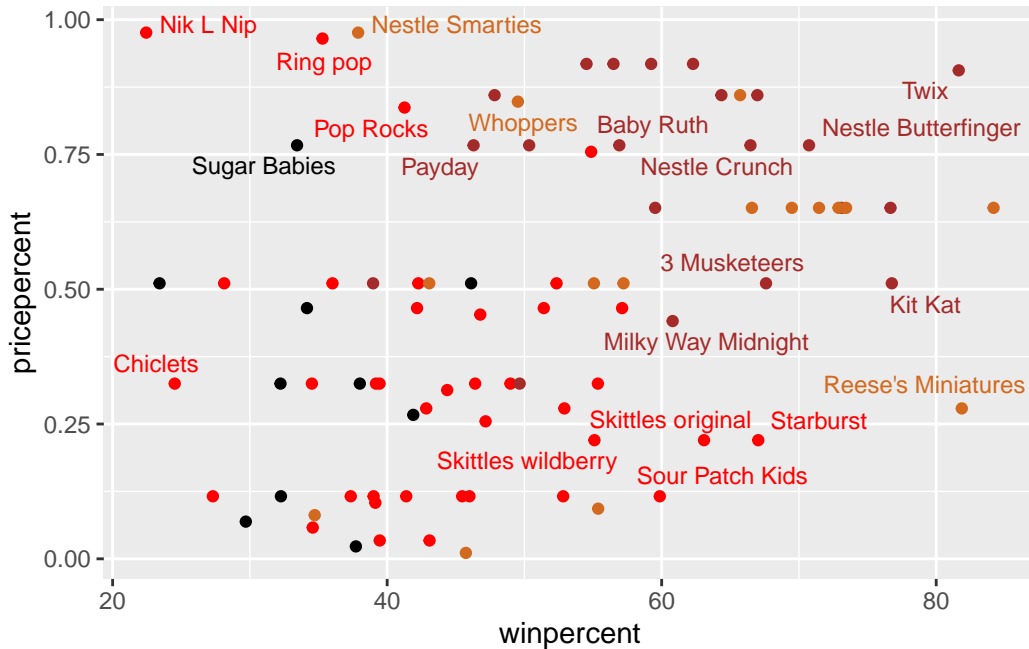
A plot with both variables/columns winpercent and pricepercent

```
library(ggrepel)
```

Warning: package 'ggrepel' was built under R version 4.4.3

```
my_cols[as.logical(candy$fruity)] = "red"
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?

The candy type that is the highest ranked in terms of winpercent for the least money is Reeses miniatures

Q20. 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

The 5 most expensive candies are Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, Hershey's Milk Chocolate. Out of these the least popular one is Nik L Nip

Exploring the correlation structure

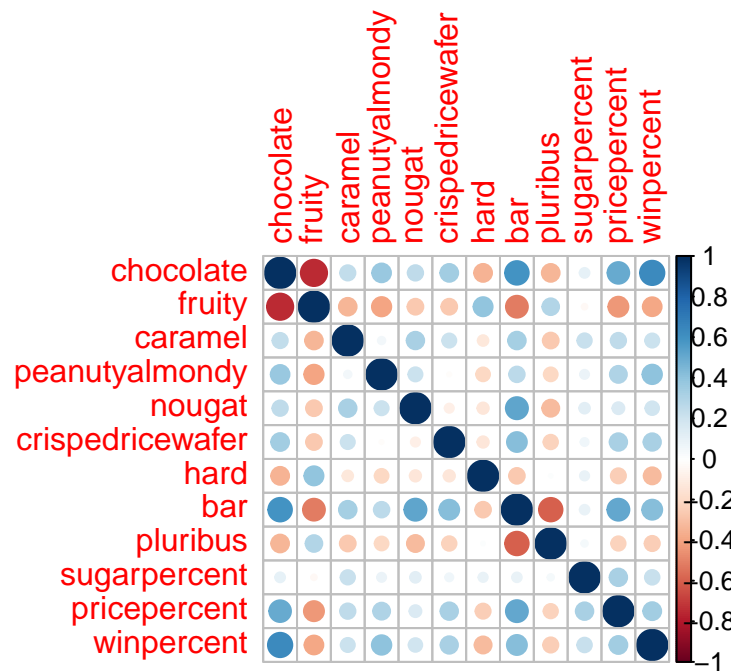
Now that we've explored the dataset a little, we'll see how the variables interact with one another. We'll use correlation and view the results with the `corrplot` package to plot a correlation matrix.

```
library(corrplot)
```

Warning: package 'corrplot' was built under R version 4.4.3

corrplot 0.95 loaded

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



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

The two variables that seem anti correlated are chocolate and fruity

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

The two variables that seem the most positively correlated are chocolate and winpercent

Principal Component Analysis

The function to use this is called `prcomp()` with an optional `scale` argument .

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

Importance of components:

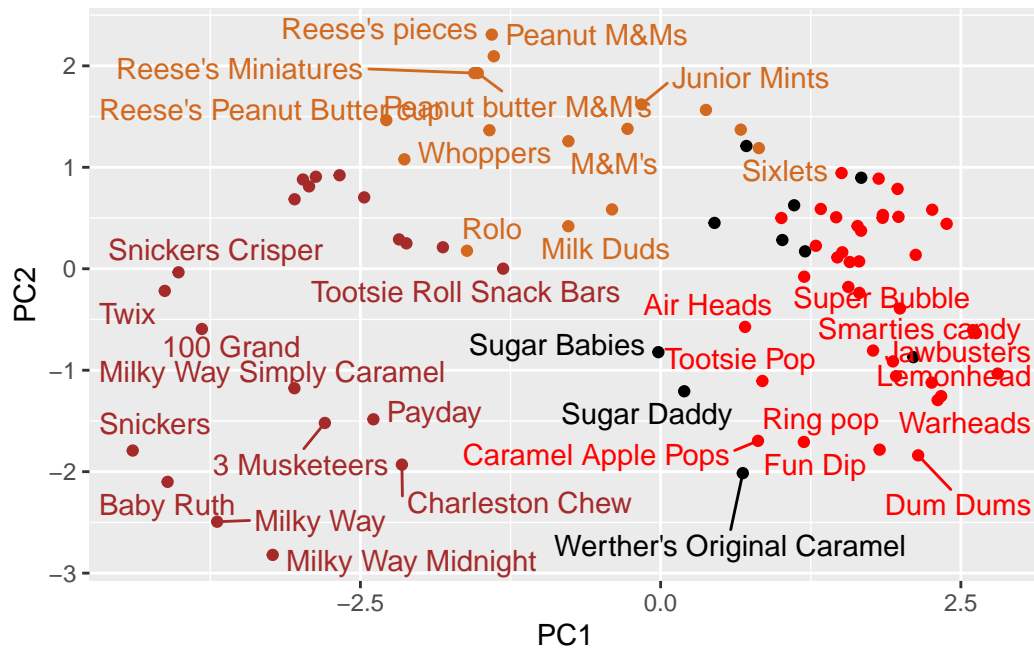
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard deviation	2.0788	1.1378	1.1092	1.07533	0.9518	0.81923	0.81530
Proportion of Variance	0.3601	0.1079	0.1025	0.09636	0.0755	0.05593	0.05539
Cumulative Proportion	0.3601	0.4680	0.5705	0.66688	0.7424	0.79830	0.85369

	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

Our main PCA results figure

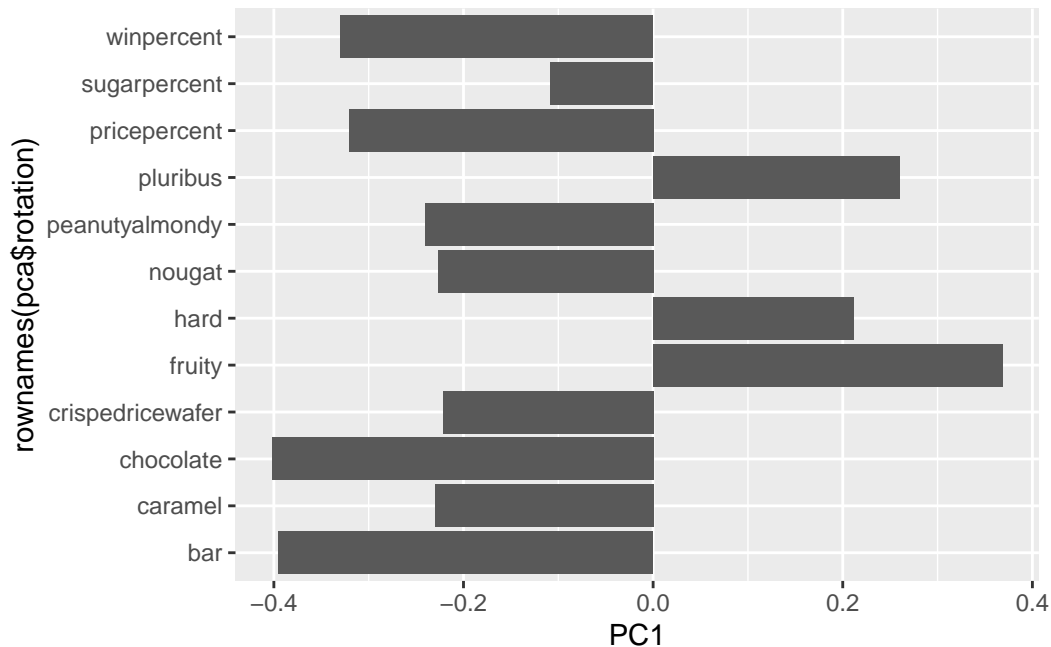
```
ggplot(pca$x)+
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=my_cols)+
  geom_text_repel(col=my_cols)
```

Warning: ggrepel: 48 unlabeled data points (too many overlaps). Consider increasing `max.overlaps`



We should also examine the loadings or contributions of the original variables to the 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?

The original variables that are picked up strongly by PC1 in the positive direction are fruity, hard, and pluribus. These makes sense as these characteristics are typically of fruity candies which usually comes in groups and are hard.

Interactive plots that can be zoomed on and “brushed” over can be made with the **plotly** package. It’s output is interactive and will not render to pdf

```
p<-ggplot(pca$x)+
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=my_cols)+
  geom_text_repel(col=my_cols)
```

```
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`

```
##plotly(p)
```