

# class10: Halloween Mini-Project

Vivian Pham

```
#install webshot for pdf rendering
#webshot::install_phantomjs()
```

## 1. Importing candy data

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

	chocolate	fruity	caramel	peanut	almond	nougat	crisped	rice	wafer
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	sugar	percent	price	percent	win	percent
100 Grand	0	1	0		0.732		0.860	66.97	173
3 Musketeers	0	1	0		0.604		0.511	67.60	294
One dime	0	0	0		0.011		0.116	32.26	109
One quarter	0	0	0		0.011		0.511	46.11	650
Air Heads	0	0	0		0.906		0.511	52.34	146
Almond Joy	0	1	0		0.465		0.767	50.34	755

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

```
dim(candy)
```

```
[1] 85 12
```

Answer: There are 85 different candy types in this dataset.

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

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

```
[1] 38
```

Answer: There are 38 fruity candy types in the dataset.

**2. What is your favorite candy?**

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

```
[1] 81.64291
```

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

```
candy["Dum Dums", ]$winpercent
```

```
[1] 39.46056
```

Answer: My favorite candy in the dataset is “Dum Dums” and its winpercent value is 39.46056.

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

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

```
[1] 76.7686
```

The winpercent value for “Kit Kat” is 76.7686.

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

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

```
[1] 49.6535
```

Answer: The winpercent value for “Tootsie Roll Snack Bars” is 49.6535.

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

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

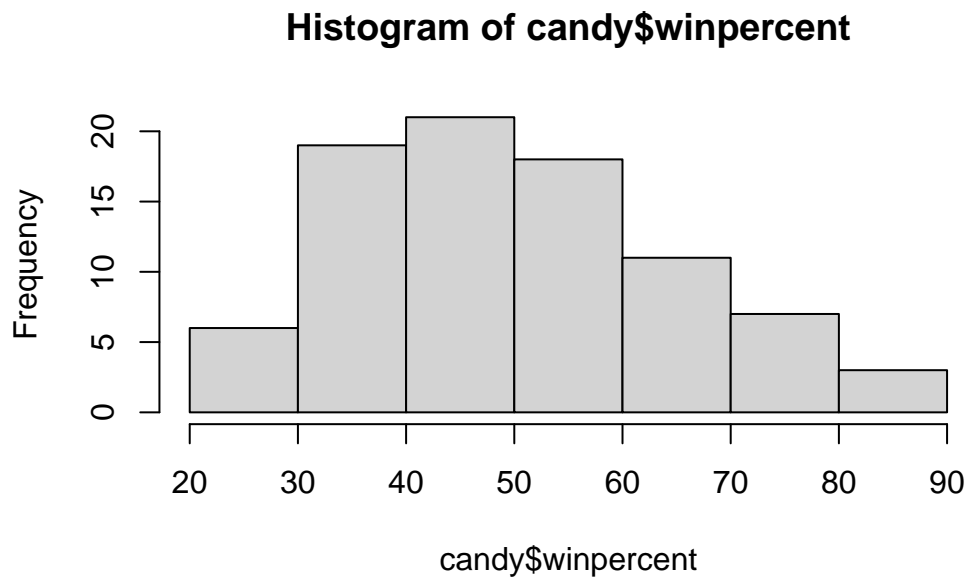
Answer: The winpercent variable looks to be on a different scale to the majority of the other variables in the dataset.

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

Answer: For the candy\$chocolate column, the zero and one represent whether the candy contains chocolate (1) or does not have chocolate (0).

**Q8. Plot a histogram of winpercent values.**

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



**Q9. Is the distribution of winpercent values symmetrical?**

Answer: No, the distribution of winpercent values is not symmetrical - the distribution is right-skewed.

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

Answer: The center of the distribution is below 50%.

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

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

```
[1] 60.92153
```

```
fruity_avg <- mean(candy$winpercent[as.logical(candy$fruity)])  
fruity_avg
```

```
[1] 44.11974
```

Answer: On average, chocolate candy is higher ranked than fruit candy.

### Q12. Is this difference statistically significant?

```
t.test(candy$winpercent[as.logical(candy$chocolate)], candy$winpercent[as.logical(candy$fr
```

Welch Two Sample t-test

```
data: candy$winpercent[as.logical(candy$chocolate)] and candy$winpercent[as.logical(candy$fr
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
```

Answer: The p-value of 2.871e-08 is less than the significance level of 0.05; this difference is statistically different.

## 3. Overall Candy Rankings

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

```
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

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

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

```
#approach 1
tail(candy[order(candy$winpercent),], n=5)
```

	chocolate	fruity	caramel	peanut	almondy	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
	crisped	rice	wafer	hard	bar	pluribus
Snickers		0	0	1		0.546
Kit Kat		1	0	1		0.313
Twix		1	0	1		0.546
Reese's Miniatures		0	0	0		0.034
Reese's Peanut Butter cup		0	0	0		0.720
	price	percent	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			

```
#approach 2
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) %>% tail(5)
```

	chocolate	fruity	caramel	peanut	almondy	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

	crisped	rice	wafer	hard bar	pluribus	sugar	percent
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

	price	percent	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

```
# I prefer approach 1 because it's much simpler as it uses base R functions.
```

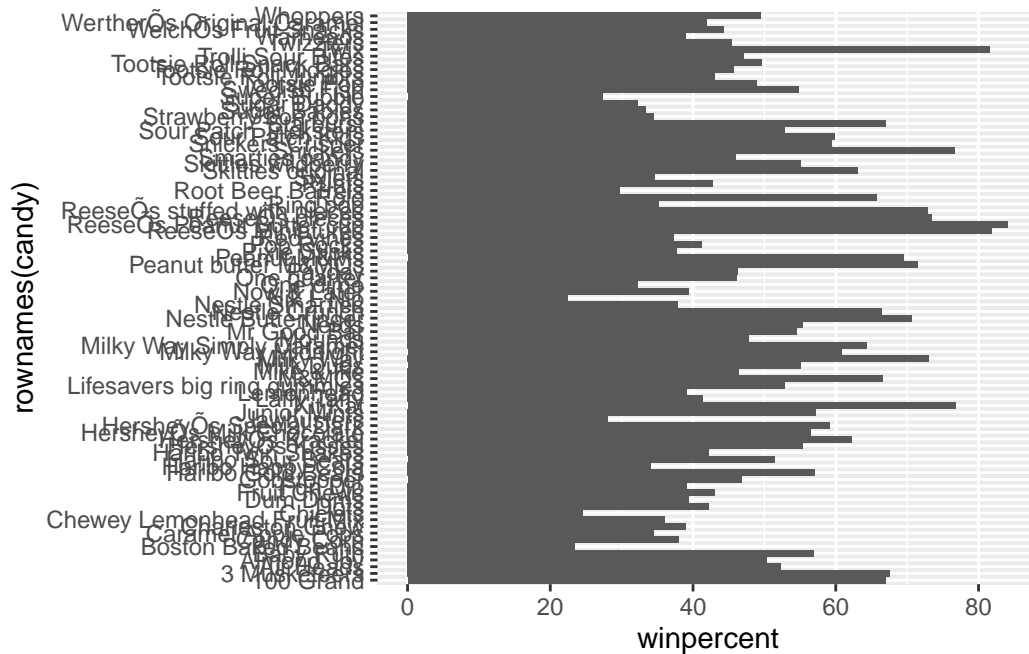
Answer: The top 5 all time favorite candy types out of this set are Snickers, Kit Kat, Twix, Reese's Miniatures, and Reese's Peanut Butter cup.



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

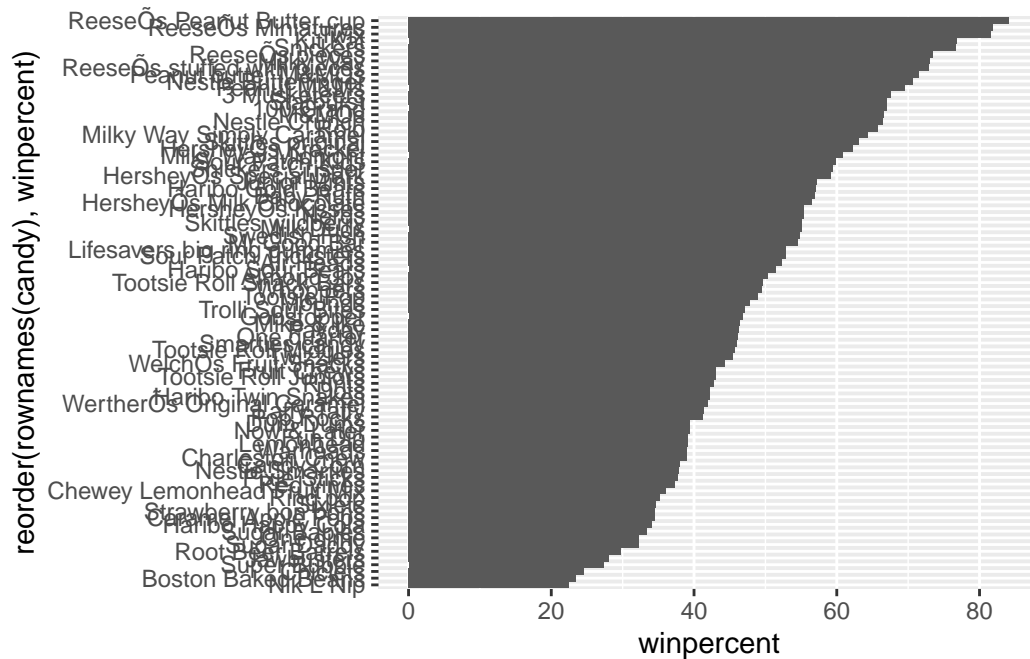
```
library(ggplot2)

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



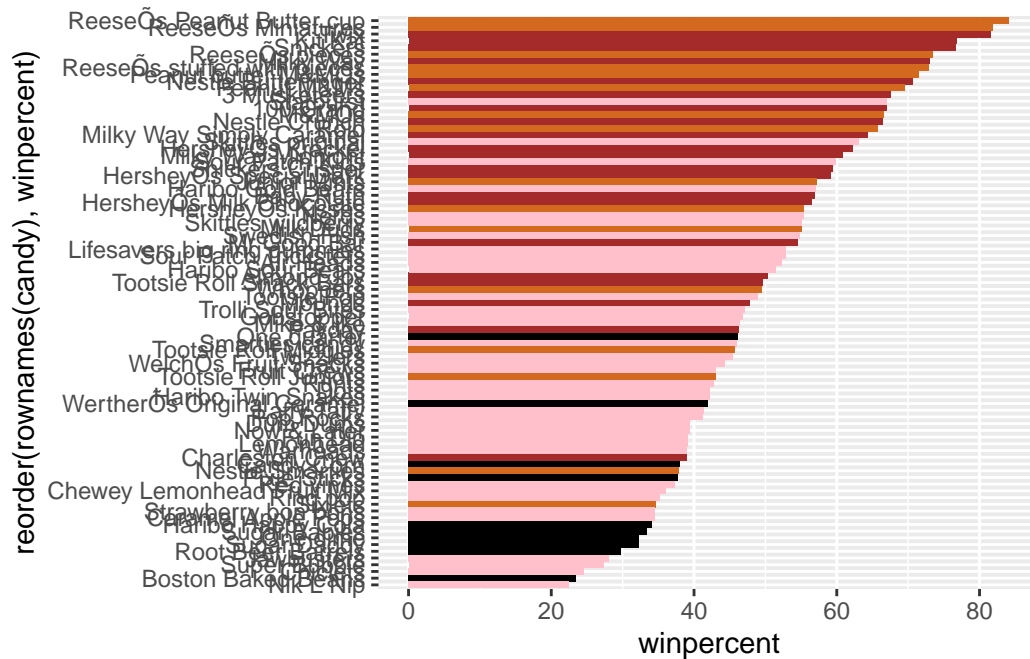
**Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?**

```
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?**

Answer: Sixlets is the worst ranked chocolate candy.

**Q18. What is the best ranked fruity candy?**

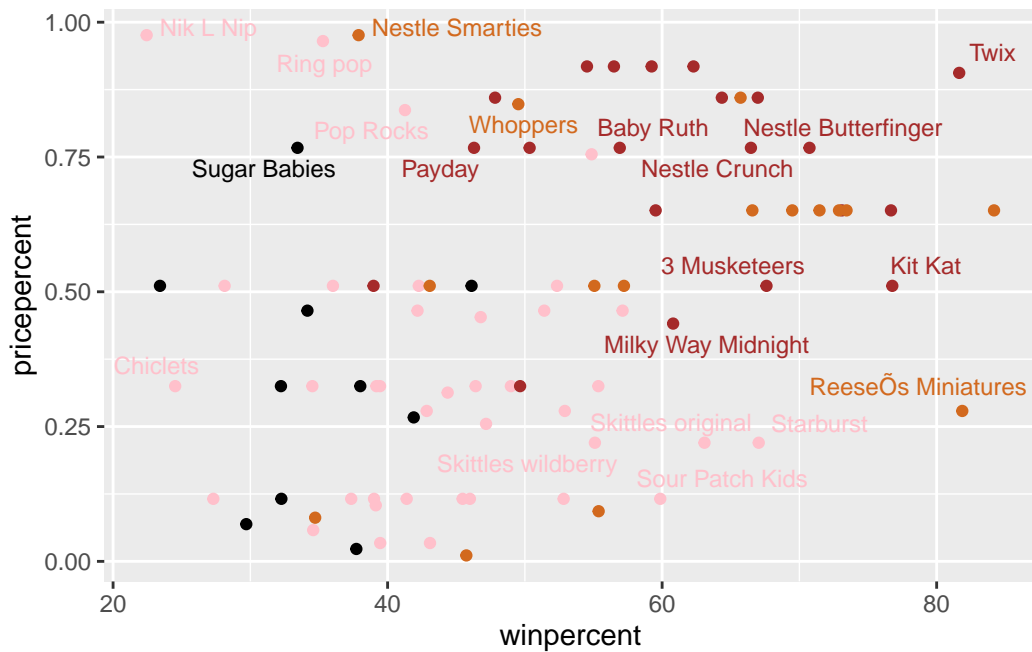
Answer: Starburst is the best ranked fruity candy.

## 4. Taking a look at pricepercent

```
library(ggrepel)

# 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=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?**

Answer: Reese's Miniatures is the highest ranked in terms of winpercent for the least money.

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

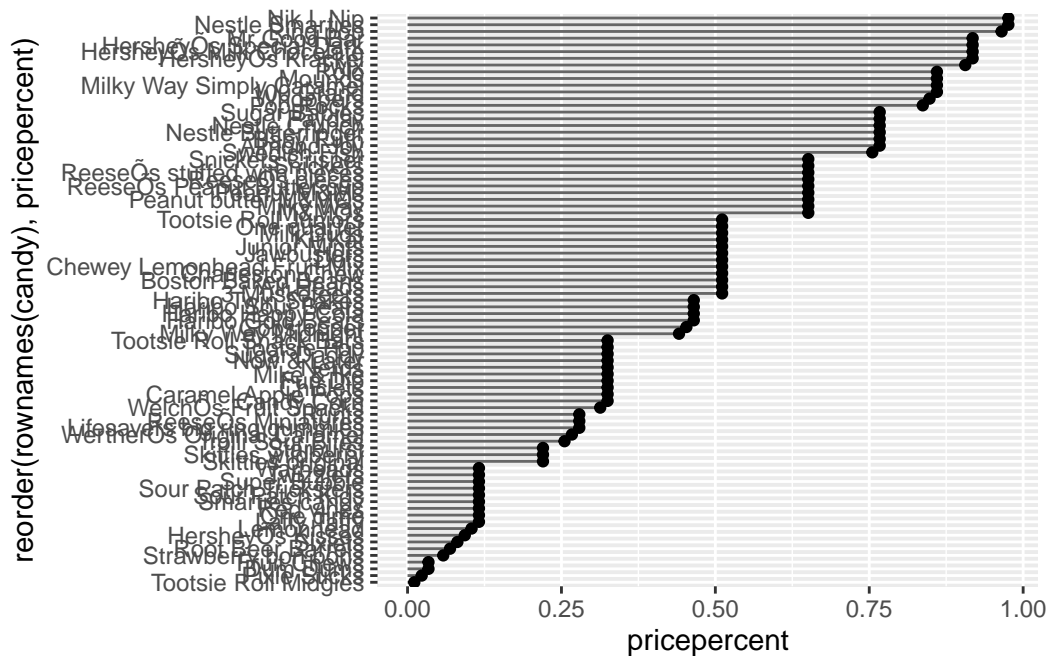
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

	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

Answer: The top 5 most expensive candy types in the dataset are Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate. Of these, Nik L Nip is the least popular.

**Q21. Make a barplot again with `geom_col()` this time using `pricepercent` and then improve this step by step, first ordering the x-axis by value and finally making a so called “dot chat” or “lollipop” chart by swapping `geom_col()` for `geom_point()` + `geom_segment()`.**

```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                    xend = 0), col="gray40") +
  geom_point()
```

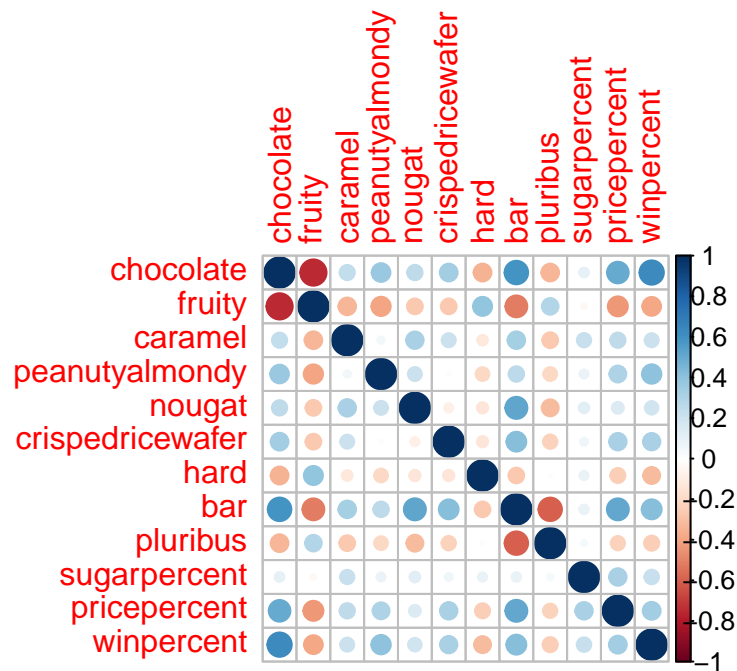


## 5. Exploring the correlation structure

```
library(corrplot)
```

corrplot 0.92 loaded

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



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

Answer: Chocolate and fruity are the two variables in this plot that are anti-correlated.

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

Answer: Chocolate and bar are the two variables in this plot that are most positively correlated.

## 6. Principal Component Analysis

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

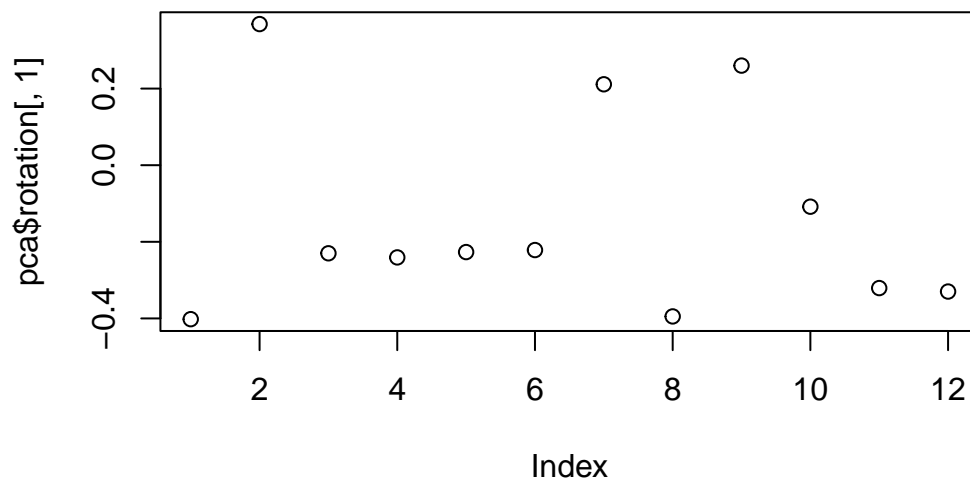
Importance of components:

	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

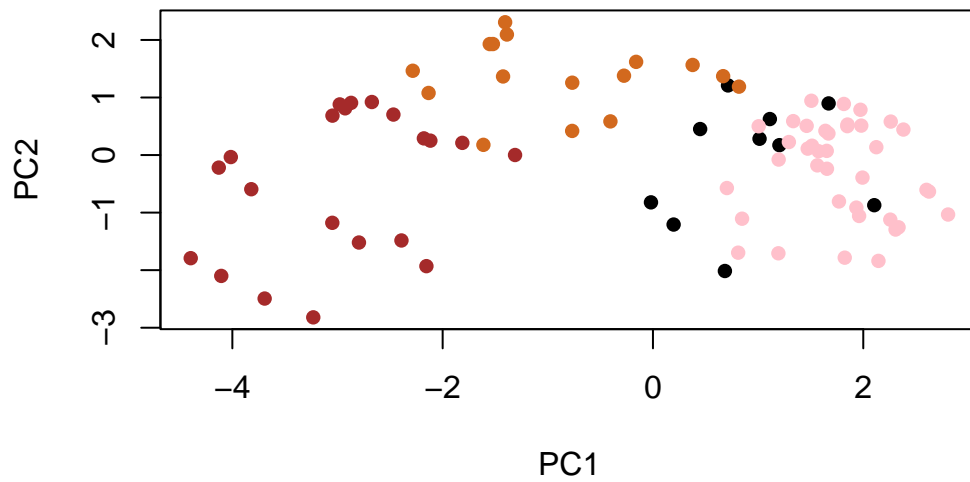
  

	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.74530	0.67824	0.62349	0.43974	0.39760
Proportion of Variance	0.04629	0.03833	0.03239	0.01611	0.01317
Cumulative Proportion	0.89998	0.93832	0.97071	0.98683	1.00000

```
plot(pca$rotation[,1])
```



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

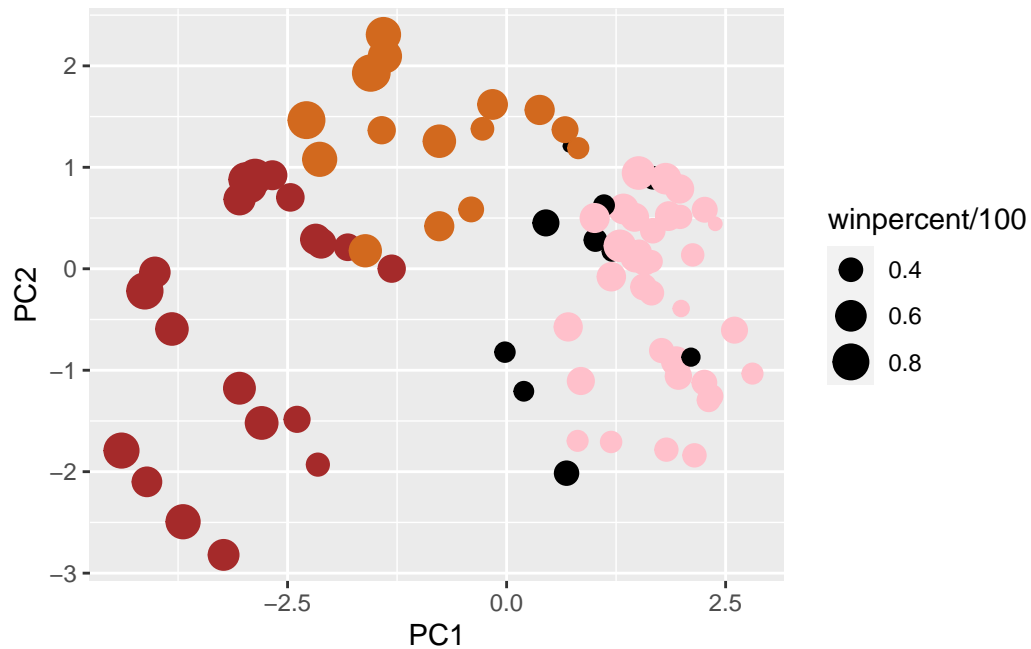


```
# Make a new data-frame with our PCA results and candy data  
my_data <- cbind(candy, pca$x[,1:3])
```

```
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





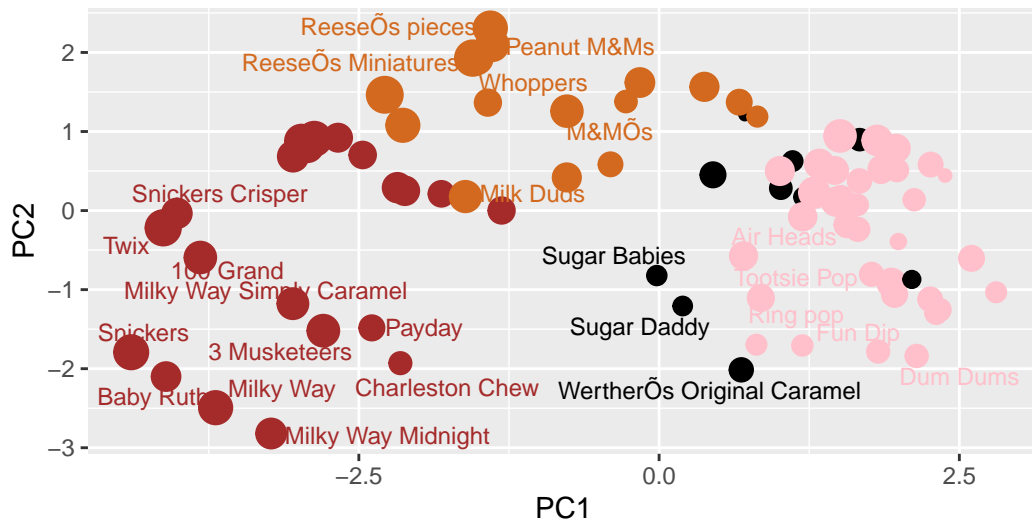
```
library(ggrepel)

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)",
        caption="Data from 538")
```

Warning: ggrepel: 60 unlabeled data points (too many overlaps). Consider increasing max.overlaps

## Halloween Candy PCA Space

Colored by type: chocolate bar (dark brown), chocolate other (light brown),



Data from 538

```
library(plotly)
```

Attaching package: 'plotly'

The following object is masked from 'package:ggplot2':

last\_plot

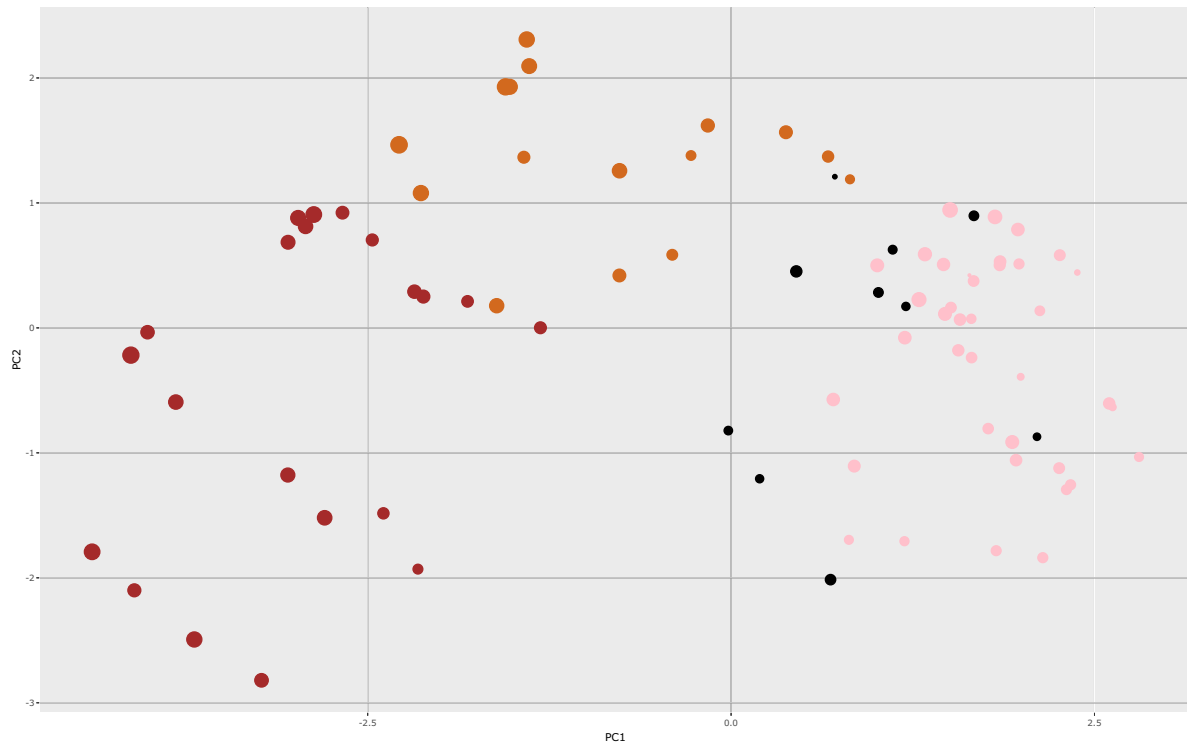
The following object is masked from 'package:stats':

filter

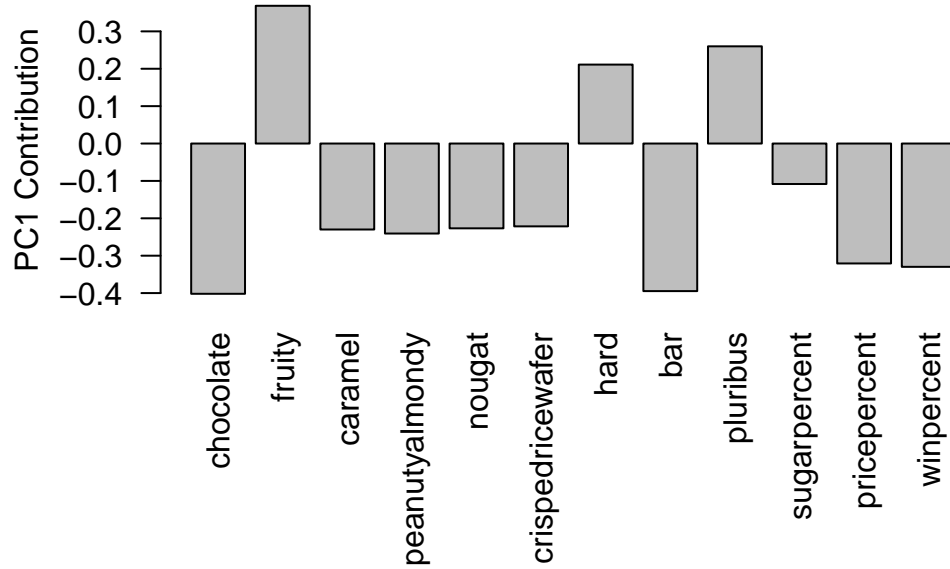
The following object is masked from 'package:graphics':

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?**

Answer: Original variables “fruity”, “hard”, and “pluribus” are picked up strongly by PC1 in the positive direction. These make sense because many people enjoy fruit-flavored candies, prefer that their candy is not melted, and like to buy candy in bulk.