

# Halloween Candy Mini-Project

Kianna

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

	chocolate	fruity	caramel	peanut	almond	nougat	crisp	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	66.97173			
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
```

There are 85 different candy types in the dataset.

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

```
sum(candy[, "fruity"])
```

```
[1] 38
```

There are 38 fruity candy types in this dataset.

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

```
[1] 81.64291
```

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

```
candy["Skittles original", ]$winpercent
```

```
[1] 63.08514
```

My favorite candy in the dataset is Skittles original. Its winpercent value is 63.08514%.

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
```

The winpercent value for Tootsie Roll Snack Bars is 49.6535%.

```
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	
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 histogram column looks to be on a different scale. There are no numbers in the column; it looks like an image of bars. Also, the `skim_variable` has no numbers in the column as it is the names of the variables themselves.

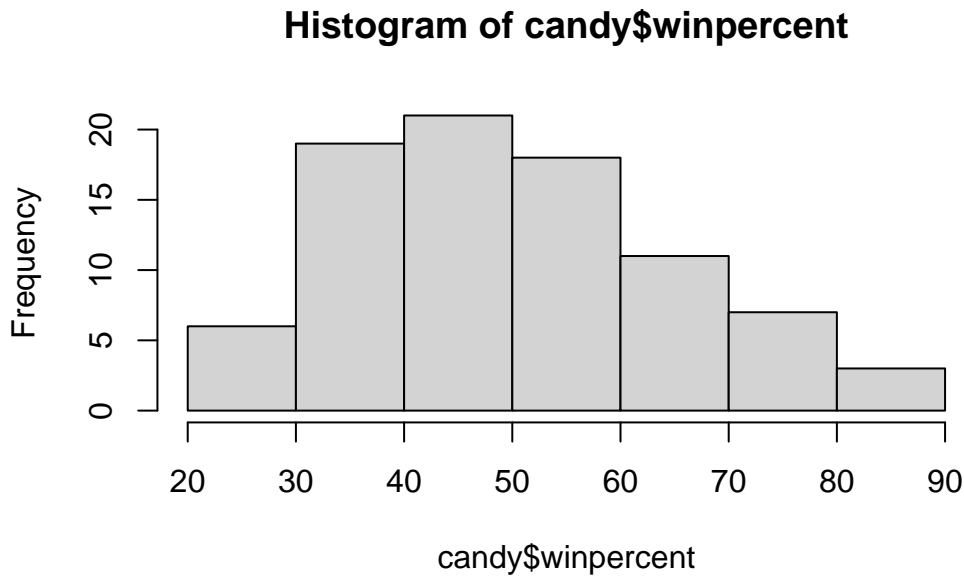
**Note: when I render the quarto document as a pdf, nothing appears in the “hist” column. However, I can see what is in the column when I open it in RStudio or render it as html.**

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

I think a 0 in this column means that candy type DOES NOT contain chocolate whereas a 1 means that candy type DOES contain chocolate.

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical?

The distribution of winpercent values is not perfectly symmetrical. It represents a bell curve, but it is definitely skewed right.

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

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

The center of distribution lies slightly above 50% at 50.31676%.

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

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

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

```
[1] 60.92153
```

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

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

```
[1] 44.11974
```

Chocolate candy is on average ranked higher than fruity candy. The average winpercent value for chocolate candy is 60.92153% and for fruit candy is 44.11974%.

Q12. Is this difference statistically significant?

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

Welch Two Sample t-test

```
data: chocolate and fruity
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
```

The reported p value when performing a two sample t-test between the winpercent values for chocolate and fruity candy was 2.871e-08. This value is below a p-value of 0.05, so we can reject the null hypothesis and conclude there is a statistically significant difference between how chocolate and fruity candy is ranked.

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

```
head(candy[order(candy$winpercent), ], n=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	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
	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.

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

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

	chocolate	fruity	caramel	peanut	almond	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					

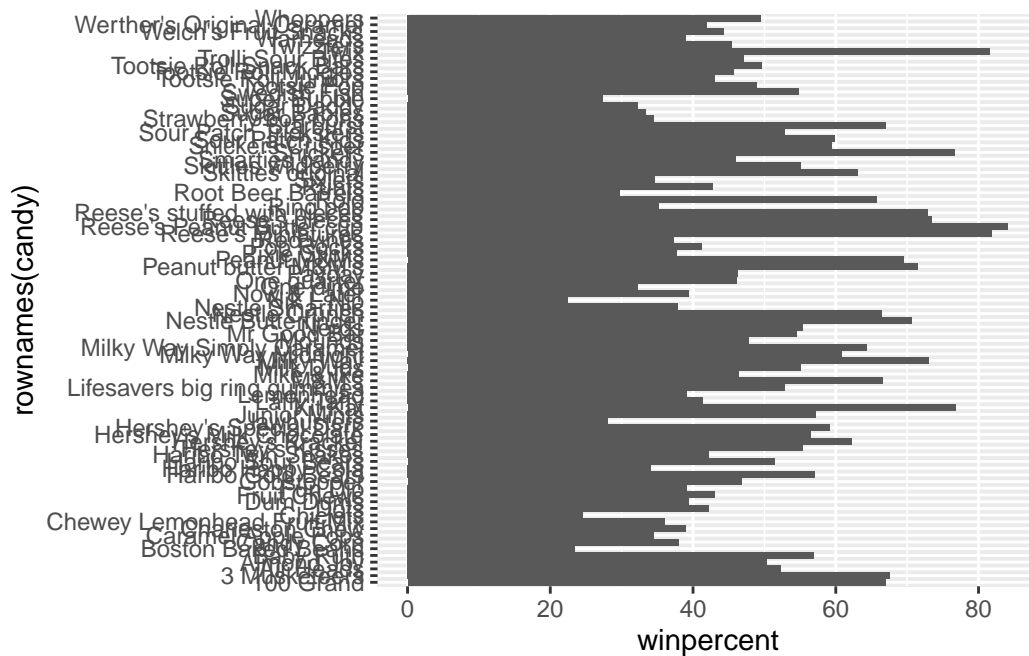
Snickers	0.651	76.67378
Kit Kat	0.511	76.76860
Twix	0.906	81.64291
Reese's Miniatures	0.279	81.86626
Reese's Peanut Butter cup	0.651	84.18029

The top 5 all time favorite candy types are Snickers, Kit Kat, Twix, Reese's Miniatures, and Reese's Peanut Butter cups.

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

```
library(ggplot2)

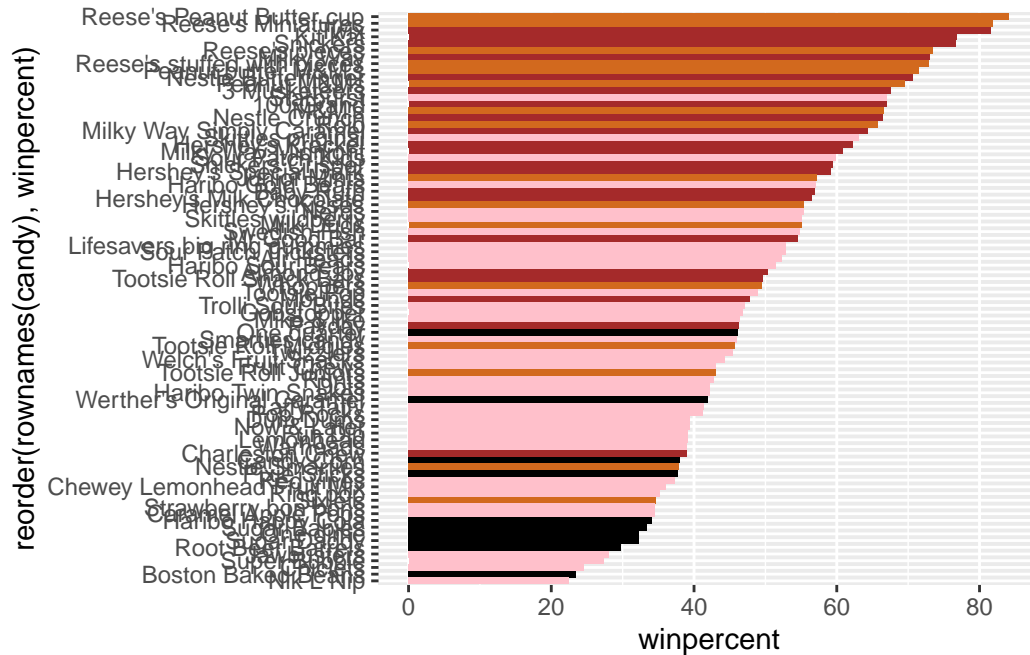
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col()
```



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







Q17. What is the worst ranked chocolate candy?

The worst ranked chocolate candy is Sixlets.

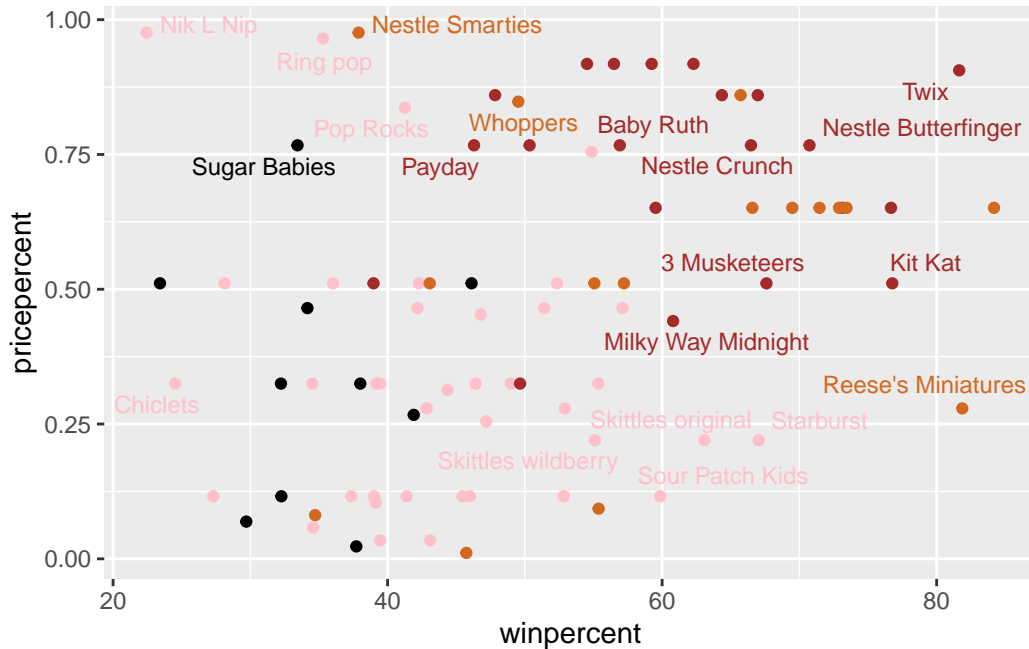
Q18. What is the best ranked fruity candy?

The best ranked fruity candy is Starburst.

```
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?

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

	pricepercent	winpercent
Tootsie Roll Midgies	0.011	45.73675
Pixie Sticks	0.023	37.72234
Dum Dums	0.034	39.46056
Fruit Chews	0.034	43.08892
Strawberry bon bons	0.058	34.57899

The candy with the lowest price percent but highest winpercent is tootsie roll midgies.

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

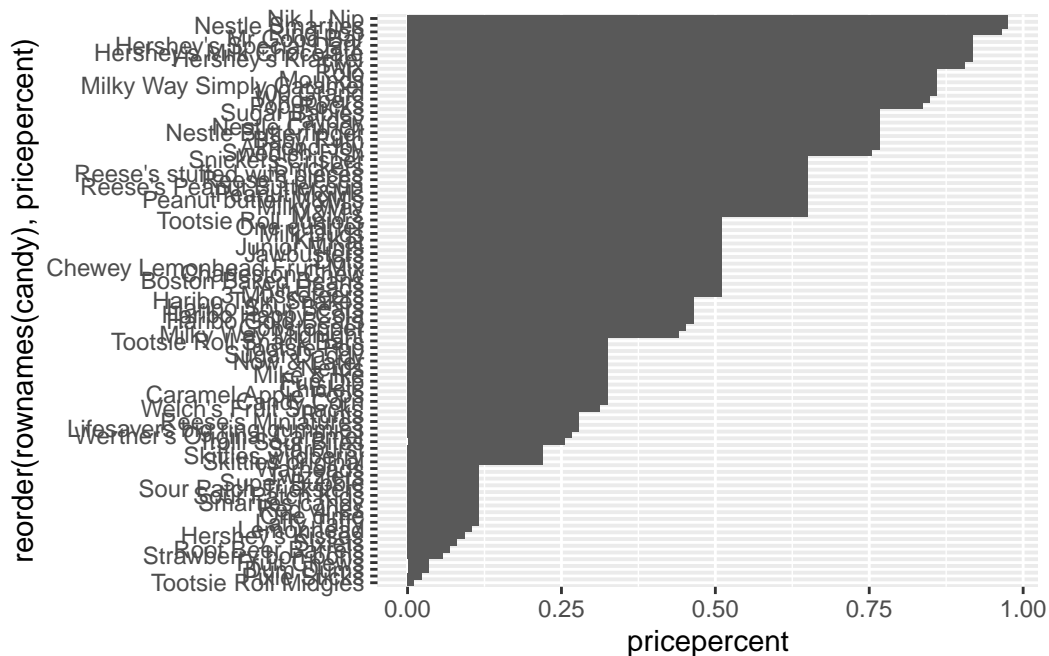
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

	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

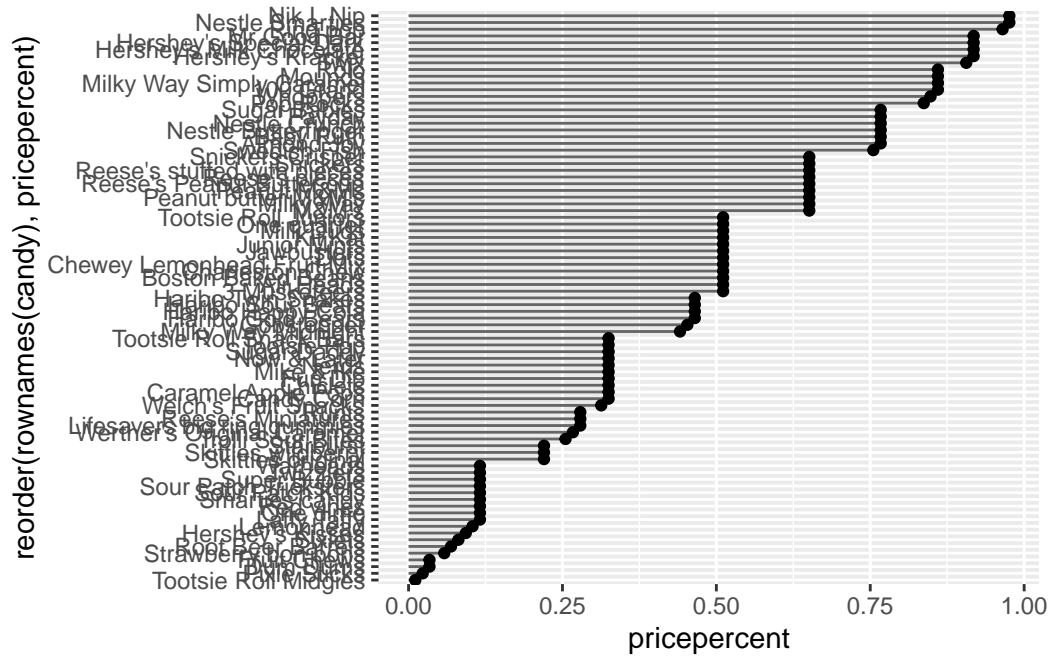
These are the top 5 most expensive candy types as they have the highest pricepercent values.  
The least popular one is Nik L Nip.

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_col()
```



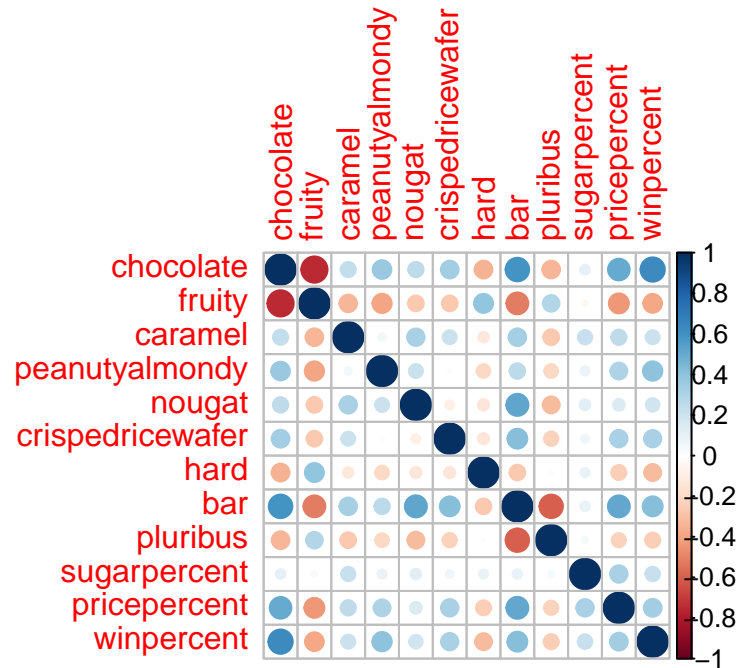
```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                  xend = 0), col="gray40") +
  geom_point()
```



```
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)?

The two variables most anti-correlated are fruity and chocolate.

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

The two variables most positively correlated are bar and chocolate, though it's hard to tell whether winpercent and chocolate have a higher positive correlation.

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

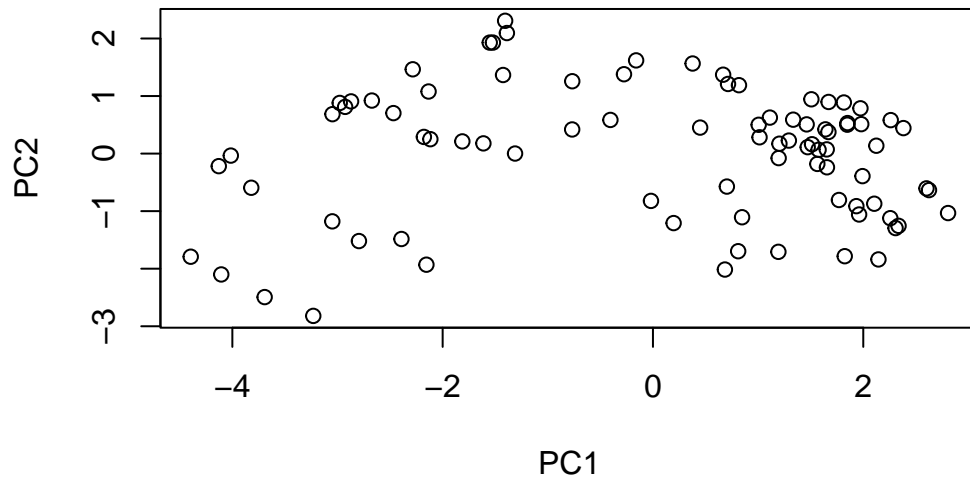
Importance of components:

	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

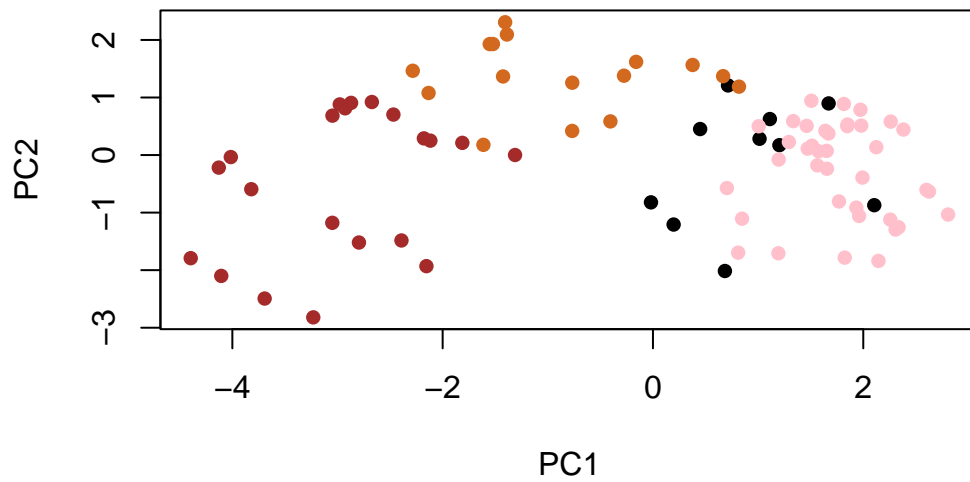
  

	PC8	PC9	PC10	PC11	PC12
Standard deviation	0.74530	0.67824	0.62349	0.43974	0.39760
Proportion of Variance	0.04629	0.03833	0.03239	0.01611	0.01317
Cumulative Proportion	0.89998	0.93832	0.97071	0.98683	1.00000

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



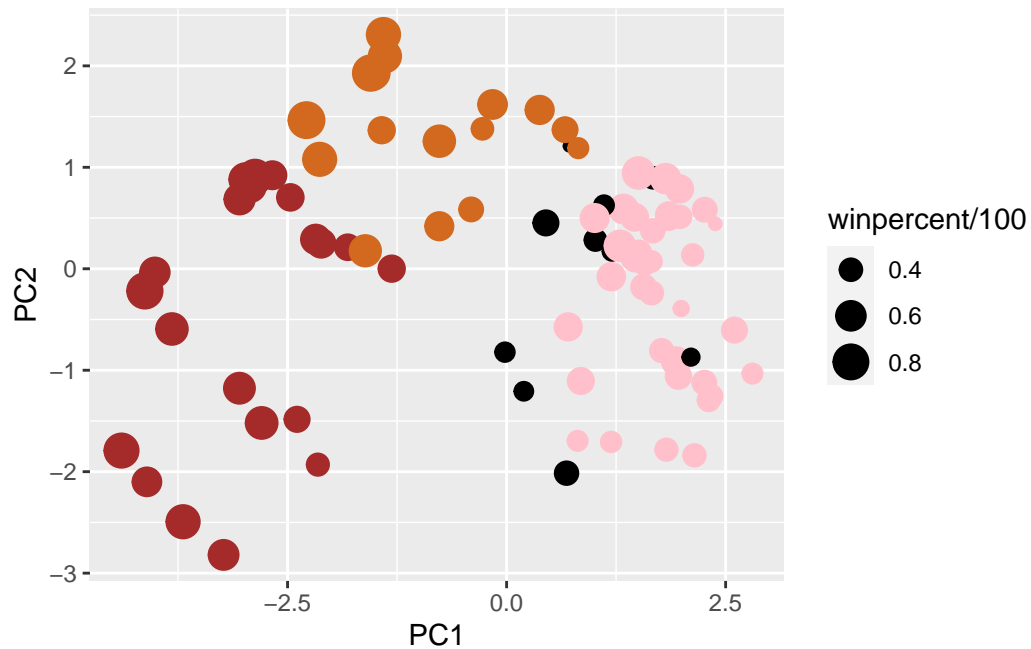
```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



```
# 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
```



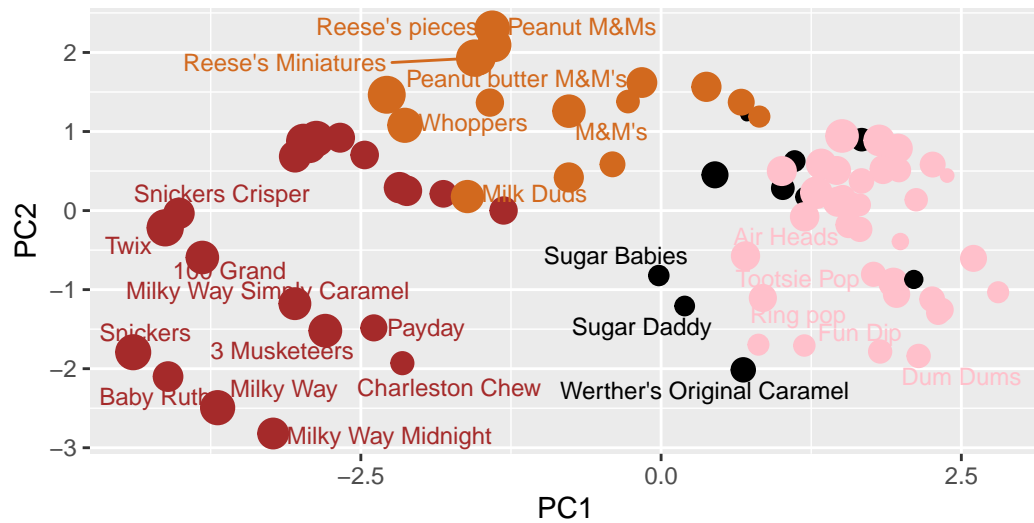
```
p + geom_text_repel(size=3.3, col=my_cols, max.overlaps = 7) +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
        subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown)",
        caption="Data from 538")
```

Warning: ggrepel: 59 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

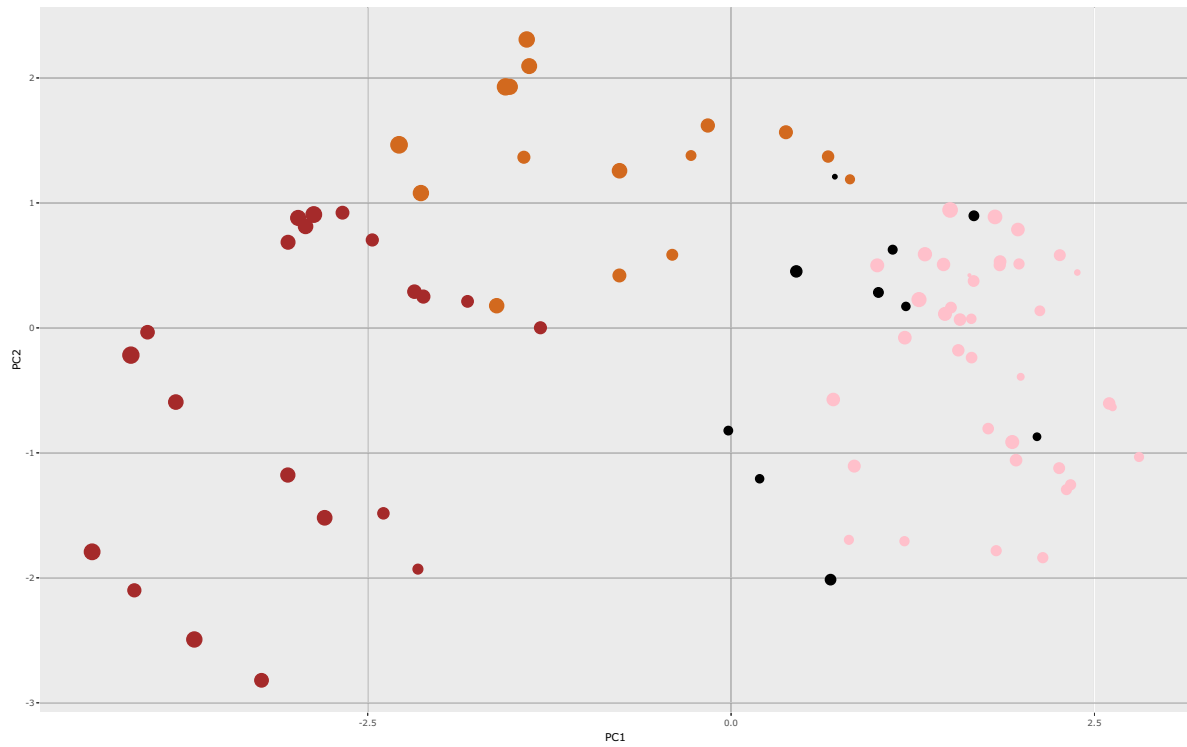
The following object is masked from 'package:stats':

filter

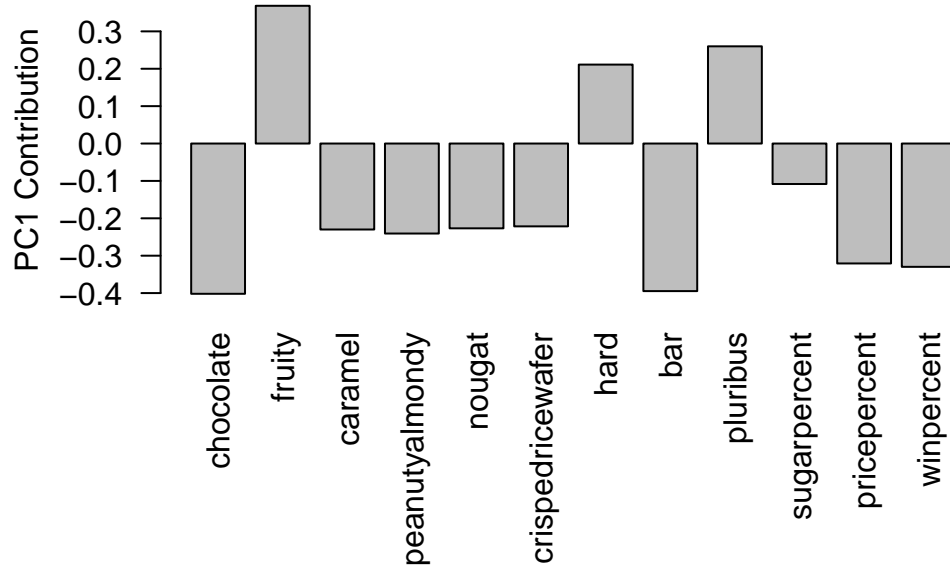
The following object is masked from 'package:graphics':

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?

The variables picked up strongly by PC1 in the positive direction are fruity, hard, and pluribus. This makes sense to me because on the correlation matrix, those variable are positive correlated with each other. Also, in my experience the fruity candy I've come across with are hard and come in a box of multiple candies, like skittles.