ClassCandy

Ksenia Mekhonoshina

Table of contents

Background

In todays mini project, we wil analyze candy data with exploratory graphics, basic statistics, correlation analysis and principal component analysis methods we have been learning.

Data Import

```
candy_file <- read.csv("Candy Mini-Project Data.txt")
candy <- read.csv("Candy Mini-Project 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
```

What is the dataset?

Let's analyze our data.

Q1. How many different candy types are in this dataset?

```
nrow(candy)
```

```
[1] 85
```

There are 85 candy types in this dataset

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

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

```
[1] 38
```

There are 38 fruity candy types

Fav candy

We can find out the percentage of people who prefer this candy over another randomly chosen candy by using `winpercent`:

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

```
[1] 81.64291
```

> Q3. What is your favorite candy (other than Twix) in the dataset and what is it's winpercent value?

Kit Kat, and the winpercent is 76.7686

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

Opps, it is 76.7686. We got this using `candy["Kit Kat", ]$winpercent` and r in front of the code

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

49.653503. Not that many ppl like it

```
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?

N_missing and complete_rate are either 0 or 1

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

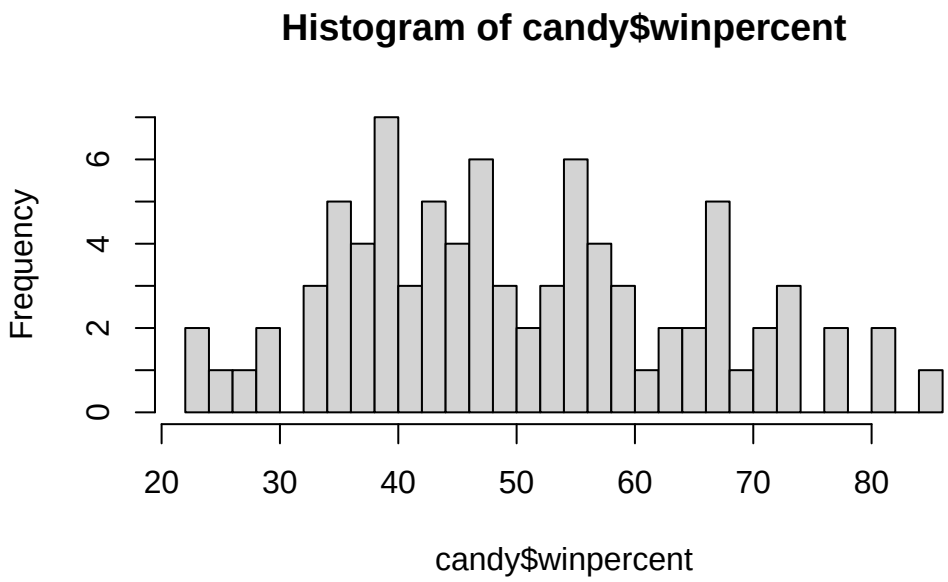
Whether there was any data input in the data set

Exploratory analysis

Q8. Plot a histogram of winpercent values using both base R an ggplot2.

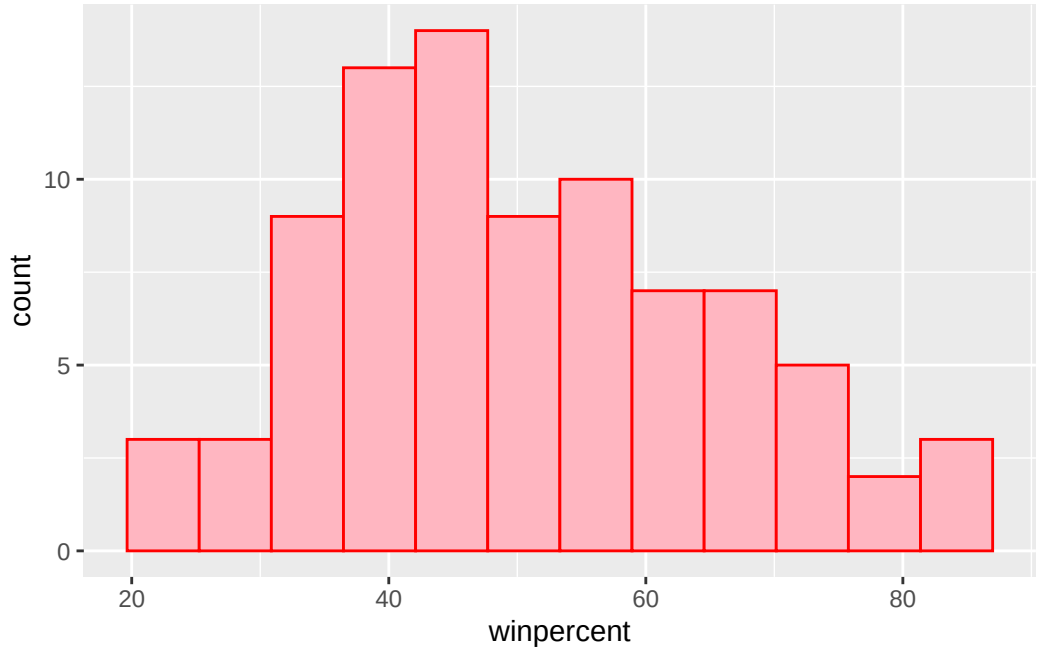
Lets do base R first:

```
hist(candy$winpercent, breaks=25)
```



Now, let's do the ggplot:

```
library(ggplot2)
ggplot(candy) +
     aes(winpercent)+
     geom_histogram(bins=12, fill="lightpink", col="red")
```



Q9. Is the distribution of winpercent values symmetrical?

No, it is not (doesn't look like it)

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

It is above, we got it by using `mean`

```
mean(candy$winpercent)
```

```
[1] 50.31676
```

Not that good to use with non-normally distributed data

```
summary(candy)
```

```
   chocolate          fruity           caramel        peanutyalmondy
 Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
 Mean   :0.4353   Mean   :0.4471   Mean   :0.1647   Mean   :0.1647
 3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
```

```
 Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
    nougat        crispedricewafer       hard             bar
 Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :0.0000
 1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:0.0000
 Median :0.00000   Median :0.00000   Median :0.0000   Median :0.0000
 Mean   :0.08235   Mean   :0.08235   Mean   :0.1765   Mean   :0.2471
 3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:0.0000   3rd Qu.:0.0000
 Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :1.0000
   pluribus       sugarpercent     pricepercent      winpercent
 Min.   :0.0000   Min.   :0.0110   Min.   :0.0110   Min.   :22.45
 1st Qu.:0.0000   1st Qu.:0.2200   1st Qu.:0.2550   1st Qu.:39.14
 Median :1.0000   Median :0.4650   Median :0.4650   Median :47.83
 Mean   :0.5176   Mean   :0.4786   Mean   :0.4689   Mean   :50.32
 3rd Qu.:1.0000   3rd Qu.:0.7320   3rd Qu.:0.6510   3rd Qu.:59.86
 Max.   :1.0000   Max.   :0.9880   Max.   :0.9760   Max.   :84.18
```

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

First, we gotta find all chocolate candy in the dataset -> extract their winpercent values -> calc the mean -> find alll fruit candy -> find their winpercent -> calc their mean

Steps 1-3:

```
choc.candy <- candy[candy$chocolate==1,]
choc.win <- choc.candy$winpercent
mean(choc.win)
```

```
[1] 60.92153
```

Let's do the same for the fruity candy (steps 3-6):

```
fruit.candy <- candy[candy$fruity==1,]
fruit.win <- fruit.candy$winpercent
mean(fruit.win)
```

```
[1] 44.11974
```

Sp, the chocolate candy is higher ranked than fruit candy (I agree)

Q12. Is this difference statistically significant?

```
t.test(choc.win, fruit.win)
```

```
    Welch Two Sample t-test

data:  choc.win and fruit.win
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
```

So, the difference between the values is statisticlly significant

Overall Candy Rankings

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

```
inds <- order(candy$winpercent)
head(candy[inds, ], 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, Jawbusters > Q14. What are the top 5 all time favorite candy types out of this set?

We do the same - but we meed top 5 candies

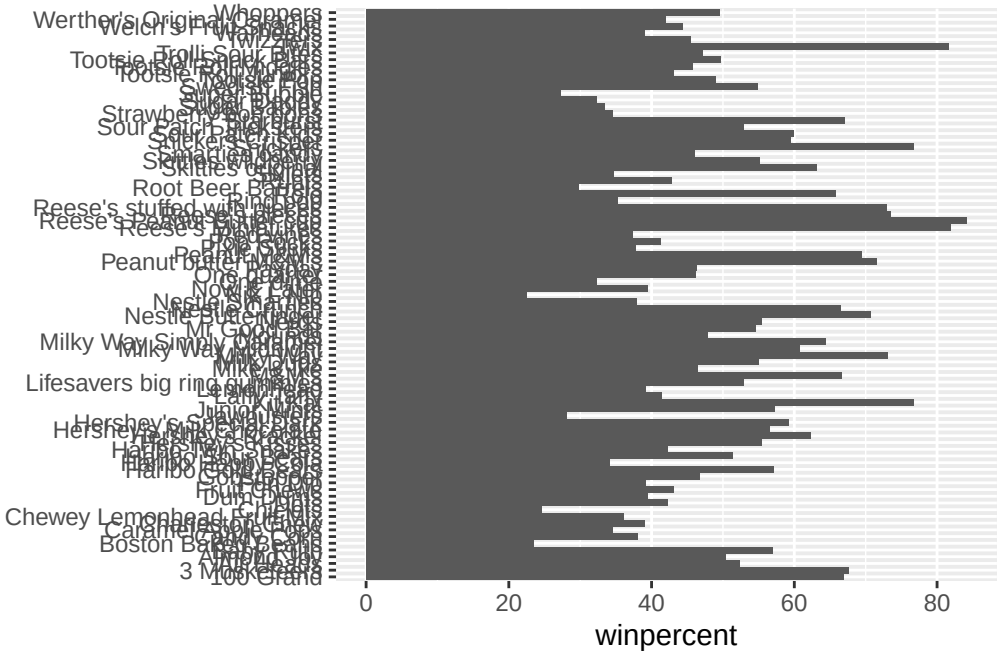
```
inds_max <- rev(inds)
head(candy[inds_max, ], 5)
```

```
                        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
```

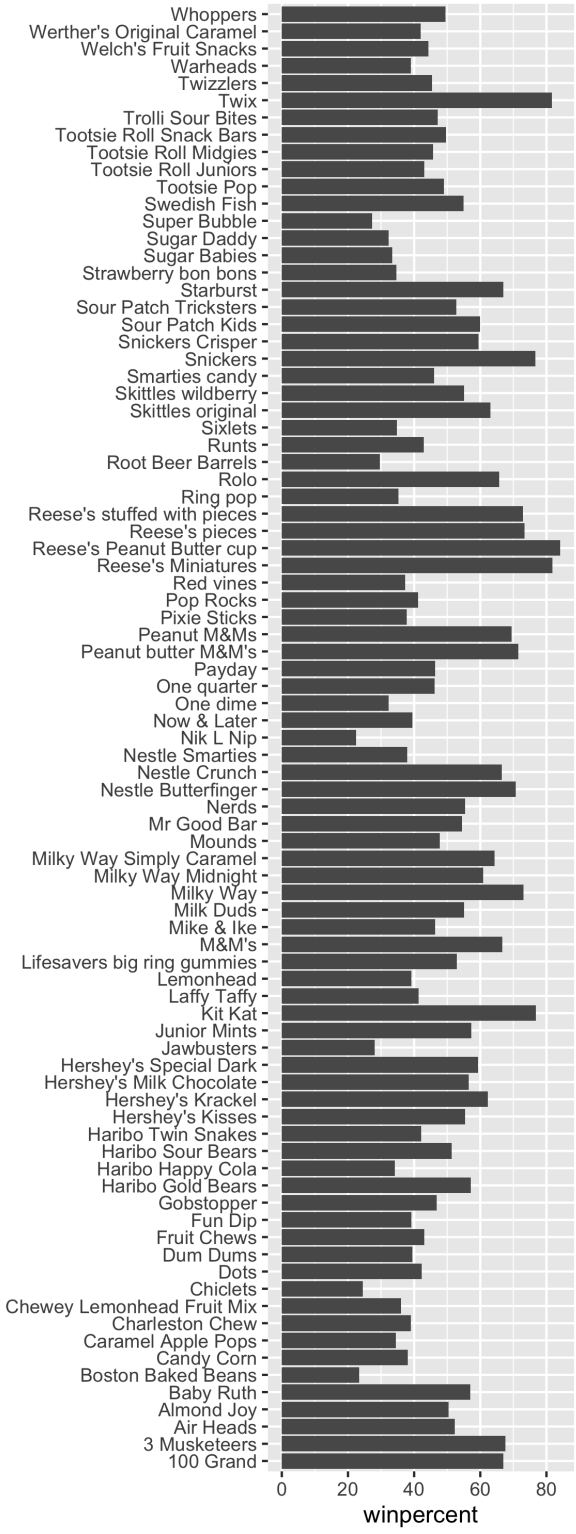
We got: reese's peanut butter cup, reese's miniatures, twix, kit kat, snickers

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

```
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col() +
  ylab("")
```

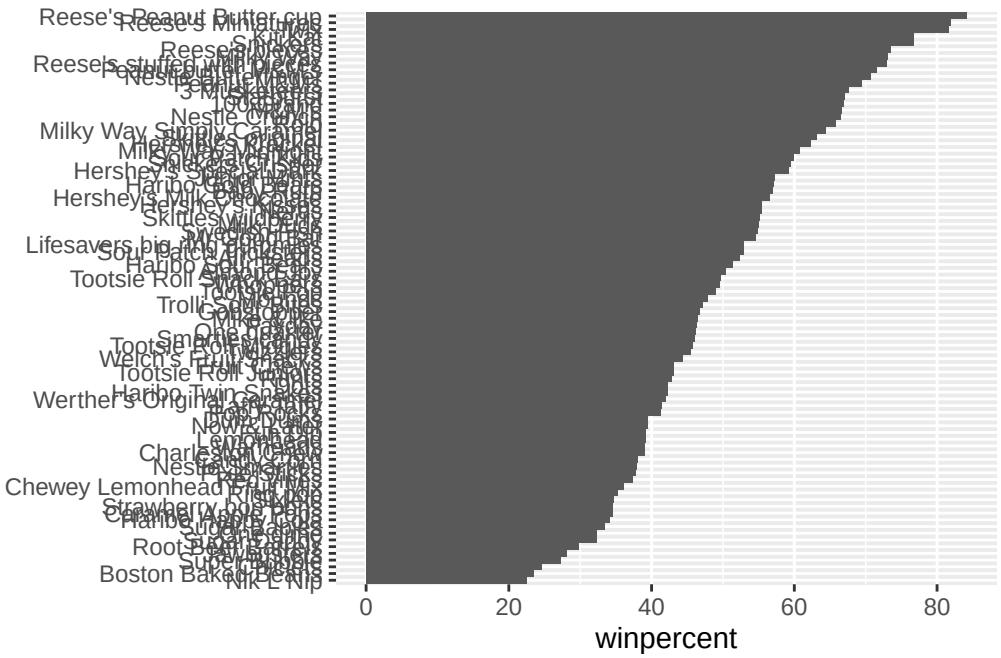



```
ggsave("barplot1.png",  height=10, width=4)
```


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

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() +
  ylab("")
```




```
ggsave("barplot2.png",  height=10, width=4)
```

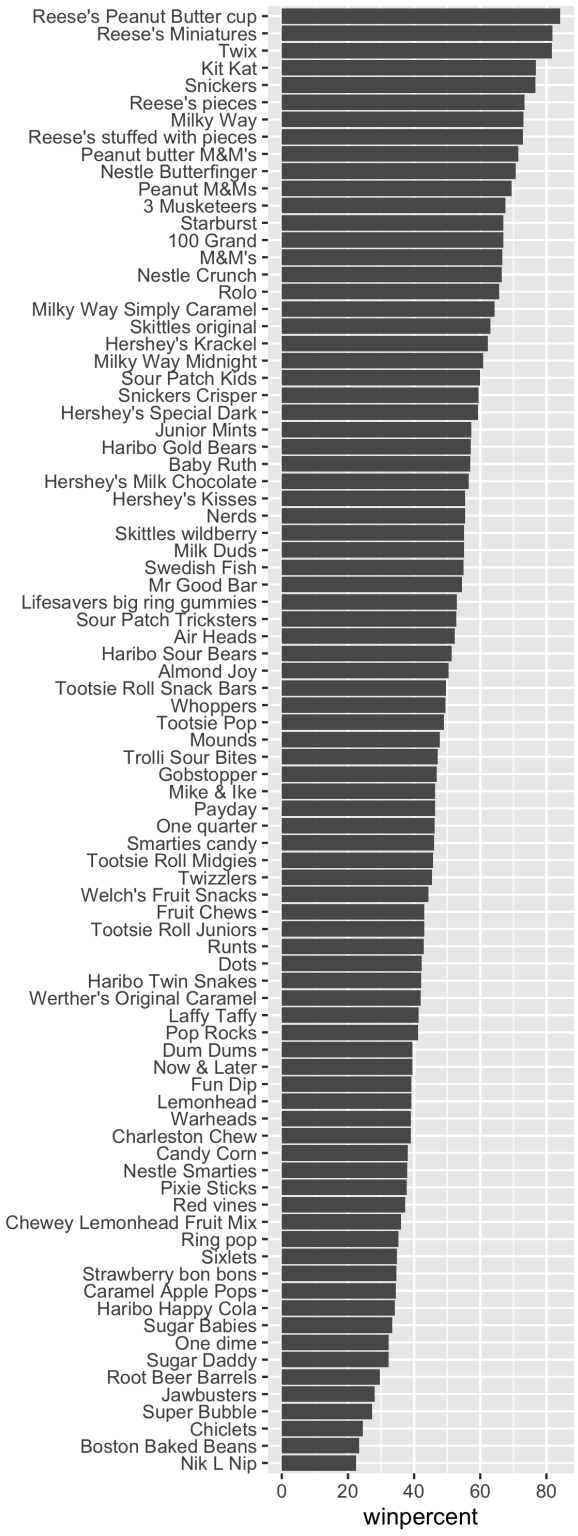
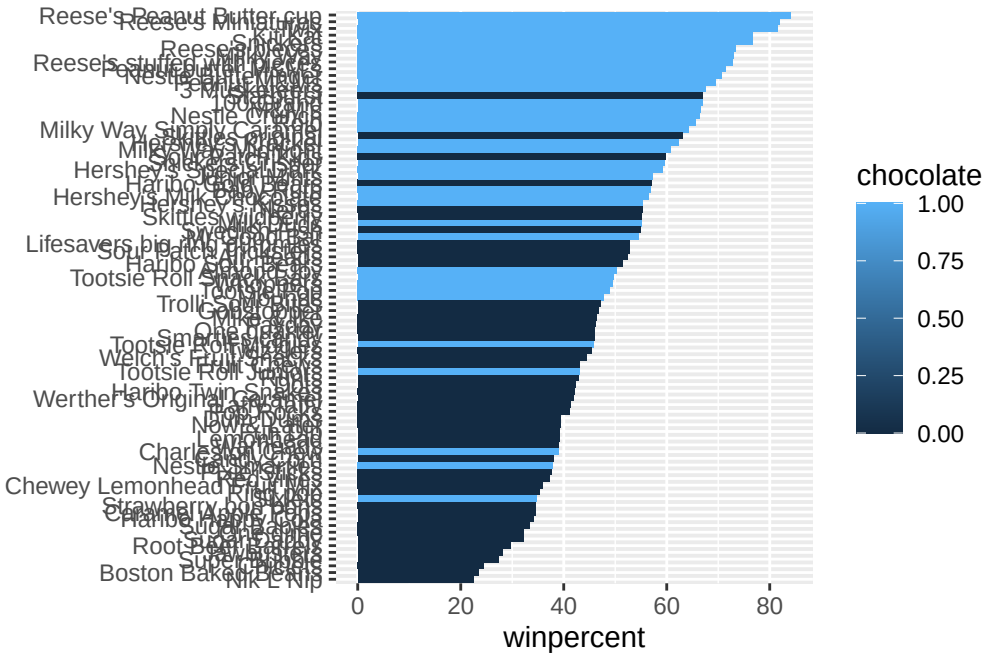


Figure 1: My second barplot for q16

Time to add some useful color

Color by chart:

```
ggplot(candy) +
  aes(winpercent,
      reorder ( rownames(candy), winpercent),
      fill=chocolate) +
  geom_col() +
  ylab("")
```



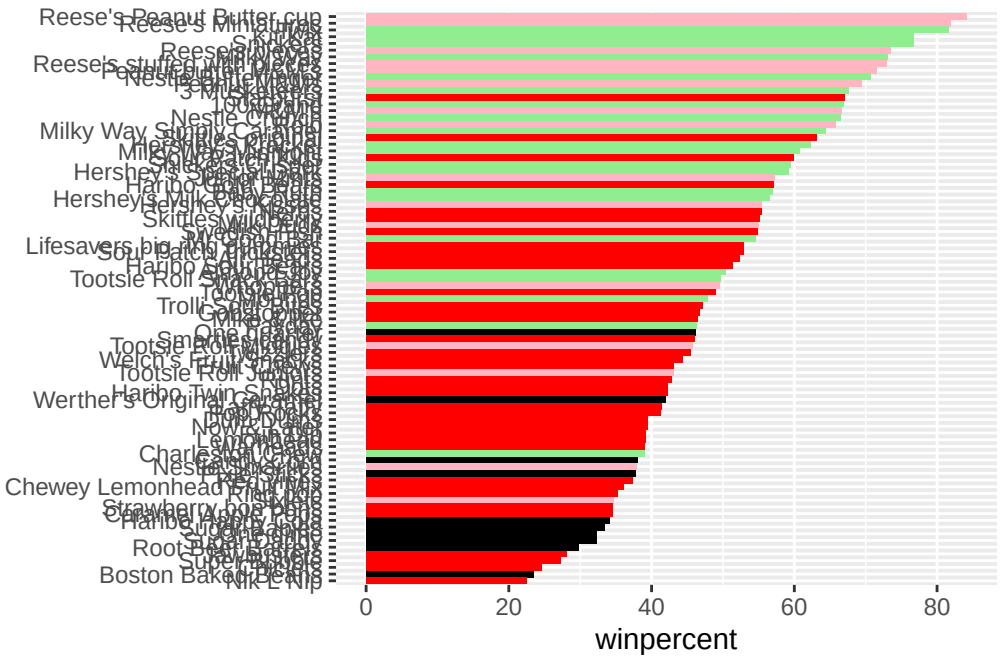
I want custom colors so lets do it ourselves

```
my_cols <- rep("black", nrow(candy))
my_cols[candy$chocolate==1] <- "lightpink"
my_cols[candy$bar==1] <- "lightgreen"
my_cols[candy$fruity==1] <- "red"
my_cols
```

```
 [1] "lightgreen" "lightgreen" "black"      "black"      "red"
 [6] "lightgreen" "lightgreen" "black"      "black"      "red"
[11] "lightgreen" "red"        "red"        "red"        "red"
```

```
 [16] "red"         "red"         "red"         "red"         "black"
 [21] "red"         "red"         "lightpink"   "lightgreen"  "lightgreen"
 [26] "lightgreen"  "red"         "lightpink"   "lightgreen"  "red"
 [31] "red"         "red"         "lightpink"   "lightpink"   "red"
 [36] "lightpink"   "lightgreen"  "lightgreen"  "lightgreen"  "lightgreen"
 [41] "lightgreen"  "red"         "lightgreen"  "lightgreen"  "red"
 [46] "red"         "lightgreen"  "lightpink"   "black"       "red"
 [51] "red"         "lightpink"   "lightpink"   "lightpink"   "lightpink"
 [56] "red"         "lightpink"   "black"       "red"         "lightpink"
 [61] "red"         "red"         "lightpink"   "red"         "lightgreen"
 [66] "lightgreen"  "red"         "red"         "red"         "red"
 [71] "black"       "black"       "red"         "red"         "red"
 [76] "lightpink"   "lightpink"   "lightgreen"  "red"         "lightgreen"
 [81] "red"         "red"         "red"         "black"       "lightpink"
```

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



Q17. What is the worst ranked chocolate candy?

Nik L Lip > Q18. What is the best ranked fruity candy

Starburst

Taking a look at pricepercent

```
my_cols[candy$fruity==1]
```

```
 [1] "red" "red" "red" "red" "red" "red" "red" "red" "red" "red" "red" "red"
[13] "red" "red" "red" "red" "red" "red" "red" "red" "red" "red" "red" "red"
[25] "red" "red" "red" "red" "red" "red" "red" "red" "red" "red" "red" "red"
[37] "red" "red"
```

```
ggplot(candy) +
  aes(x=winpercent, y=pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text(col=my_cols, size=3.3, max.overlaps = 5)
```

```
Warning in geom_text(col = my_cols, size = 3.3, max.overlaps = 5): Ignoring
unknown parameters: `max.overlaps`
```

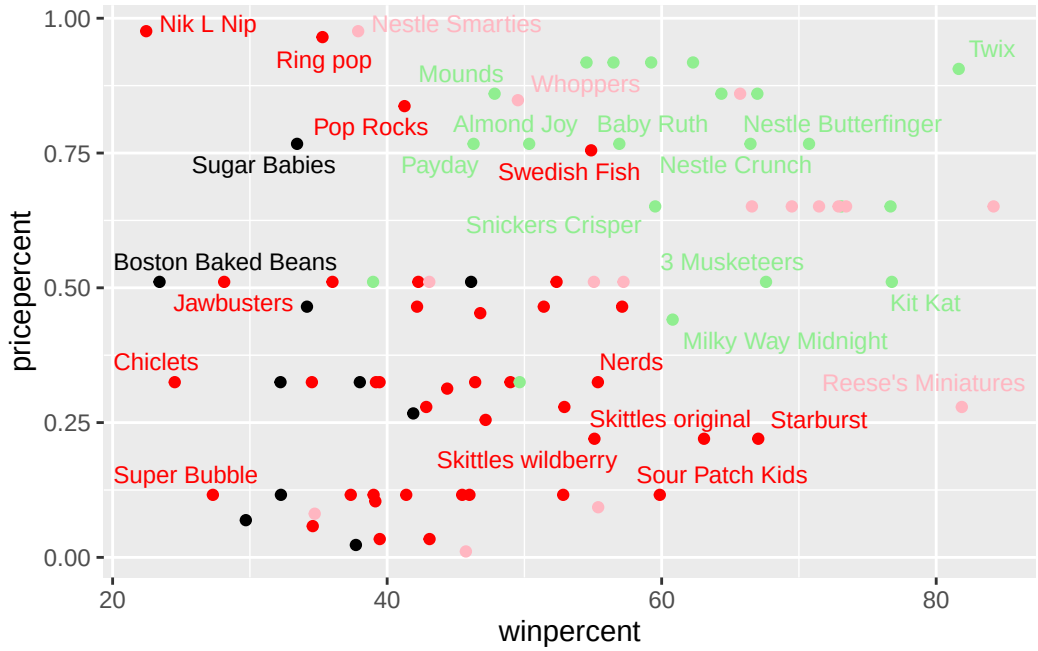
We can use the **ggrepel** package for better label placement:

```
library(ggrepel)

ggplot(candy) +
  aes(x=winpercent, y=pricepercent, label=rownames(candy)) +
  geom_point(col=my_cols) +
  geom_text_repel(col=my_cols, size=3.3, max.overlaps = 7, size=3.3)
```

```
Warning: Duplicated aesthetics after name standardisation: size
```

```
Warning: ggrepel: 57 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



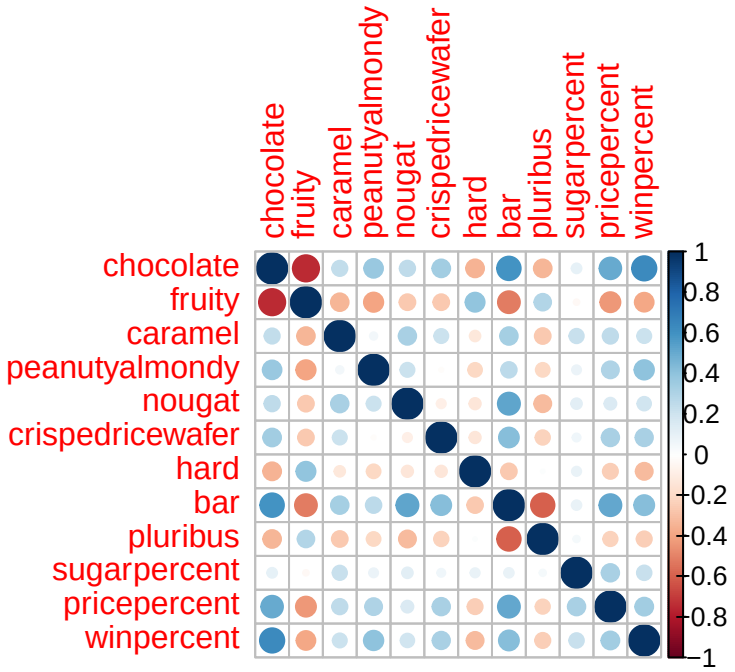
Exploring the correlation structure

Pearson correlation values range from -1 to +1

```
library(corrplot)
```

```
corrplot 0.95 loaded
```

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



PCA

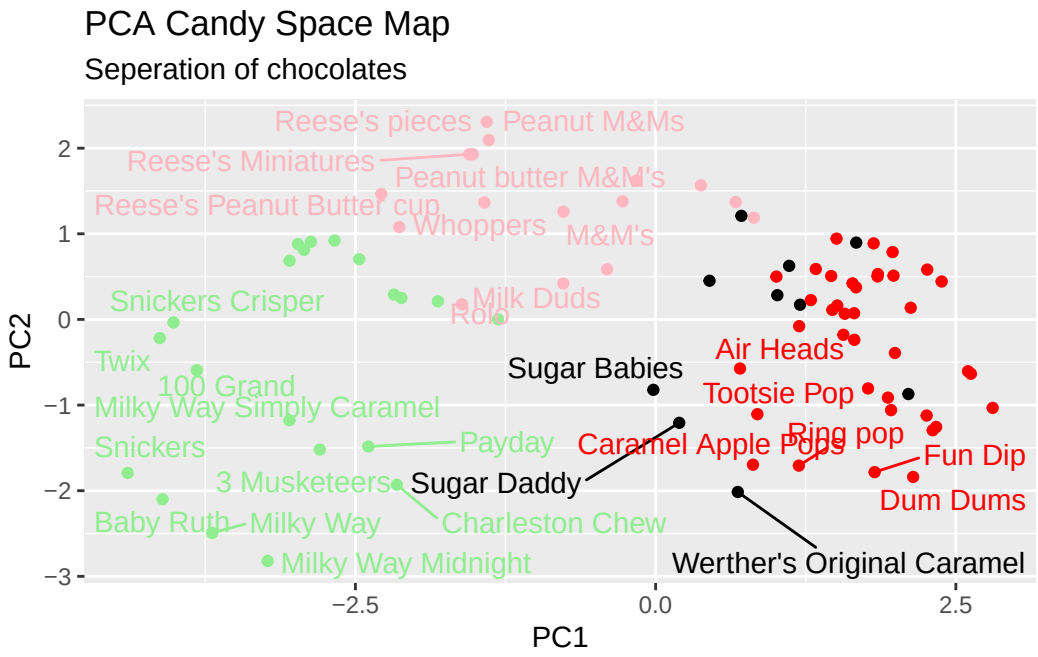
```
pca<-prcomp(candy, scale=T)
summary(pca)
```

```
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
```

The main results figure: PCA score plot

```
ggplot(pca$x)+
  aes(PC1, PC2, label=rownames(pca$x))+
  geom_point(col=my_cols)+
  geom_text_repel(col=my_cols)+
  labs(title="PCA Candy Space Map",
       subtitle = "Seperation of chocolates")
```

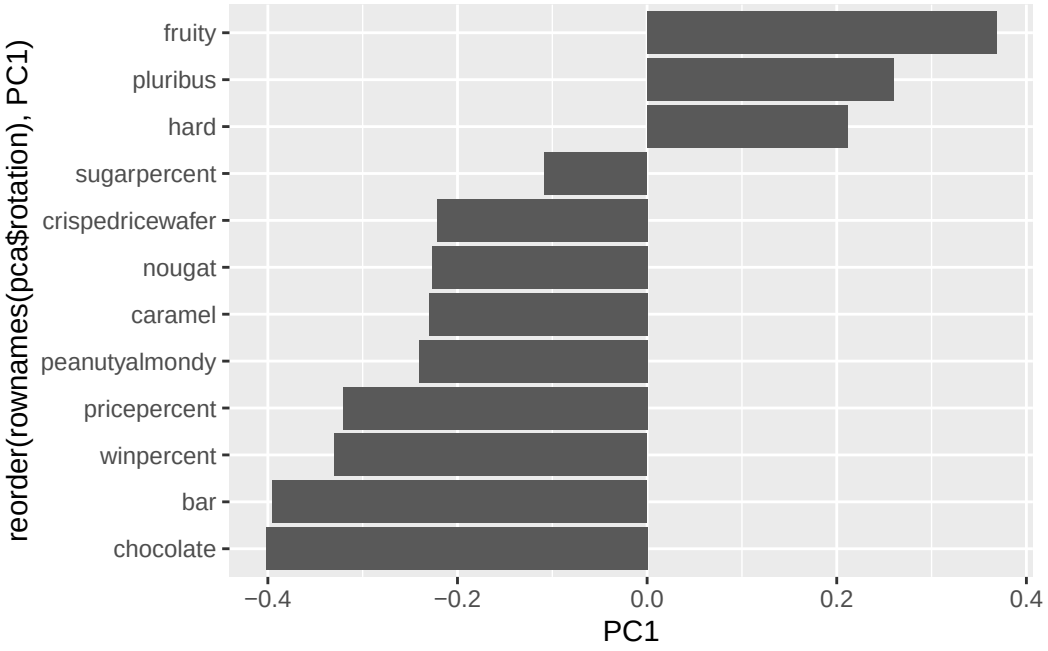
```
Warning: ggrepel: 56 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



Q24. Complete the code to generate the loadings plot above. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you? Where did you see this relationship highlighted previously?

The "loadings" plot for PC1

```
ggplot(pca$rotation)+
  aes(PC1,
      reorder(rownames(pca$rotation), PC1))+
  geom_col()
```



Q25. Based on your exploratory analysis, correlation findings, and PCA results, what combination of characteristics appears to make a "winning" candy? How do these different analyses (visualization, correlation, PCA) support or complement each other in reaching this conclusion?

We would want to have chocolate, fruity candies. But not chocolate-fruity camdies. Also chocolate bars would be cool.