Lab 8 Halloween Mini-Project

AUTHOR
Tiffany Chin 15700705

Import candy data

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

```
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

```
#there are 38 fruity candy types.
```

What is your favorite candy? At this exact moment, lemonhead! Q3. What is your favorite candy in the dataset and what is it's `winpercent` value?

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

```
[1] 39.14106
```

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)
```

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	▇▇▇▇▇

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

`winpercent` variable looks to be different.

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

Zero represents that it is not chocolate, while one means it is!

```
skim(candy$chocolate)
```

Data summary

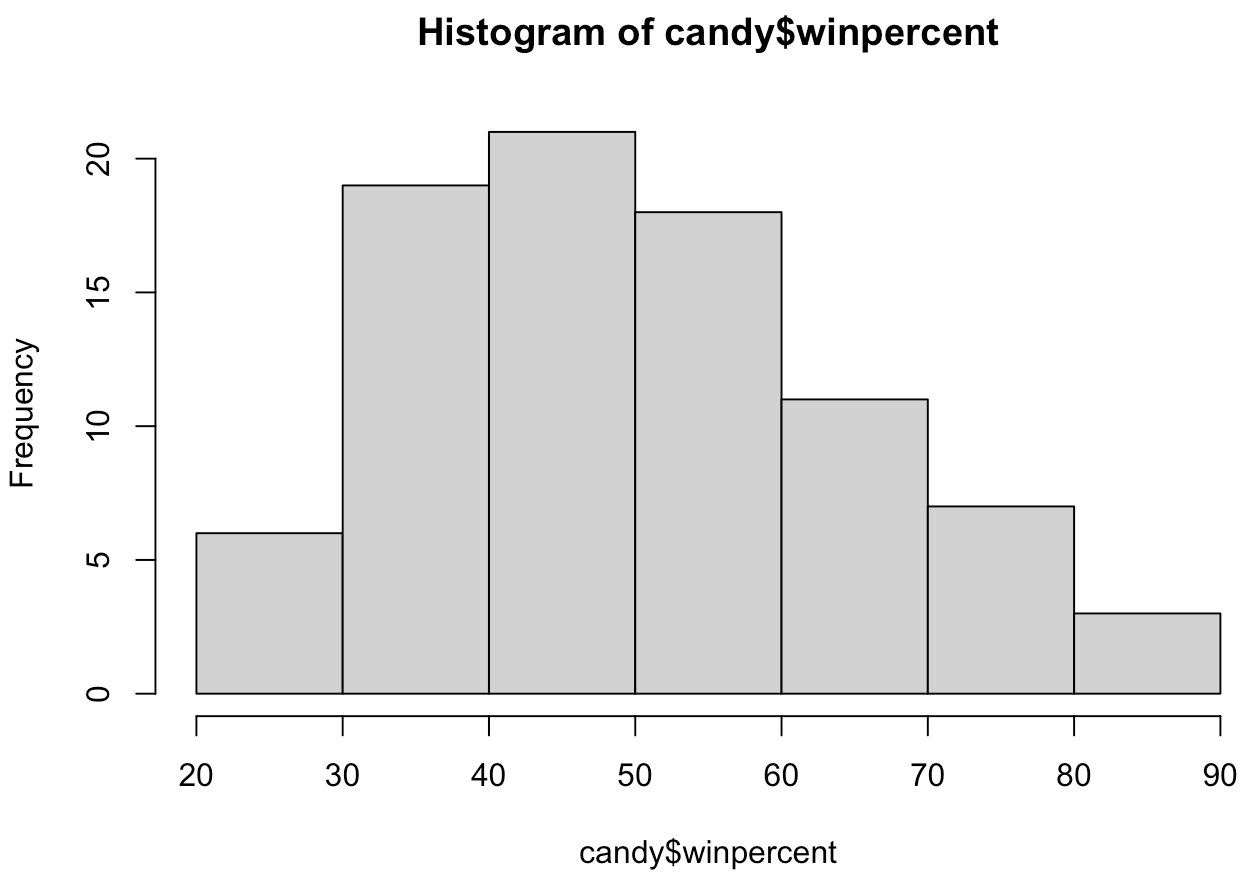
Name	candy$chocolate
Number of rows	85
Number of columns	1
_____	
Column type frequency:	
numeric	1
_____	
Group variables	None

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
data	0	1	0.44	0.5	0	0	0	1	1	▇▁▁▁▇

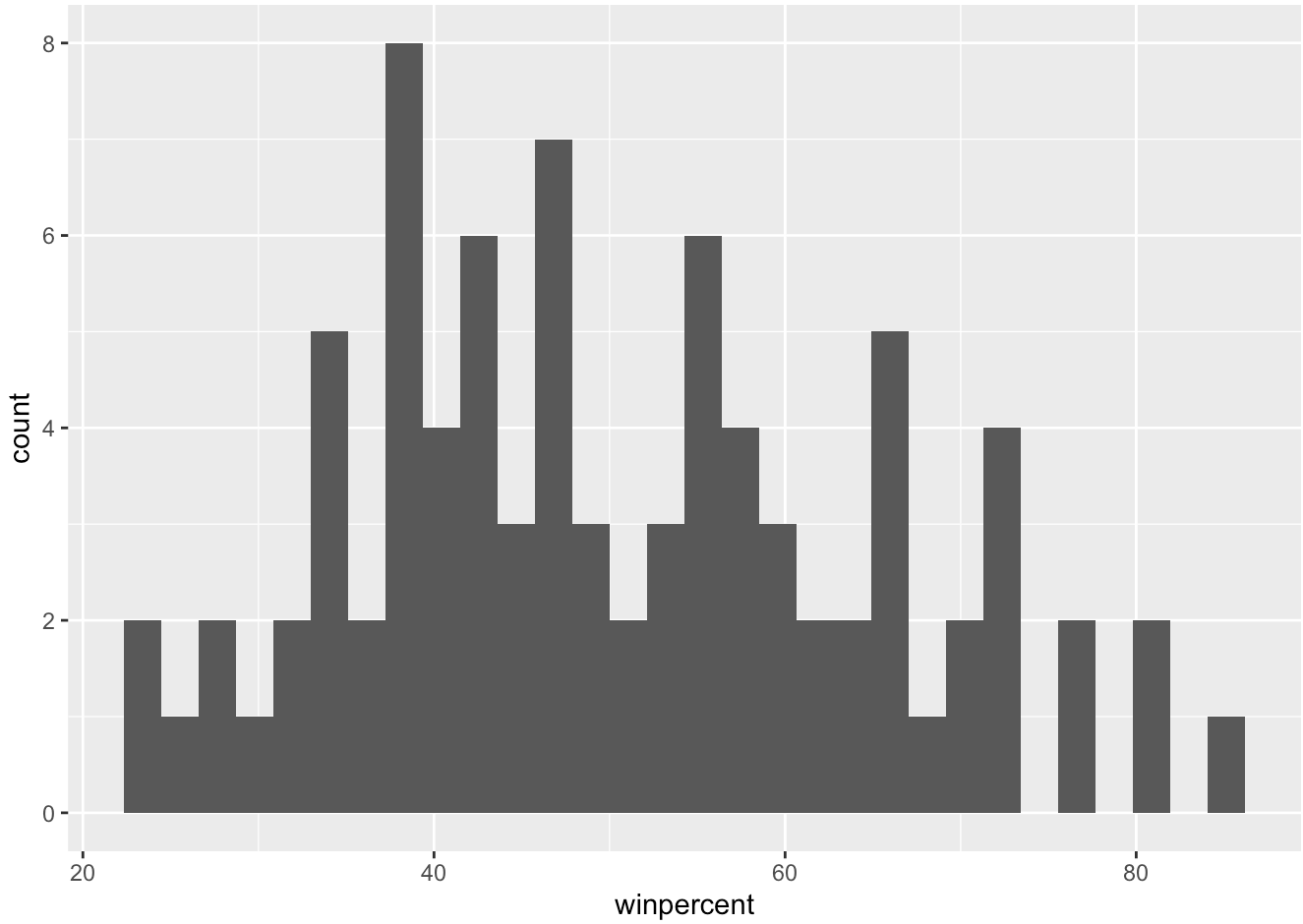
Q8. Plot a histogram of `winpercent` values

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




```
library(ggplot2)
ggplot(candy, aes(winpercent)) +
  geom_histogram()
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

It is not symmetrical.

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

It is below 50%.

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

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

```
[1] 60.92153
```

```
fruit
```

```
[1] 44.11974
```

```
as.logical(chocolate > fruit)
```

```
[1] TRUE
```

```
#Chocolate is higher ranked than fruit candy
```

Q12. Is this difference statistically significant?

```
choc <- (candy$winpercent[as.logical(candy$chocolate)])
froot <- (candy$winpercent[as.logical(candy$fruity)])
t.test(choc, froot, alternative = c("two.sided"))
```

```
    Welch Two Sample t-test

data:  choc and froot
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
```

```
#Yes, statistically significant!
```

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
```

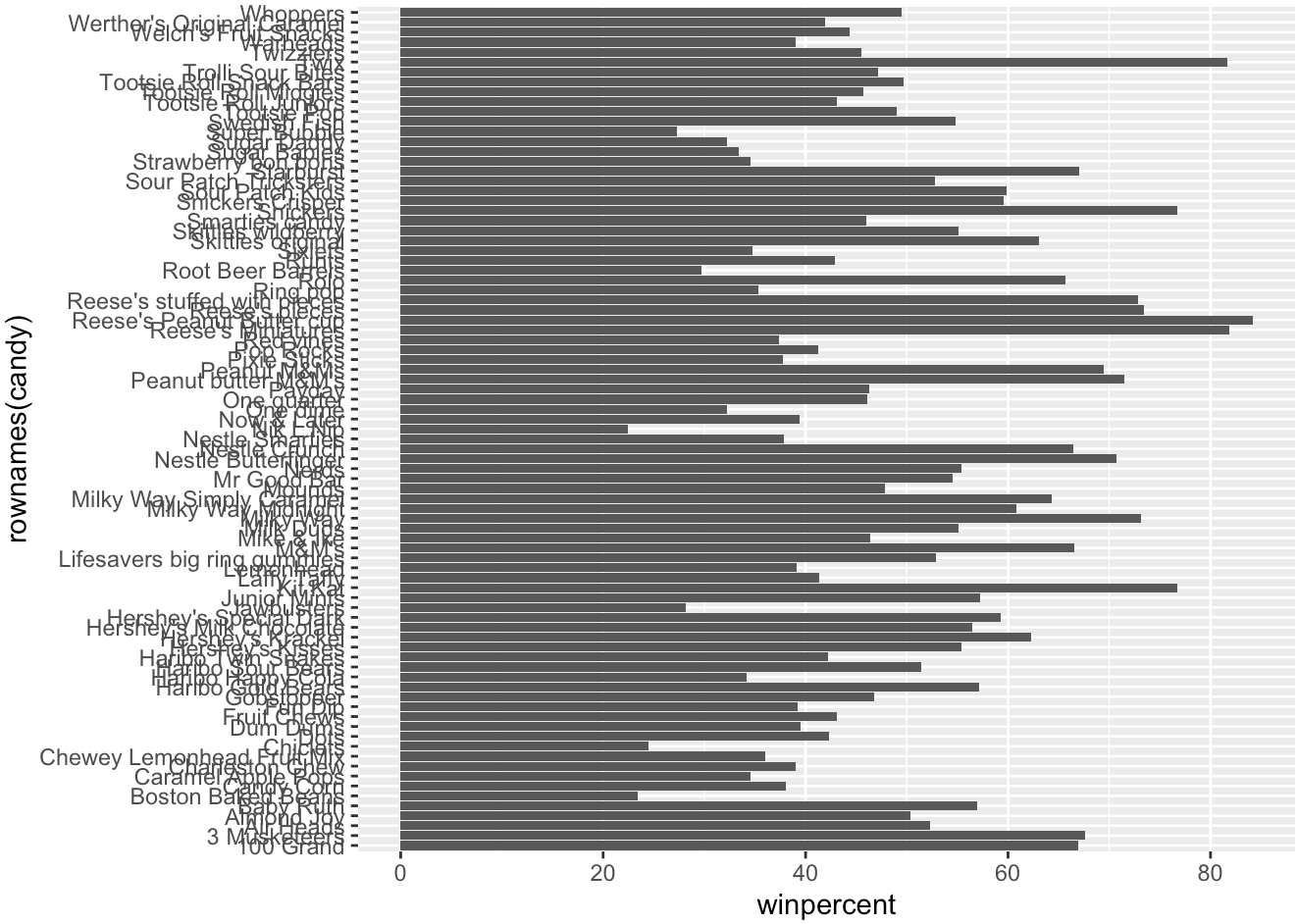
Q14. What are the top 5 all time favorite candy types out of this set?

```
head(candy[order(candy$winpercent, decreasing = T),], n =5)
```

```
                        chocolate fruity caramel peanutyalmondy 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
                        crispedricewafer hard bar pluribus sugarpercent
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
                        pricepercent 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
```

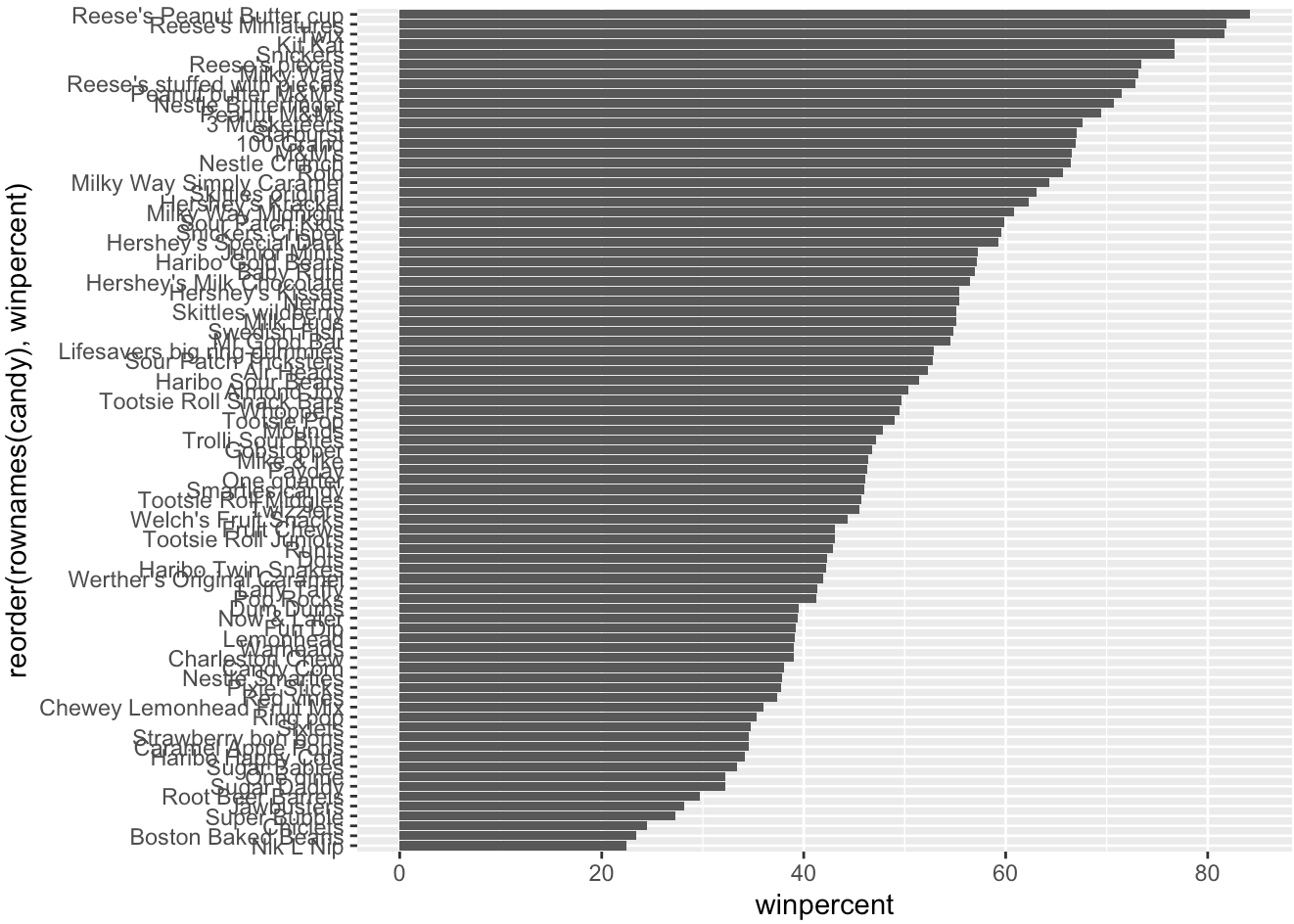
Q15. Make a first barplot of candy ranking based on `winpercent` values.

```
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()
```

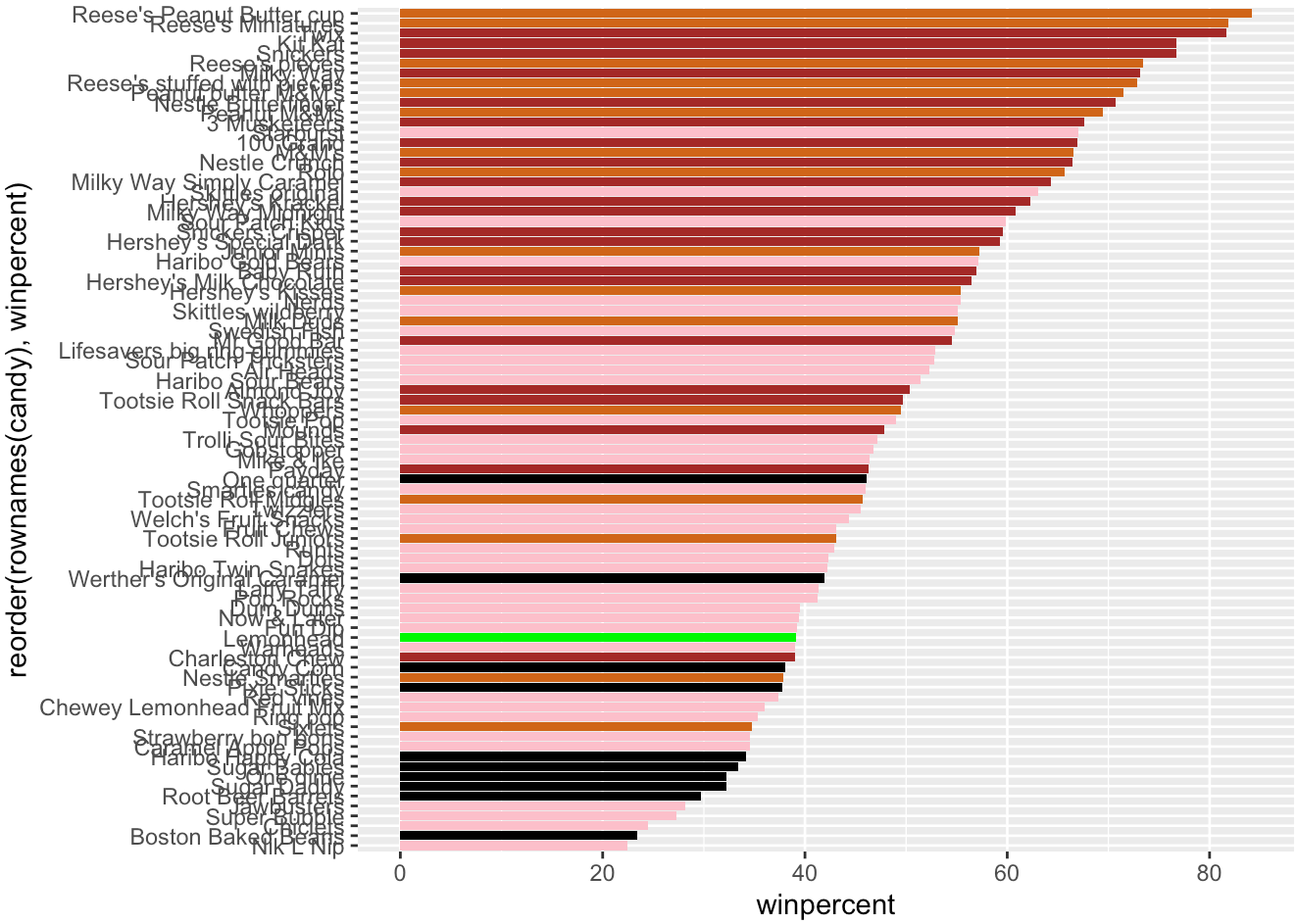


Add some useful color And color your favorite candy

```
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"
my_cols[row.names(candy) == "Lemonhead"] = "green"
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] "green"     "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"
```

```
ggplot(candy, aes(winpercent, reorder(rownames(candy), winpercent))) +
  geom_col(fill = my_cols)
```



Q17. What is the worst ranked chocolate candy?

Sixlets

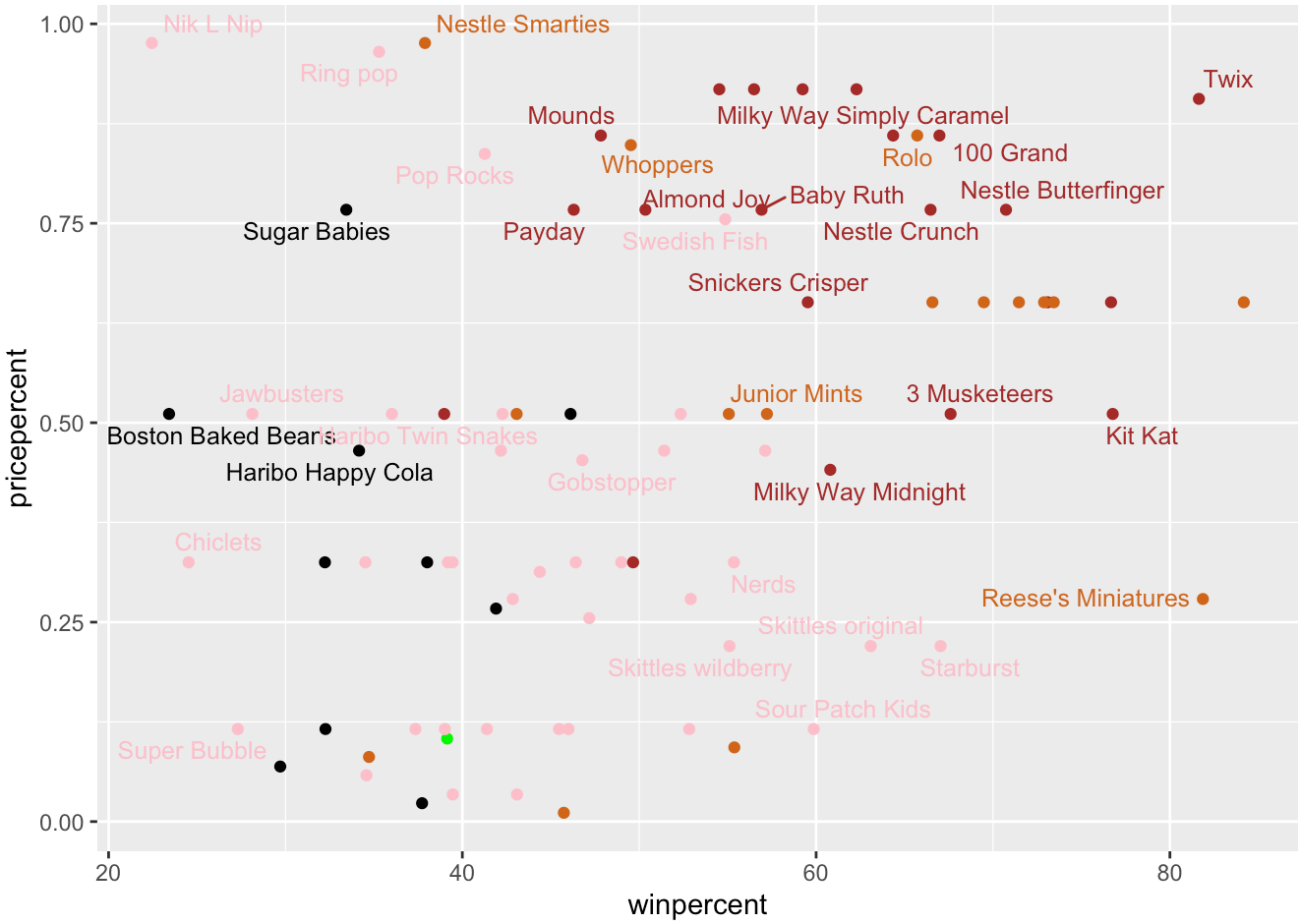
Q18. What is the best ranked fruity candy?

Nik L Nip

Taking a look at pricepercent - candy's price

```
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: 50 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?

Fruity seems pretty good. If considering individual candies, Reese's Miniatures have the best price per ranking.

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)
colnames(candy)
```

```
 [1] "chocolate"        "fruity"           "caramel"          "peanutyalmondy"
 [5] "nougat"           "crispedricewafer" "hard"             "bar"
 [9] "pluribus"         "sugarpercent"     "pricepercent"     "winpercent"
```

```
#col 11 and 12 are pricepercent and winpercent
head( candy[ord,c(11,12)], n=5 )
```

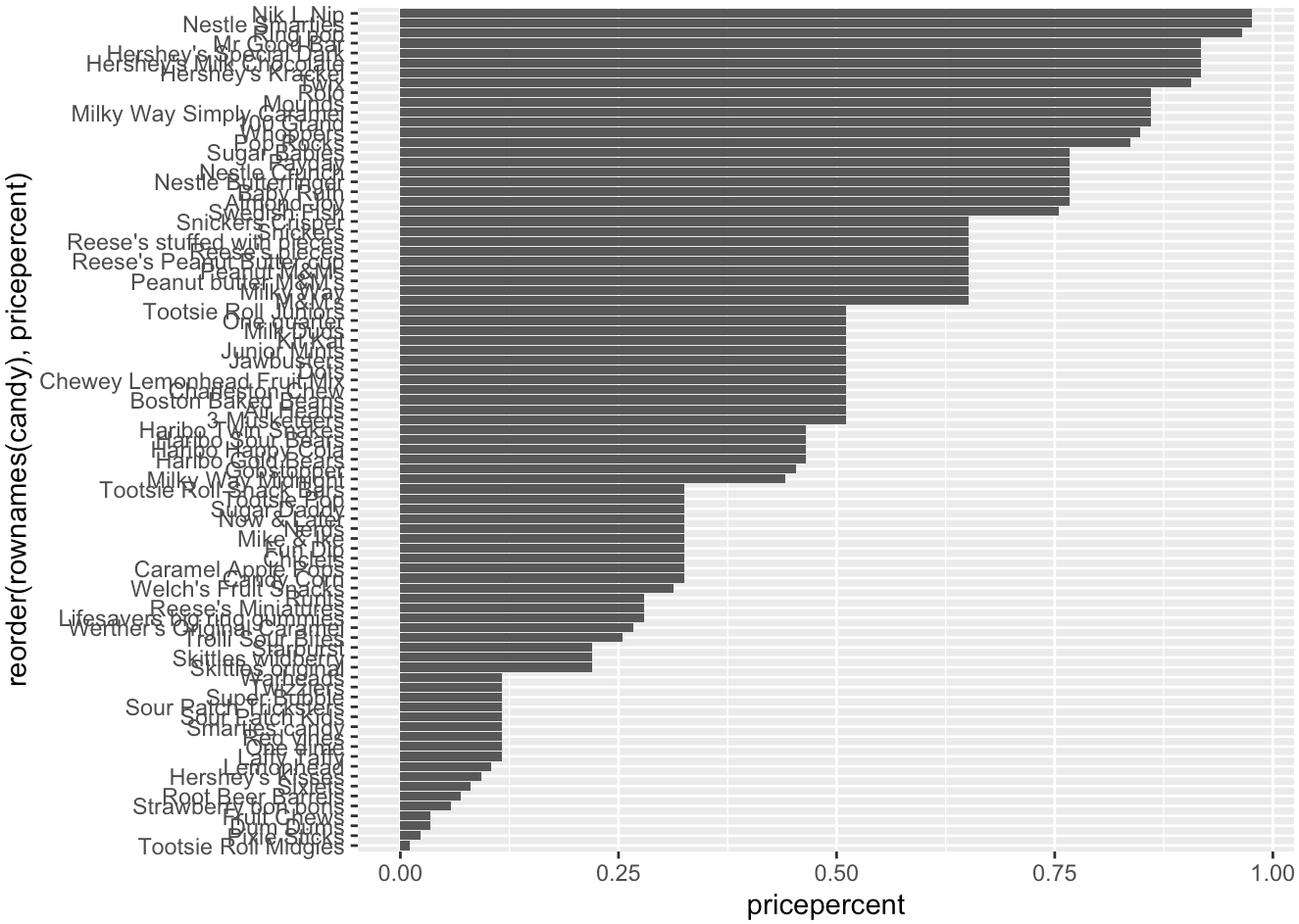
```
                    pricepercent winpercent
Nik L Nip                  0.976   22.44534
Nestle Smarties            0.976   37.88719
Ring pop                   0.965   35.29076
```

```
Hershey's Krackel             0.918   62.28448
Hershey's Milk Chocolate      0.918   56.49050
```

```
#Nik L Nip is most expensive yet least liked!!
```

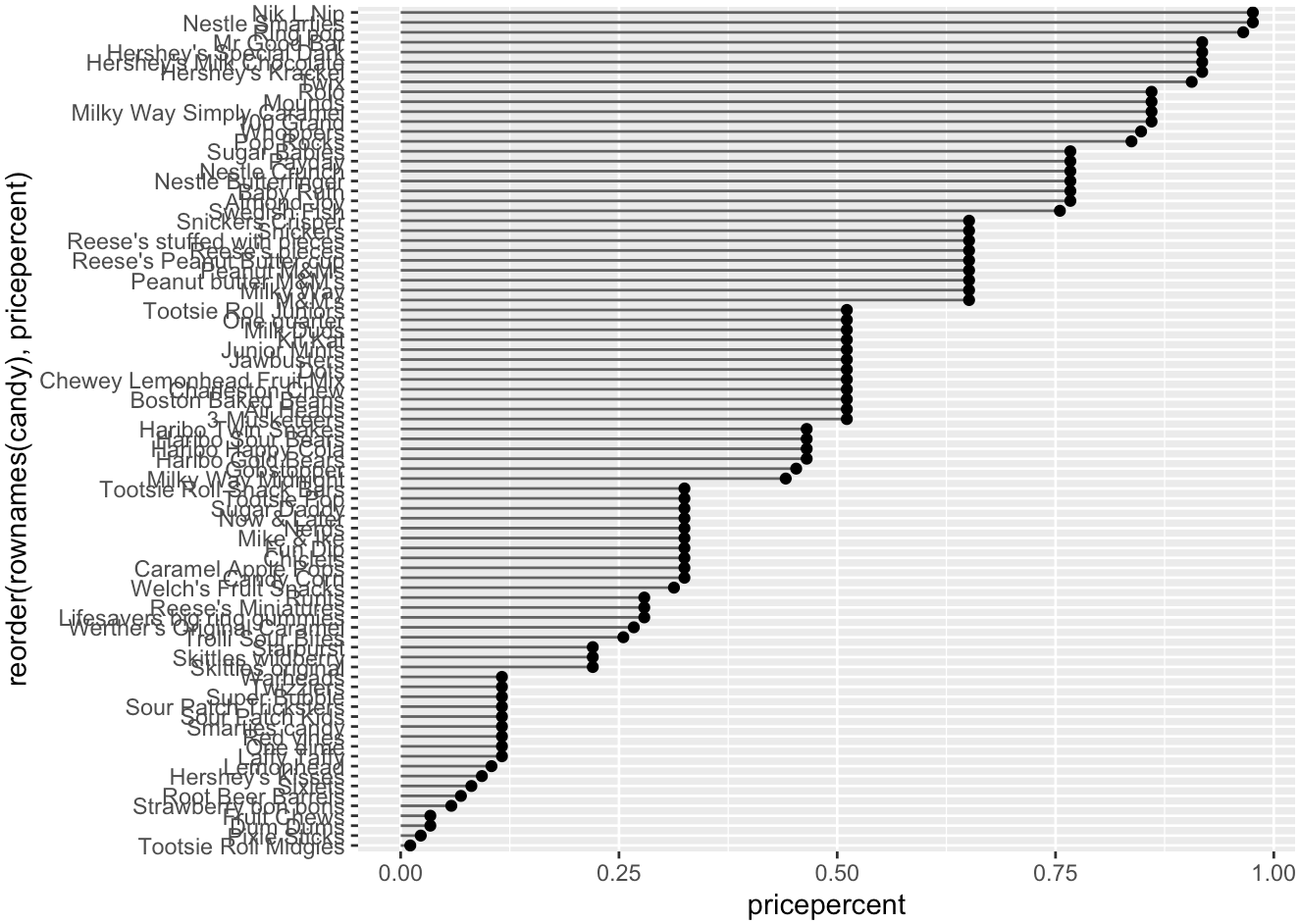
Q21. Make a barplot again with `geom_col()` this time using `pricepercent` and then improve this step by step, first ordering the x-axis by value and finally making a so called "dot chat" or "lollipop" chart by swapping `geom_col()` for `geom_point() + geom_segment()`.

```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_col()
```



Now make a lollipop chart!

```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```

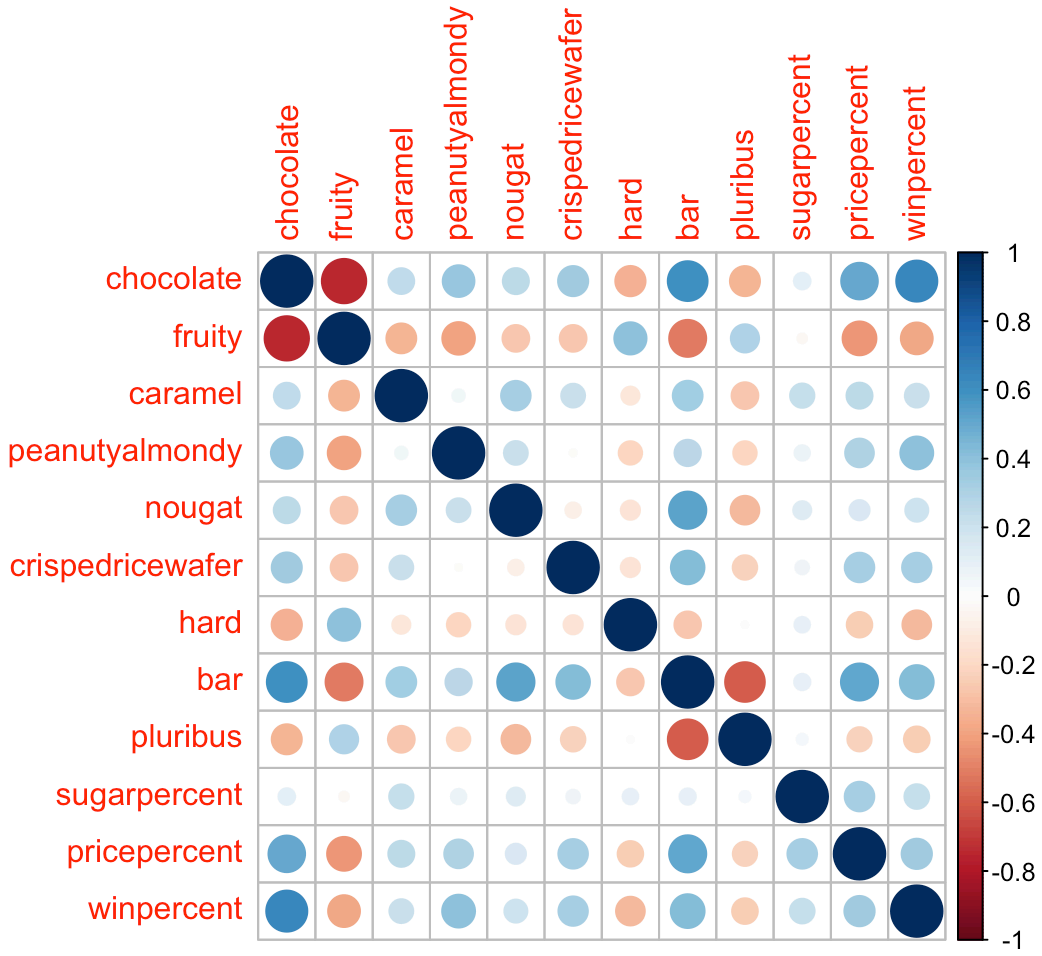


Explore the correlation structure and plot a correlation matrix

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



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

Fruity and chocolate are anti-correlated. Another strong anti-correlation is pluribus and bar.

Q23. Similarly, what two variables are most positively correlated? **Chocolate and bar** and **chocolate and winpercent** are both pretty positively correlated!

My favorite snack is jaffa cakes, so yes on fruity choc :)

Principal Component Analysis

Apply PCA using `prcomp()`

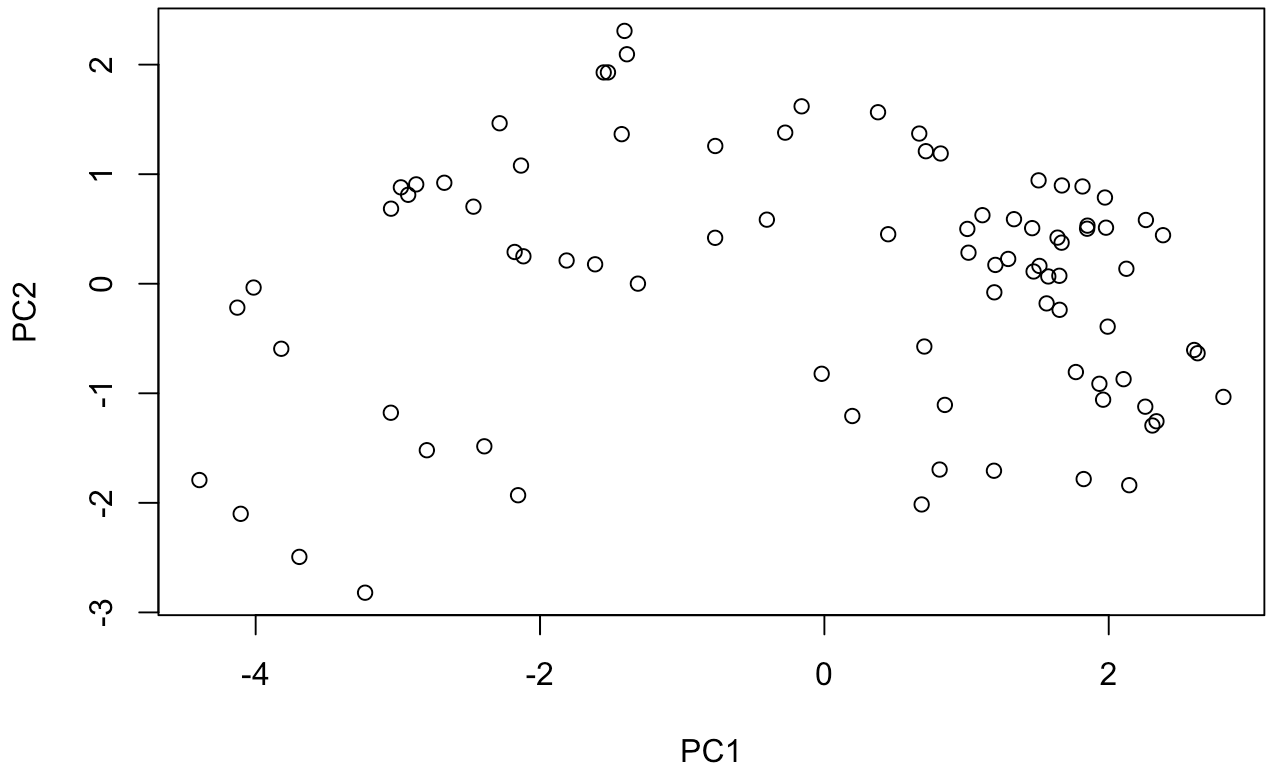
```
#we want scale = TRUE because winpercent is on a very different scale than the rest
pca <- prcomp(candy, scale=T)
summary(pca)
```

```
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
```

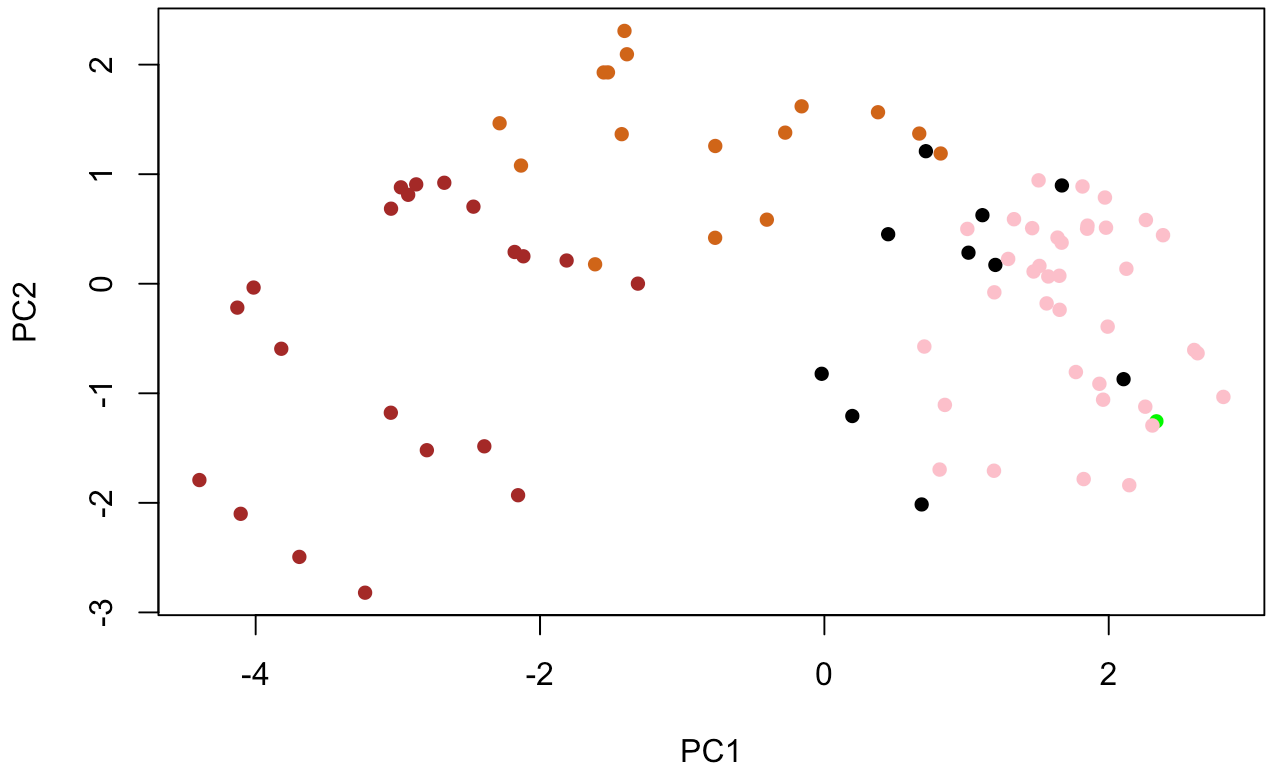
```
Proportion of Variance 0.04629 0.03833 0.03239 0.01611 0.01317
Cumulative Proportion  0.89998 0.93832 0.97071 0.98683 1.00000
```

Plot our main PCA score plot of PC1 vs PC2

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

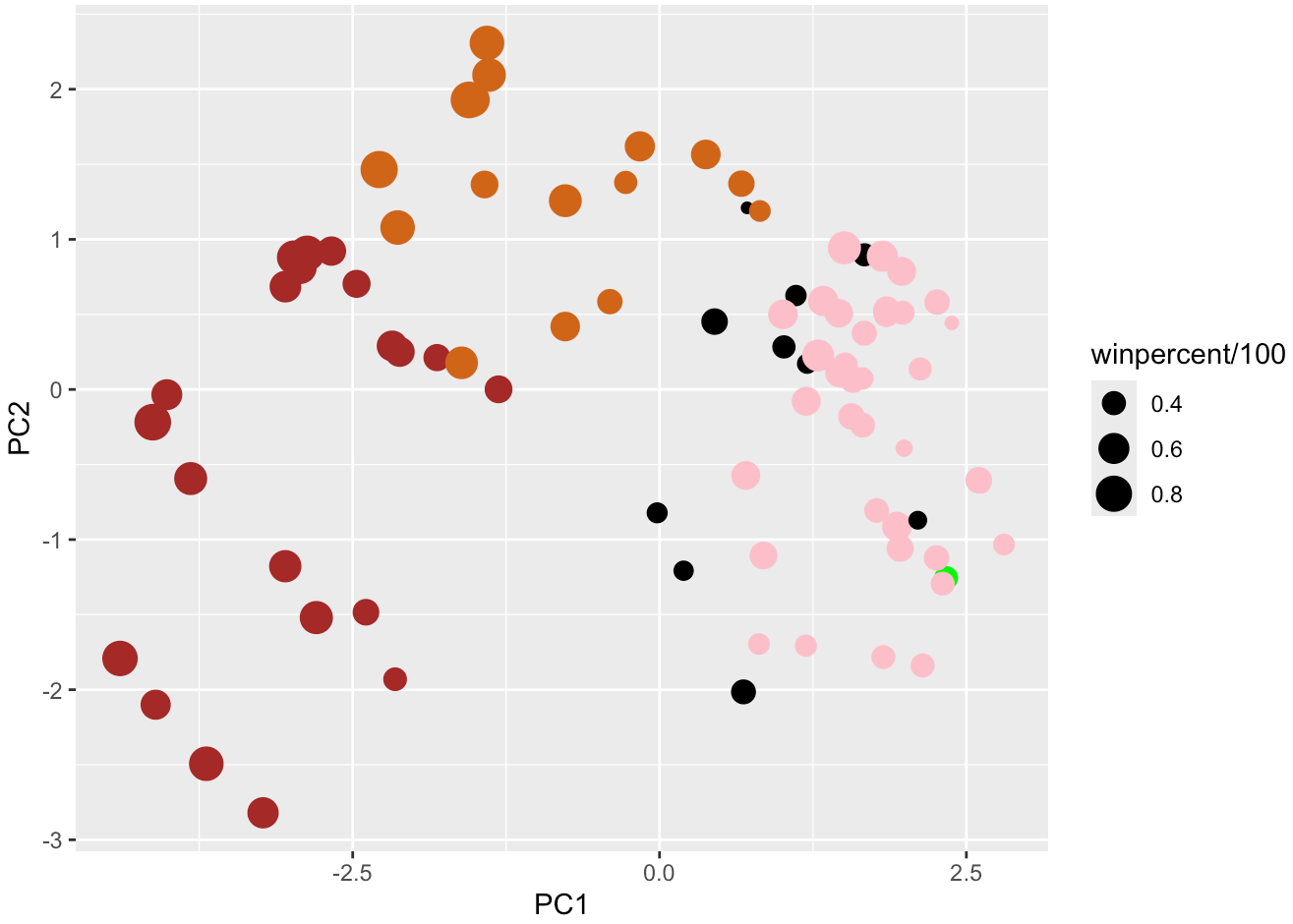



```
plot(pca$x[,1:2], col = my_cols, pch=16)
```



Plot some nicer plots with ggplot2

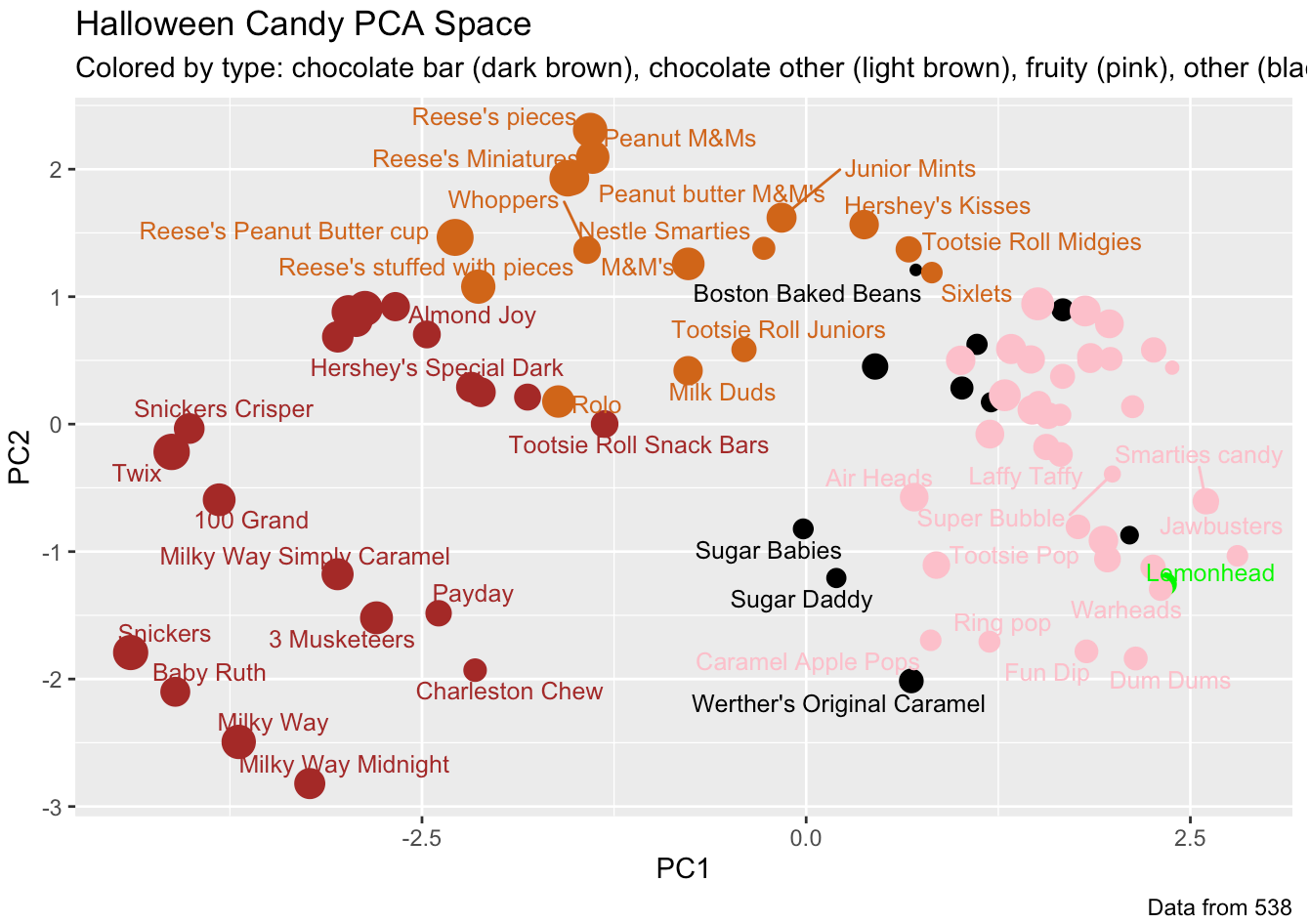
```
my_data <- cbind(candy, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



We can also use **ggrepel** package with `ggrepel::geom_text_repel()` to help label our plot with non-overlapping text.

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

```
Warning: ggrepel: 39 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

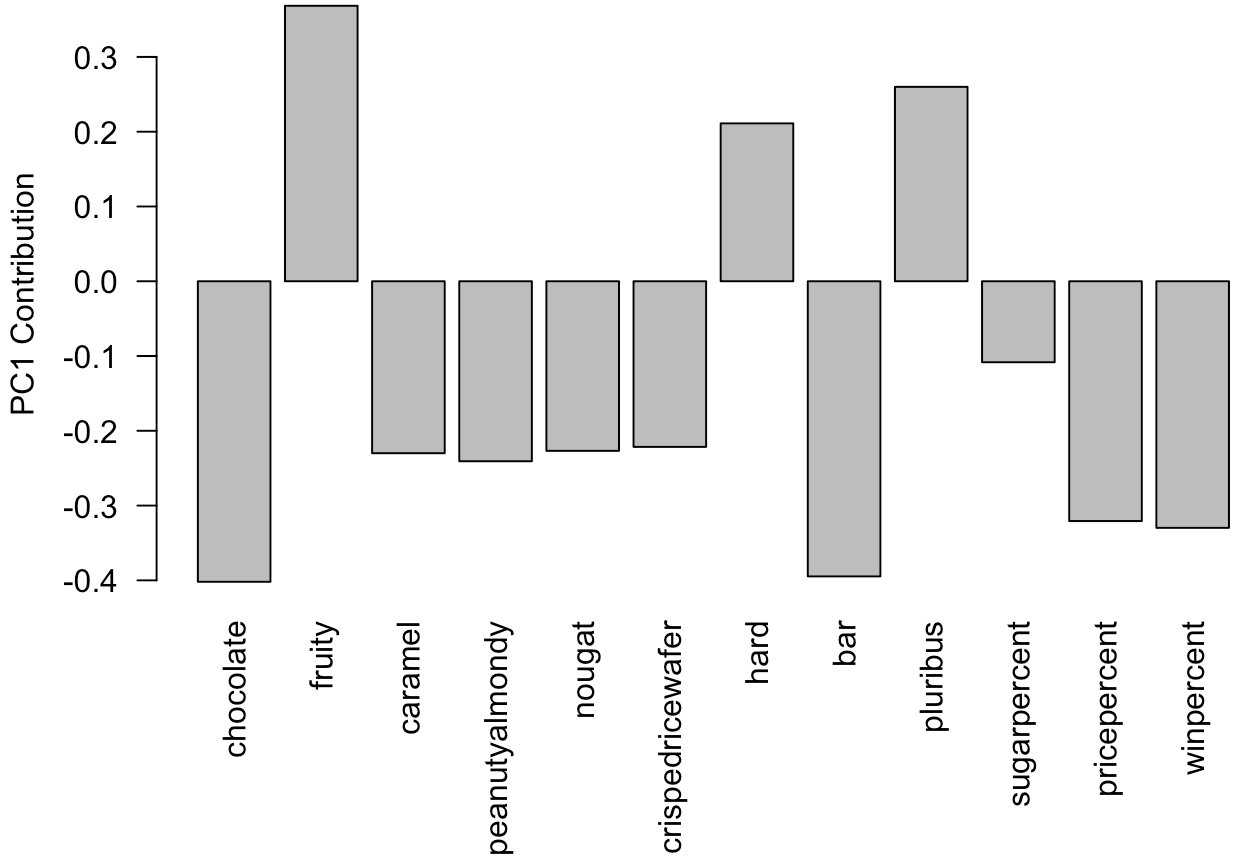


We can also use **plotly** which can generate an interactive plot!

I won't do it though since it won't render well.

Instead, let's take a look at our PCA using a bar plot.

```
#parameterssss
par(mar=c(8,4,2,2))
#plotttt
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?

Fruity, hard, and pluribus are picked up by PC1 in the positive direction. In our original plot we made of PC1 vs PC2 our data, we see the fruity candies clustered in the bottom right quadrant showing that it correlates with higher PC1 values. I think this makes sense because if we look at our correlation plot, we can see that fruity, hard, and pluribus only seem to correlate with each other, while the other variables like chocolate and bar correlate more with each other.