

Class 9 : Halloween Candy Mini Project

Jaimy Chung (A16366976)

Here we analyze a candy dataset from the 538 website.

```
candy <- read.csv ("candy-data.txt", 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?

```
dim(candy)
```

```
[1] 85 12
```

85 different types of candy

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

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

```
[1] 38
```

There are 38 fruity candy

Data Exploration

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

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

```
[1] 81.64291
```

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
```

```
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	

Q. What is the least liked candy in the dataset - lowest winpercent

```
x <- c(5, 3, 4, 1)
sort(x)
```

```
[1] 1 3 4 5
```

```
order(x)
```

```
[1] 4 2 3 1
```

```
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

	crisp	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	

	win	percent
Nik L Nip	22.44	534
Boston Baked Beans	23.41	782
Chiclets	24.52	499
Super Bubble	27.30	386
Jawbusters	28.12	744
Root Beer Barrels	29.70	369

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

Table 3: 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	

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

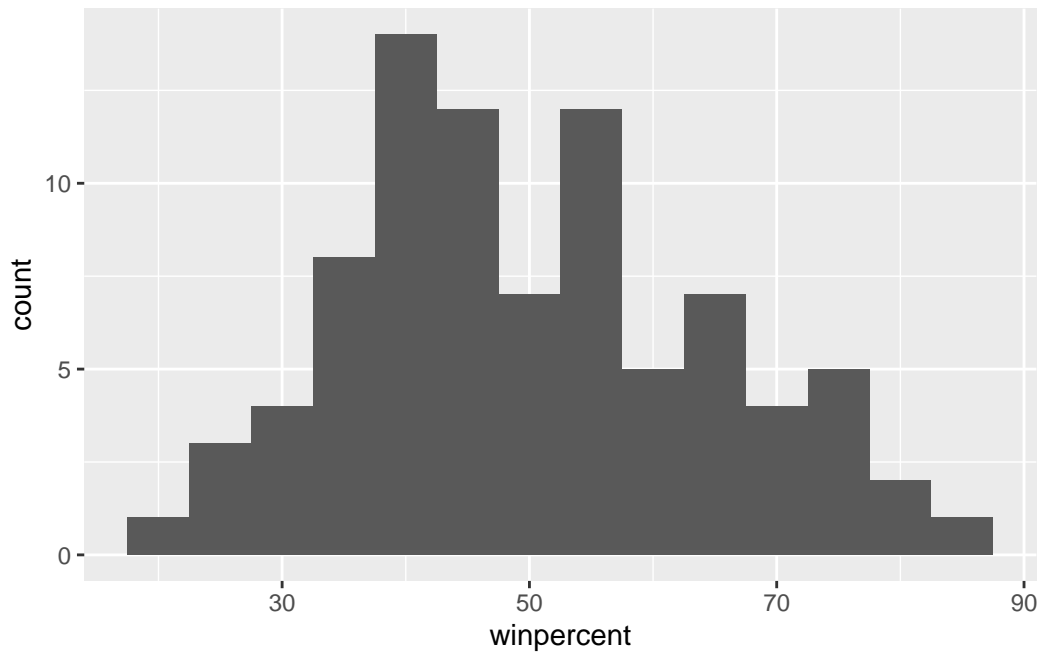
The winpercent is relatively high compared to the other numbers in the dataset.

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

0 means it's a chocolate, 1 means it's not a chocolate.

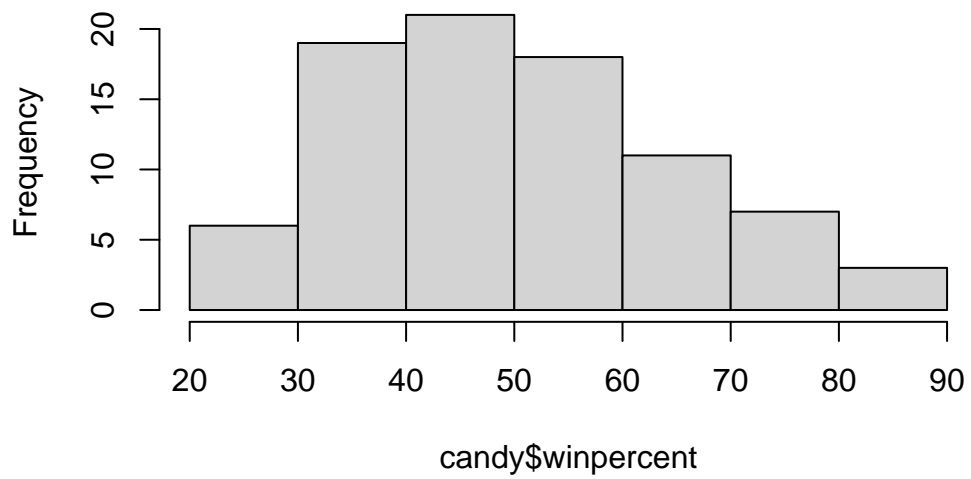
Q8. Plot a histogram of winpercent values

```
library(ggplot2)
ggplot(candy) +
  aes(winpercent) +
  geom_histogram(binwidth=5)
```



```
hist(candy$winpercent, breaks=8)
```

Histogram of candy\$winpercent



Q9. Is the distribution of winpercent values symmetrical?

No

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

Below

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

First find all chocolate candy and their \$winpercent values

Next summarize these values into one number

Then do the same for fruit candy and compare the numbers.

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

```
[1] 60.92153
```

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

```
[1] 44.11974
```

On average, chocolate candy is higher.

Q12. Is this difference statistically significant?

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

Welch Two Sample t-test

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

Based off of the p-value which is super low, the difference is statistically significant.

```
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

	crisp	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

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

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

```
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

	crisp	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

	price	percent	winpercent
Reese's Peanut Butter cup	0.651		84.18029

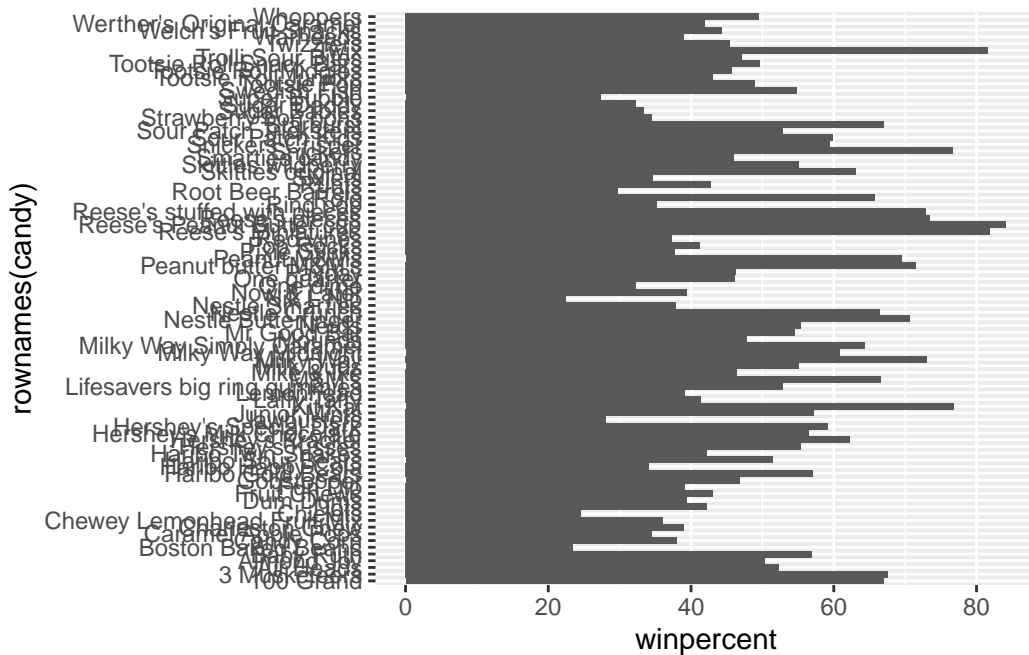
Reese's Miniatures	0.279	81.86626
Twix	0.906	81.64291
Kit Kat	0.511	76.76860
Snickers	0.651	76.67378

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

The top 5 are Reese's Peanut Butter cups, Reese's minatures, twix, kit kat, and snickers.

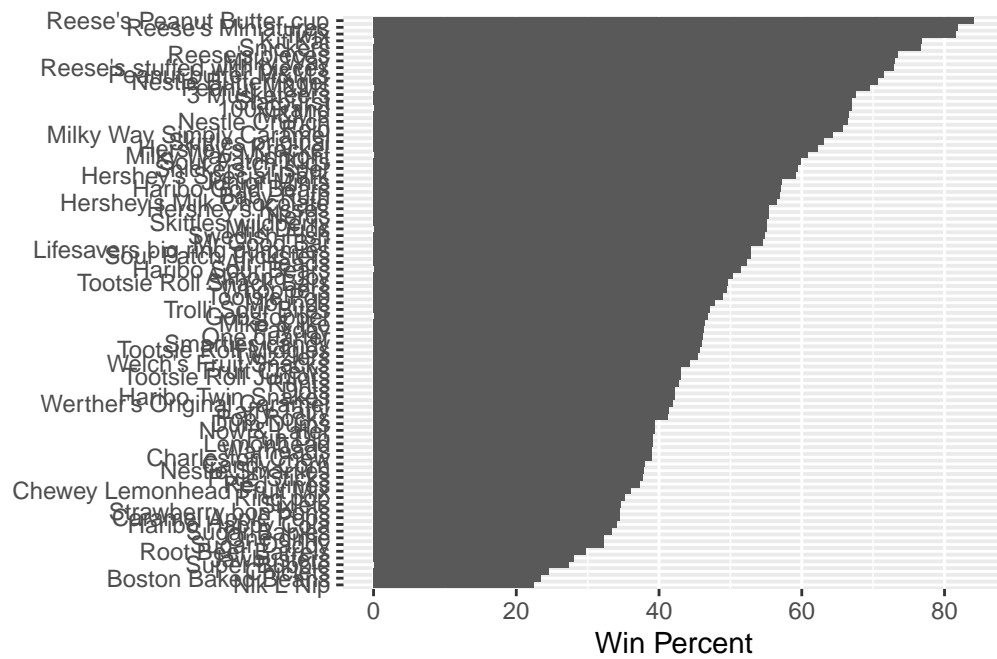
Q15. Make a first barplot of candy ranking based on winpercent values.

```
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() +
  labs(x="Win Percent", y="")
```



```
ggsave('barplot1.png', width=7, height=10)
```

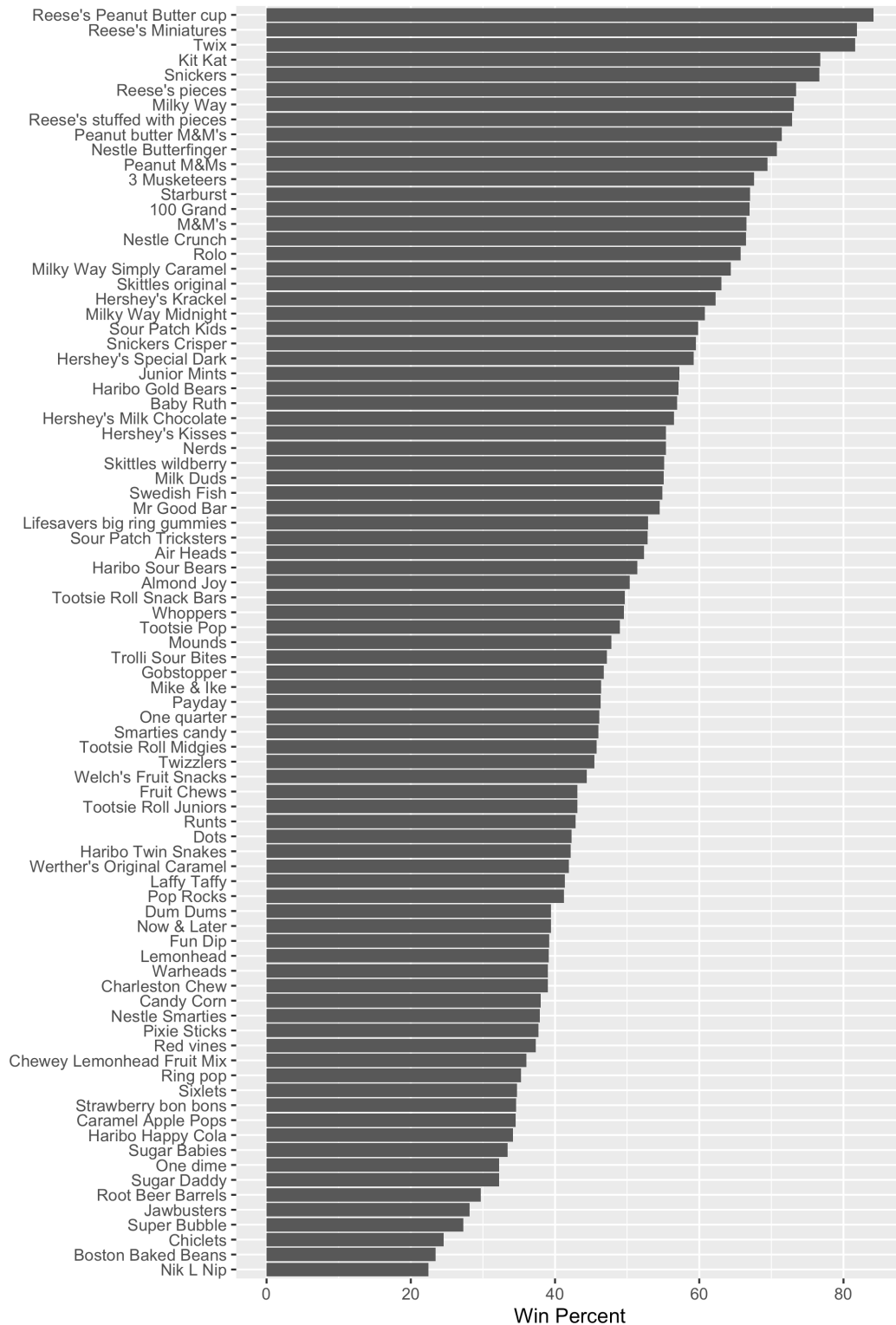


Figure 1: A plot with better aspect ratio

