Class 10: Halloween Project

Shitian Li (PID: A13294481)

10/31/2021

Import 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
```

What does this dataset contain?

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
```

Favorite candy?

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

```
## [1] 81.64291
```

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

```
candy["Haribo Gold Bears", ]$winpercent
```

```
## [1] 57.11974
```

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
```

Skim:

```
# install.packages("skimr")
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	

Use `skim()` and answer the following:

Q6. Is there any variable/column that looks to be on a different scale to the majority of the other columns in the dataset?

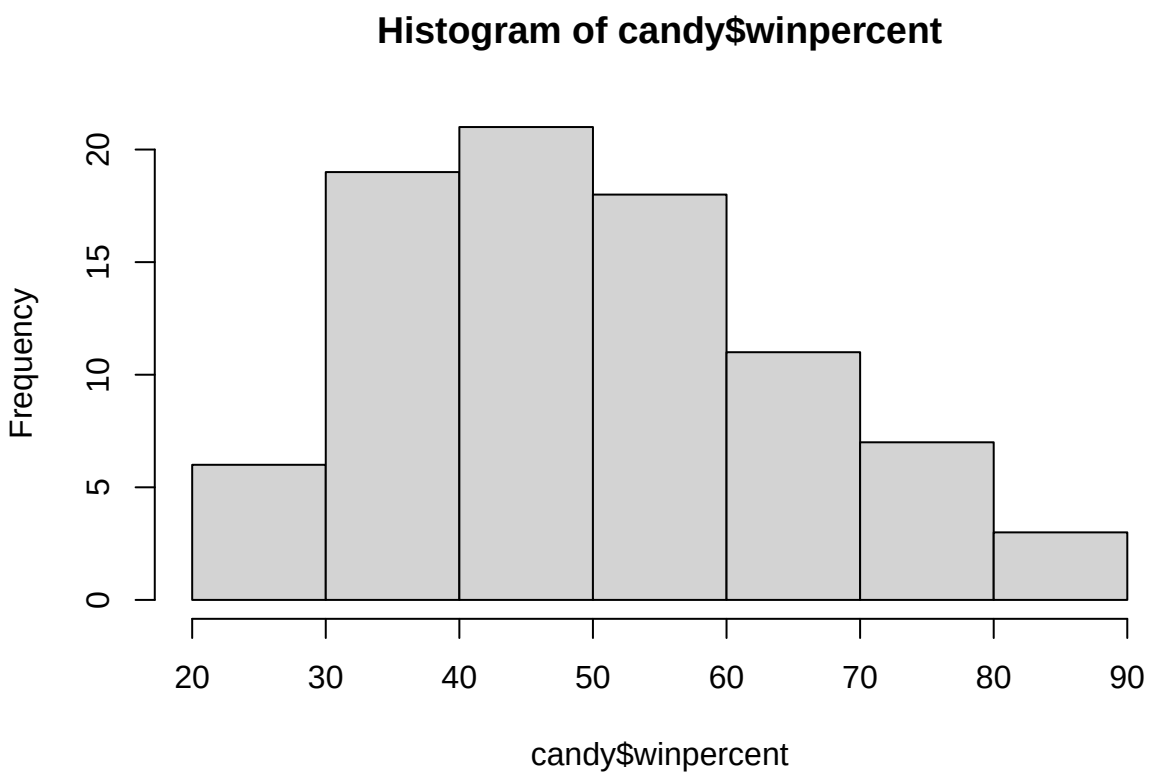
Winpercent is in a different scale to majority since it has a very high STD.

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

0: does not contain chocolate 1: contains chocolate

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical?

Yes.

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

Below.

```
median(candy$winpercent)
```

```
## [1] 47.82975
```

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

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

t.test(candy.choco.type, candy.fruit.type)
```

```
##
##  Welch Two Sample t-test
##
## data:  candy.choco.type and candy.fruit.type
## 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
```

As a result, we see that chocolate is higher ranked than fruity type.

Q12. Is this difference statistically significant?

Yes, with $p = 2.871e\text{-}08$.

Overall candy rankings

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
```

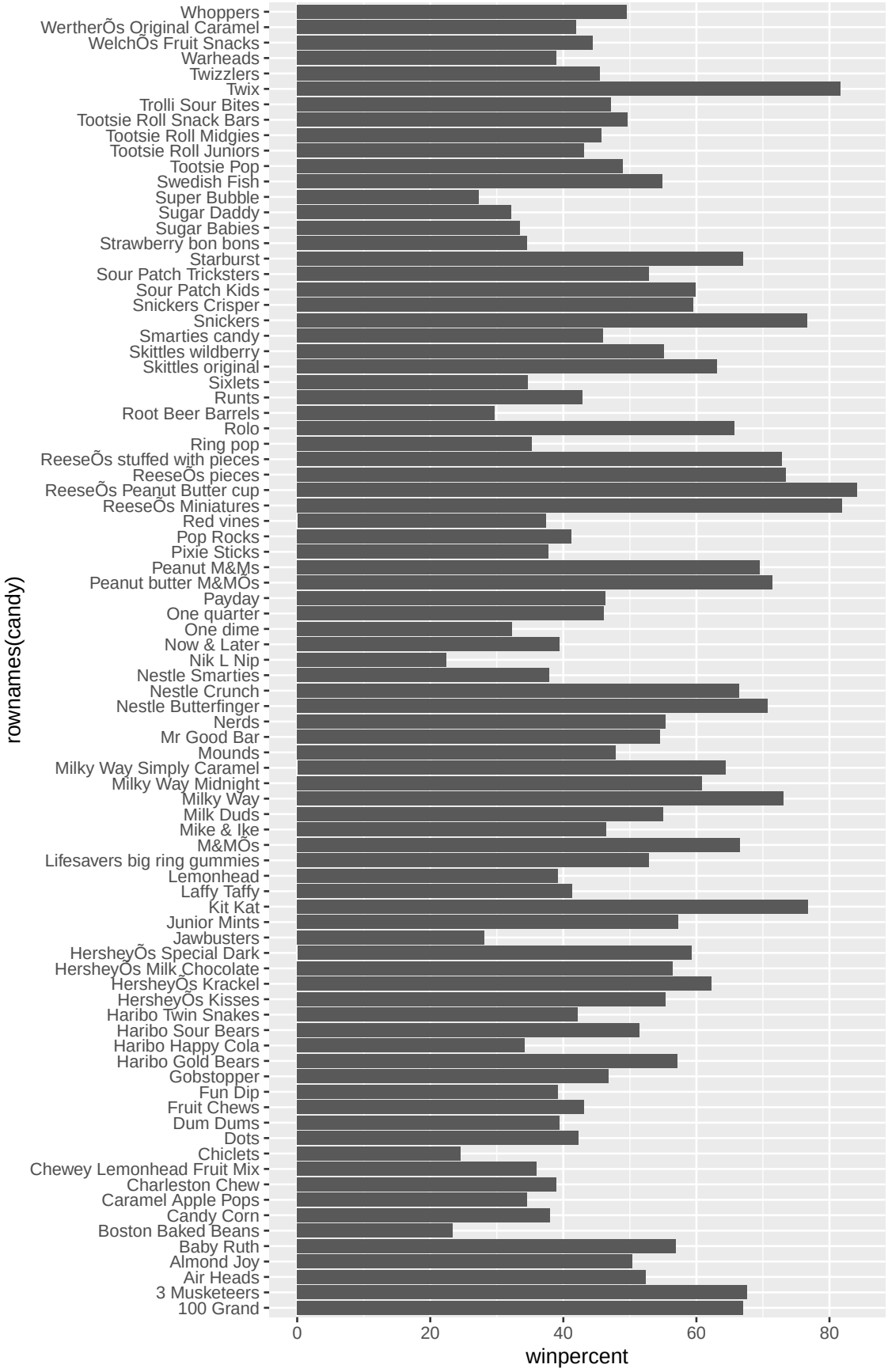
Q14. What are the top 5 all time favorite candy types out of this set?

```
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
```

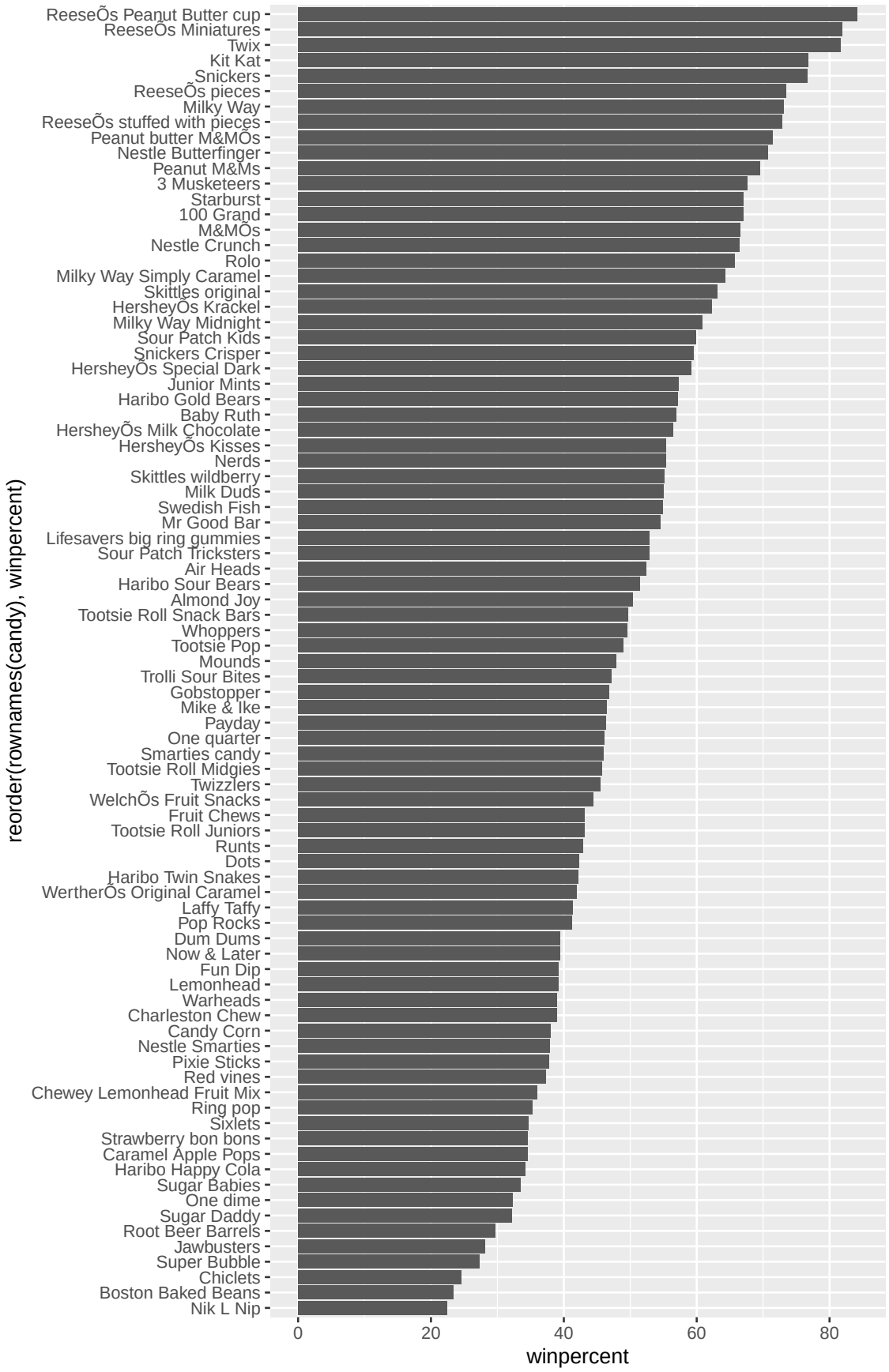
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)

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


rownames(candy)
Whoppers
WertherÕs Original Caramel
WelchÕs 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
ReeseÕs stuffed with pieces
ReeseÕs pieces
ReeseÕs Peanut Butter cup
ReeseÕs Miniatures
Red vines
Pop Rocks
Pixie Sticks
Peanut M&Ms
Peanut butter M&MÕs
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&MÕs
Lifesavers big ring gummies
Lemonhead
Laffy Taffy
Kit Kat
Junior Mints
Jawbusters
HersheyÕs Special Dark
HersheyÕs Milk Chocolate
HersheyÕs Krackel
HersheyÕs 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?

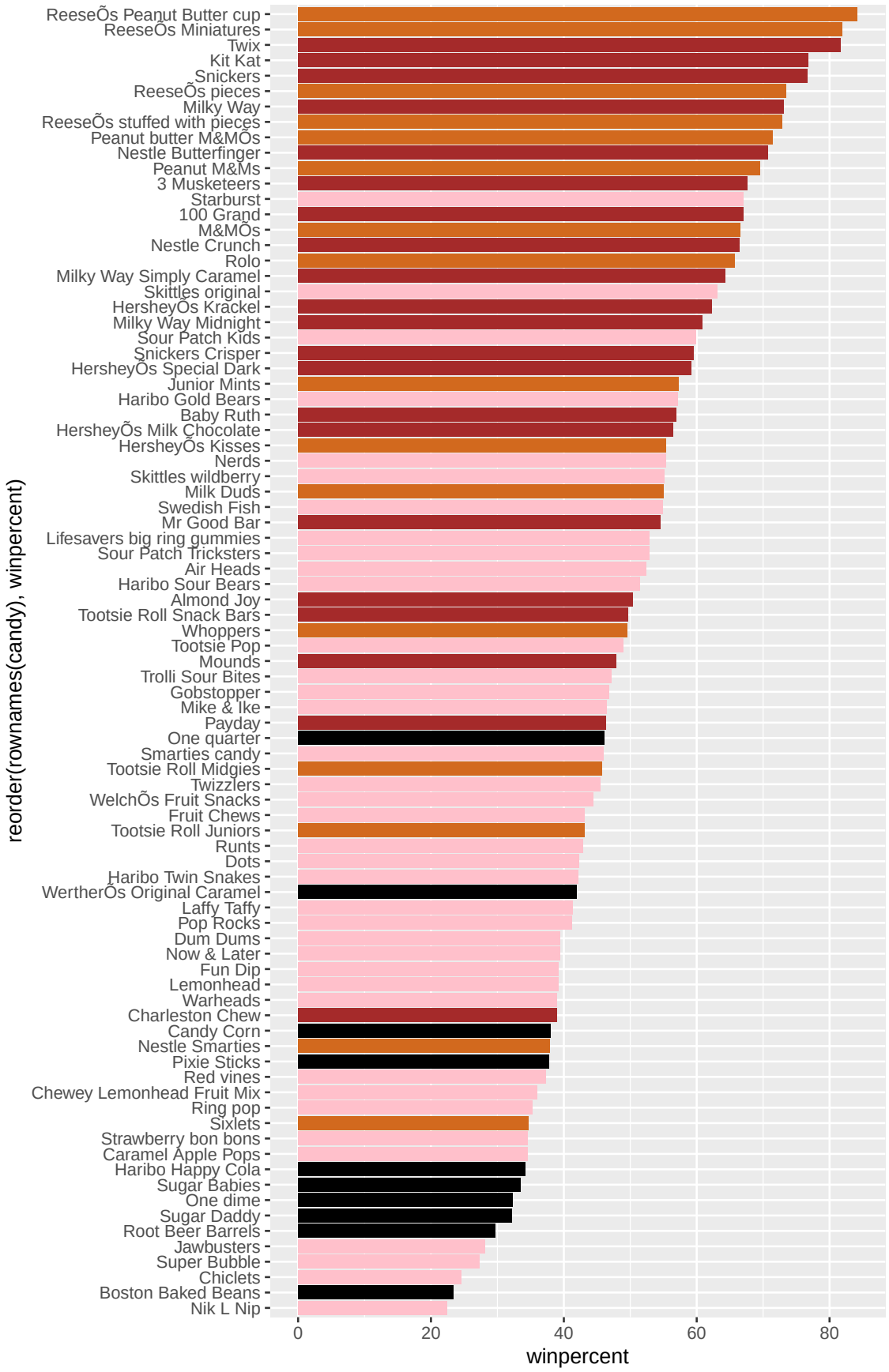
```
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col()
```


ReeseÕs Peanut Butter cup
ReeseÕs Miniatures
Twix
Kit Kat
Snickers
ReeseÕs pieces
Milky Way
ReeseÕs stuffed with pieces
Peanut butter M&MÕs
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&MÕs
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
HersheyÕs Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
HersheyÕs Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
HersheyÕs Milk Chocolate
HersheyÕs 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
WelchÕs Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
WertherÕs 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

Set up 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)
```


ReeseÕs Peanut Butter cup
ReeseÕs Miniatures
Twix
Kit Kat
Snickers
ReeseÕs pieces
Milky Way
ReeseÕs stuffed with pieces
Peanut butter M&MÕs
Nestle Butterfinger
Peanut M&Ms
3 Musketeers
Starburst
100 Grand
M&MÕs
Nestle Crunch
Rolo
Milky Way Simply Caramel
Skittles original
HersheyÕs Krackel
Milky Way Midnight
Sour Patch Kids
Snickers Crisper
HersheyÕs Special Dark
Junior Mints
Haribo Gold Bears
Baby Ruth
HersheyÕs Milk Chocolate
HersheyÕs 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
WelchÕs Fruit Snacks
Fruit Chews
Tootsie Roll Juniors
Runts
Dots
Haribo Twin Snakes
WertherÕs 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

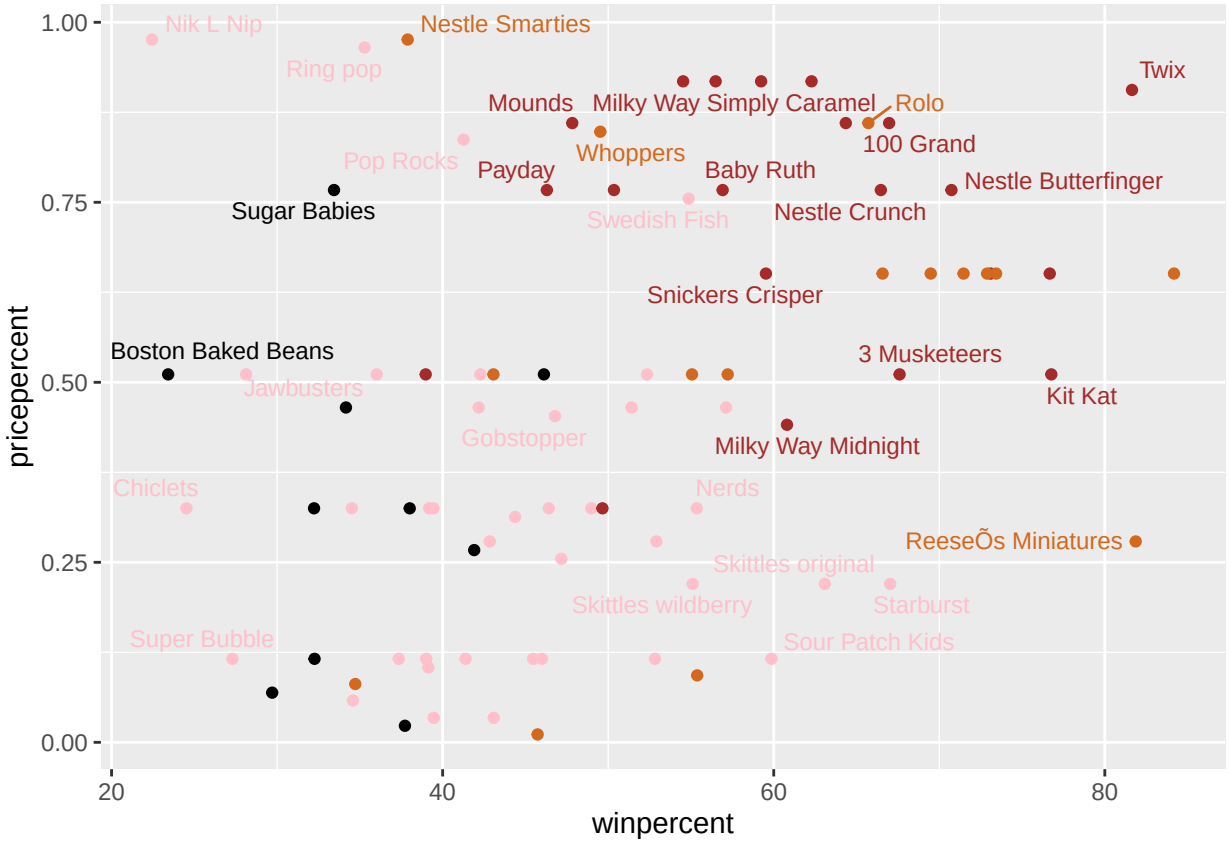
Q18. What is the best ranked fruity candy?

starbusts

Pricepercent

```
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: 54 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 miatures

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

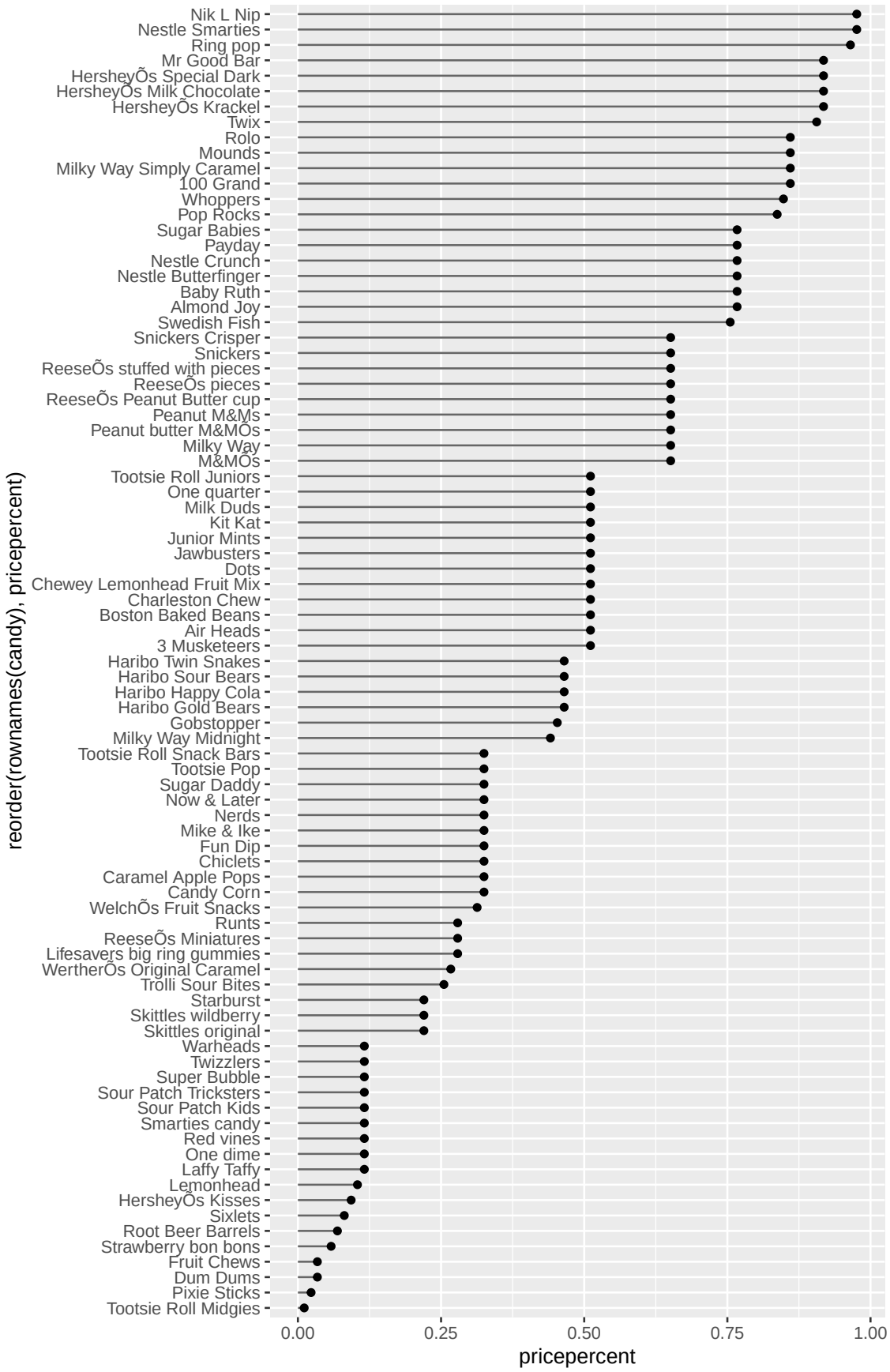
Nik L Nip

```
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
```

opitional 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()`.

```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```

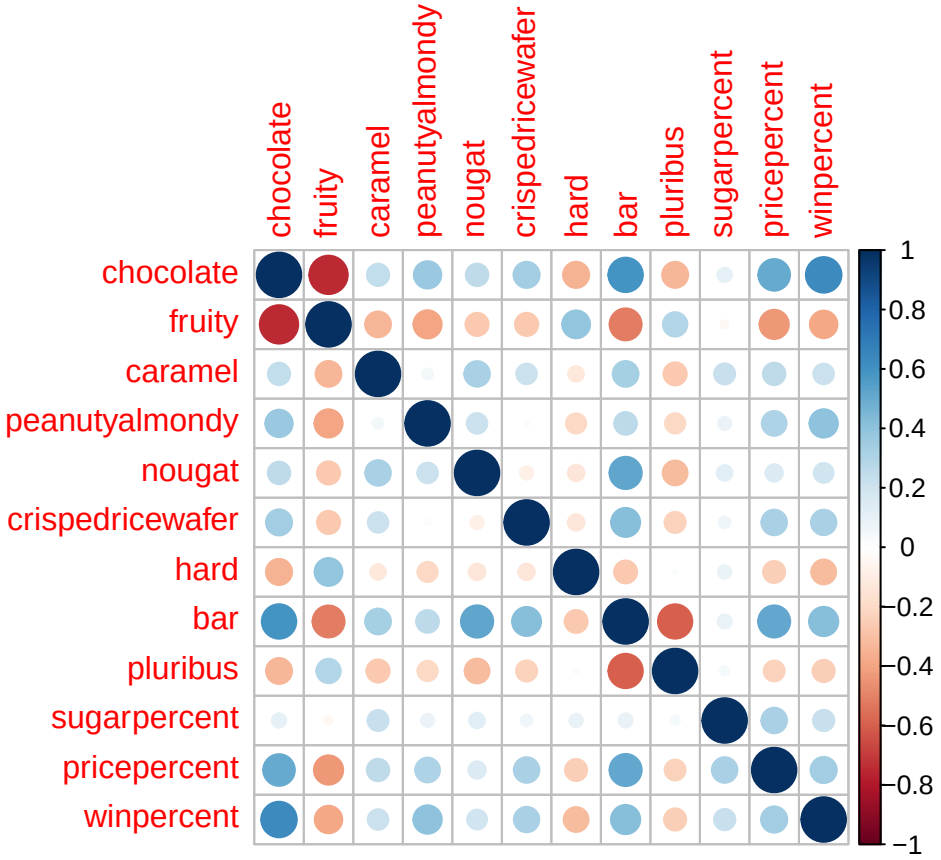

Nik L Nip
Nestle Smarties
Ring pop
Mr Good Bar
HersheyÕs Special Dark
HersheyÕs Milk Chocolate
HersheyÕs Krackel
Twix
Rolo
Mounds
Milky Way Simply Caramel
100 Grand
Whoppers
Pop Rocks
Sugar Babies
Payday
Nestle Crunch
Nestle Butterfinger
Baby Ruth
Almond Joy
Swedish Fish
Snickers Crisper
Snickers
ReeseÕs stuffed with pieces
ReeseÕs pieces
ReeseÕs Peanut Butter cup
Peanut M&Ms
Peanut butter M&MÕs
Milky Way
M&MÕs
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
WelchÕs Fruit Snacks
Runts
ReeseÕs Miniatures
Lifesavers big ring gummies
WertherÕs 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
HersheyÕs 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

Correlation

```
# install.packages("corrplot")
library(corrplot)
```

```
## corrplot 0.90 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?

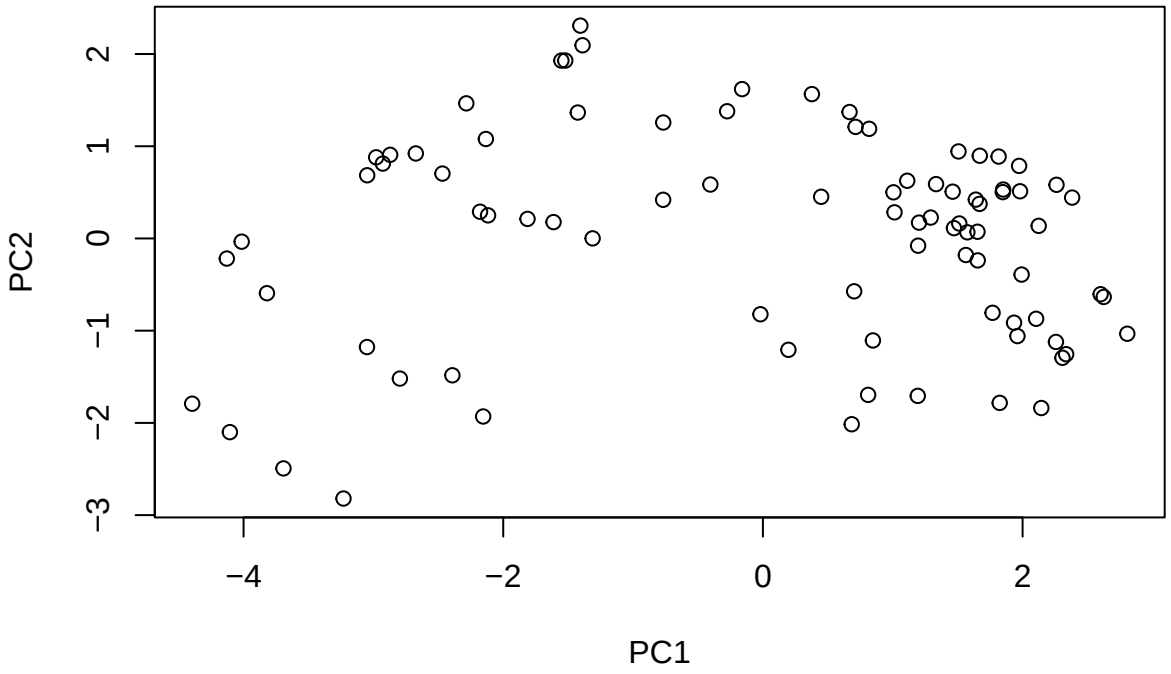
Winpercent and Chocolate

PCA

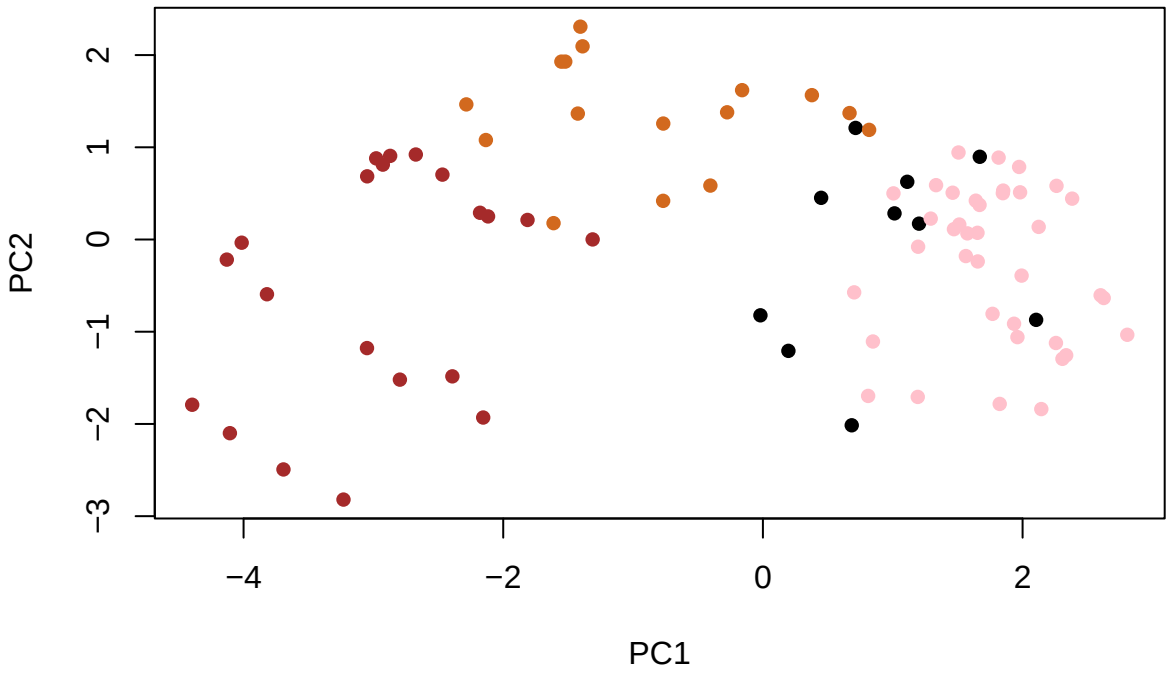
```
pca <- prcomp(candy, scale = T)
summary(pca)
```

```
## 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
```

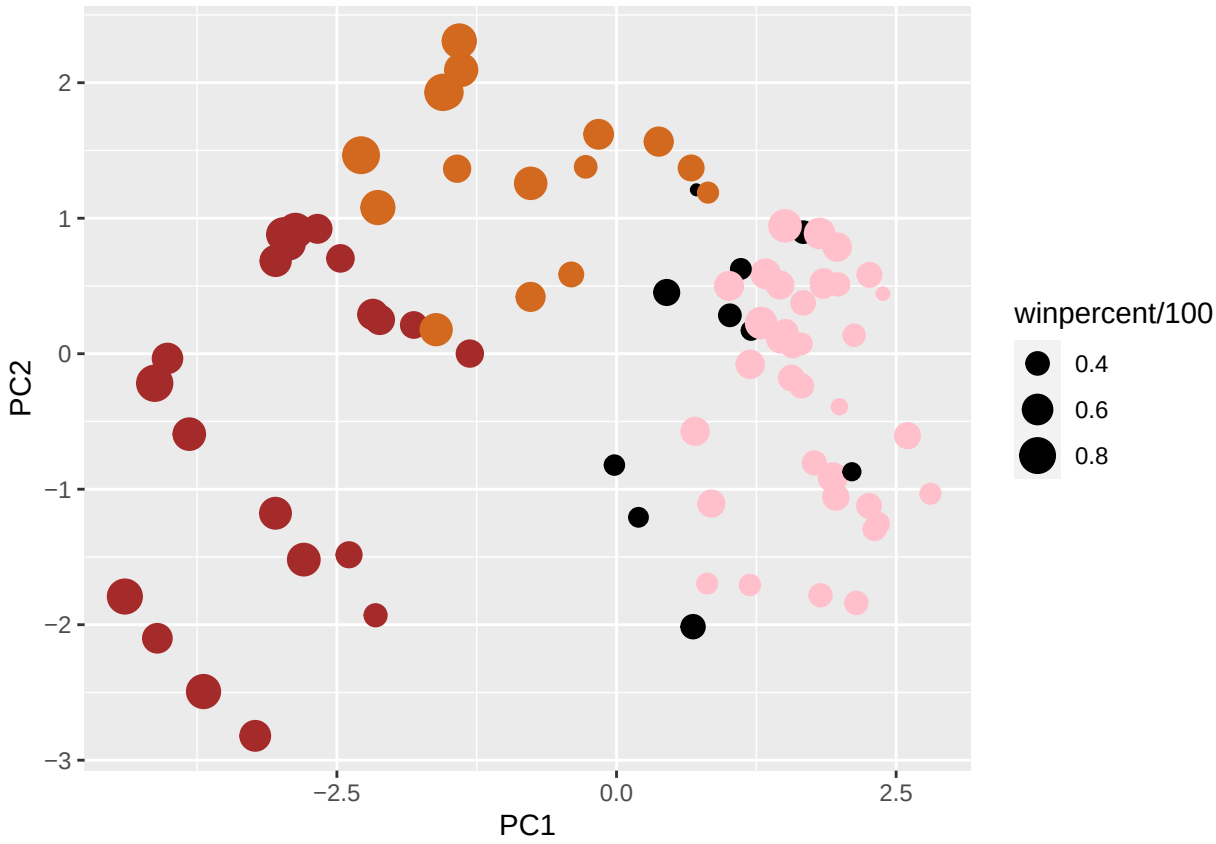
```
plot(pca$x[,1:2])
```




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

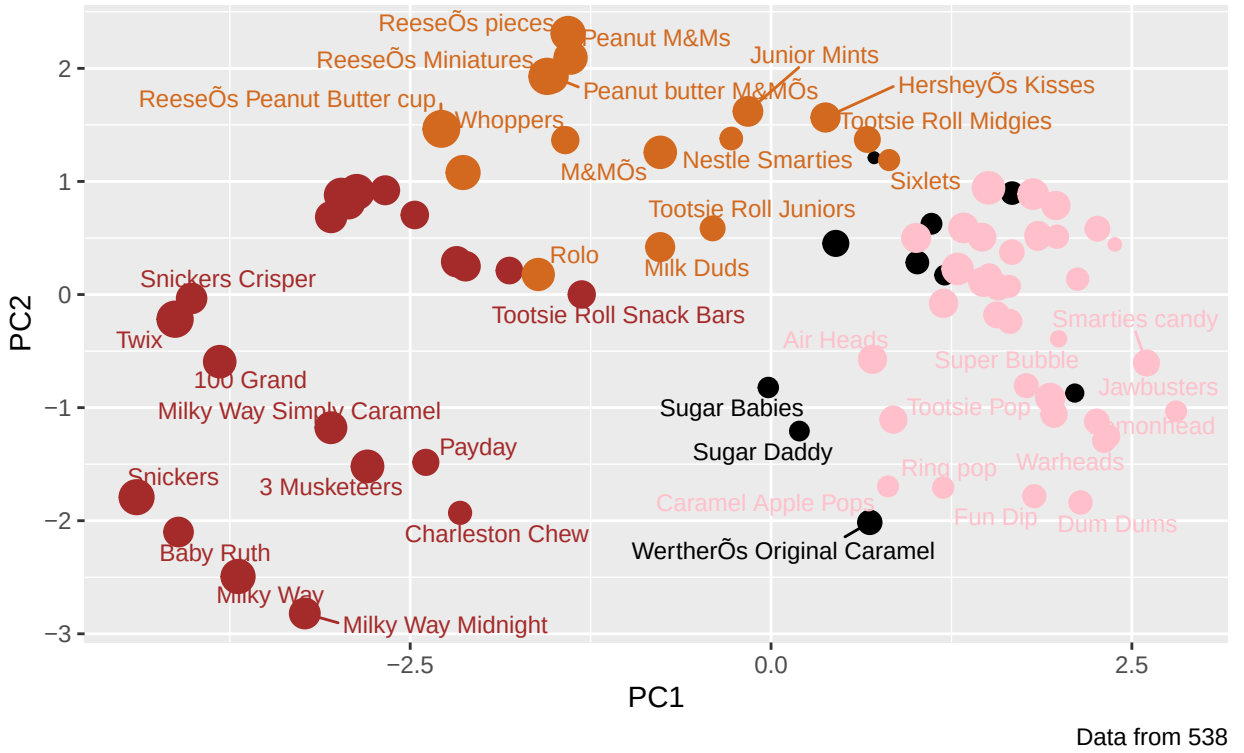



```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])

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
```



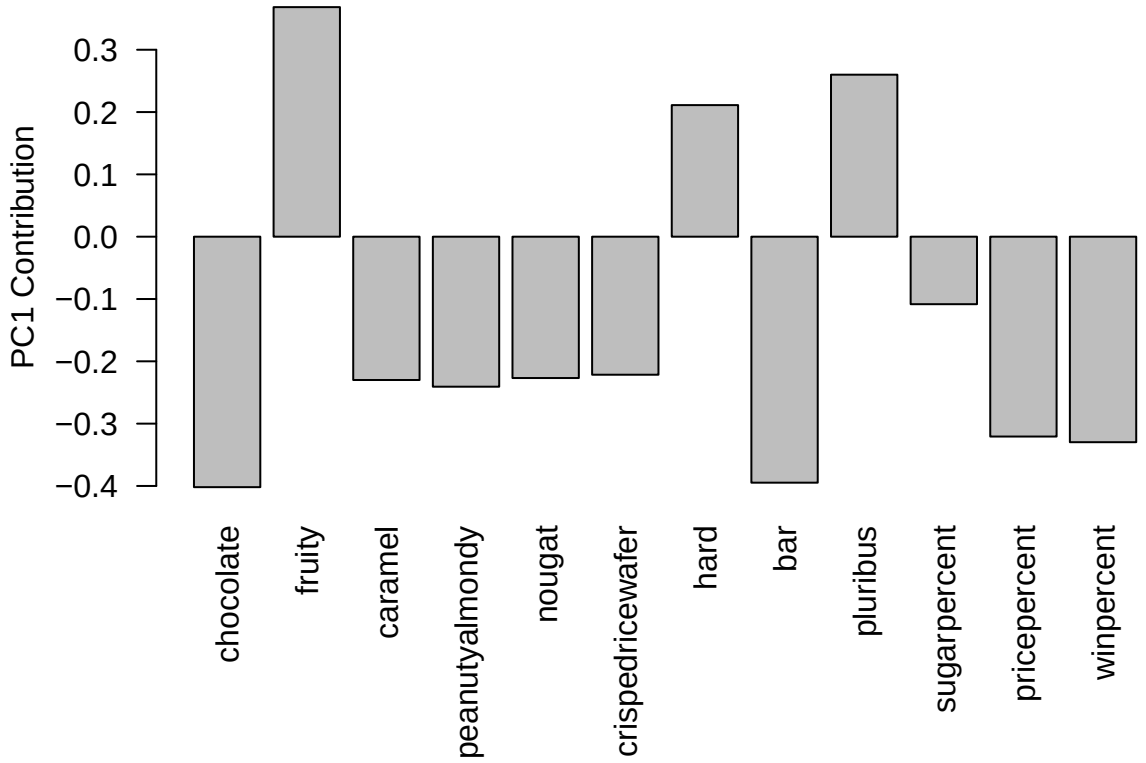

```
library(ggrepel)

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 brown), fruity (re
       caption="Data from 538")
```

```
## Warning: ggrepel: 44 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
ggplotly(p)
```

```
par(mar=c(8,4,2,2))
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, pluribus, and hard. Yes, as fruity candies are usually hard nad pluribus, and not the other components.