HalloweenMiniProject

AUTHOR
Jinsung

Halloweeen Mini Project

Data Implication

First, read the data off from Github pro. Load it with 'read.csv()' and inspect.

```
candy_file <- "candy-data.txt"
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?

```
ncol(candy)
```

```
[1] 12
```

There are 12 different types of candies in the dataset.

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

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

```
[1] 38
```

There are 38 fruity candy types in the dataset.

Favorite Candy

Winpercent is percentage of people who prefer this over others. Winpercent of the particular candy can be found in the dataset. For example, Twix's win% would be:

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

```
[1] 81.64291
```

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

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

```
[1] 76.7686
```

Win% of Kit Kat is 76.7686%.

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

As stated above, my favorite candy, Kit Kat, has win% value of 76.7686.

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

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

```
[1] 49.6535
```

Tootsie Rol Snack Bars has win% of 49.6535. Not so popular one is it?

'skim()' function is useful for having overview of the dataset.

```
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	
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 seem to have the different scale to the rest of the columns.

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

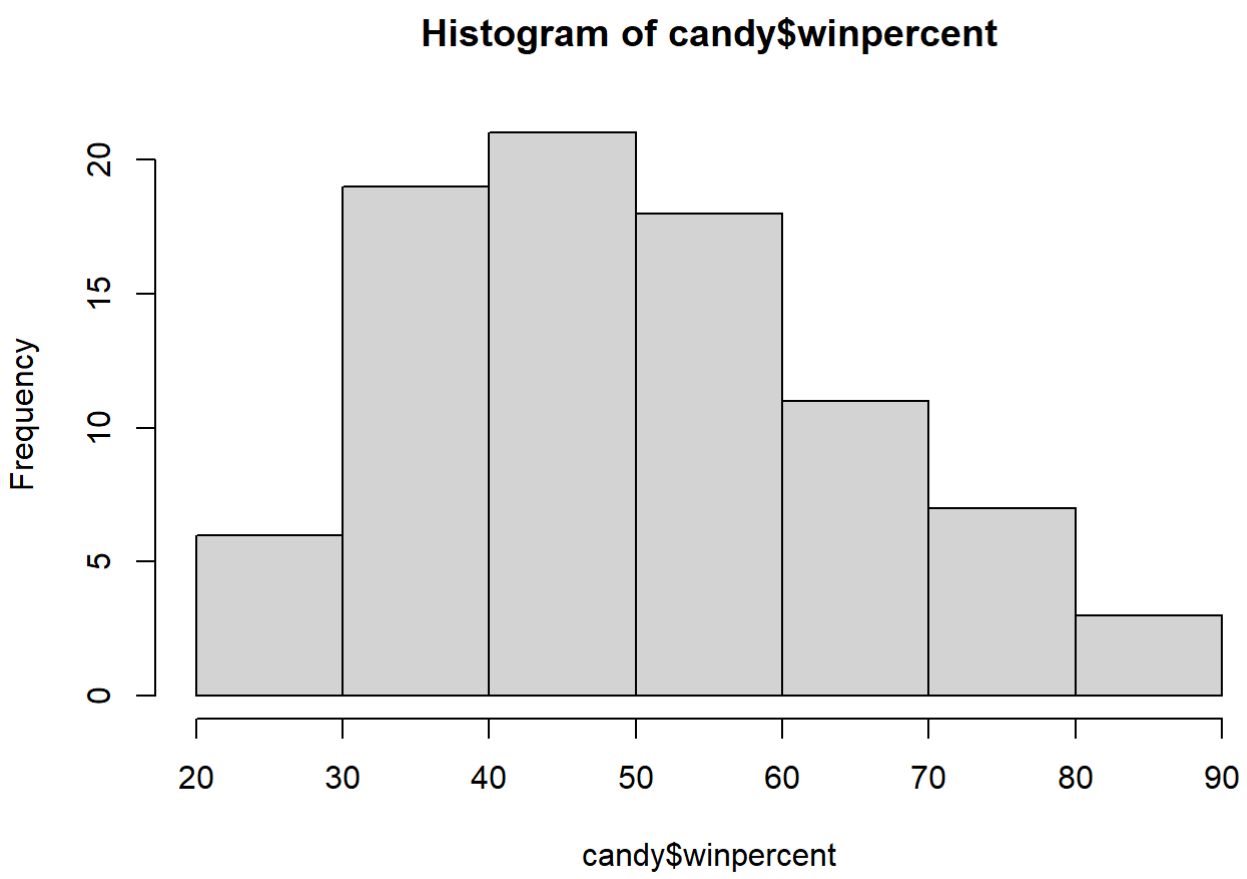
```
candy$chocolate
```

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

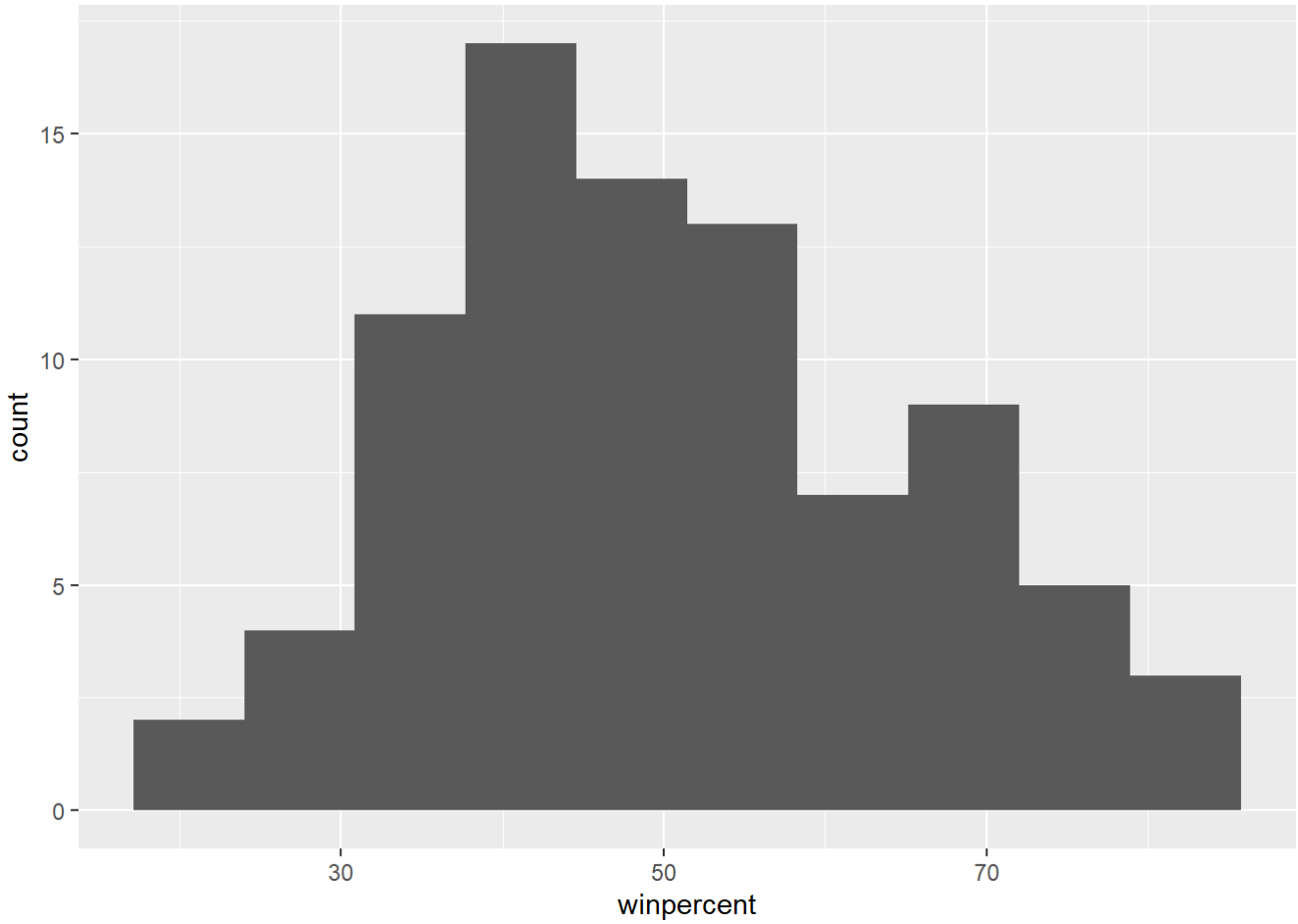
0 and 1 in the 'candy$chocolate'column represent value of logical notation TRUE or FALSE, each being 1 and 0.

> Q8. Plot a histogram of winpercent values

```
hist(candy$winpercent, breaks = 8)
```




```
library(ggplot2)
ggplot(candy) + aes(winpercent) + geom_histogram(bins=10)
```



> Q9. Is the distribution of winpercent values symmetrical?

Histogram of winpercent does not have symmetrical distribution and rather has left skewed shape.

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

Center of the distribution is slightly below 50%, between 40-50%.

> 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
```

On average, chocolate candies are higher ranked than fruity candies.

> Q12. Is this difference statistically significant?

```
chocolate - fruit
```

```
[1] 16.80179
```

```
t.test(candy$winpercent[as.logical(candy$chocolate)], candy$winpercent[as.logical(candy$fruity
```

```
    Welch Two Sample t-test

data:  candy$winpercent[as.logical(candy$chocolate)] and
candy$winpercent[as.logical(candy$fruity)]
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
```

Difference in mean value of their winpercentage is 16.8%. With the p-value of 2.871e-08, the difference is statistically significant.

Overall Candy Rankings

Using the'order()' and 'head()' together allows us to sort the dataset by certain category. You can also use 'arrange()' function with 'head()' to yield same result.

```
ranking <- candy[order(candy$winpercent),]
head(candy[order(candy$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
```

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

Five least liked candies are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, and Jawbusters.

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

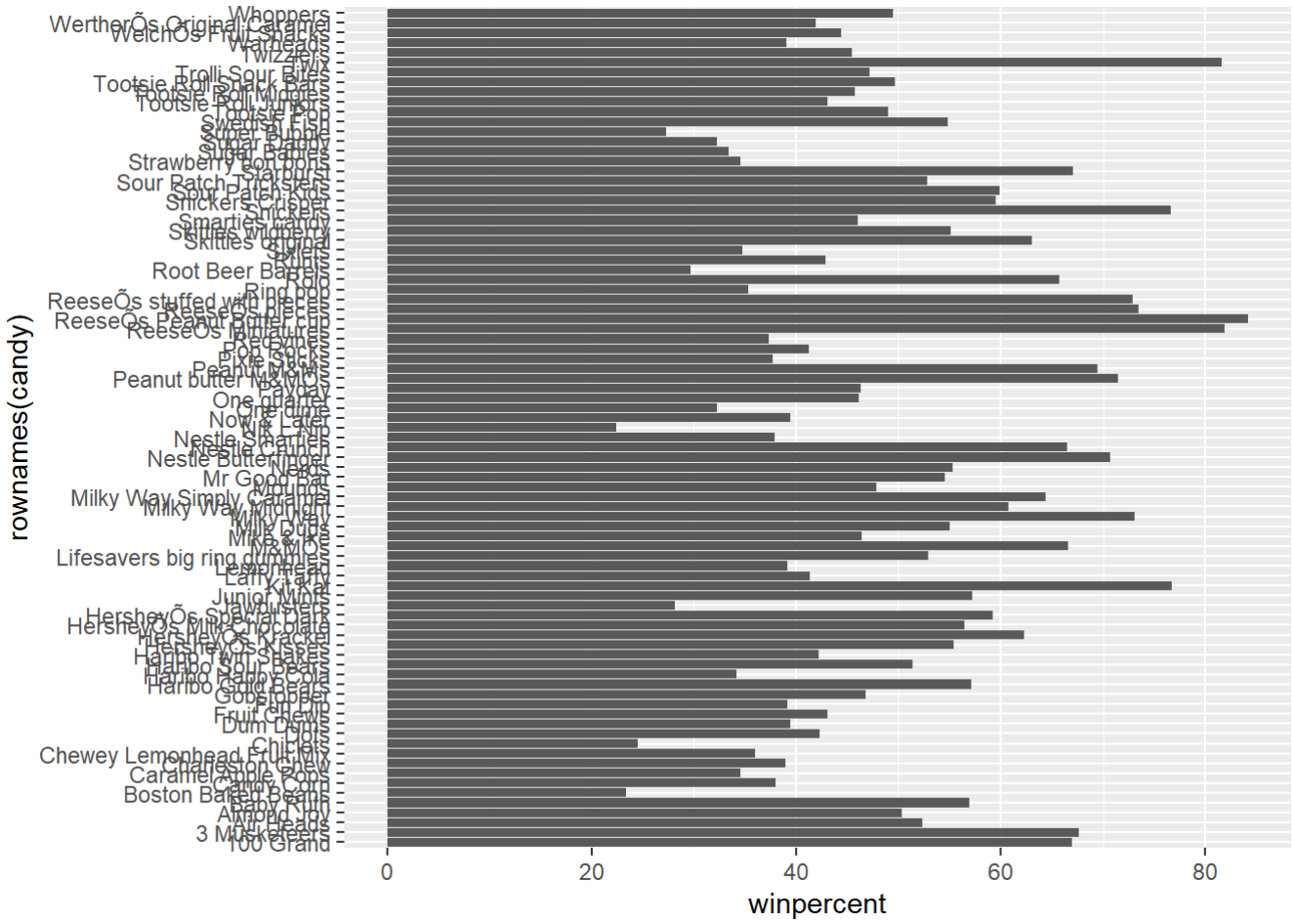
```
tail(candy[order(candy$winpercent),], n=5)
```

```
                        chocolate fruity caramel peanutyalmondy nougat
Snickers                        1      0       1              1      1
Kit Kat                         1      0       0              0      0
Twix                            1      0       1              0      0
ReeseÕs Miniatures              1      0       0              1      0
ReeseÕs Peanut Butter cup       1      0       0              1      0
                        crispedricewafer hard bar pluribus sugarpercent
Snickers                               0    0   1        0        0.546
Kit Kat                                1    0   1        0        0.313
Twix                                   1    0   1        0        0.546
ReeseÕs Miniatures                     0    0   0        0        0.034
ReeseÕs Peanut Butter cup              0    0   0        0        0.720
                        pricepercent winpercent
Snickers                       0.651   76.67378
Kit Kat                        0.511   76.76860
Twix                           0.906   81.64291
ReeseÕs Miniatures             0.279   81.86626
ReeseÕs Peanut Butter cup      0.651   84.18029
```

Top 5 favorite candies are Snickers, Kit Kat, Twix, Reeseos Miniatures, and Reeseos Peanut Butter cup.

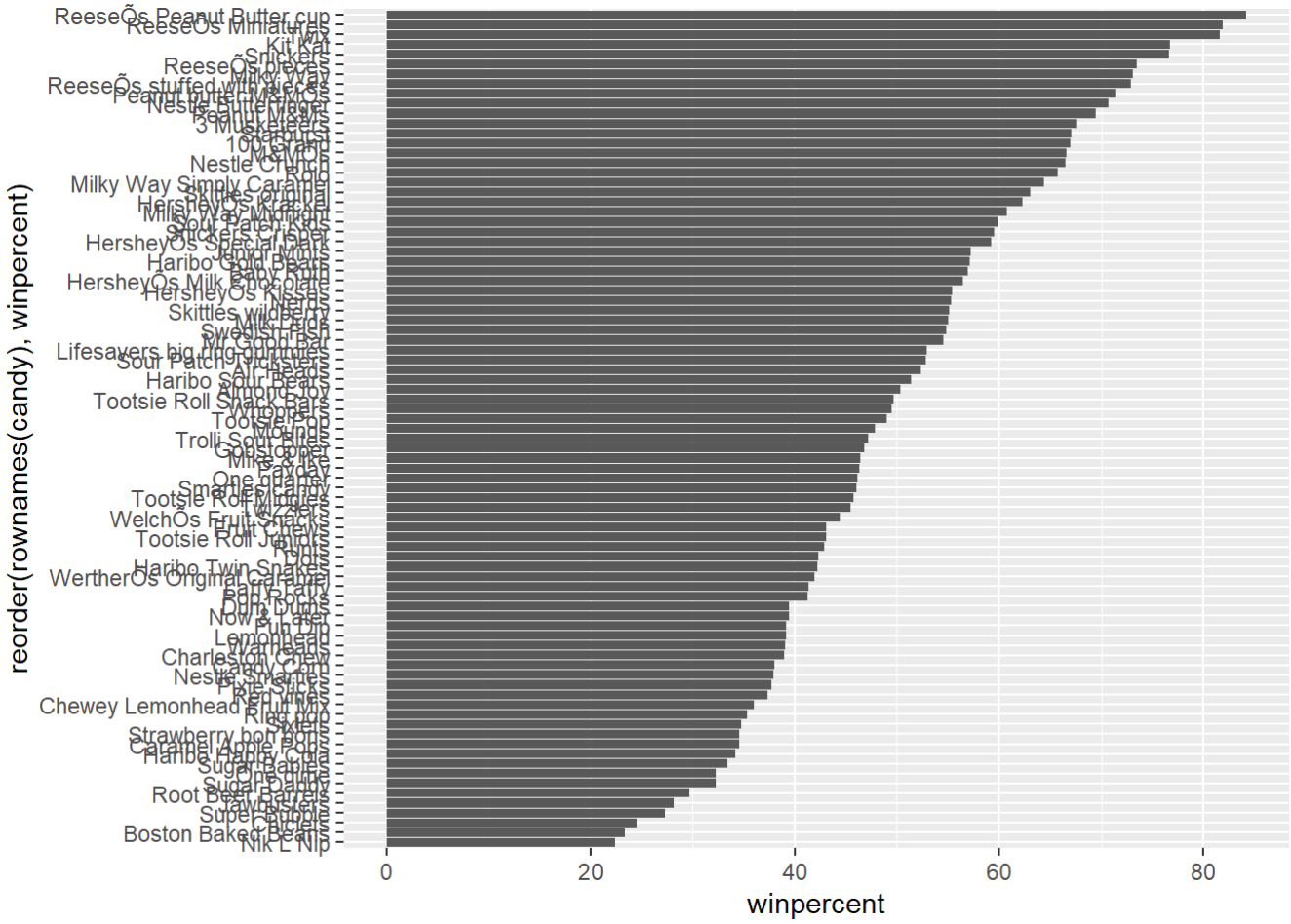
> Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_col()
```



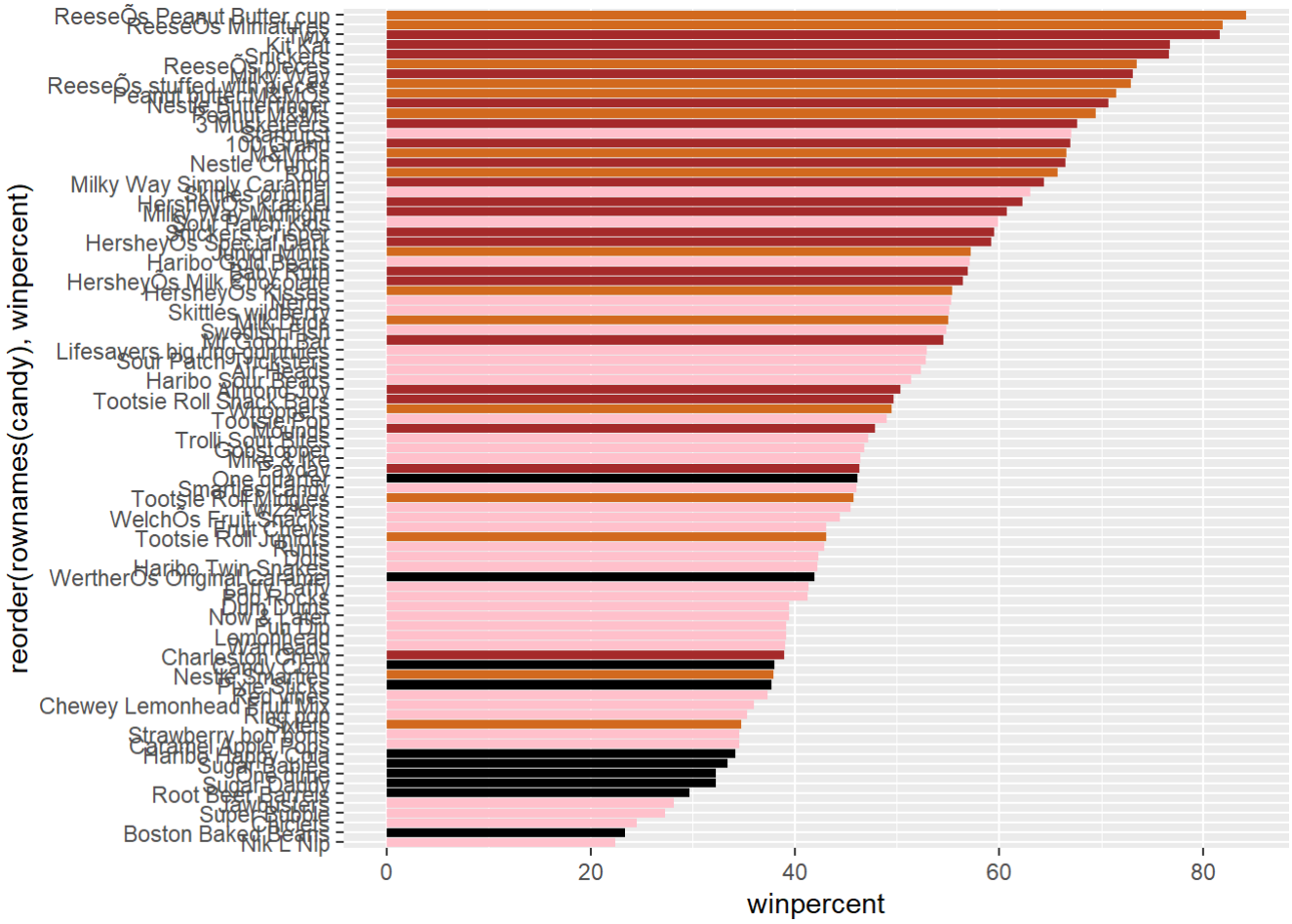
> Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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



Now add colors to differentiate between candy types!

```
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)
```




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

```
Saving 7 x 5 in image
```

> Q17. What is the worst ranked chocolate candy?

Worst ranked chocolate candy is Charleston Chew.

> Q18. What is the best ranked fruity candy?

Best ranked fruity candy is Starburst.

Price Percent

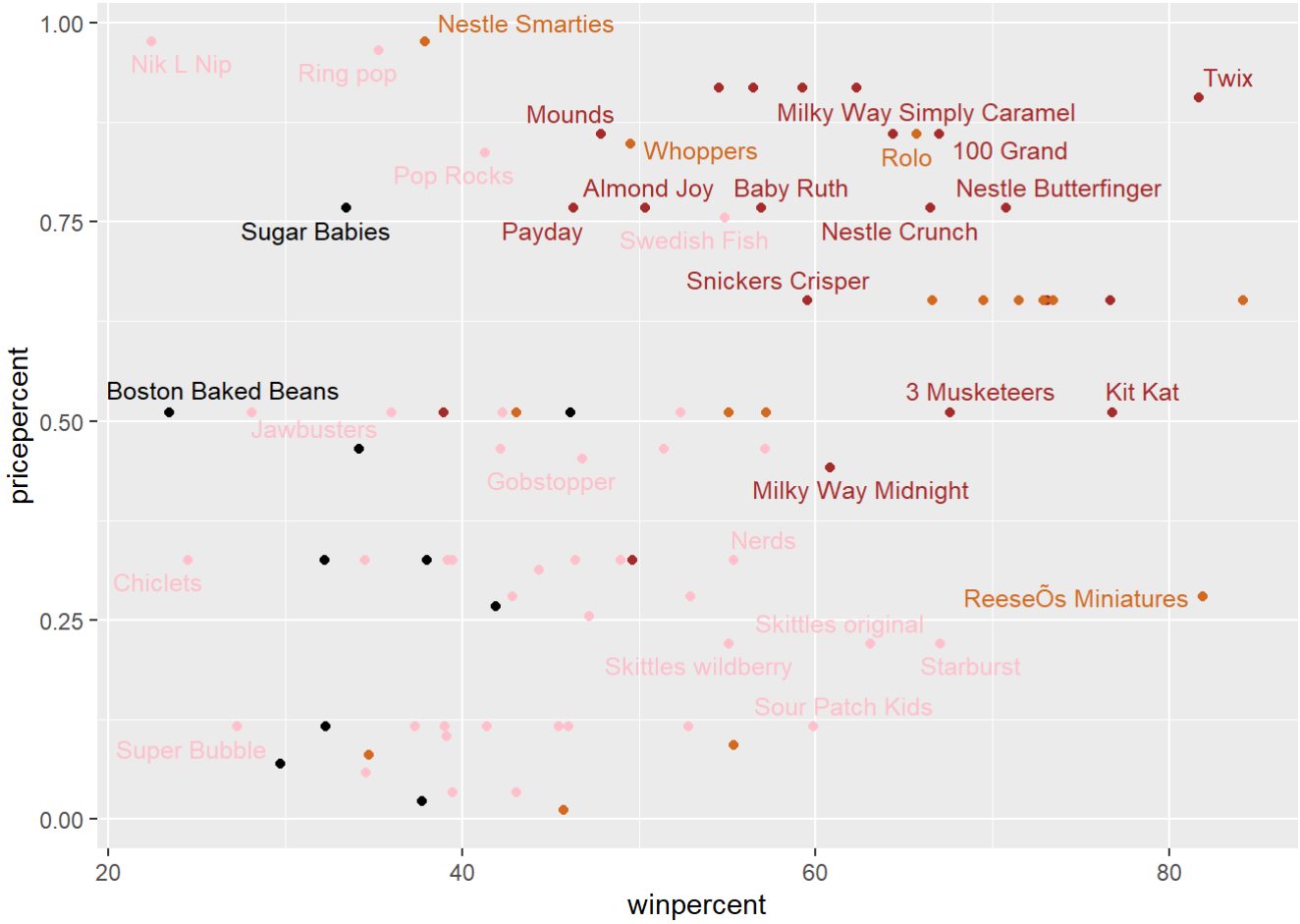
Price percent is barometer for candies' value for money. Best candy for the least money can be found with winpercent vs the pricepercent plot.

'geom_text_repel()' allow us to make sure the labels are not overlapping.

```
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: 53 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?

Best candy with win% and price% is Reeseos's Miniatures at the right bottom corner.

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

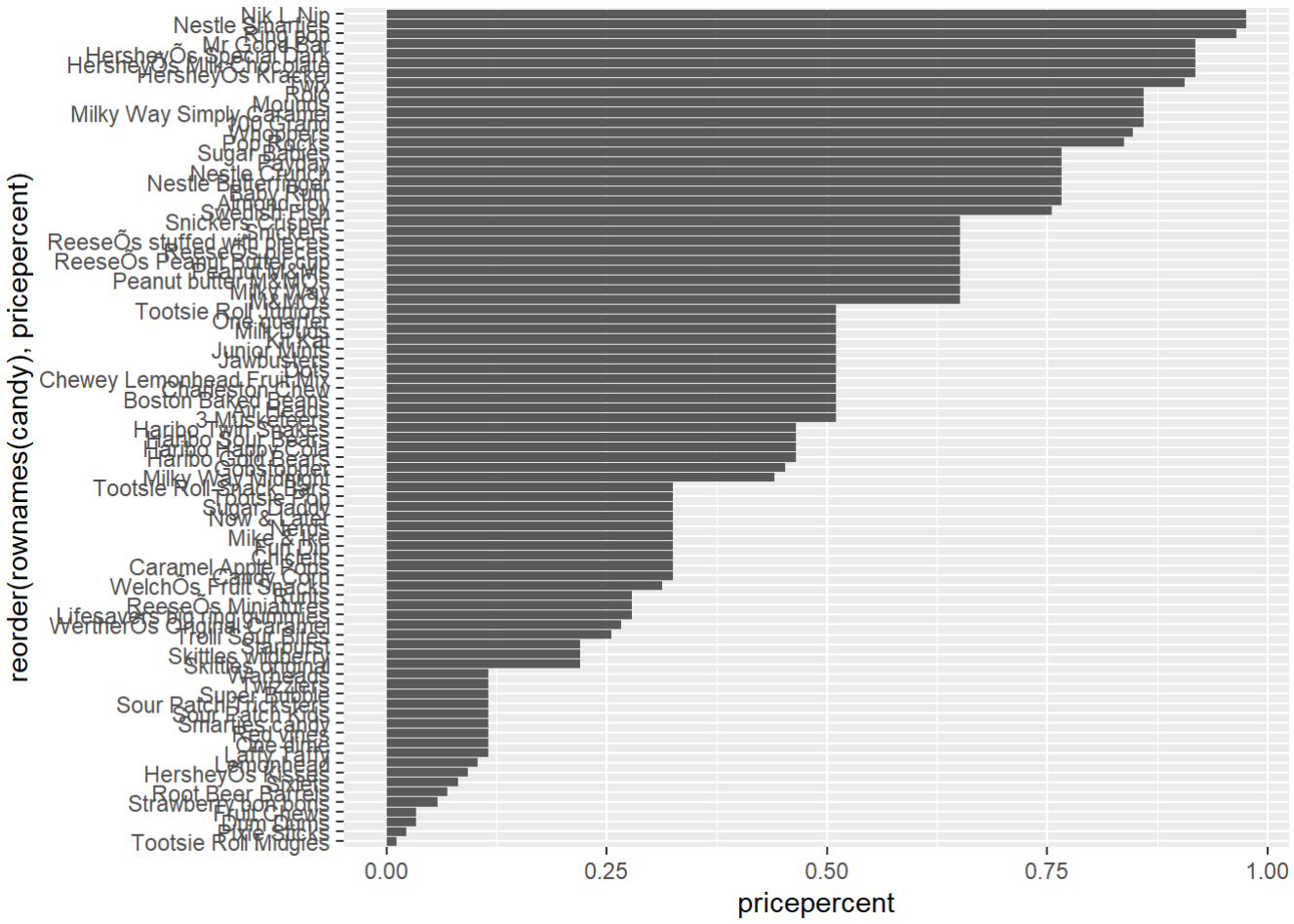
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head(candy[ord,c(11,12)], n=5 )
```

	pricepercent	winpercent
Nik L Nip	0.976	22.44534
Nestle Smarties	0.976	37.88719
Ring pop	0.965	35.29076
HersheyÕs Krackel	0.918	62.28448
HersheyÕs Milk Chocolate	0.918	56.49050

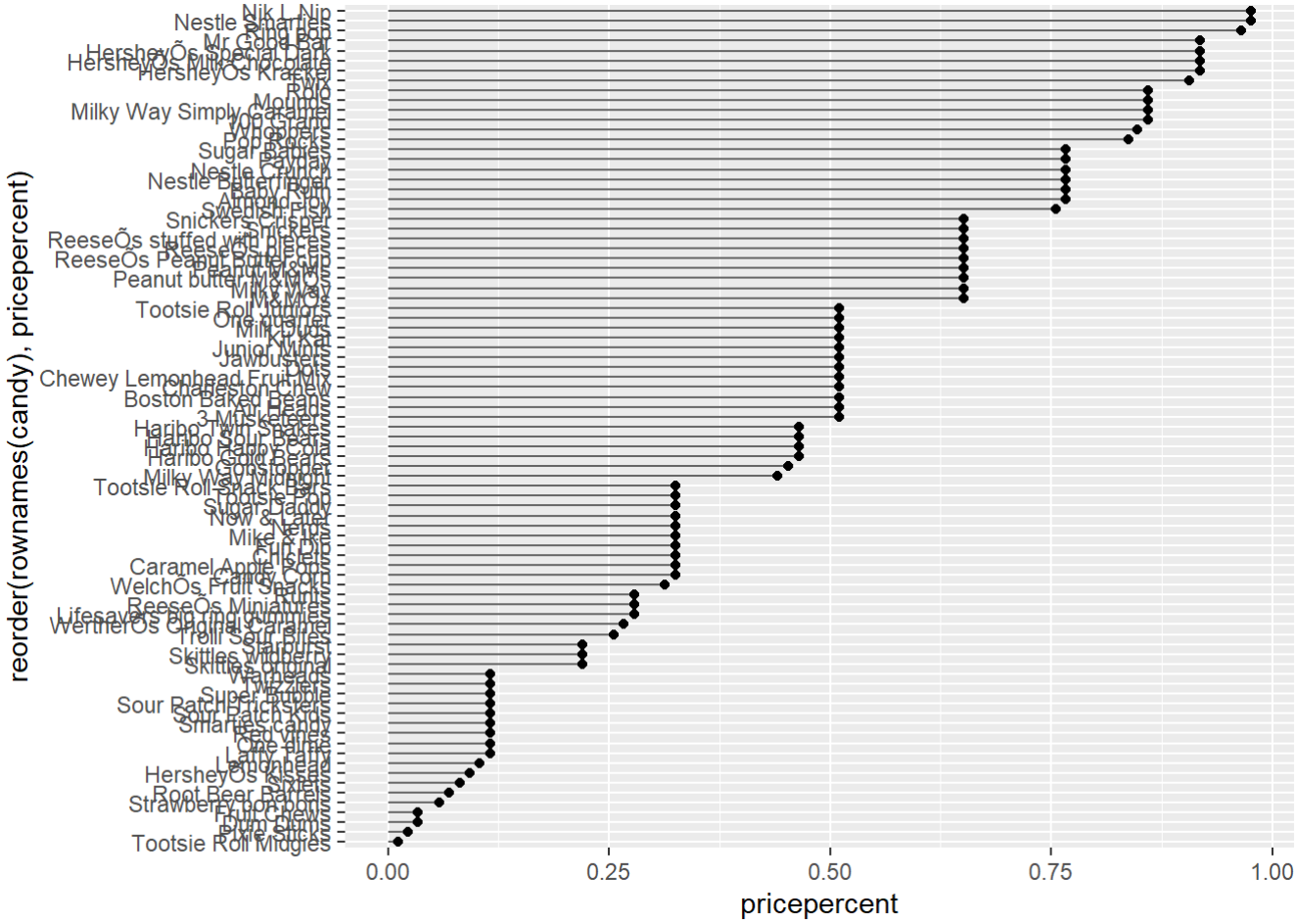
Shown above is the top 5 most expensive candies and amongst them, Nik L Nip is the least popular kind.

> 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()
```




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



We can notice that the quite a lot of them share the same prices in each group.

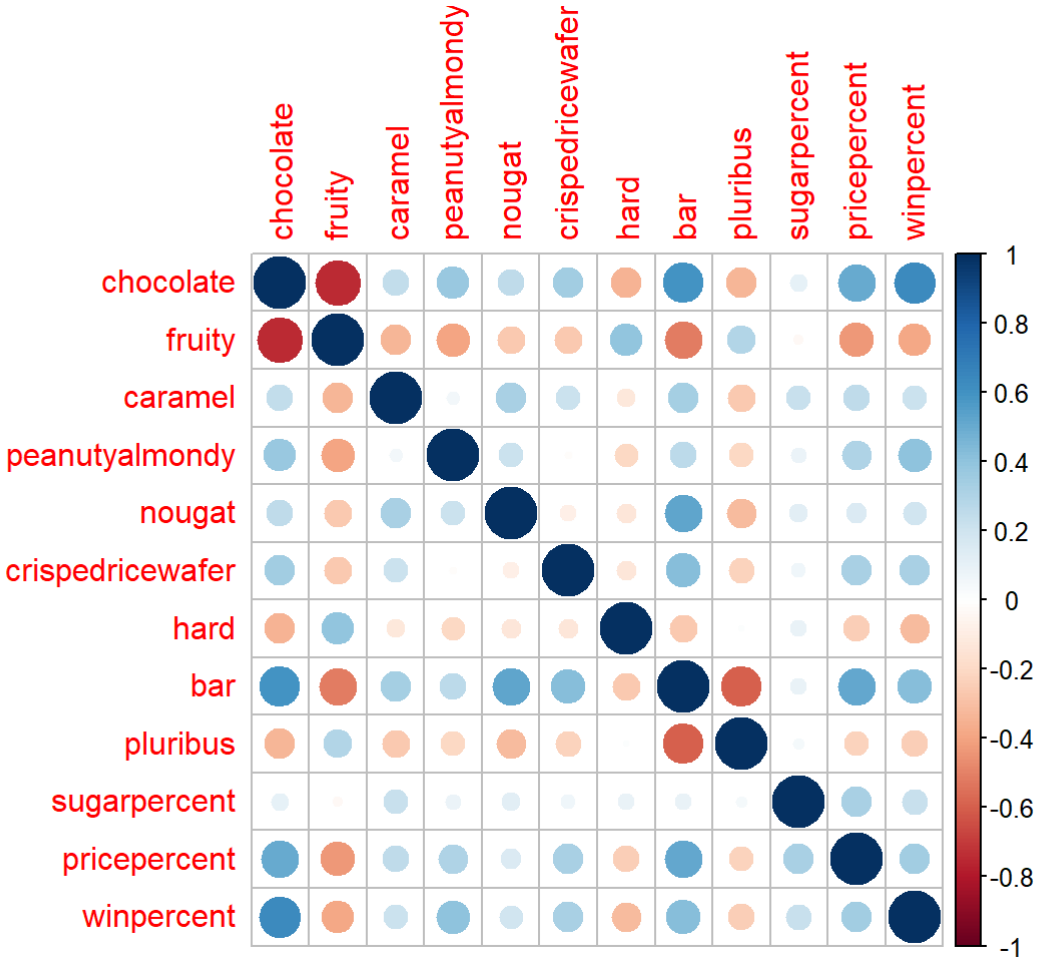
Exploring Correlation Data

To see correlation, we use corrplot package.

```
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 variables are most strongly anti-correlated because no body likes fruity and chocolaty candies.

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

The positively correlated variables are chocolate and winpercent. Chocolate for the win!

PCA

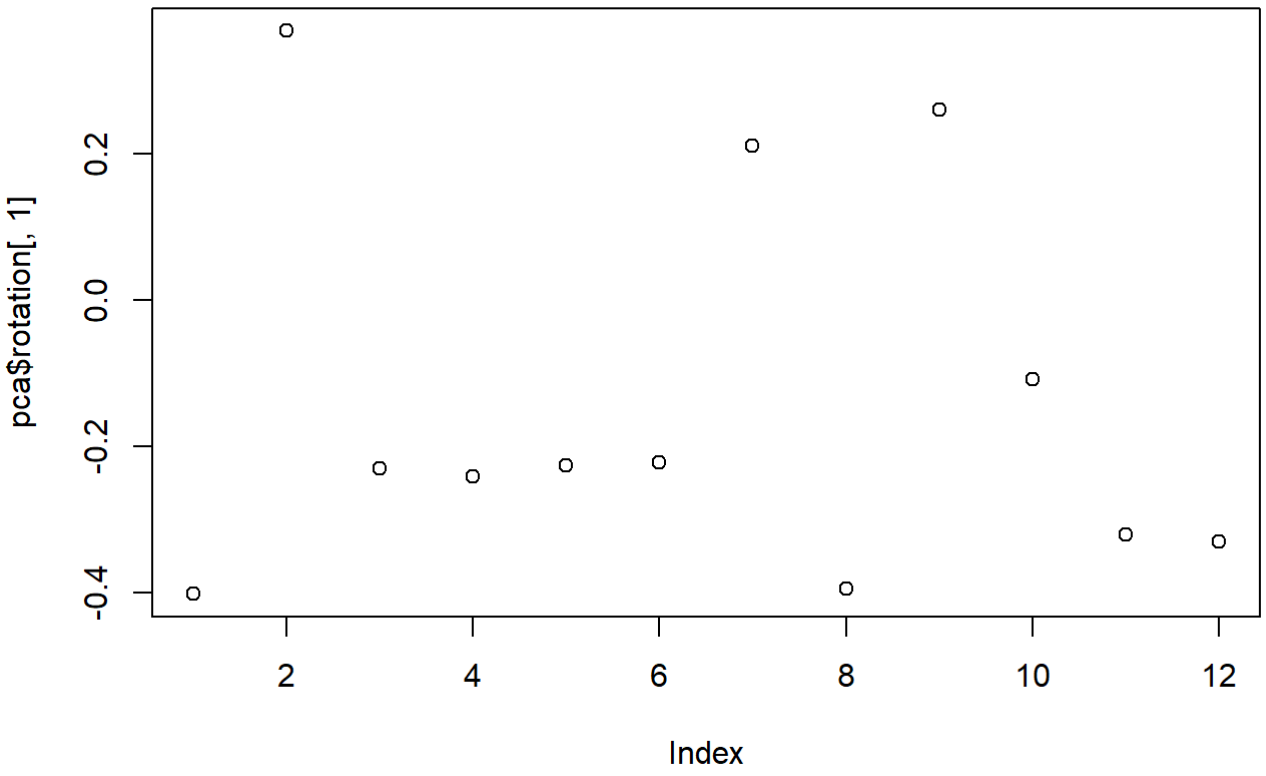
PCA using the prcom() function to our candy dataset remembering to set the scale=TRUE argument.

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

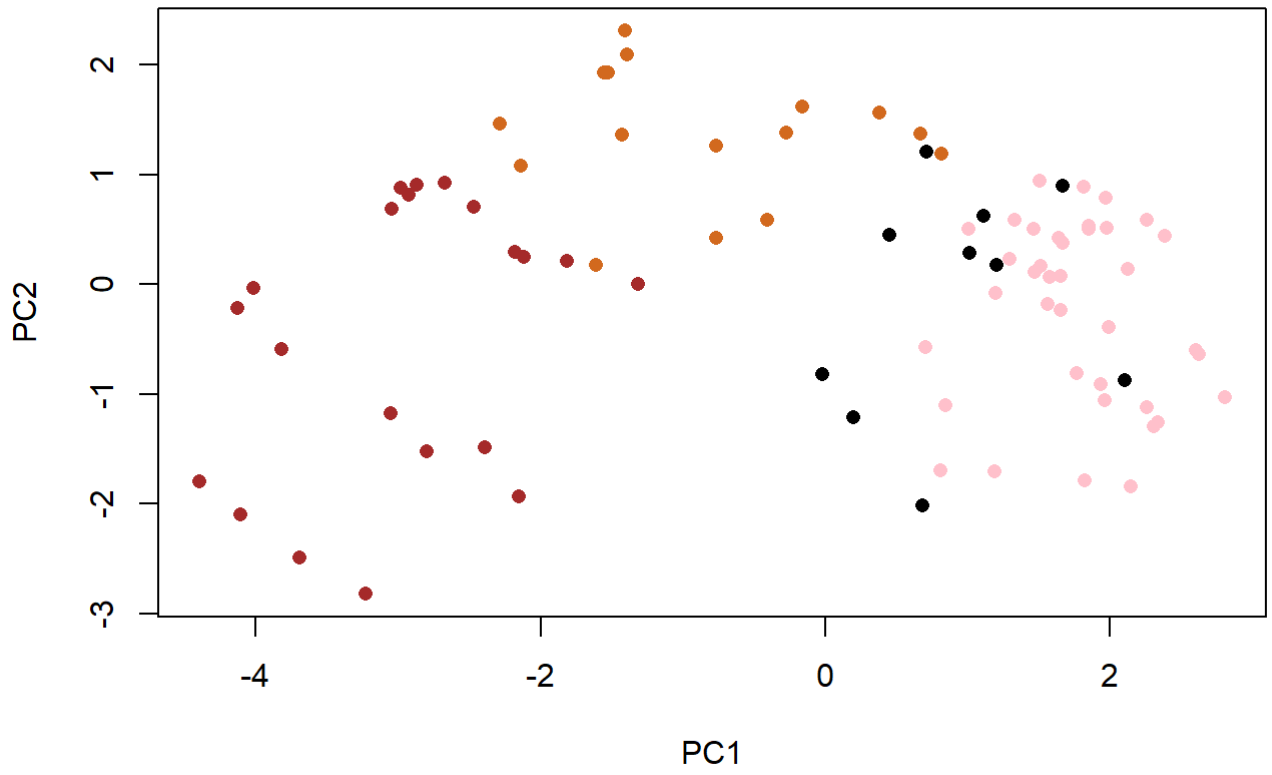
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 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$rotation[,1])
```

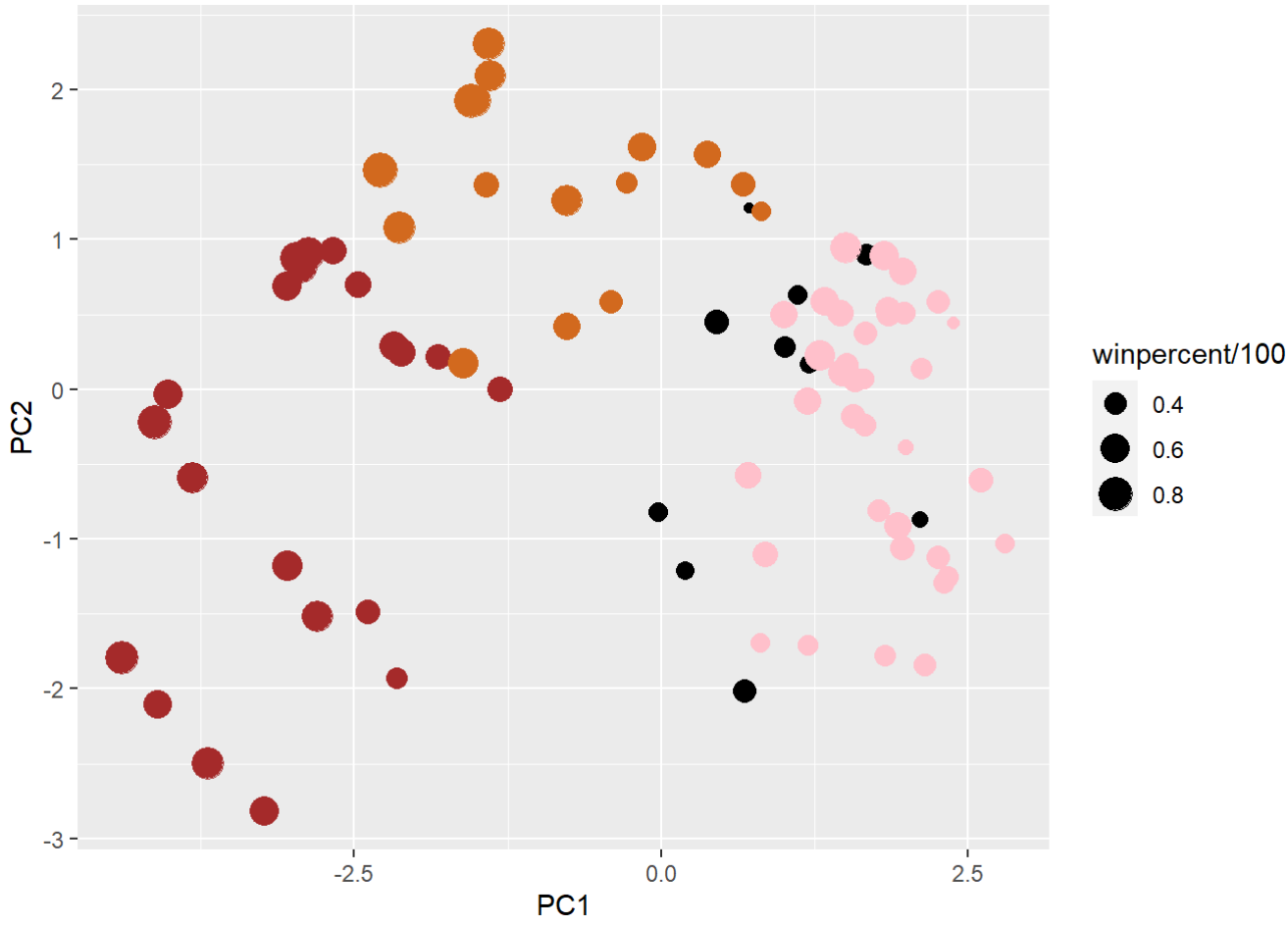



```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



To use PCA data in ggplot2, we have to make a new data.frame that can be used as an input for ggplot2.

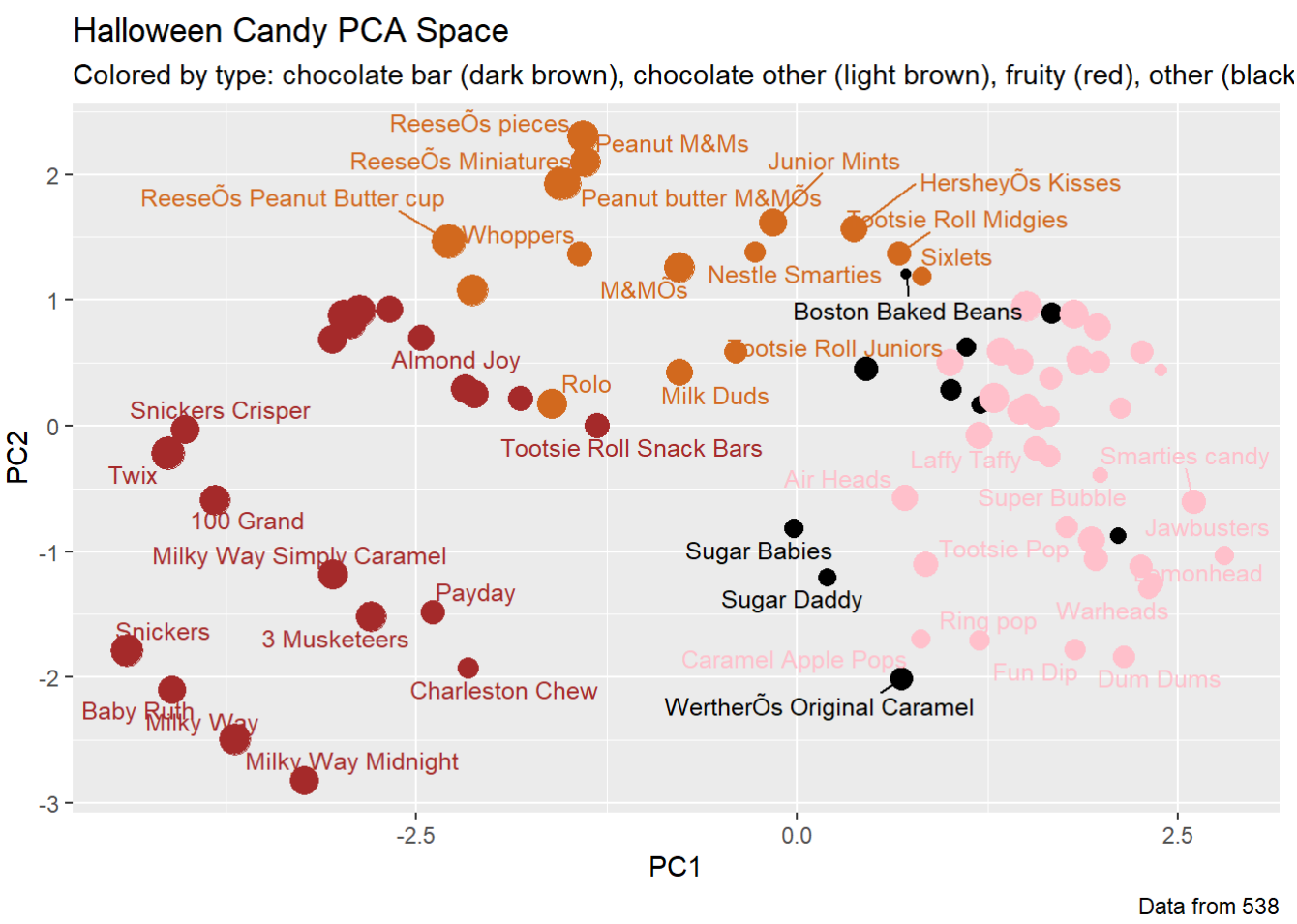
```
my_data <- cbind(candy, pca$x[,1:3])
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```



Again we can use the ggrepel package and the function ggrepel::geom_text_repel() without overlapping labels.

```
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), f
       caption="Data from 538")
```

```
Warning: ggrepel: 41 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```

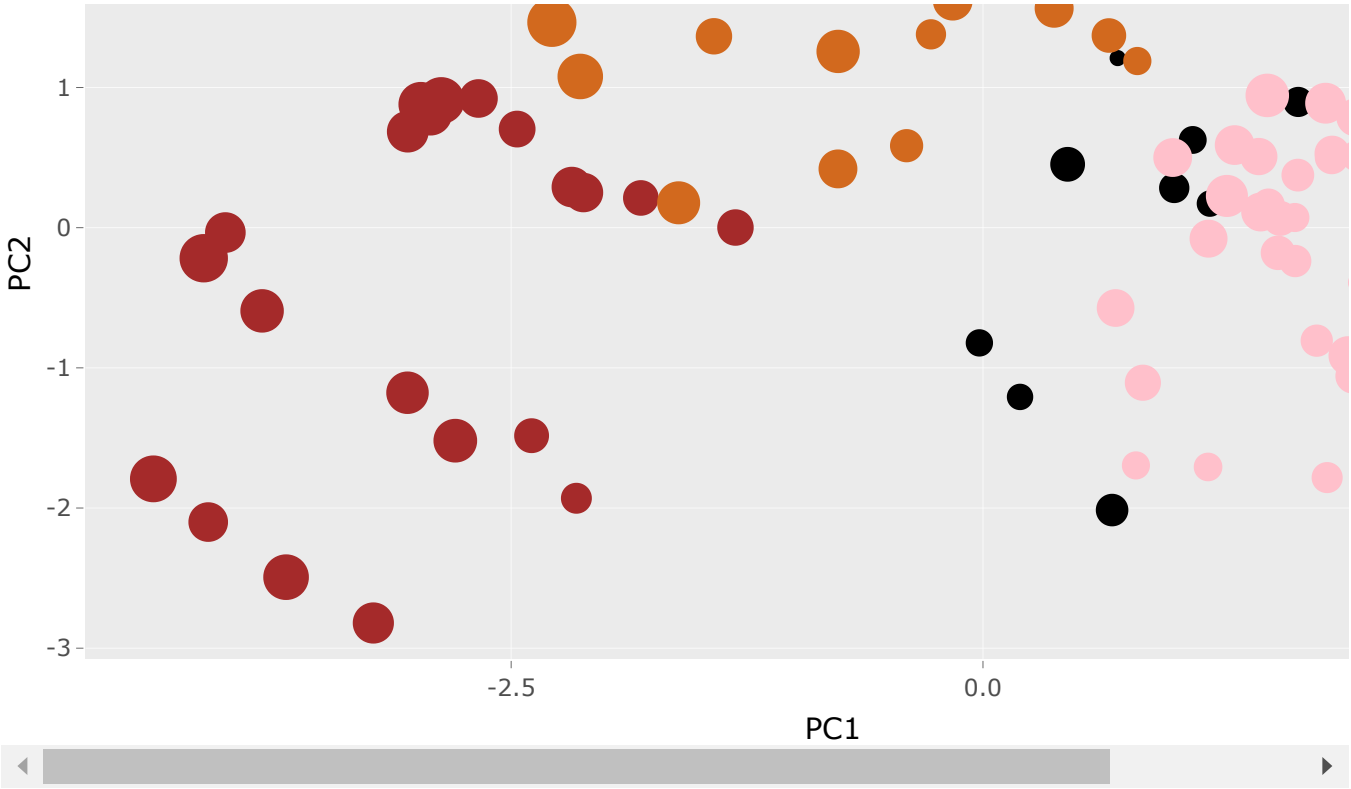


You can change 'max.overlaps' to allow more overlapping values or pass the ggplot object p to plotly to generate interactive plot.

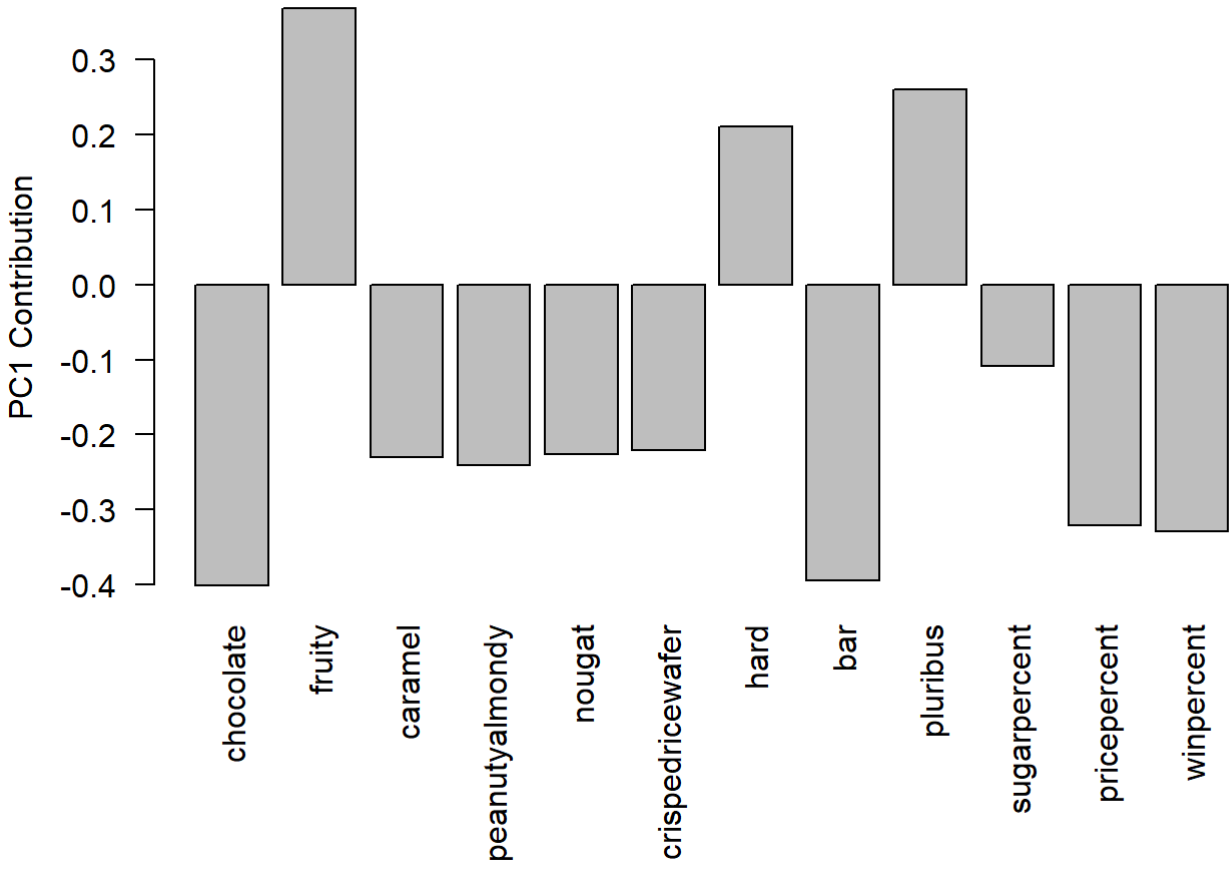
```
library(plotly)
```

```
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)
```




```
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?

PC1 in positive direction picked up fruity, hard and pluribus. This makes sense as it dispaly the same variables that were positively correlated in the 'corrplot' above. It is also visible in PCA plot as well, mostly on the right side of the graph. There are lot of hard fruity candies that come in a bag of multiple small packagings.