Halloween Mini-Proj

Nickolas Beam

```r
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?

```r
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

Winpercent

The most interesting variables in the dataset is `winpercentt`. For a given candy this value is the percentage of people who prefer this candy over another randomly chosen candy

Q3. What is your favorite candy in the dataset and what is it's winpercent value?

```
rownames(candy)
```

```
 [1] "100 Grand"                "3 Musketeers"
 [3] "One dime"                 "One quarter"
 [5] "Air Heads"                "Almond Joy"
 [7] "Baby Ruth"                "Boston Baked Beans"
 [9] "Candy Corn"               "Caramel Apple Pops"
[11] "Charleston Chew"          "Chewey Lemonhead Fruit Mix"
[13] "Chiclets"                 "Dots"
[15] "Dum Dums"                 "Fruit Chews"
[17] "Fun Dip"                  "Gobstopper"
[19] "Haribo Gold Bears"        "Haribo Happy Cola"
[21] "Haribo Sour Bears"        "Haribo Twin Snakes"
[23] "HersheyÕs Kisses"         "HersheyÕs Krackel"
[25] "HersheyÕs Milk Chocolate" "HersheyÕs Special Dark"
[27] "Jawbusters"               "Junior Mints"
[29] "Kit Kat"                  "Laffy Taffy"
[31] "Lemonhead"                "Lifesavers big ring gummies"
[33] "Peanut butter M&MÕs"      "M&MÕs"
[35] "Mike & Ike"               "Milk Duds"
[37] "Milky Way"                "Milky Way Midnight"
[39] "Milky Way Simply Caramel" "Mounds"
[41] "Mr Good Bar"              "Nerds"
[43] "Nestle Butterfinger"      "Nestle Crunch"
[45] "Nik L Nip"                "Now & Later"
[47] "Payday"                   "Peanut M&Ms"
[49] "Pixie Sticks"             "Pop Rocks"
[51] "Red vines"                "ReeseÕs Miniatures"
[53] "ReeseÕs Peanut Butter cup" "ReeseÕs pieces"
[55] "ReeseÕs stuffed with pieces" "Ring pop"
[57] "Rolo"                     "Root Beer Barrels"
[59] "Runts"                    "Sixlets"
[61] "Skittles original"        "Skittles wildberry"
[63] "Nestle Smarties"          "Smarties candy"
[65] "Snickers"                 "Snickers Crisper"
```

```
[67] "Sour Patch Kids"          "Sour Patch Tricksters"
[69] "Starburst"                "Strawberry bon bons"
[71] "Sugar Babies"             "Sugar Daddy"
[73] "Super Bubble"             "Swedish Fish"
[75] "Tootsie Pop"              "Tootsie Roll Juniors"
[77] "Tootsie Roll Midgies"     "Tootsie Roll Snack Bars"
[79] "Trolli Sour Bites"        "Twix"
[81] "Twizzlers"                "Warheads"
[83] "WelchÕs Fruit Snacks"     "WertherÕs Original Caramel"
[85] "Whoppers"
```

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

```
[1] 55.1037
```

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
```

A useful function from the skimr package

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

Table 1: Data summary

Name	candy
Number of rows	85
Number of columns	12

Table 1: Data summary

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	

Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent, breaks=20)
```

Histogram of candy$winpercent
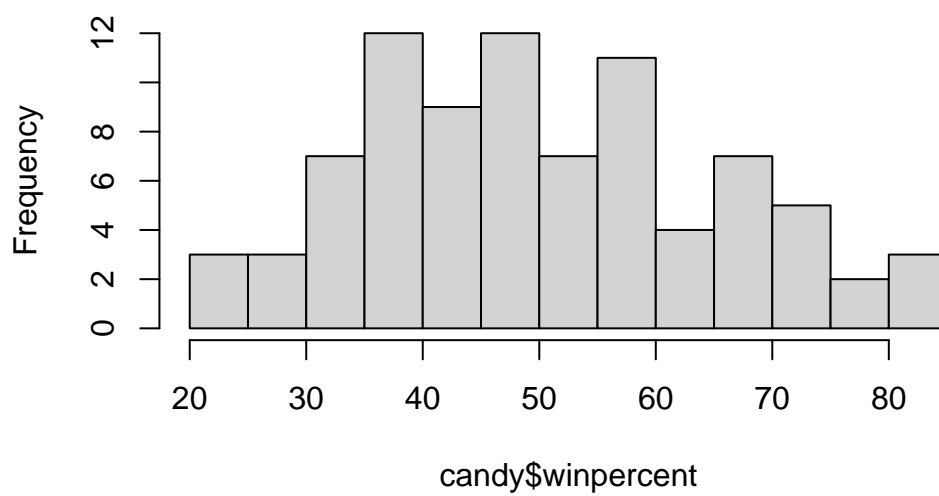

```
library(ggplot2)
ggplot(candy) + aes(winpercent) + geom_histogram(bins=10, col="red", fill="forest green")
```
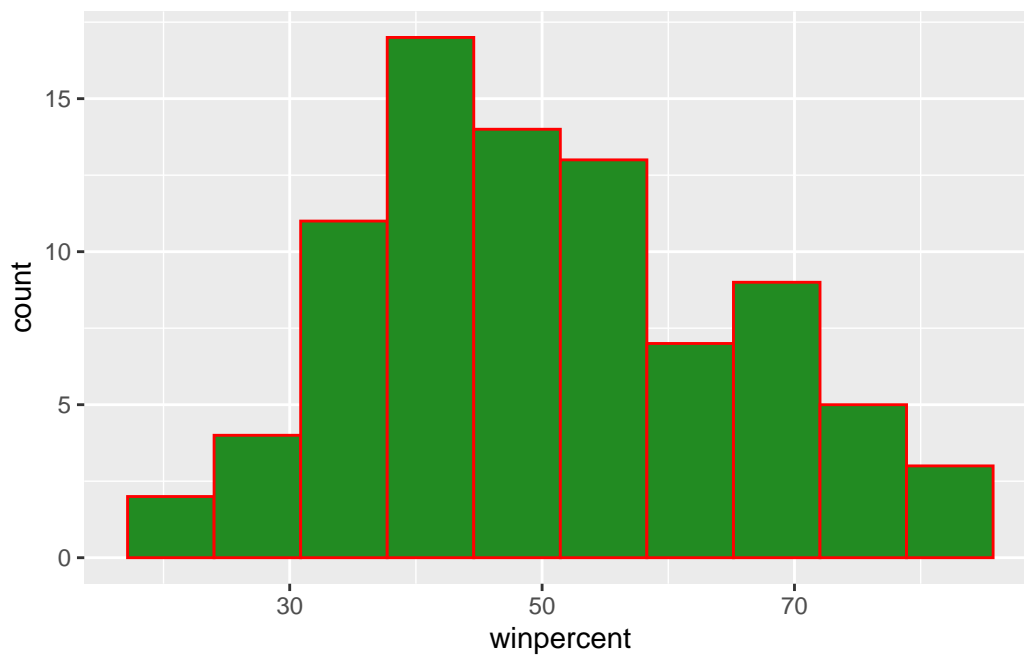
Q11. On average is chocolate candy higher or lower ranked than fruit candy?

```r
chocolate.inds <- as.logical(candy$chocolate)
choc.win <- candy[chocolate.inds,]$winpercent

fruity.inds <- as.logical(candy$fruity)
fruity.win <- candy[fruity.inds,]$winpercent

mean(choc.win)
```

```
[1] 60.92153
```

```r
mean(fruity.win)
```

```
[1] 44.11974
```

3. Candy Ranking

First setup some colors for different candy types

```r
my_cols=rep("black", nrow(candy))
my_cols
```

```
 [1] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[10] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[19] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[28] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[37] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[46] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[55] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[64] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[73] "black" "black" "black" "black" "black" "black" "black" "black" "black"
[82] "black" "black" "black" "black"
```

```r
my_cols[as.logical(candy$chocolate)] = "chocolate"
my_cols[as.logical(candy$bar)] = "brown"
my_cols[as.logical(candy$fruity)] = "pink"
my_cols
```

```
 [1] "brown"     "brown"     "black"     "black"     "pink"      "brown"
 [7] "brown"     "black"     "black"     "pink"      "brown"     "pink"
[13] "pink"      "pink"      "pink"      "pink"      "pink"      "pink"
[19] "pink"      "black"     "pink"      "pink"      "chocolate" "brown"
[25] "brown"     "brown"     "pink"      "chocolate" "brown"     "pink"
[31] "pink"      "pink"      "chocolate" "chocolate" "pink"      "chocolate"
[37] "brown"     "brown"     "brown"     "brown"     "brown"     "pink"
[43] "brown"     "brown"     "pink"      "pink"      "brown"     "chocolate"
[49] "black"     "pink"      "pink"      "chocolate" "chocolate" "chocolate"
[55] "chocolate" "pink"      "chocolate" "black"     "pink"      "chocolate"
[61] "pink"      "pink"      "chocolate" "pink"      "brown"     "brown"
[67] "pink"      "pink"      "pink"      "pink"      "black"     "black"
[73] "pink"      "pink"      "pink"      "chocolate" "chocolate" "brown"
[79] "pink"      "brown"     "pink"      "pink"      "pink"      "black"
[85] "chocolate"
```
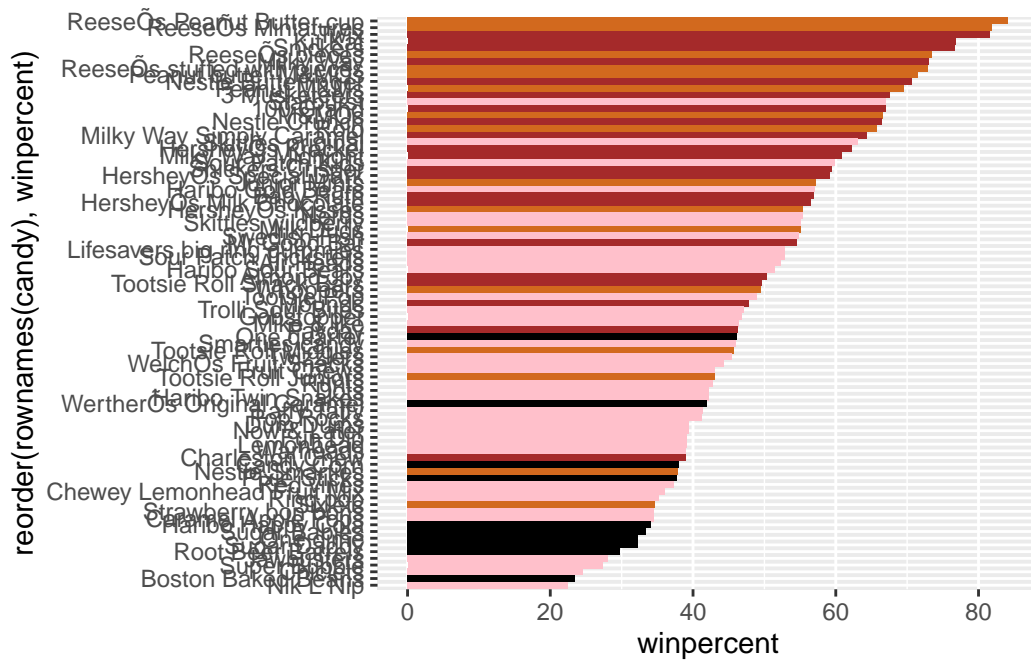
```r
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy), winpercent))+
  geom_col(fill = my_cols)
```



```r
ggsave("tmp.png")
```

```
Saving 5.5 x 3.5 in image
```

Now, for the first time, using this plot we can answer questions like: > Q17. What is the worst ranked chocolate candy?

> 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?

```r
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
```
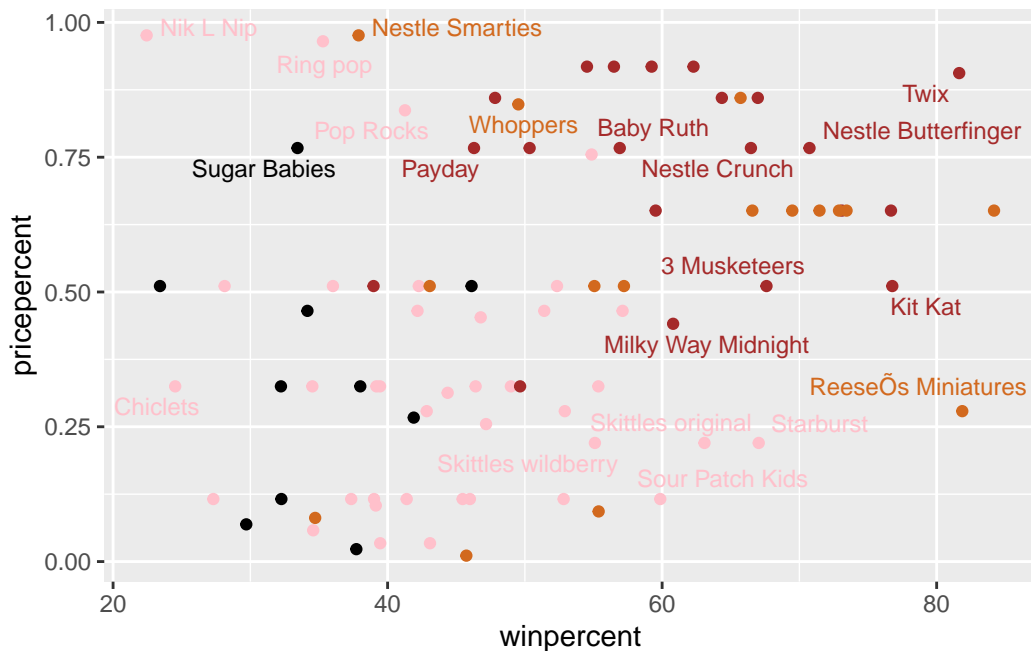

The labels on this plot are too squished and overlapping. I am going to turn to the ggrepl package and the `geom_text_repel()` function to help avoid overlapping labels.

```
library(corrplot)
```

corrplot 0.92 loaded

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