class09

AUTHOR

Jaewon Kim

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

candy = read.csv(candy_files, 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
```

Row is candies, while column is criteria. Therefore, there are 85 types of candies.

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

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

```
[1] 38
```

Candies that are fruity has 1 in cell. Therefore, there are 38 fruity candies.

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

```
candy["Skittles original", ]$winpercent
```

```
[1] 63.08514
```

My favorite candy is Skittles, and its winpercent is 63.085%.

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

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

```
[1] 76.7686
```

Winpercent of Kit Kat is 76.769%.

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

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

```
[1] 49.6535
```

Winpercent of Tootsie Roll Snack Bars is 49.654%.

```
#install.packages("skimr")
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? - p0 to p100 of chocolate to pluribus is either zero or one. However, column values for sugarpercent to winpercent are in continuous numeric value between zero to one or zero to hundred (because they're percent data). Therefore, variable chocolate$_{\text{pluribus/sugarpercent}}$winpercent have different scale for column p0~p100.

Q7. What do you think a zero and one represent for the candy$chocolate column? - Zero means that corresponding candies doesn't contain chocolate.

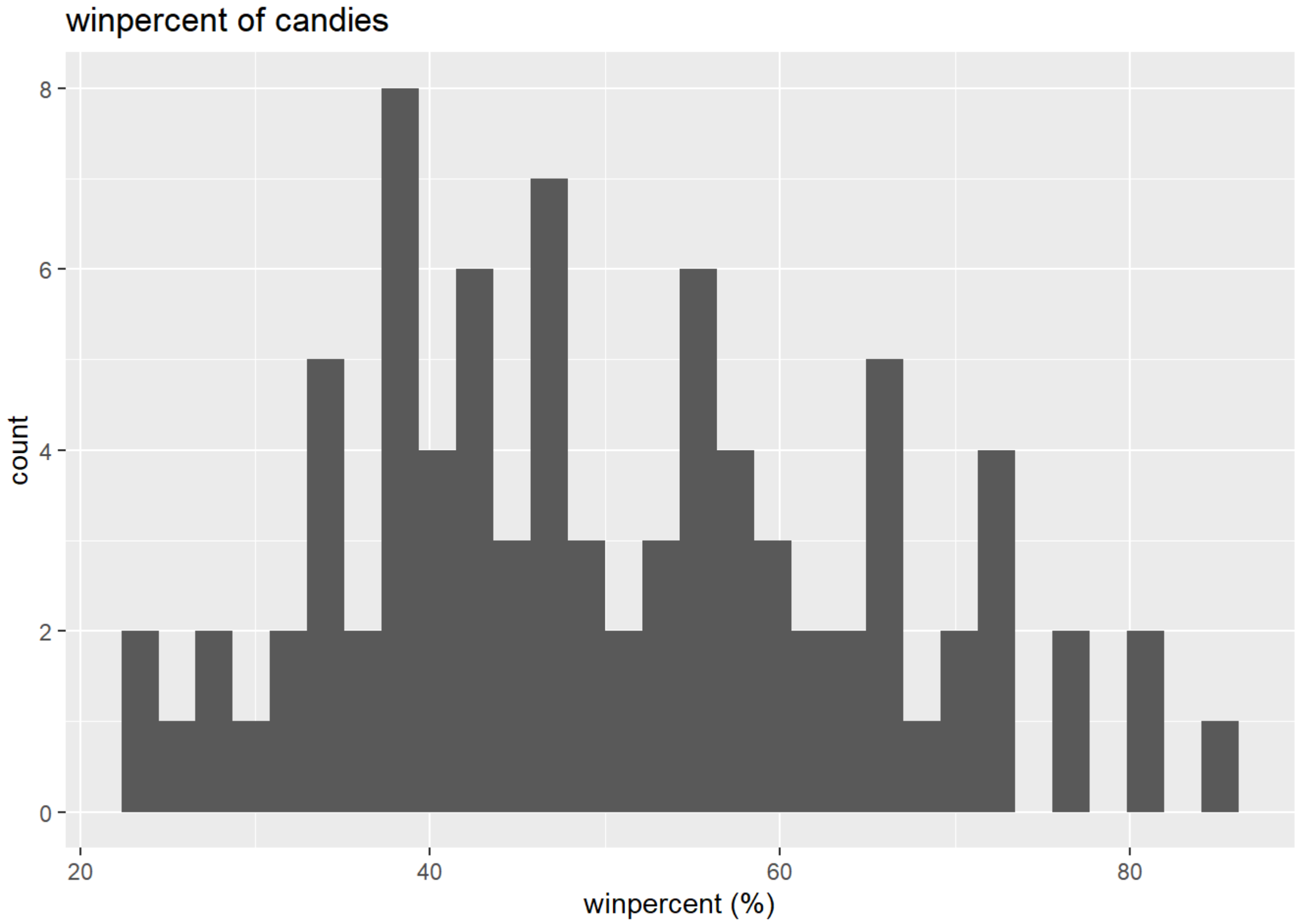
Q8. Plot a histogram of winpercent values

```r
library(ggplot2)
library(dplyr)
```

```

다음의 패키지를 부착합니다: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```r
ggplot(candy, aes(x = candy$winpercent)) +
  geom_histogram() +
  xlab("winpercent (%)") +
  ylab("count") +
  ggtitle("winpercent of candies")
```

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



Q9. Is the distribution of winpercent values symmetrical? - No, distribution shows right skewness.

Q10. Is the center of the distribution above or below 50%? - Since distribution is positive skew, center of the distribution is below 50%.

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

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

```
[1] 60.92153
```

```r
mean(can)
```

```
[1] 44.11974
```

Average winpercent of chocolate candy is 60.922%, while fruity candy is 44.120%. Therefore, chocolate candy is ranked higher than fruit candy.

Q12. Is this difference statistically significant?

```r
t.test(choc, can)
```

```

	Welch Two Sample t-test
```

```
data:  choc and can
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 
```

Calculated t value was 6.25, and corresponding p-value at 68.882 degree of freedom was less than 0.001. Low p-value fails to reject null hypothesis, hence, there is no statistically significant evidence that tow groups are significantly different. Therefore, no

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

```
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
```

From least favorite, 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?

```
candy %>%
  arrange(desc(winpercent)) %>%
  head(5)
```

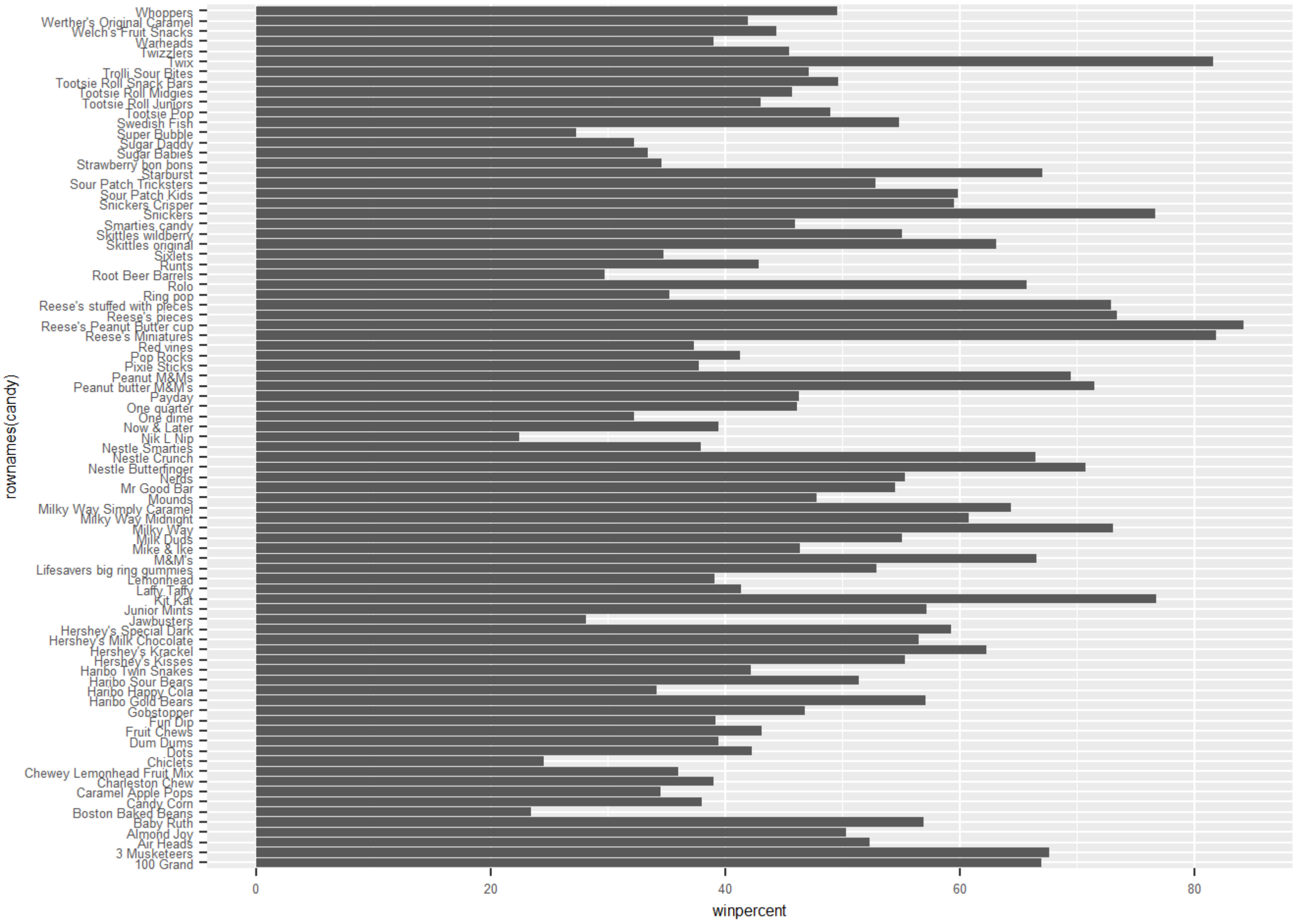
```
                          chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat                           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
```

From most favorite, five all time favorite candies are Reese’s Peanut Butter cup, Reese’s Miniatures, Twix, Kit Kat, and Snickers

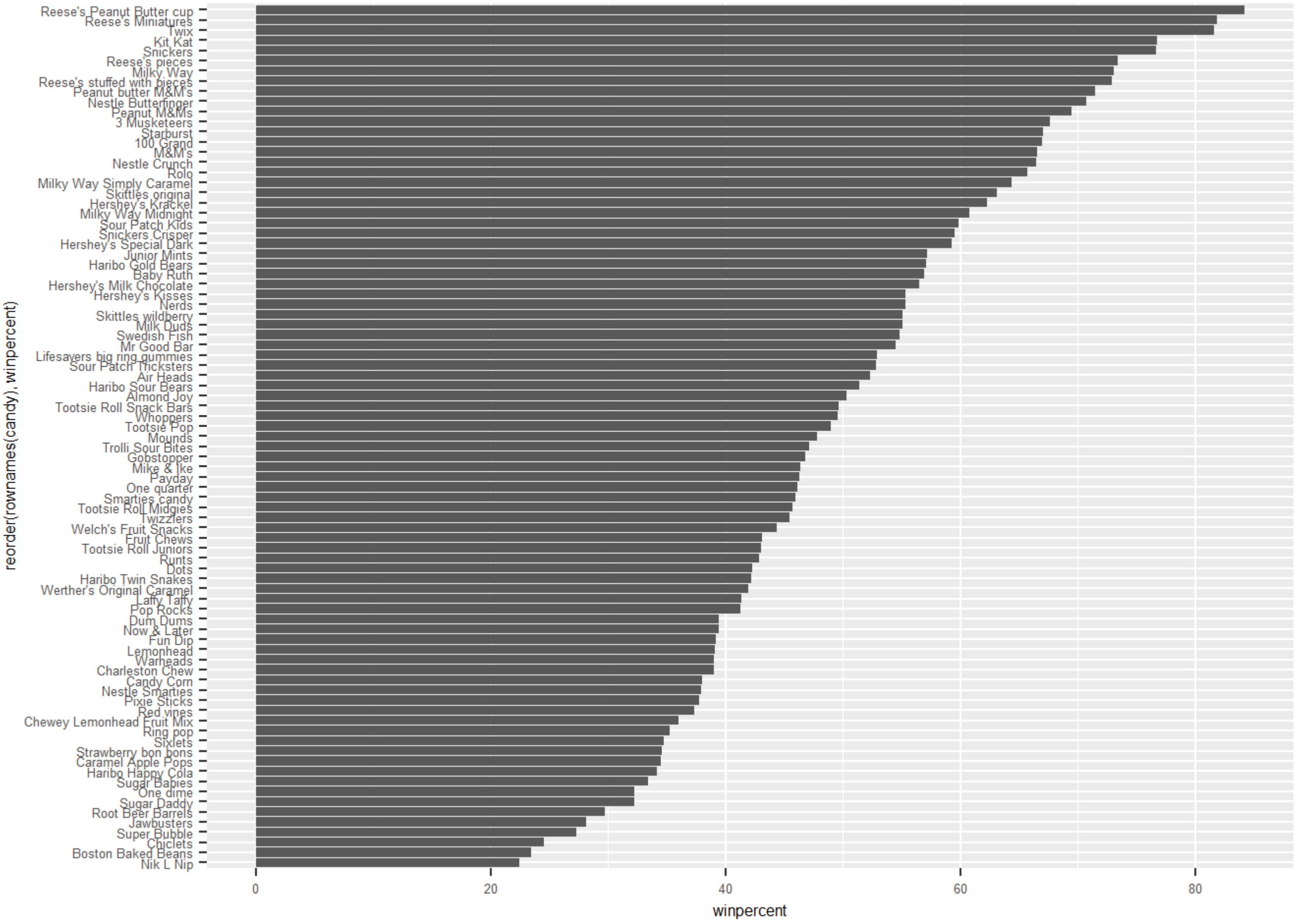
Q15. Make a first barplot of candy ranking based on winpercent values.

```
ggplot(candy) +
  aes(x = rownames(candy), y = winpercent) +
  geom_col() +
  coord_flip() +
  theme(text = element_text(size = 6))
```

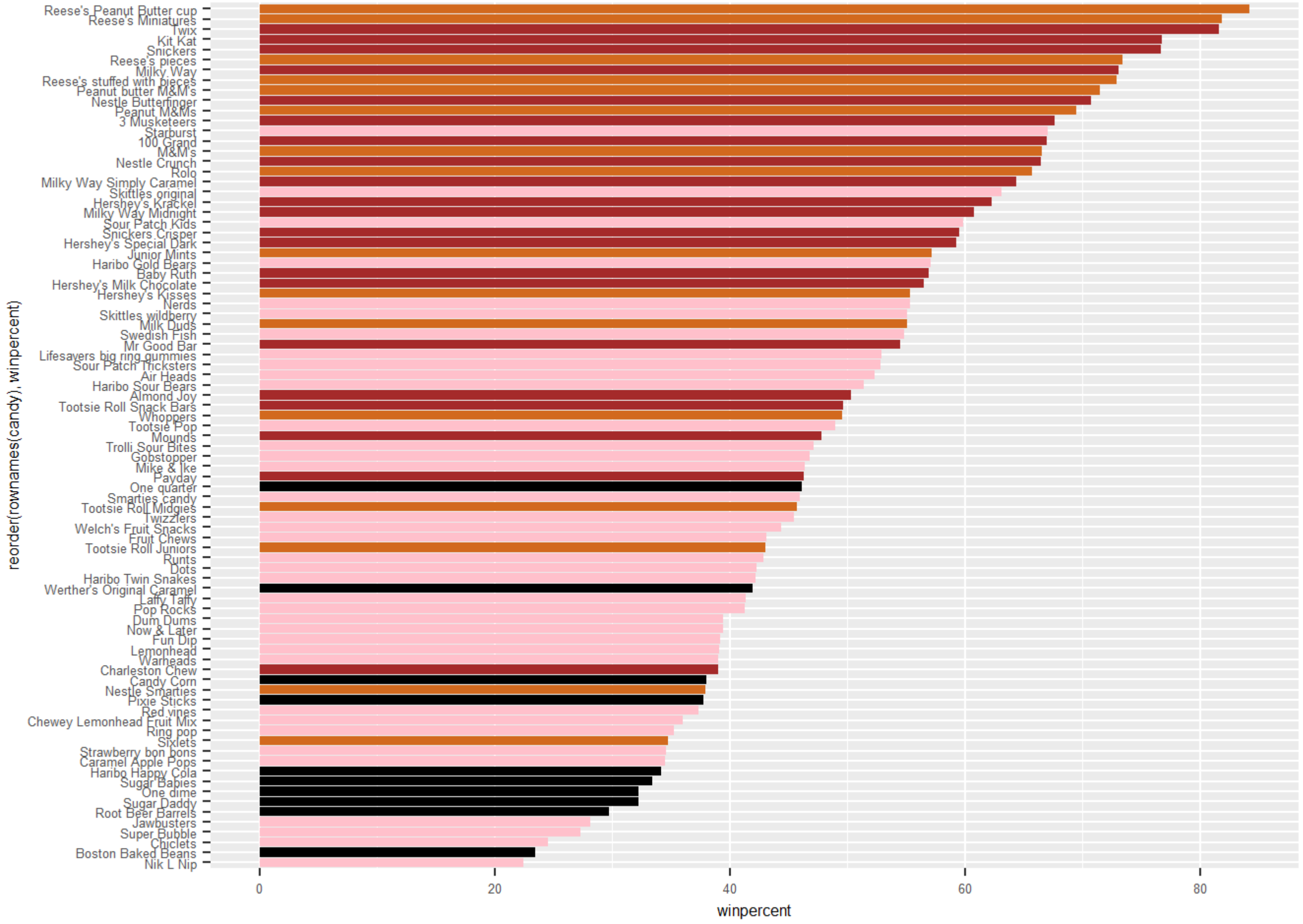


Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
p<- ggplot(candy) +
  aes(x = reorder(rownames(candy),winpercent), y = winpercent) +
  geom_col() +
  coord_flip() +
  theme(text = element_text(size = 6))
p
```



```
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"

p +
  geom_col(fill = my_cols)
```



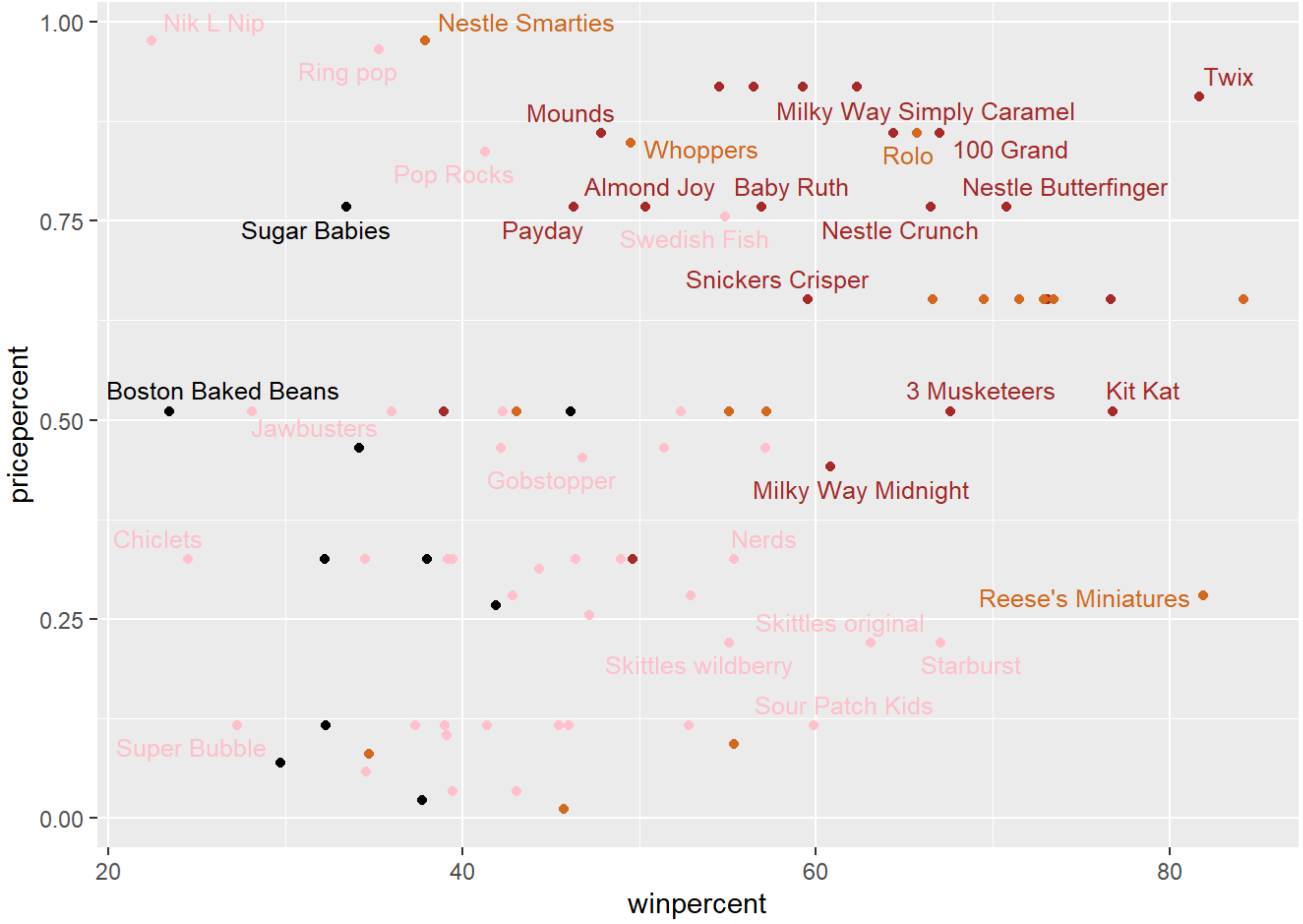
Now, for the first time, using this plot we can answer questions like: - Q17. What is the worst ranked chocolate candy? Lowest bar with chocolate color is Sixlets. Hence, sixlets is the worst ranked chocolate candy.

- Q18. What is the best ranked fruity candy? Highest bar with pink color is Starburst. Hence, Starburst is the best ranked fruity candy.

plot privepercent vs. winpercent

```
#install.packages("ggrepel")
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 = 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?

pricepercent shows percentile rank of unit price, and winpercent shows percentage of wins. Therefore, highest winpercent with least money (lowest pricepercent) is Reese's Miniatures.

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

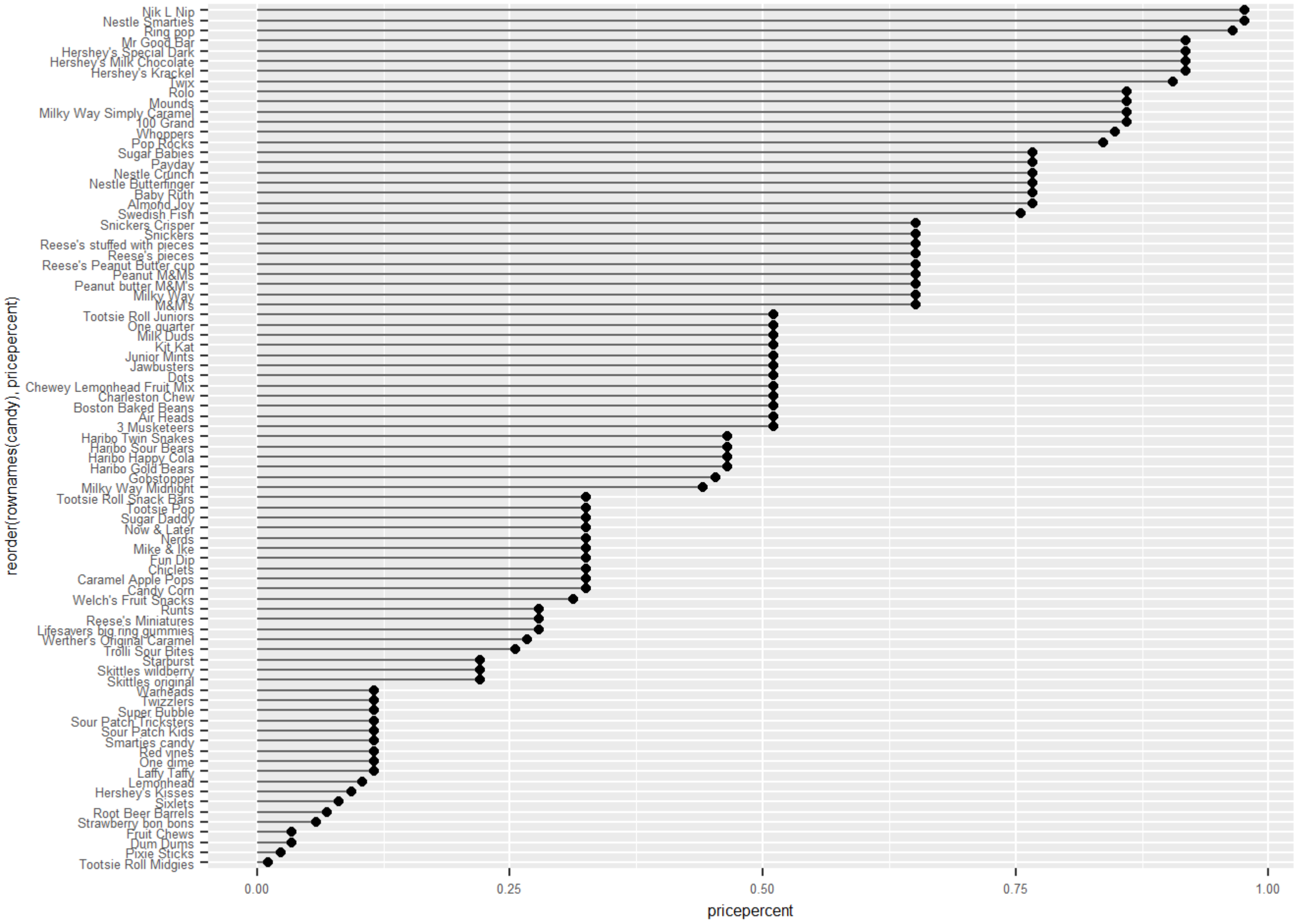
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

```
                         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
```

Five most expensive candies are Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel, and Hershey's Milk Chocolate. Least popular among them is Nik L Nip.

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().

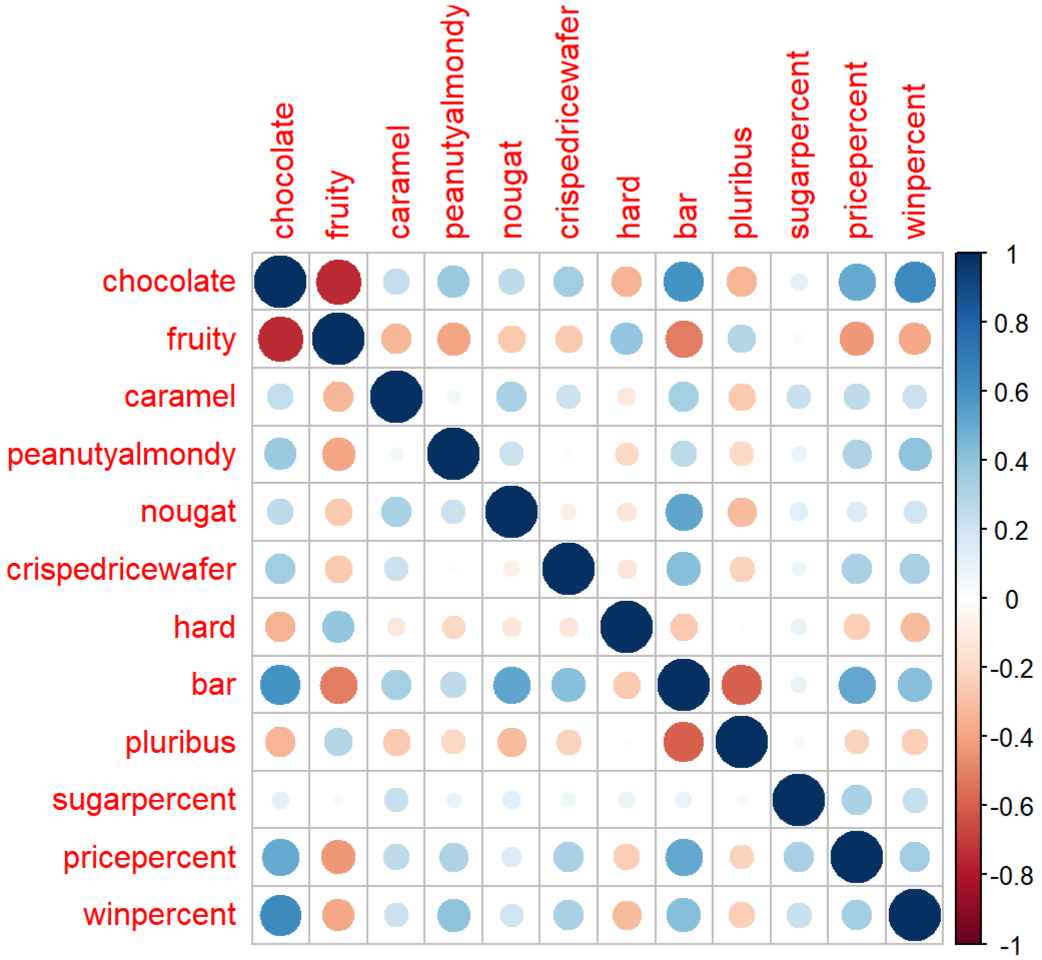
```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
  geom_point() +
  theme(text = element_text(size = 6))
```



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

Fruity candy and Chocolate candies are most anti-correlated.

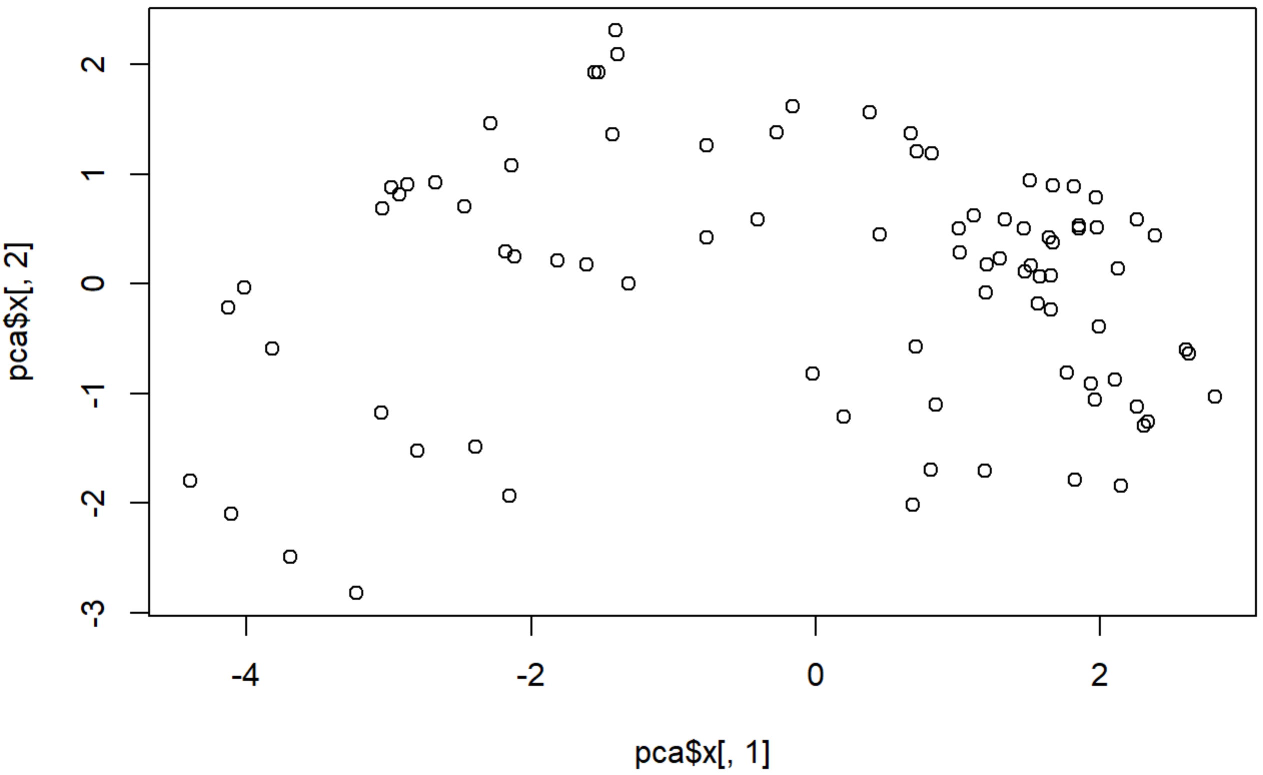
Q23. Similarly, what two variables are most positively correlated?

Besides diagonal, Chocolate and bar, winpercent and chocolate are most positively correlated.

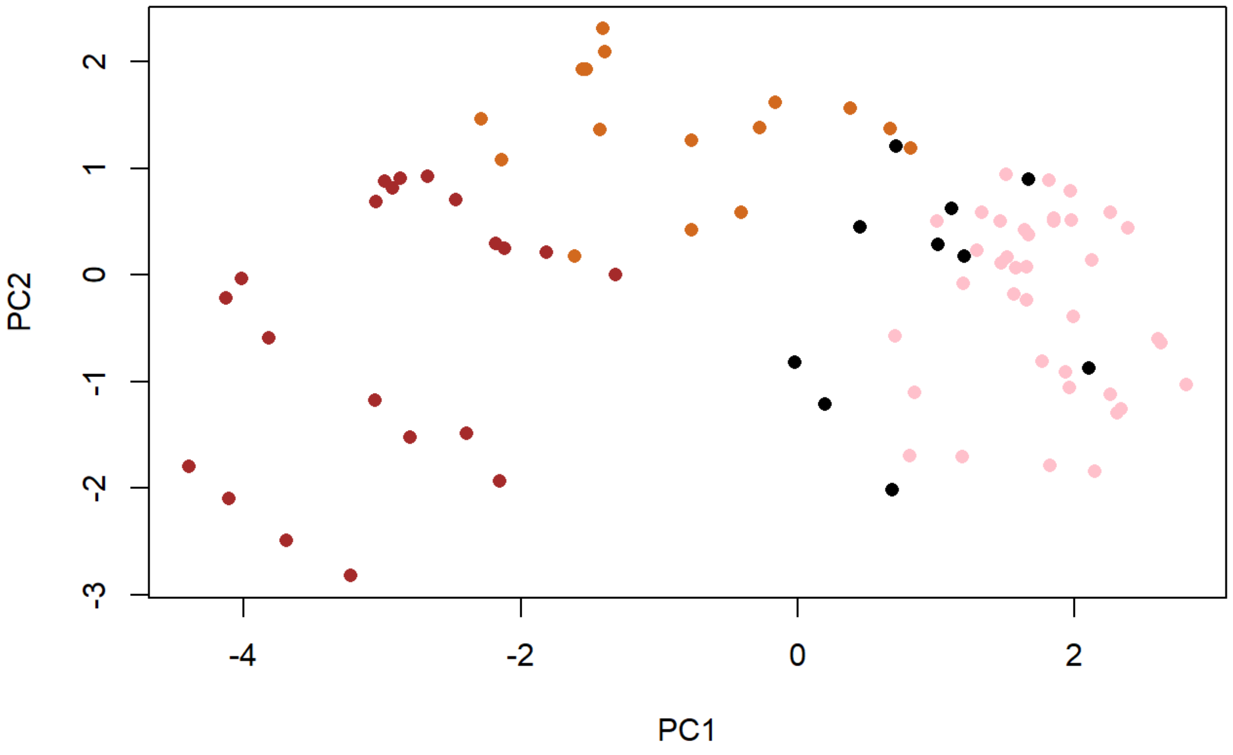
```r
pca <- prcomp(candy, scale = TRUE)
summary(pca)
```

```
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
```

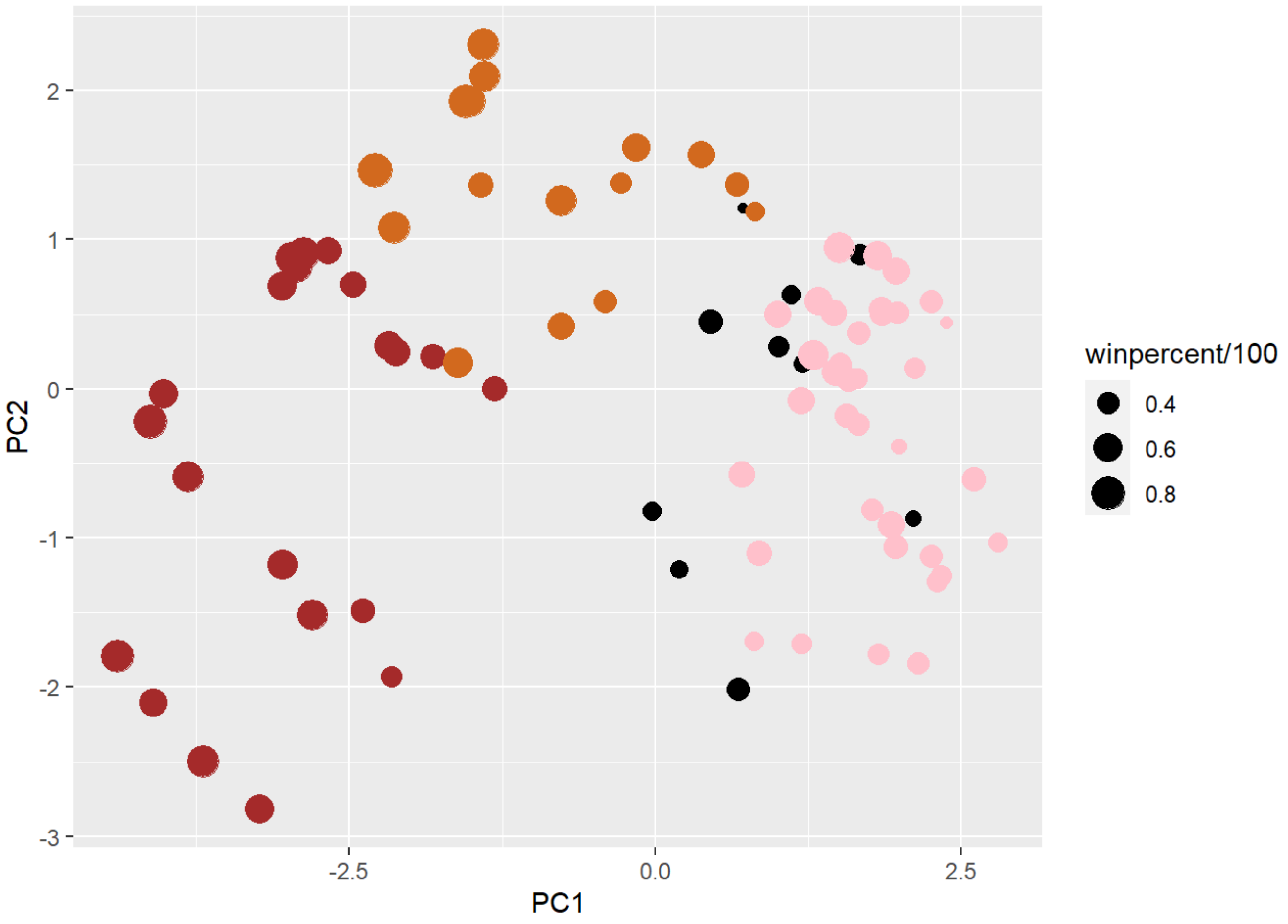
```r
plot(pca$x[ ,1], pca$x[ ,2])
```



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

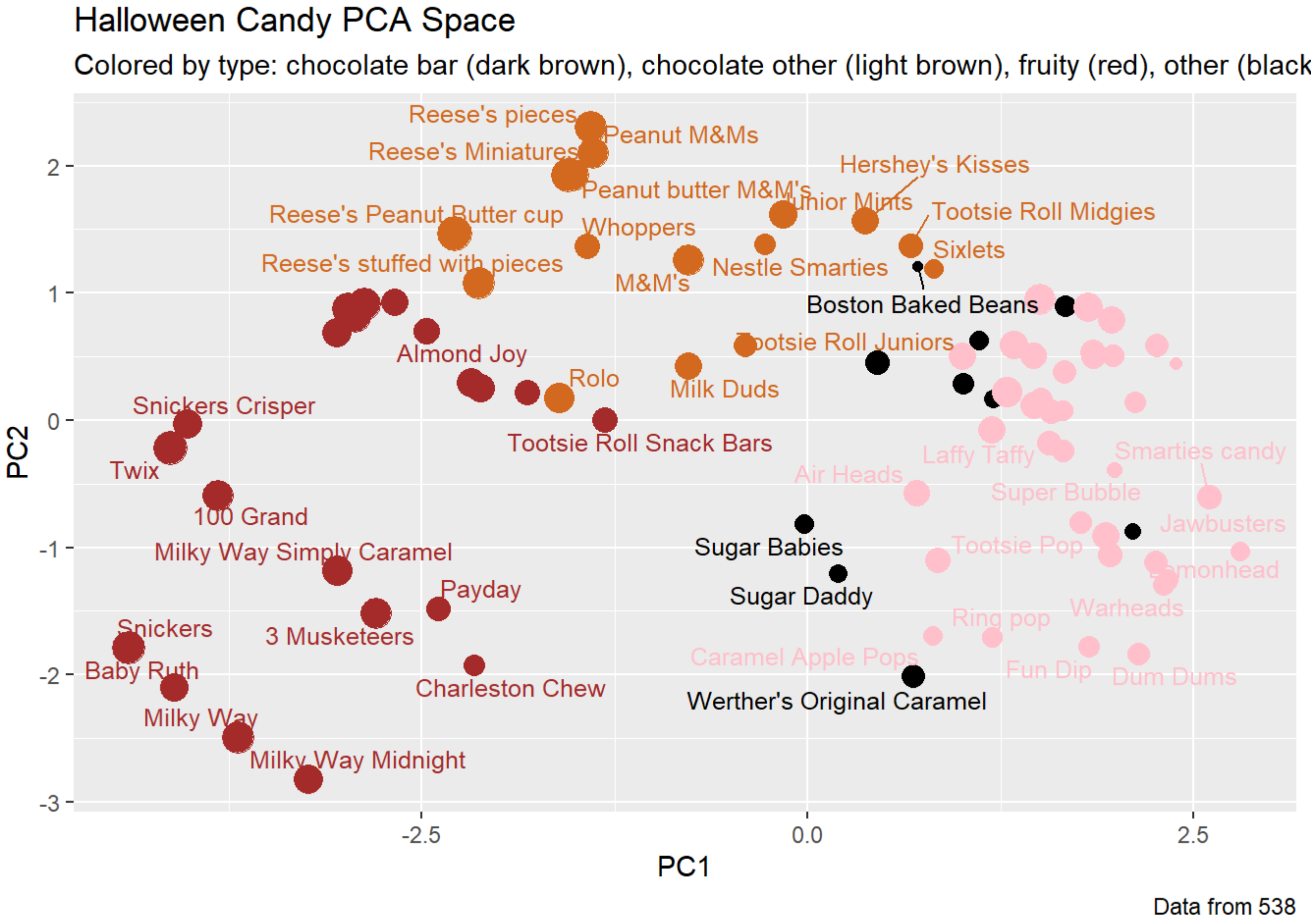


```
my_data <- cbind(candy, pca$x[,1:3])

pp <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

pp
```



```
pp +
  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
```

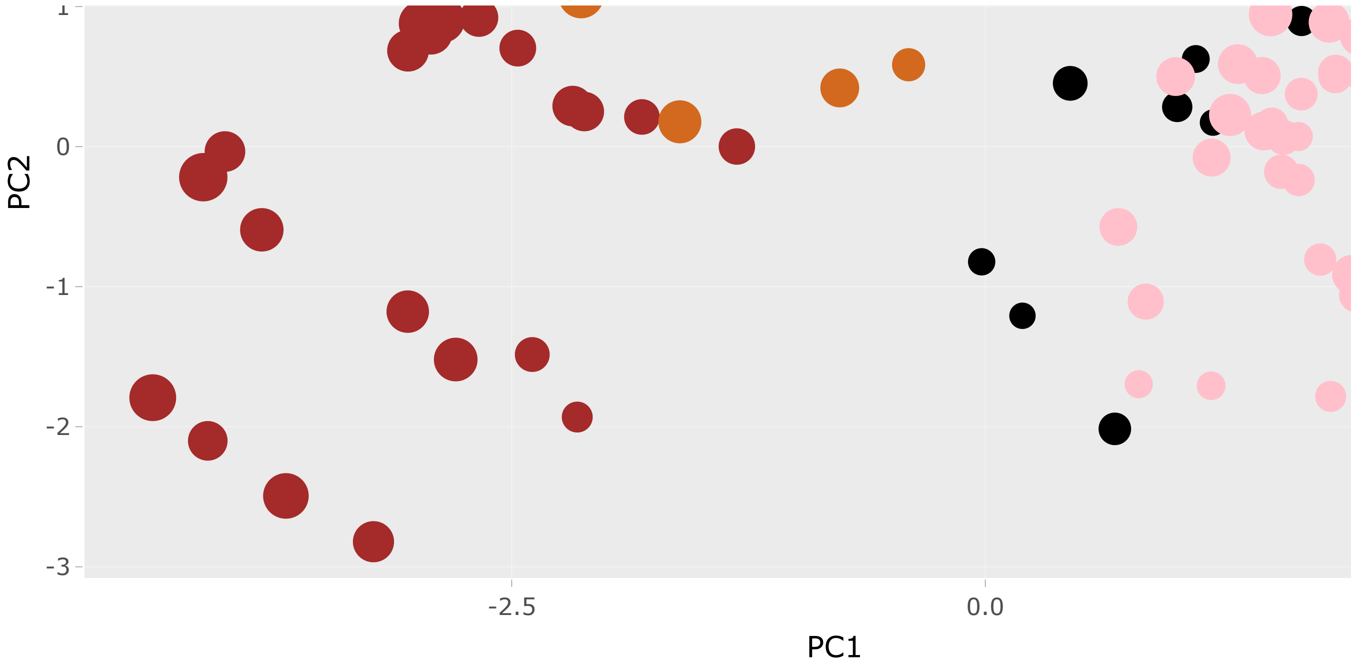
```
Warning: ggrepel: 40 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



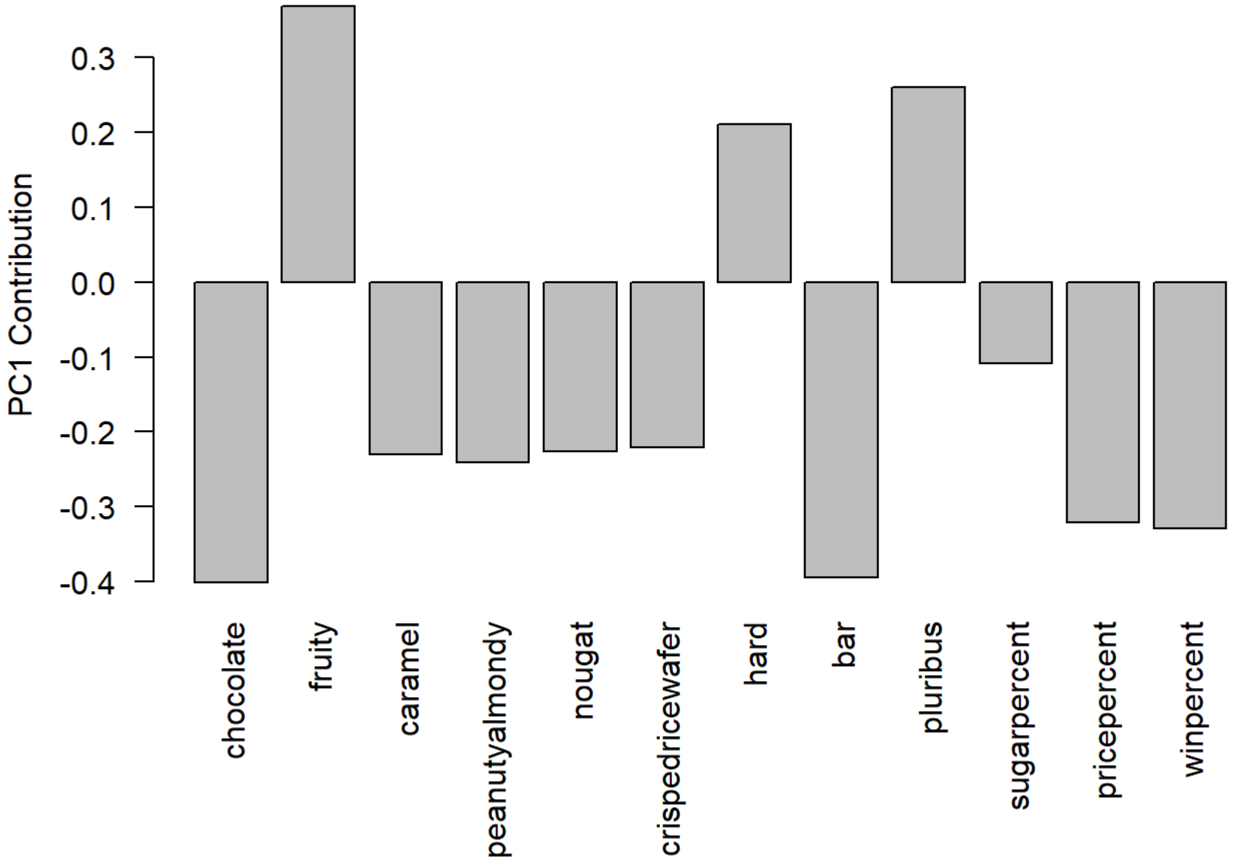
```
#install.packages("plotly")
library(plotly)
```

```
다음의 패키지를 부착합니다: '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(pp)
```



```
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, hard, and pluribus are strongly affecting PC1 into positive direction. This result make sense because all three variables are either zero or one, where candies must align to trend or have huge deviation (when each variables are plotted).