Halloween Mini-Project

Varun Durai

1. Importing candy data:

```
candy_file <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-r

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?

```
dim(candy)
```

```
[1] 85 12
```

By observing the dimensions of the dataset, we can see that there are 85 rows, and thus 85 different types of candy.

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

```
sum(candy$fruity == 1)
```

```
[1] 38
```

By summing all the instances of the fruity column when it is equal to 1, we can find that there are 38 fruity candy types.

2. What is your favorite candy?

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

```
candy["Almond Joy", ]$winpercent
```

```
[1] 50.34755
```

The favorite candy is Almond Joy, and the function above tells us that its winpercent value is 50.34755

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

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

```
[1] 76.7686
```

76.7686 is the winpercent of Kit Kat

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

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

```
[1] 49.6535
```

49.6535 is the winpercent of Tootsie Roll Snack Bars

```
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?

From the skim results, winpercent seems to be on a different scale from the other variables since it displays advanced stats with values between 0-100 while the advanced stats of the other variables range from 0-1.

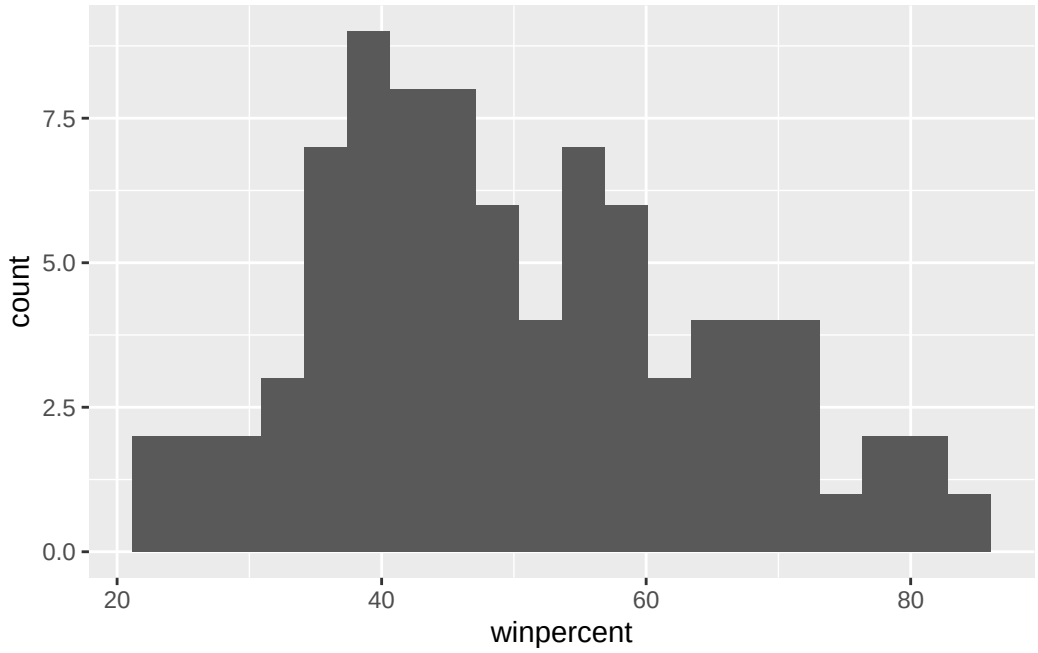
Q7. What do you think a zero and one represent for the candy$chocolate column?

A zero indicates that the candy does not contain chocolate, while a one indicates that it does.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)
```

```
ggplot(candy, aes(x = winpercent)) +
  geom_histogram(bins = 20)
```



Q9. Is the distribution of winpercent values symmetrical?

The distribution of the winpercent values is not symmetrical.

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

The center of the distribution seems to be slightly below 50%

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

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

```
[1] 44.11974
```

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

```
[1] 60.92153
```

On average, chocolate candy is higher ranked.

Q12. Is this difference statistically significant?

```
t.test(candy$winpercent[as.logical(candy$fruit)])
```

```
    One Sample t-test

data:  candy$winpercent[as.logical(candy$fruit)]
t = 26.498, df = 37, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 40.74612 47.49337
sample estimates:
mean of x
 44.11974
```

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

```
    One Sample t-test

data:  candy$winpercent[as.logical(candy$chocolate)]
t = 28.926, df = 36, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 56.65009 65.19297
sample estimates:
mean of x
 60.92153
```

Both have extremely small p-values so the data is likely statistically significant.

3. Overall Candy Ranking

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

```
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) %>% head(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
```

Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble and Jawbusters are the least liked types in the set.

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

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

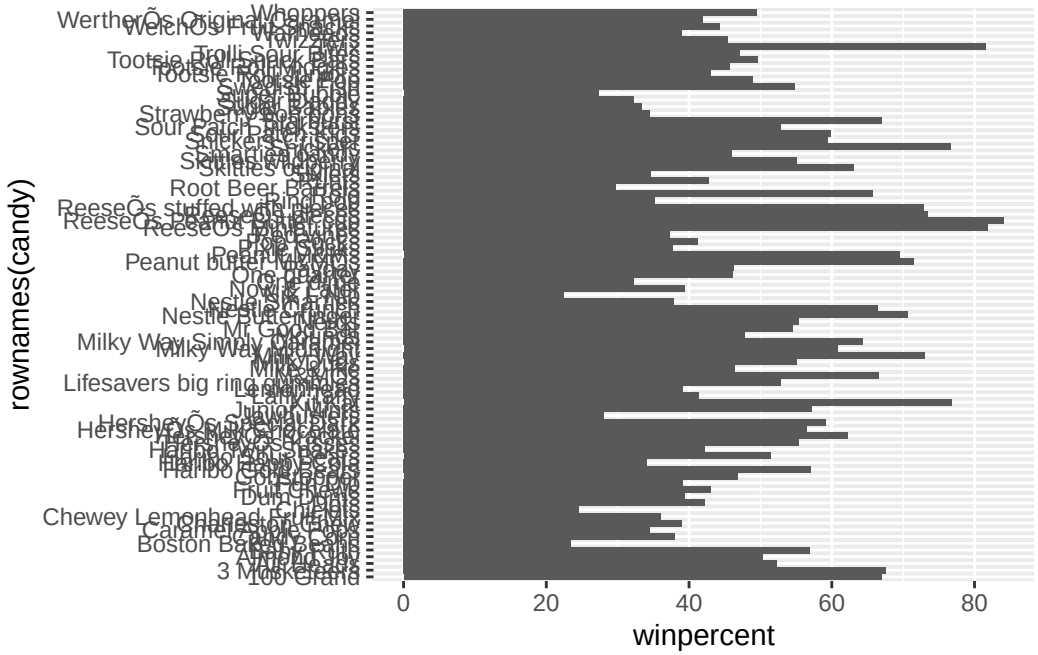
```
                         chocolate fruity caramel peanutyalmondy nougat
ReeseÕs Peanut Butter cup         1      0       0              1      0
ReeseÕs Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           1      0       0              0      0
Snickers                          1      0       1              1      1
                         crispedricewafer hard bar pluribus sugarpercent
ReeseÕs Peanut Butter cup                0    0   0        0        0.720
ReeseÕs 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
ReeseÕs Peanut Butter cup        0.651   84.18029
ReeseÕs 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 types out of this set are Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat and Snickers.

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

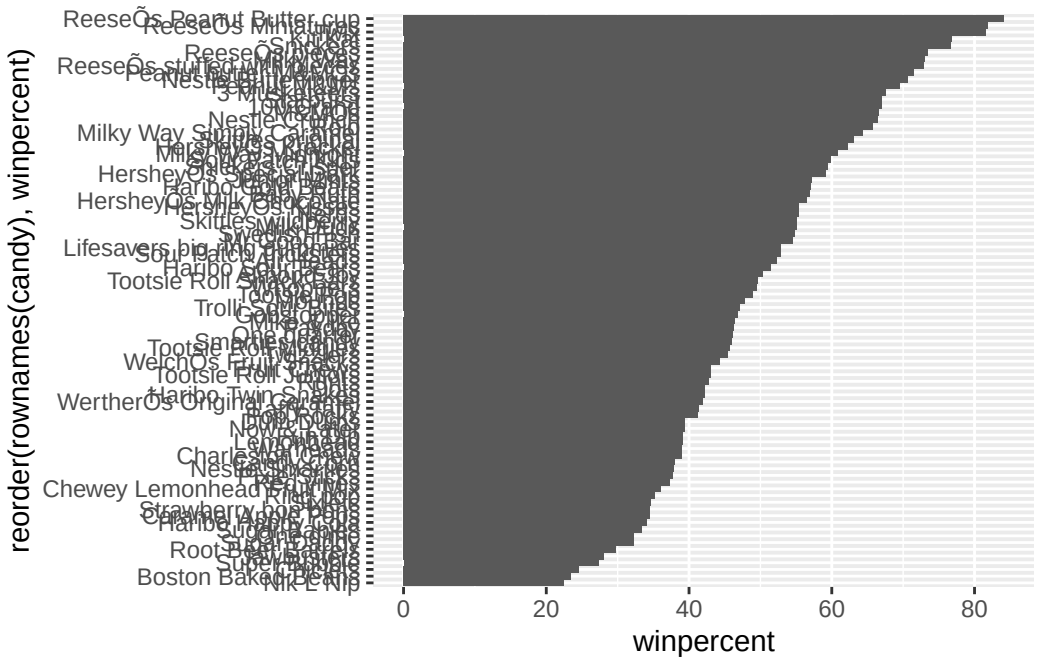
```
library(ggplot2)

ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_bar(stat = 'identity')
```



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

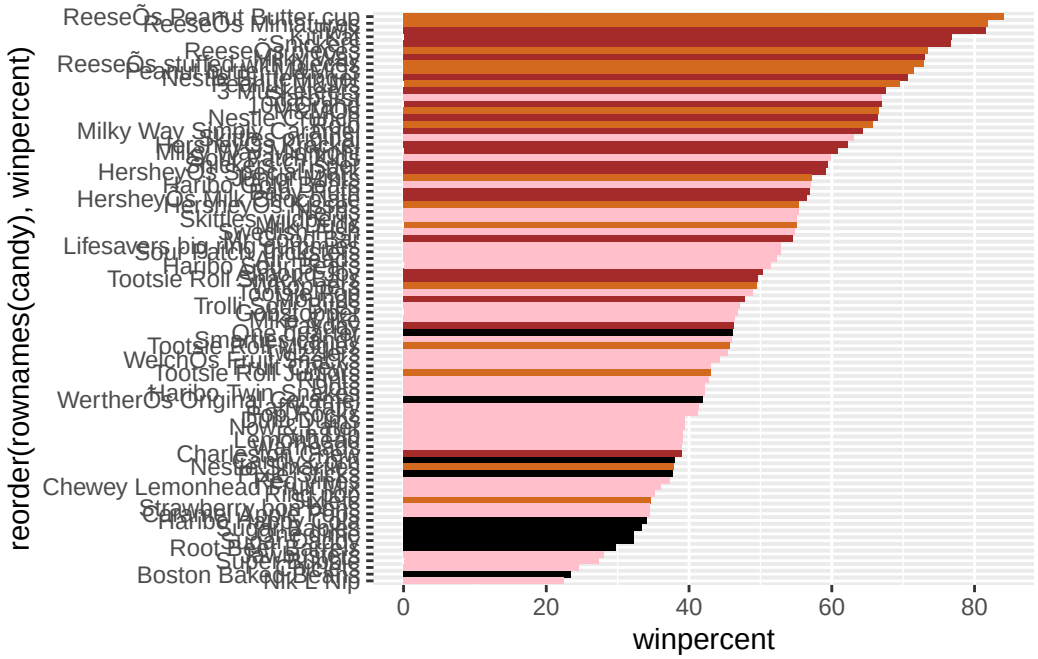
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_bar(stat = 'identity')
```




```
#Adding Color,

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?

Sixlet is the worst ranked chocolate candy.

Q18. What is the best ranked fruity candy?

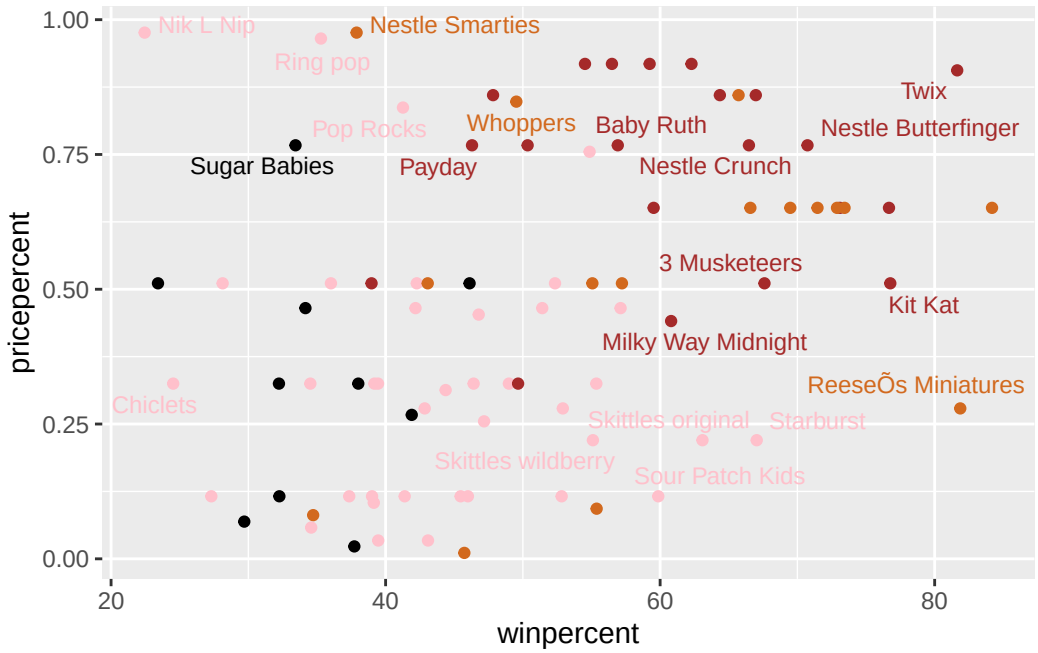
Starburst is the best ranked fruity candy.

4. Taking a look at pricepercent

```
library(ggrepel)

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?

Reese's Minatures probably offer the best bang for you buck.

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
HersheyÕs Krackel              0.918   62.28448
HersheyÕs Milk Chocolate       0.918   56.49050
```

So, Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel and Hershey's Milk Chocolate are the most expensive.