Class09: Candy Analysis Mini Project

Yi-Hung Lee (PID:A16587141)

Today we will examine the Halloween candy dataset

1. Importing candy data

```
candy_file <- "candy-data.csv"

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?

There are 85 candy in the dataset

```
nrow(candy)
```

```
[1] 85
```

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

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

```
[1] 38
```

2. What is your favorate candy?

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

```
[1] 81.64291
```

```r
## or

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

```
[1] 81.64291
```

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

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

```
[1] 76.7686
```

> Q4. What is the winpercent value for “Kit Kat”?

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

```
[1] 76.7686
```

> Q5. What is the winpercent value for “Tootsie Roll Snack Bars”?

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

```
[1] 49.6535
```

```r
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 hist seems to be different

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

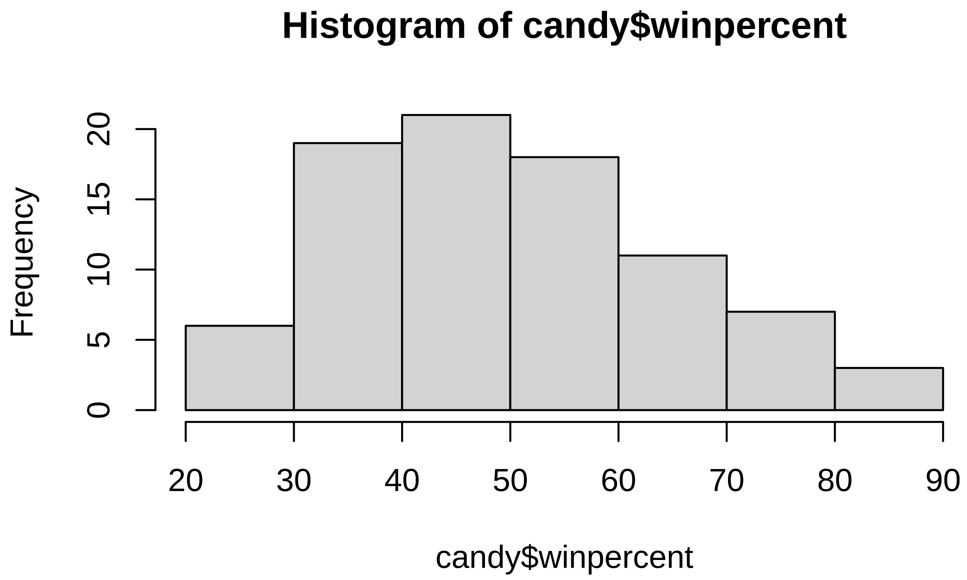
```r
candy$chocolate
```

```
 [1] 1 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 0 1 1 0 0 0 1 1 0 1 1 1
[39] 1 1 1 0 1 1 0 0 0 1 0 0 0 1 1 1 1 0 1 0 0 1 0 0 1 0 1 1 0 0 0 0 0 0 0 0 1 1
[77] 1 1 0 1 0 0 0 0 1
```

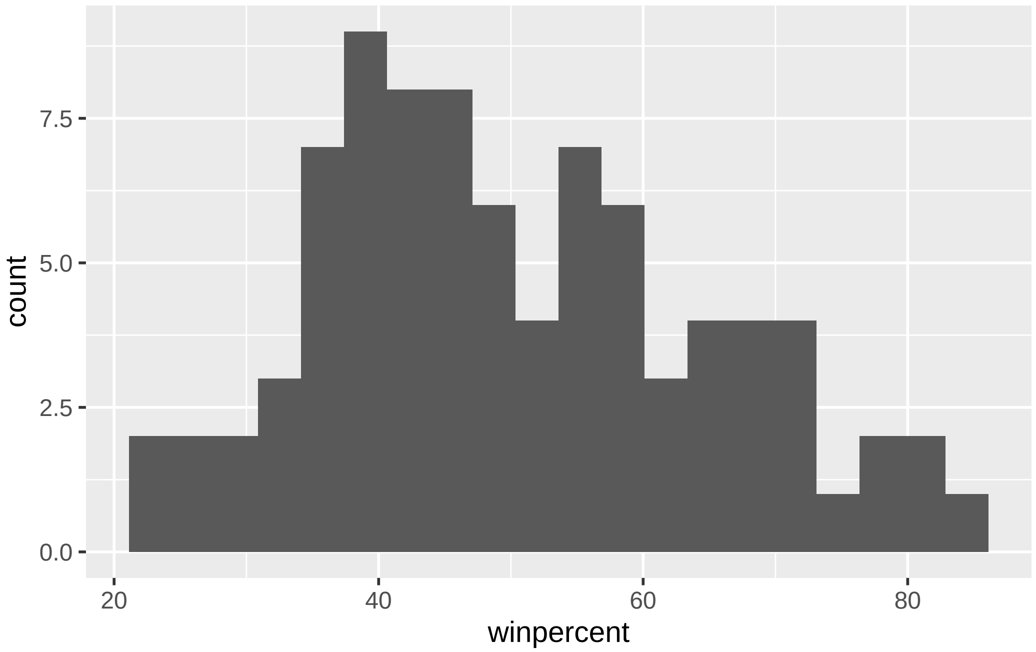
It means whether the specific candy contains chocolate

> Q8. Plot a histogram of winpercent values

```
library(ggplot2)
hist(candy$winpercent)
```



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



Q9. Is the distribution of winpercent values symmetrical?

No

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

Less than 50%

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

```
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.14   47.83   50.32   59.86   84.18 
```

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

mean(chocolate_winpercent)
```

```
[1] 60.92153
```

```
mean(fruity_winpercent)
```

```
[1] 44.11974
```

Chocolate is higher ranked than the fruity ones.

> Q12. Is this difference statistically significant?

```
pVal <- t.test(chocolate_winpercent, fruity_winpercent)$p.value
```

It is TRUE to say this difference is statistically significant

3. Overall Candy Rankings

```
head(candy[order(candy$winpercent), ])
```

```
                   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
Root Beer Barrels          0      0       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
Root Beer Barrels                 0    1   0        1        0.732        0.069
                   winpercent
Nik L Nip            22.44534
Boston Baked Beans   23.41782
Chiclets             24.52499
Super Bubble         27.30386
Jawbusters           28.12744
Root Beer Barrels    29.70369
```

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

```r
candy_order <- candy[order(candy$winpercent), ]
head(row.names(candy_order), n = 5)
```

```
[1] "Nik L Nip"          "Boston Baked Beans" "Chiclets"
[4] "Super Bubble"       "Jawbusters"
```

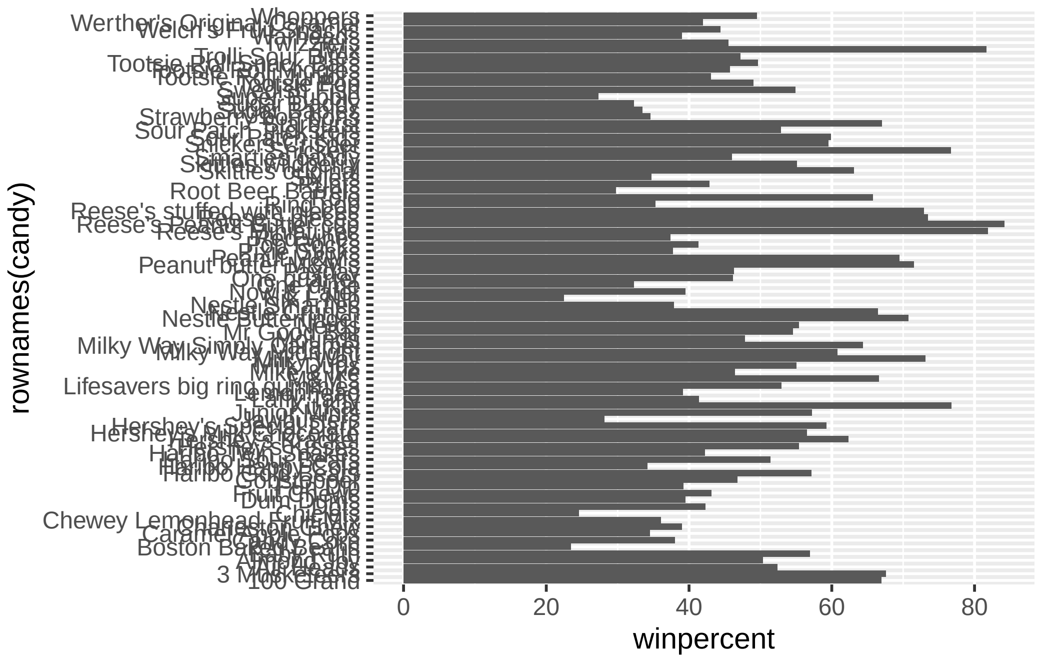
Q14. What are the top 5 all time favorite candy types out of this set?

```r
tail(row.names(candy_order), n = 5)
```

```
[1] "Snickers"                  "Kit Kat"
[3] "Twix"                      "Reese's Miniatures"
[5] "Reese's Peanut Butter cup"
```

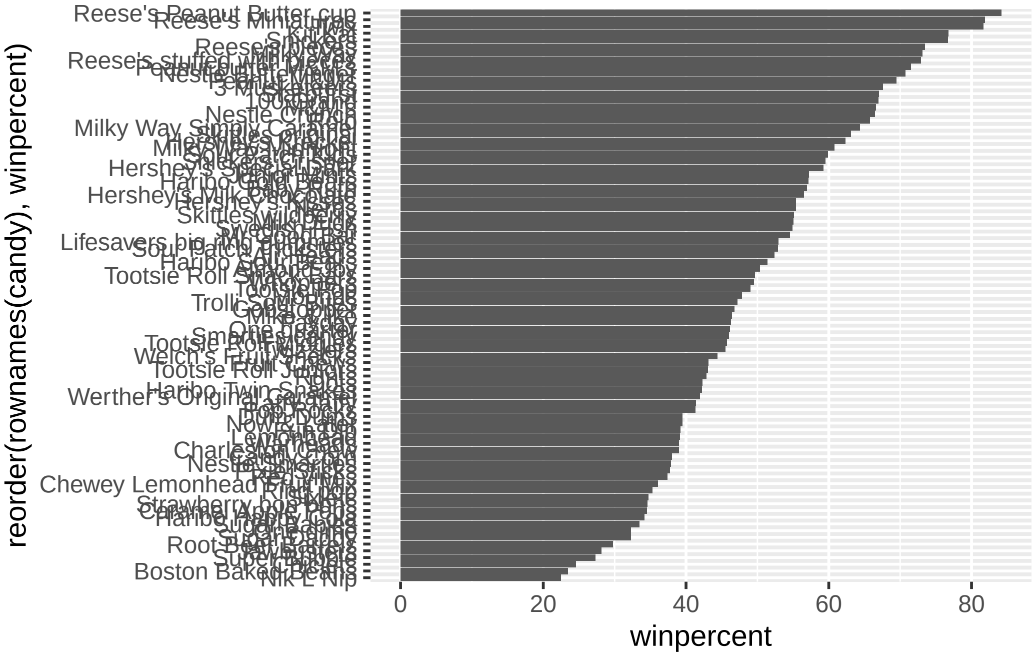
Q15. Make a first barplot of candy ranking based on winpercent values.

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



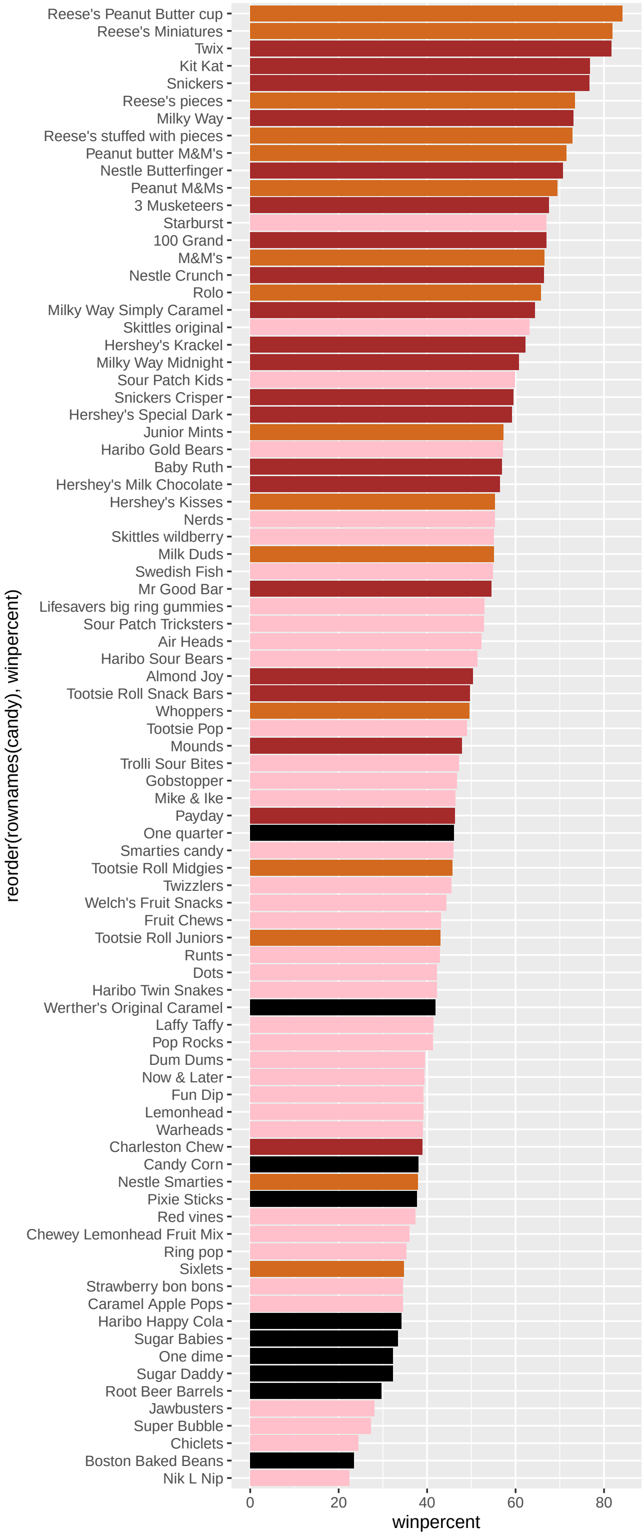
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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



```r
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"
```

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


Reese's Peanut Butter cup
Reese's Miniatures
Twix
Kit Kat
Snickers
Reese's pieces
Milky Way
Reese's stuffed with pieces
Peanut butter M&M's
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&M's
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
Hershey's Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
Hershey's Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
Hershey's Milk Chocolate
Hershey's Kisses
Nerds
Skittles wildberry
Milk Duds
Swedish Fish
Mr Good Bar
Lifesavers big ring gummies
Sour Patch Tricksters
Air Heads
Haribo Sour Bears
Almond Joy
Tootsie Roll Snack Bars
Whoppers
Tootsie Pop
Mounds
Trolli Sour Bites
Gobstopper
Mike & Ike
Payday
One quarter
Smarties candy
Tootsie Roll Midgies
Twizzlers
Welch's Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
Werther's Original Caramel
Laffy Taffy
Pop Rocks
Dum Dums
Now & Later
Fun Dip
Lemonhead
Warheads
Charleston Chew
Candy Corn
Nestle Smarties
Pixie Sticks
Red vines
Chewey Lemonhead Fruit Mix
Ring pop
Sixlets
Strawberry bon bons
Caramel Apple Pops
Haribo Happy Cola
Sugar Babies
One dime
Sugar Daddy
Root Beer Barrels
Jawbusters
Super Bubble
Chiclets
Boston Baked Beans
Nik L Nip
reorder(rownames(candy), winpercent)
0
20
40
60
80
winpercent

Q17. What is the worst ranked chocolate candy?

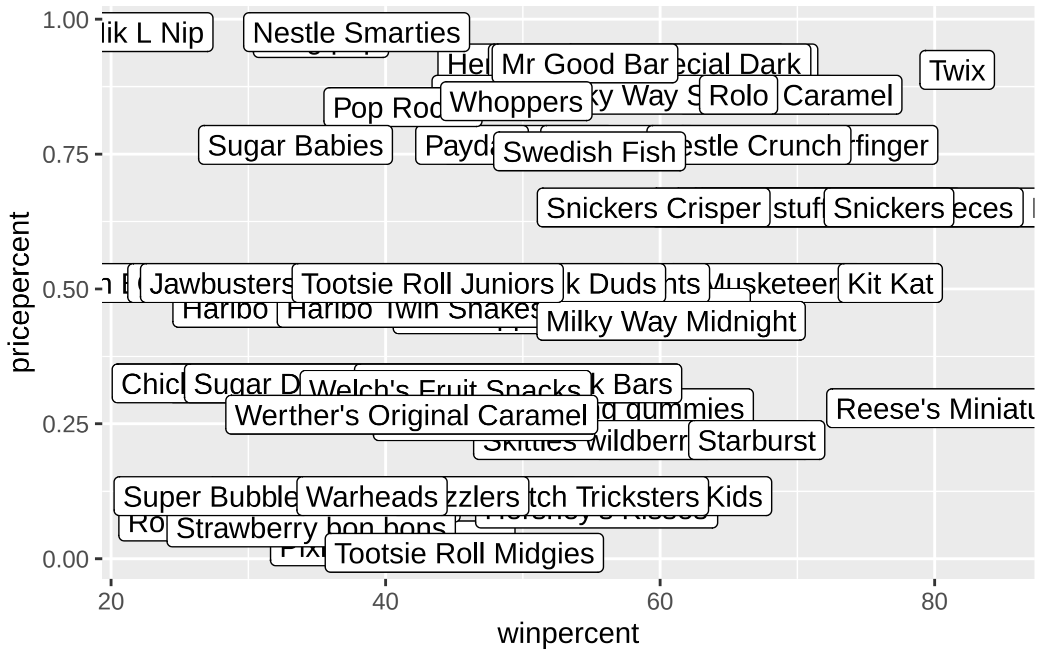
Sixlets

Q18. What is the best ranked fruity candy?

Starburst

4. Taking a look at pricepercent

```r
ggplot(candy, aes(winpercent, pricepercent, label = rownames(candy))) +
  geom_point(col = my_cols) +
  geom_label()
```

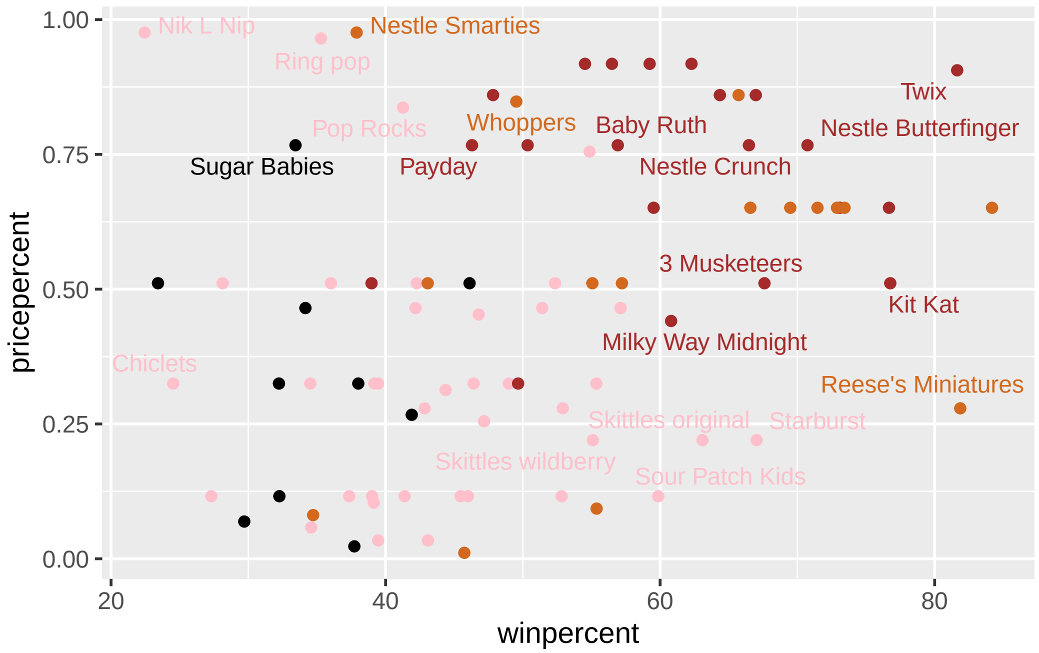


Too MANY LABELLLLSSS!!!!!

```r
library(ggrepel)

ggplot(candy, aes(winpercent, pricepercent, label = rownames(candy))) +
  geom_point(col = my_cols) +
  geom_text_repel(col = my_cols, size = 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?

Super Bubble

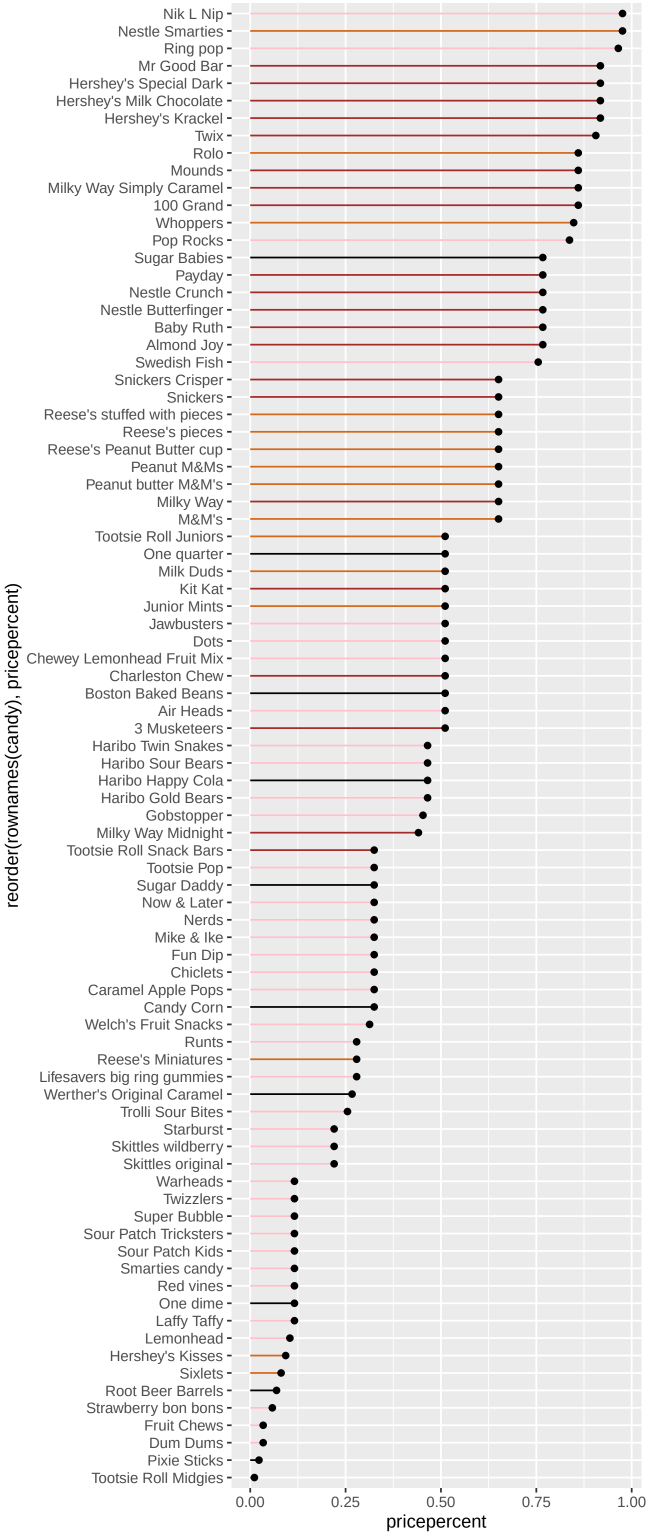
> Q20. What are the top 5 most expensive candy types in the dataset and of these which is the least popular?

```
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
```

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_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col=my_cols) +
    geom_point()
```

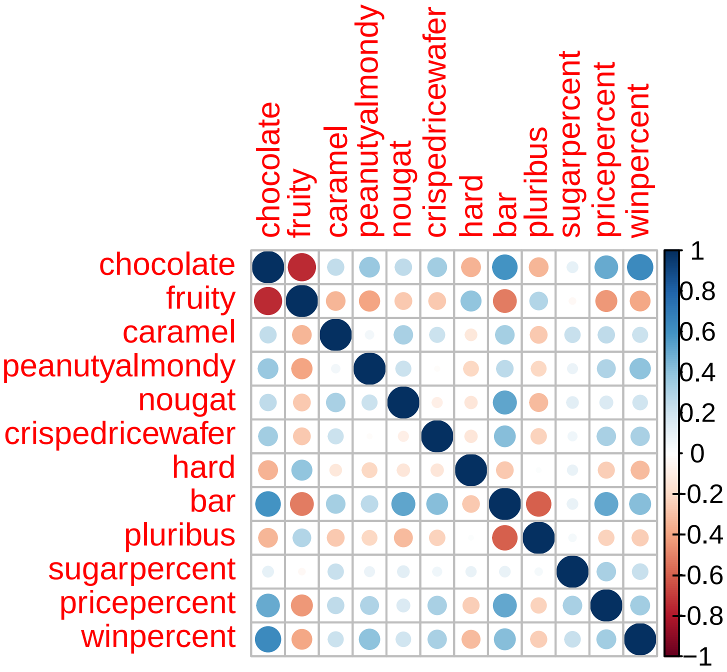

Nik L Nip
Nestle Smarties
Ring pop
Mr Good Bar
Hershey's Special Dark
Hershey's Milk Chocolate
Hershey's Krackel
Twix
Rolo
Mounds
Milky Way Simply Caramel
100 Grand
Whoppers
Pop Rocks
Sugar Babies
Payday
Nestle Crunch
Nestle Butterfinger
Baby Ruth
Almond Joy
Swedish Fish
Snickers Crisper
Snickers
Reese's stuffed with pieces
Reese's pieces
Reese's Peanut Butter cup
Peanut M&Ms
Peanut butter M&M's
Milky Way
M&M's
Tootsie Roll Juniors
One quarter
Milk Duds
Kit Kat
Junior Mints
Jawbusters
Dots
Chewey Lemonhead Fruit Mix
Charleston Chew
Boston Baked Beans
Air Heads
3 Musketeers
Haribo Twin Snakes
Haribo Sour Bears
Haribo Happy Cola
Haribo Gold Bears
Gobstopper
Milky Way Midnight
Tootsie Roll Snack Bars
Tootsie Pop
Sugar Daddy
Now & Later
Nerds
Mike & Ike
Fun Dip
Chiclets
Caramel Apple Pops
Candy Corn
Welch's Fruit Snacks
Runts
Reese's Miniatures
Lifesavers big ring gummies
Werther's Original Caramel
Trolli Sour Bites
Starburst
Skittles wildberry
Skittles original
Warheads
Twizzlers
Super Bubble
Sour Patch Tricksters
Sour Patch Kids
Smarties candy
Red vines
One dime
Laffy Taffy
Lemonhead
Hershey's Kisses
Sixlets
Root Beer Barrels
Strawberry bon bons
Fruit Chews
Dum Dums
Pixie Sticks
Tootsie Roll Midgies
reorder(rownames(candy), pricepercent)
0.00
0.25
0.50
0.75
1.00
pricepercent

5. Exploring the correlation structure

```
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)?

Chocolate and Fruity

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

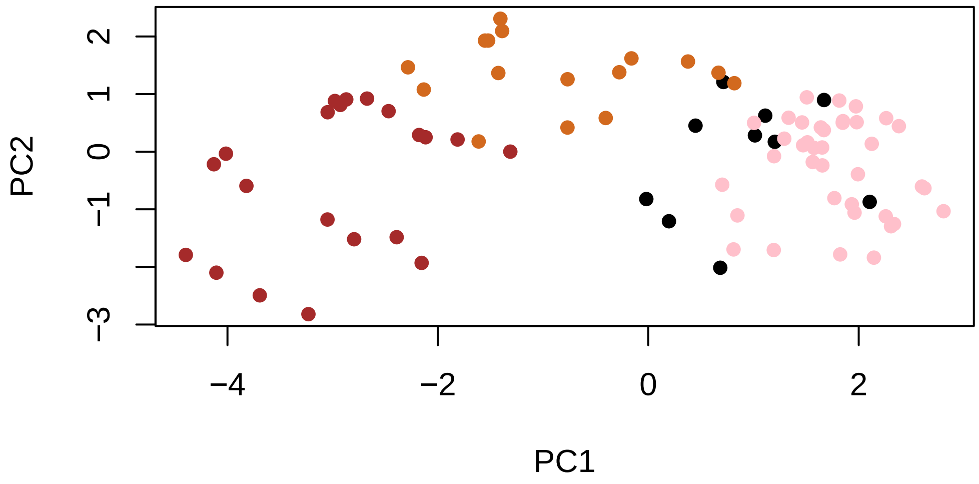
Chocolate and bar

6. Principal Component Analysis

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

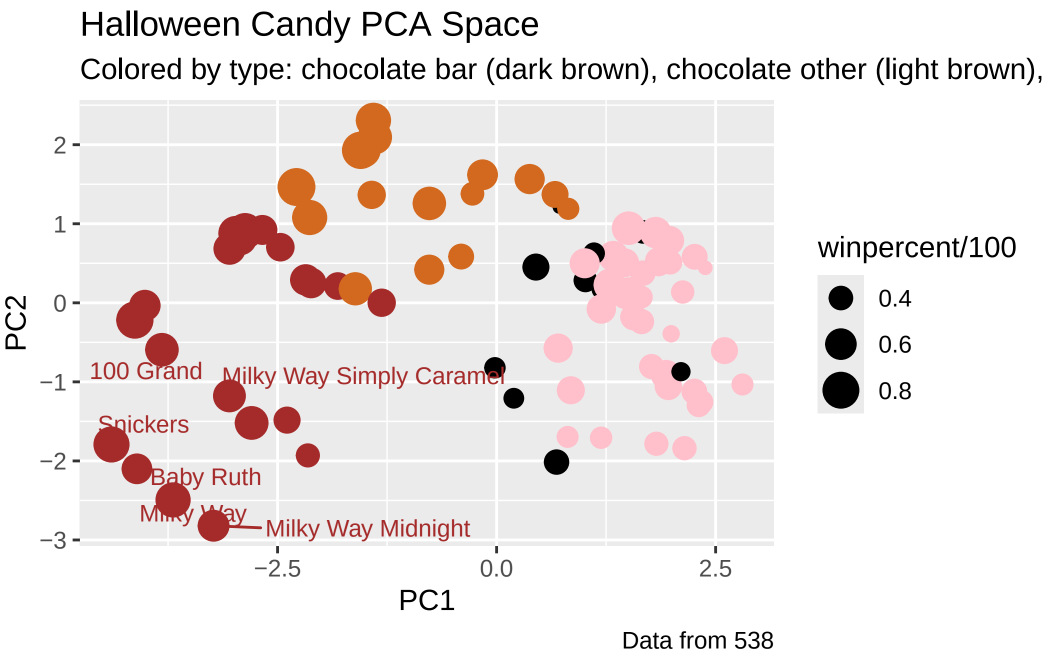


```
my_data <- cbind(candy, pca$x[, 1:3])

p <-  ggplot(my_data, aes(x = PC1, y = PC2, size = winpercent/100, label = rownames(my_dat
  geom_point(col = my_cols) +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
       caption="Data from 538") +
  geom_text_repel(col = my_cols, size = 3, max.overlaps = 3)

p
```

```
Warning: ggrepel: 79 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



```
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
```

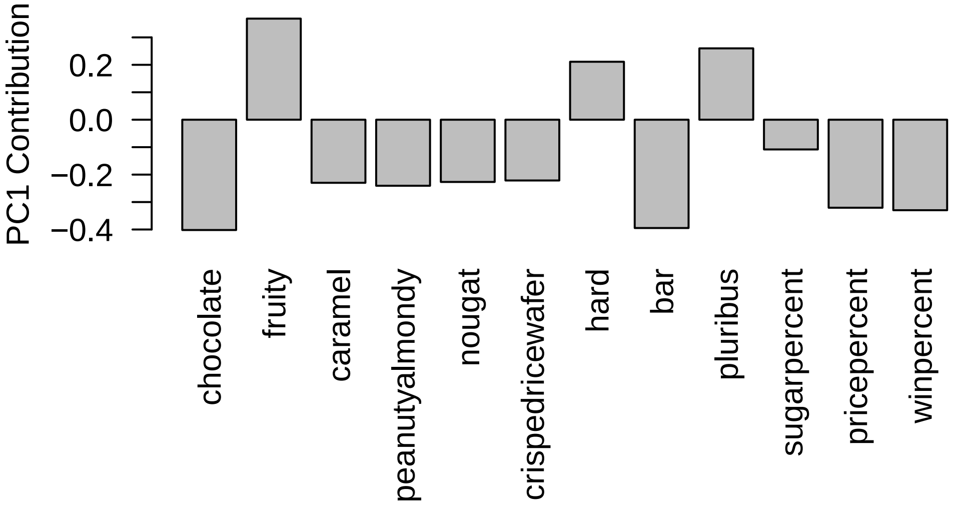
```
# ggplotly(p)
```

How do the original variables contributes to our PCs? For this we look at the loading component of our results objects

```
pca$rotation[,1]
```

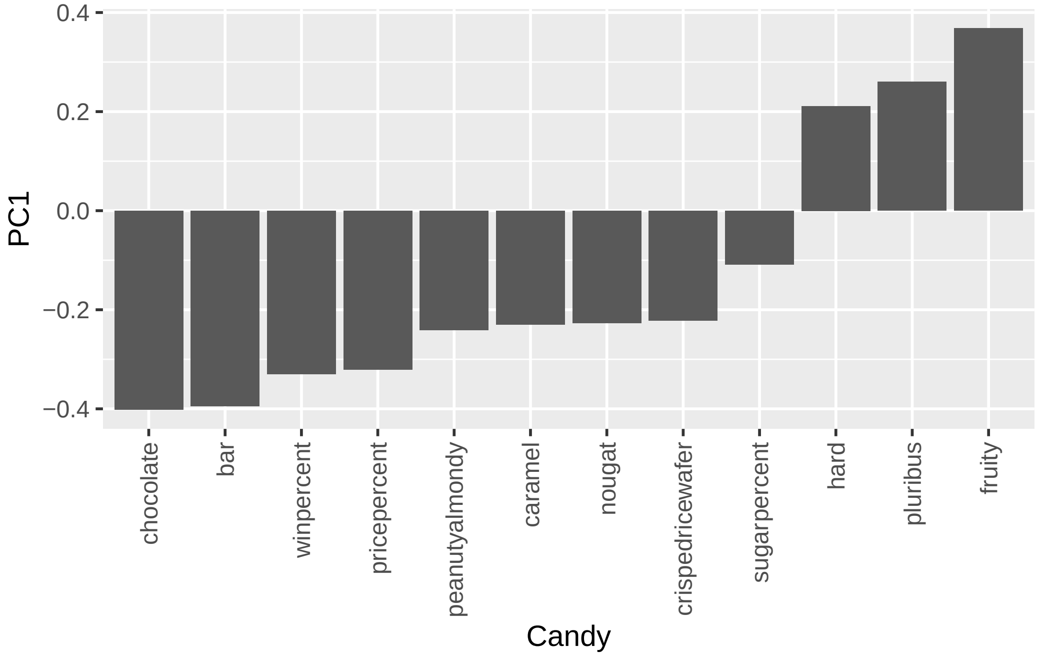
```
       chocolate           fruity          caramel   peanutyalmondy
      -0.4019466        0.3683883       -0.2299709       -0.2407155
          nougat crispedricewafer             hard              bar
      -0.2268102       -0.2215182        0.2111587       -0.3947433
        pluribus     sugarpercent     pricepercent       winpercent
       0.2600041       -0.1083088       -0.3207361       -0.3298035
```

```
par(mar = c(8, 4, 4, 2))
barplot(pca$rotation[,1], las=2, ylab="PC1 Contribution")
```



```
res <- pca$rotation

ggplot(res, aes(PC1, reorder(rownames(res), PC1))) +
  geom_col() +
  coord_flip() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))+
  ylab("Candy")
```



Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

Fruit, Pluribus and hard are all picked up in the +ve direction. It make sense from the corralation from the candy experience of eating candy.