

Class 9: Halloween Candy Mini-Project

Hyejeong Choi (PID: A16837133)

Table of contents

Data Import	1
What is your favorite candy?	2
Exploratory Analysis	3
Overall Candy Rankings	6
Winpercent VS Pricepercent	12
Correlation Structure	13
Principal Component Analysis (PCA)	15

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

Data Import

```
candy_file <- "candy-data.txt"

candy = read.csv(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

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

```
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

What is your favorite candy?

Q3. What is your favorite candy in the dataset and what is its winpercent value?

```
candy["M&M's",]$winpercent
```

```
[1] 66.57458
```

Q4. What is the winpercent value for “Kit Kat”?

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

```
[1] 76.7686
```

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

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

```
[1] 49.6535
```

Exploratory Analysis

We can use **skimr** package to get a quick overview of a given dataset. This can be useful for the first time you encounter a new dataset.

```
skimr::skim(candy)
```

Table 1: Data summary

Name	candy
Number of rows	85
Number of columns	12
<hr/>	
Column type frequency: numeric	12
<hr/>	
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?

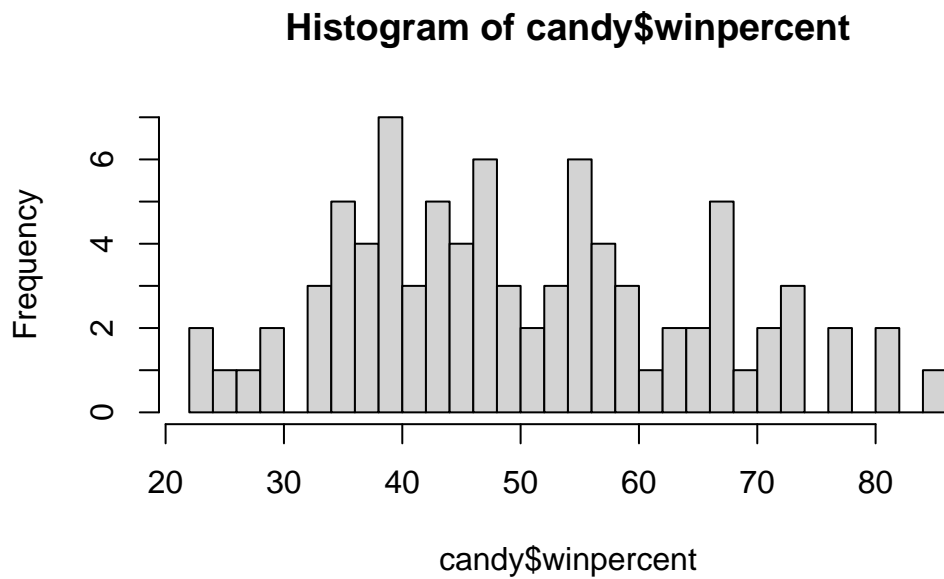
The last column `candy$winpercent` is on a different scale to all others.

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

The one represents the candy that is chocolate, and the zero represents the candy that is not chocolate.

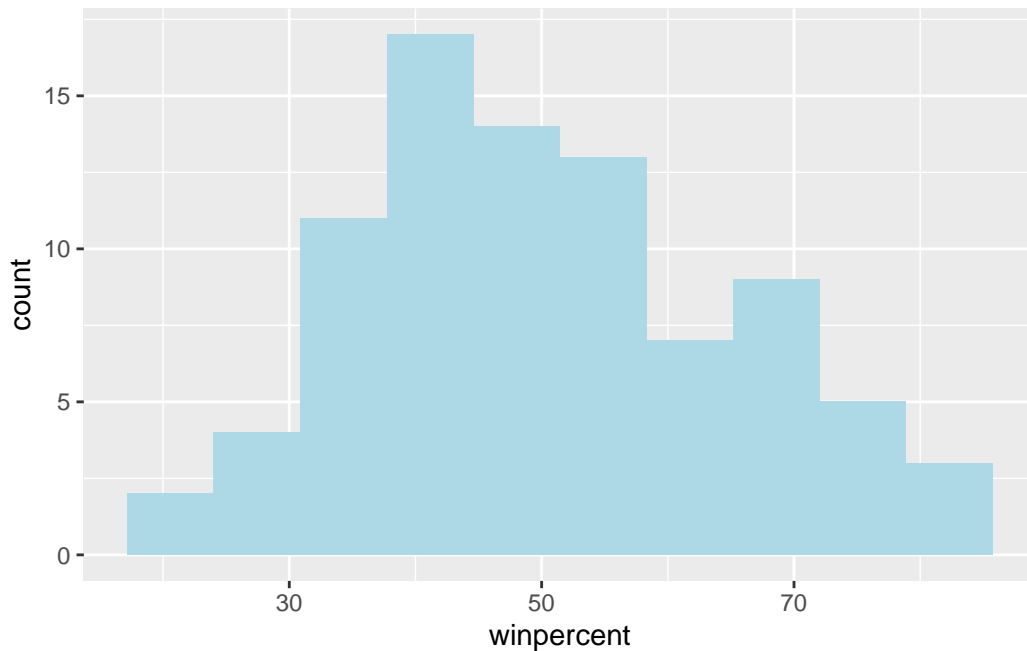
Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent, breaks=30)
```



```
library(ggplot2)

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



Q9. Is the distribution of winpercent values symmetrical?

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 below 50% because the median is 47.83%.

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

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

```
fruit.win2 <- candy[as.logical(candy$fruity),]$winpercent
mean(fruit.win2)
```

```
[1] 44.11974
```

The chocolate candy ranks higher on average than fruity candy.

Q12. Is this difference statistically significant?

```
ans <- t.test(choc.win, fruity.win)
ans
```

Welch Two Sample t-test

```
data:  choc.win and fruity.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, it is statistically significant with a p-value of 2.8713778×10^{-8} .

Overall Candy Rankings

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

There are two related functions that can help here. One is the classic `sort()` and `order()`.

```
inds <- order(candy$winpercent)
head(candy[inds,])
```

	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
Root Beer Barrels	0	0	0		0	0

	crisped	rice	wafer	hard	bar	pluribus	sugar	percent	price	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
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

The least liked candy is Nik L Nip.

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

```
head(candy[order(candy$winpercent, decreasing=TRUE),], n=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		0.720
Reese's Miniatures				0	0	0		0.034
Twix				1	0	1		0.546

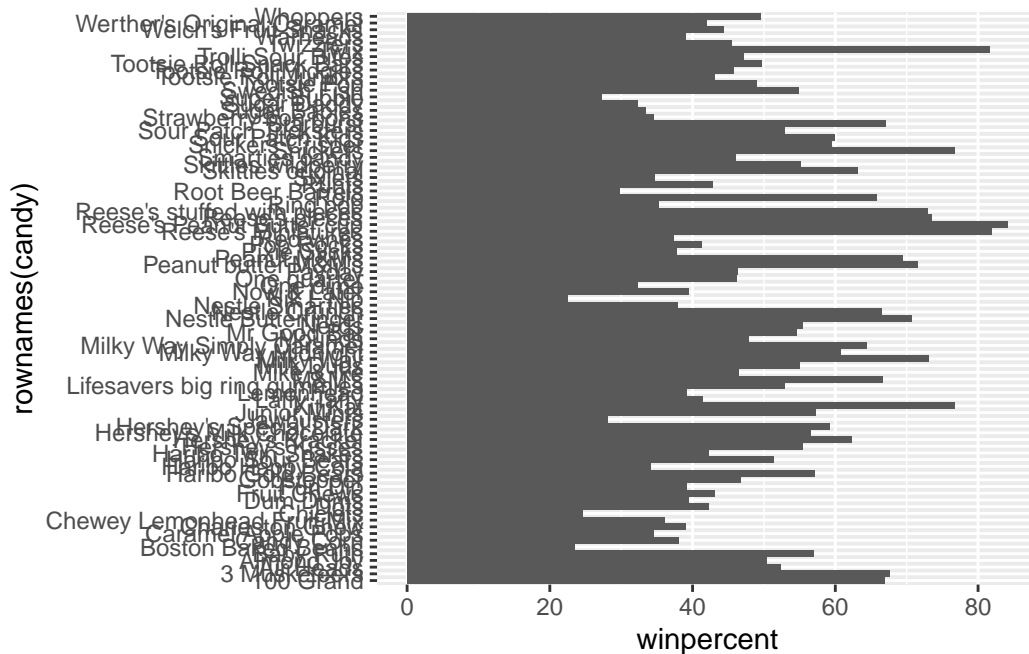
Kit Kat	1	0	1	0	0.313
Snickers	0	0	1	0	0.546

	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

Make a bar plot with ggplot and order it by winpercent values

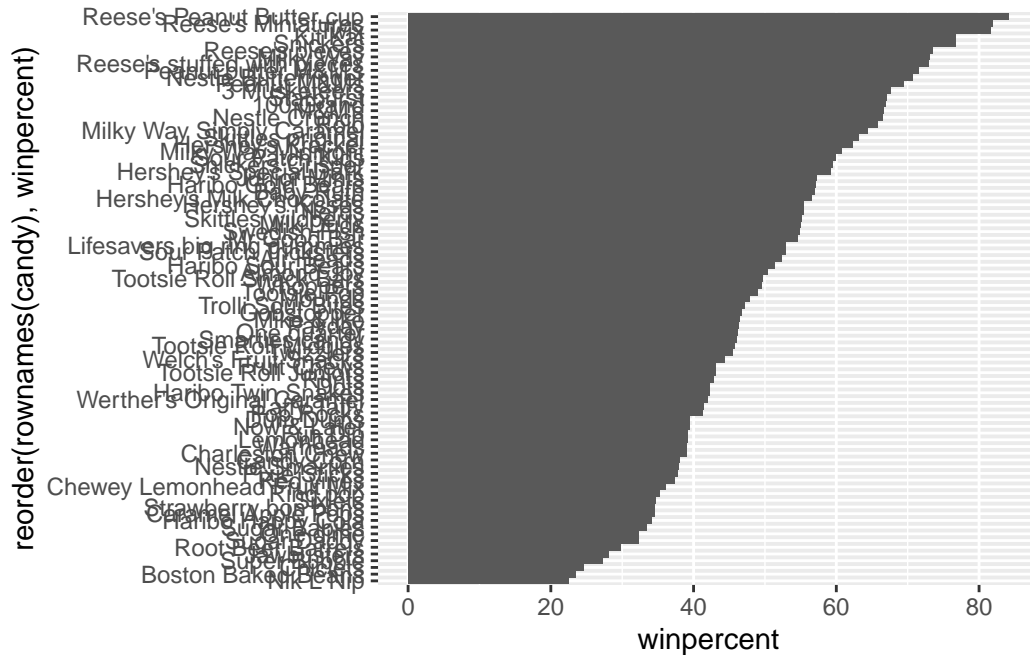
Q15. Make a first barplot of candy ranking based on winpercent values.

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



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

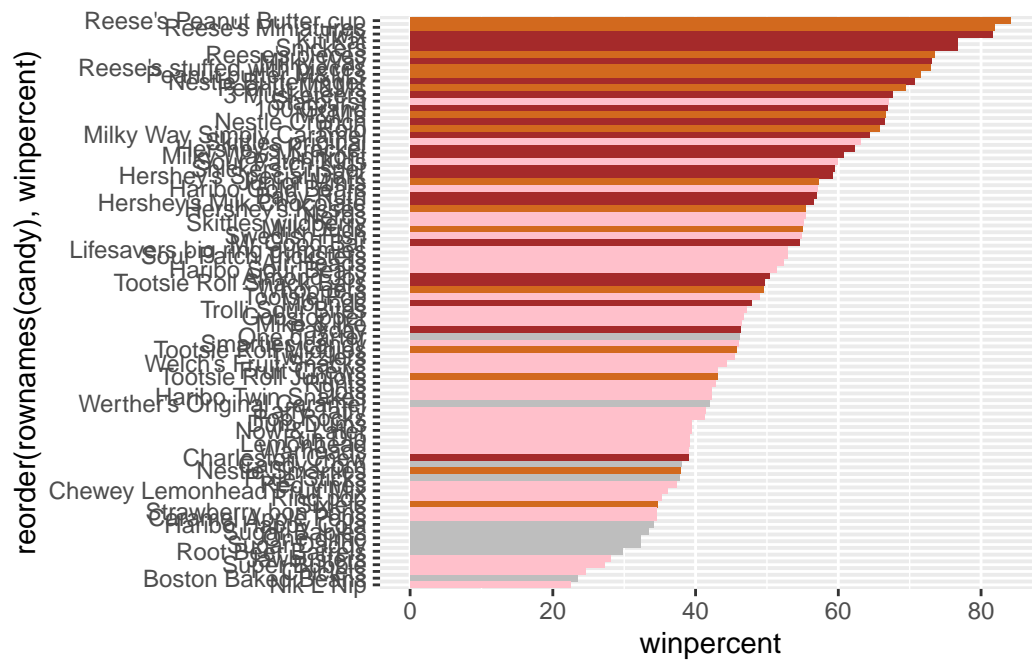
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col()
```

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

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

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



```
# ggsave("mybarplot.png", width=3, height=8)
```

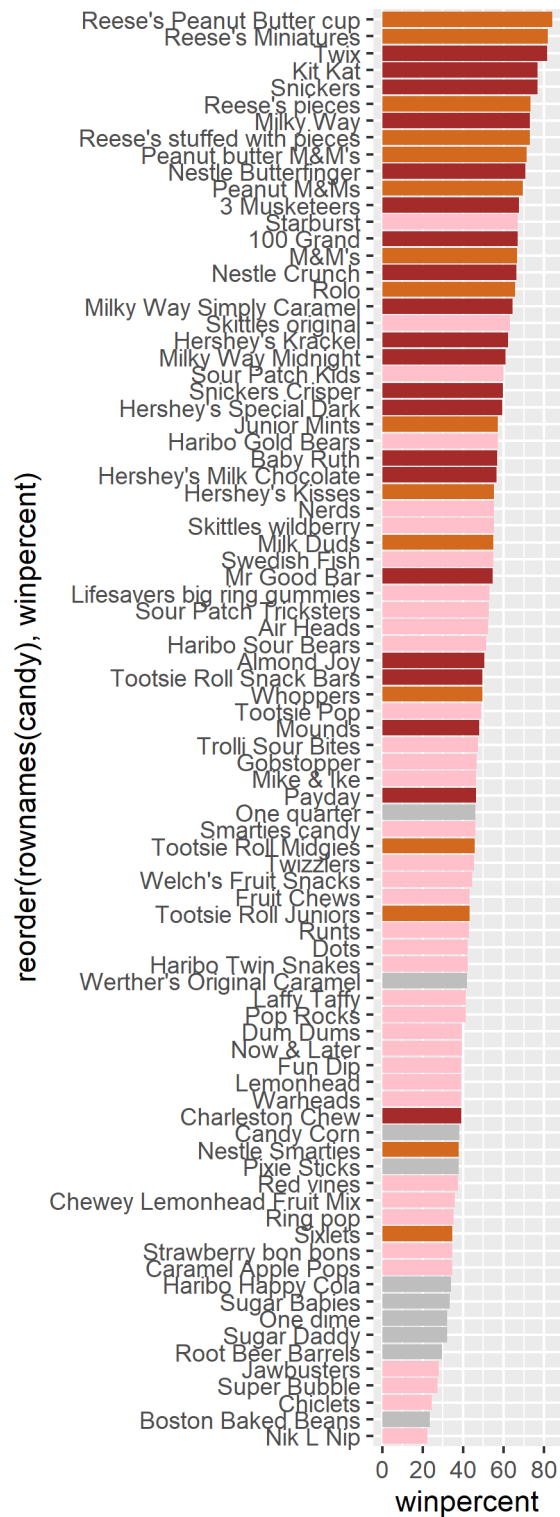


Figure 1: My silly barplot image

Q17. What is the worst ranked chocolate candy?

Sixlets

Q18. What is the best ranked fruity candy?

Starburst

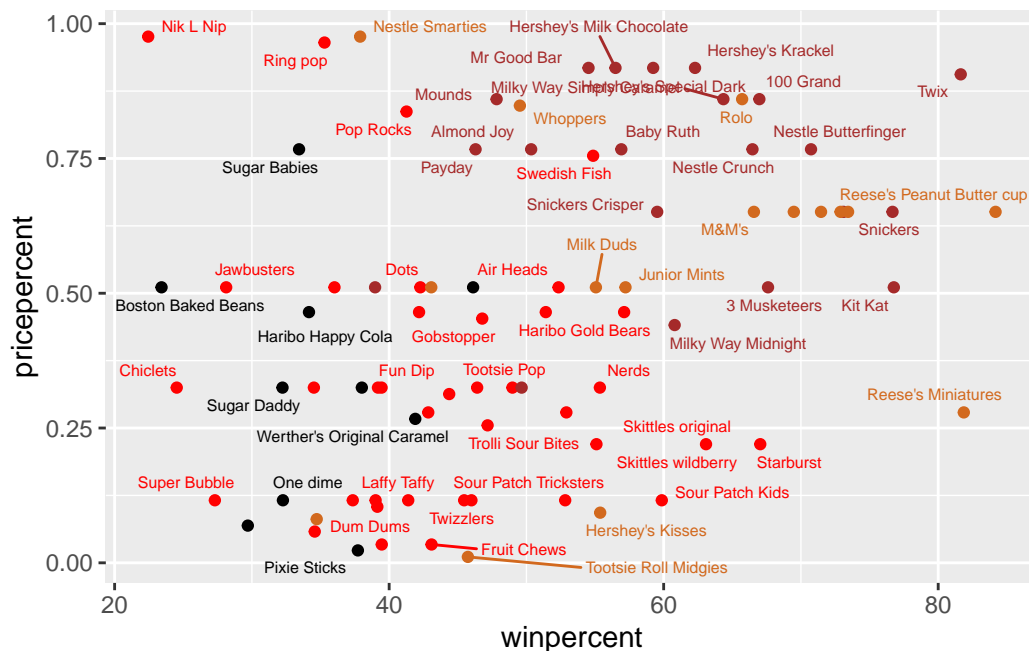
Winpercent VS Pricepercent

```
library(ggrepel)
# The pink is too washed out
my_cols <- rep("black", nrow(candy))
my_cols[as.logical(candy$chocolate)] <- "chocolate"
my_cols[as.logical(candy$fruity)] <- "red"
my_cols[as.logical(candy$bar)] <- "brown"

# 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=2.2, max.overlaps = 8)
```

Warning: ggrepel: 26 unlabeled data points (too many overlaps). Consider increasing max.overlaps



Correlation Structure

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

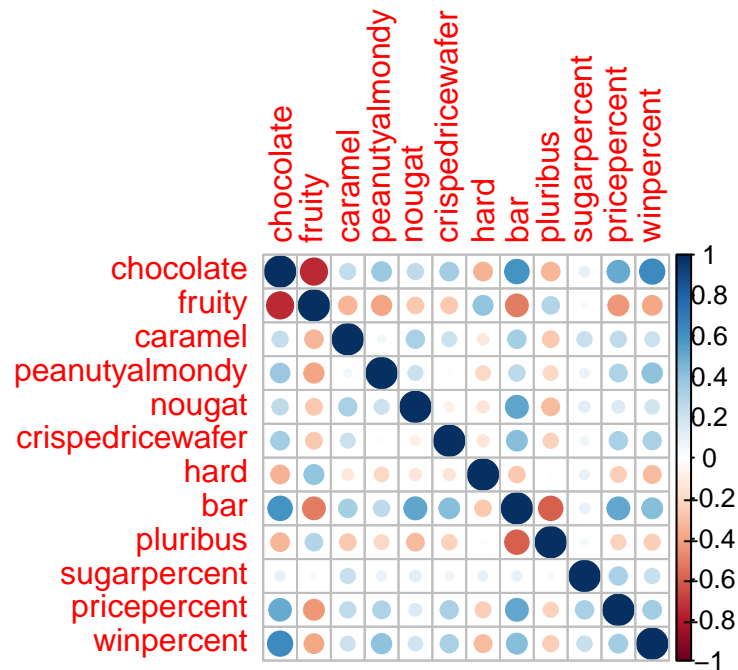
	chocolate	fruity	caramel	peanutyalmondy	nougat
chocolate	1.0000000	-0.74172106	0.24987535	0.37782357	0.25489183
fruity	-0.7417211	1.00000000	-0.33548538	-0.39928014	-0.26936712
caramel	0.2498753	-0.33548538	1.00000000	0.05935614	0.32849280
peanutyalmondy	0.3778236	-0.39928014	0.05935614	1.00000000	0.21311310
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	

```
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 negatively correlated.

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

```
[1] -0.74
```

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

Excluding the correlation between variables that are the same on both axis, the chocolate and winpercent variables have the most positive correlation.

Principal Component Analysis (PCA)

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

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

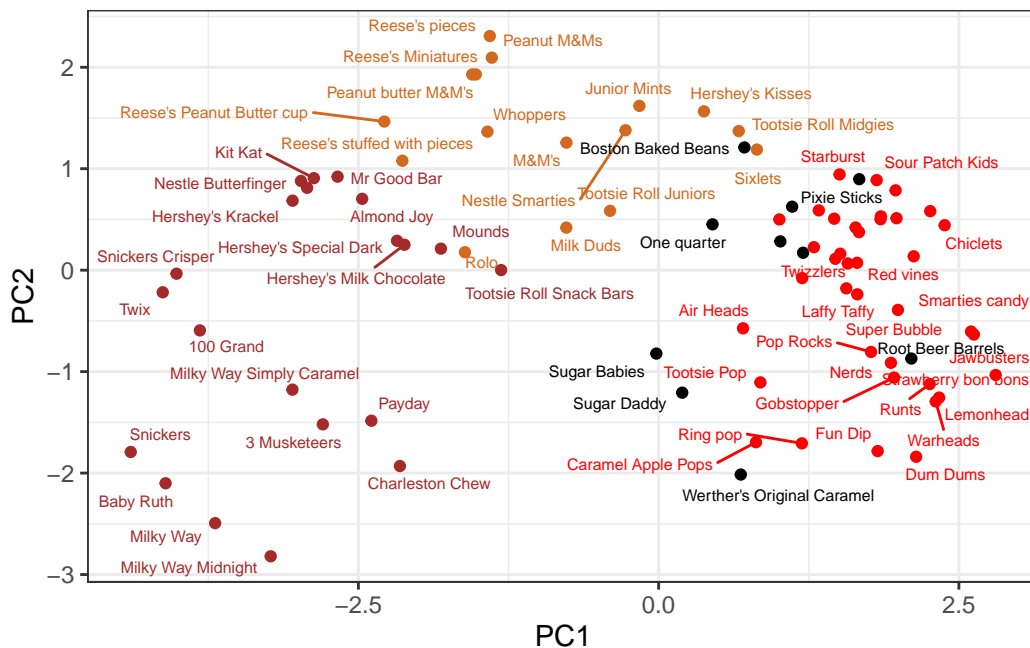
Importance of components:

	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		

First main result figure is my “PCA plot”

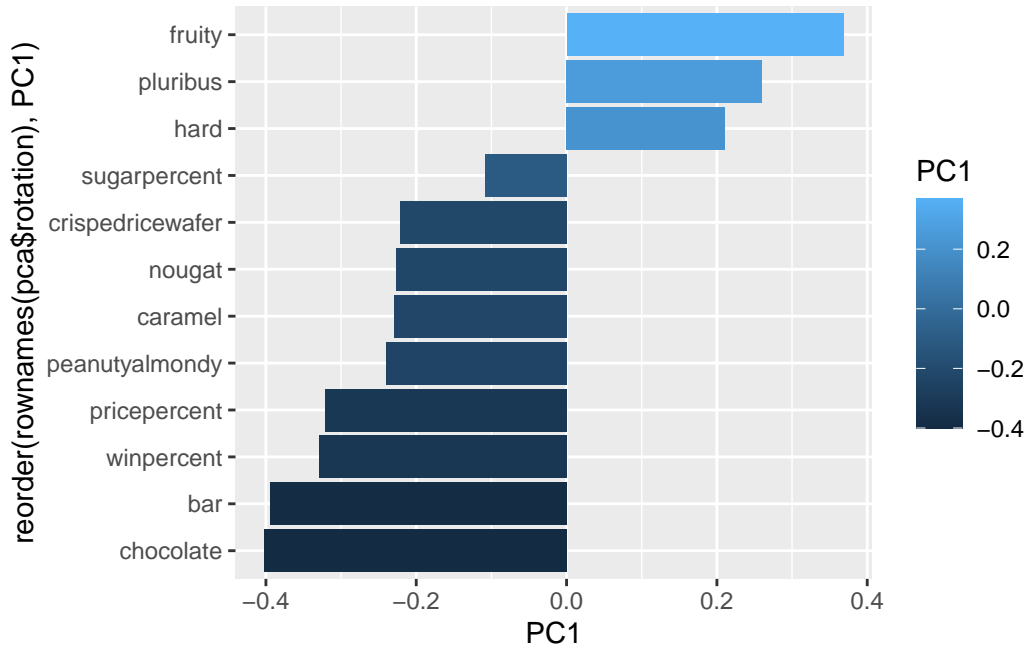
```
ggplot(pca$x) +
  aes(PC1, PC2, label=rownames(pca$x)) +
  geom_point(col=my_cols) +
  geom_text_repel(size=2, max.overlaps = 10, col=my_cols) +
  theme_bw()
```

Warning: ggrepel: 20 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.


```
ggplot(pca$rotation) +
  aes(PC1, reorder(rownames(pca$rotation), PC1), fill=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 all in the positive direction. Yes, this does make sense because these are the variables that were least likely to have a positive correlation with the variable “chocolate”.