Lab10

Matthew

2/11/2023

First

```
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?

```
dim(candy)
```

```
## [1] 85 12
```

There are 85 different candy types

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

dim(candy)

```
sum(candy$fruity,na.rm=TRUE)
```

```
## [1] 38
```

There are 38 fruity candy types in the database.

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

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

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

```
candy["Sour Patch Kids", ]$winpercent
```

```
## [1] 59.864
```

My favorite candy in the dataset is Sour Patch Kids and it has a winpercent value of 59.864%.

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

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

```
## [1] 76.7686
```

Kit Kat has a win percent of 76.7686%.

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

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

```
## [1] 49.6535
```

Tootsie Roll Snack Bars has a win percent of 49.6535%.

```
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?

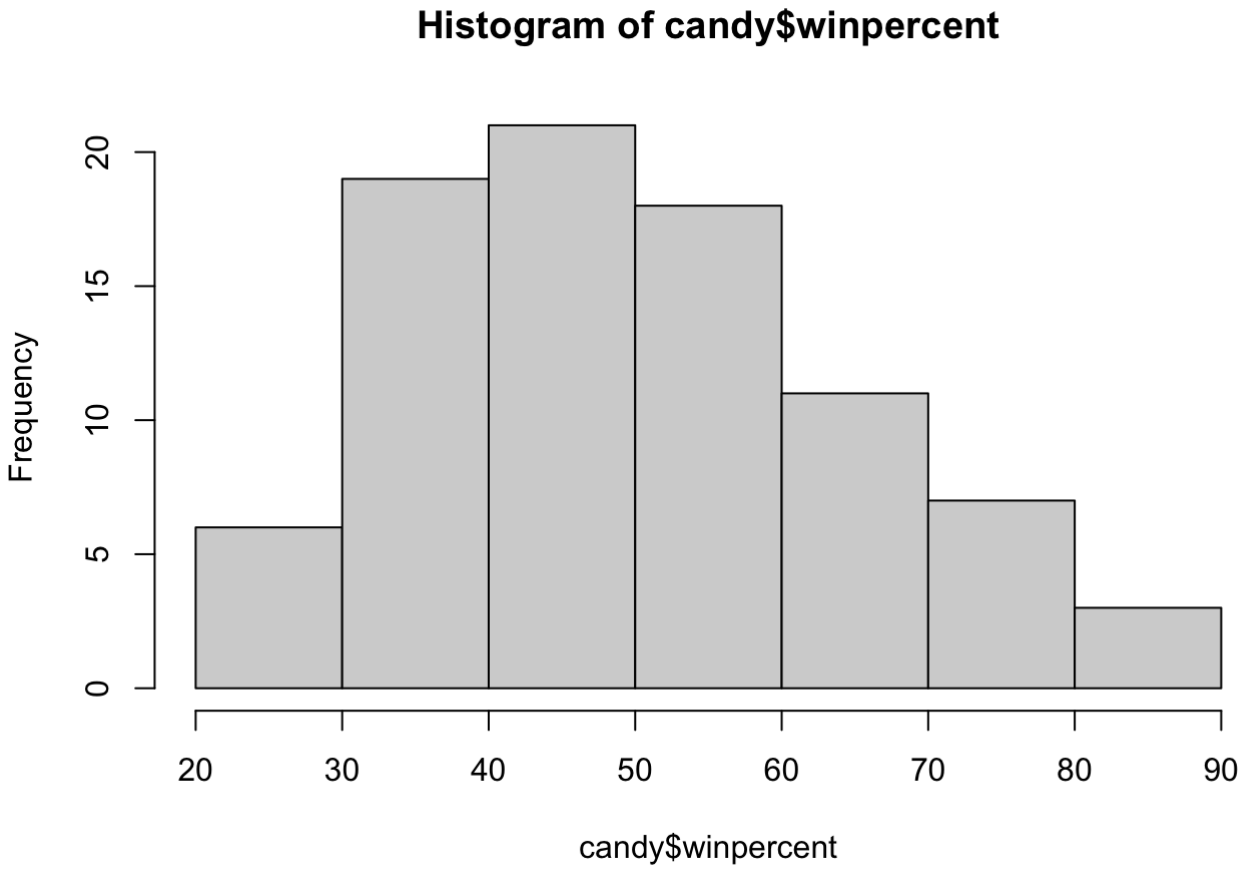
p100 because compared to all the other p value, that value has most of the values close to 1 while the other has most values with values that are 0.0.

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

If it has chocolate or not in the candy.

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical?

No, it seems like the values are more skewed to the left

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

The center of distribution seem to be below 50%.

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

```
choc_per <- candy$winpercent[as.logical(candy$chocolate)]
choc_per
```

```
##  [1] 66.97173 67.60294 50.34755 56.91455 38.97504 55.37545 62.28448 56.49050
##  [9] 59.23612 57.21925 76.76860 71.46505 66.57458 55.06407 73.09956 60.80070
## [17] 64.35334 47.82975 54.52645 70.73564 66.47068 69.48379 81.86626 84.18029
## [25] 73.43499 72.88790 65.71629 34.72200 37.88719 76.67378 59.52925 48.98265
## [33] 43.06890 45.73675 49.65350 81.64291 49.52411
```

```
mean(choc_per)
```

```
## [1] 60.92153
```

```
fruit_per <- candy$winpercent[as.logical(candy$fruity)]
fruit_per
```

```
##  [1] 52.34146 34.51768 36.01763 24.52499 42.27208 39.46056 43.08892 39.18550
##  [9] 46.78335 57.11974 51.41243 42.17877 28.12744 41.38956 39.14106 52.91139
## [17] 46.41172 55.35405 22.44534 39.44680 41.26551 37.34852 35.29076 42.84914
## [25] 63.08514 55.10370 45.99583 59.86400 52.82595 67.03763 34.57899 27.30386
## [33] 54.86111 48.98265 47.17323 45.46628 39.01190 44.37552
```

```
mean(fruit_per)
```

```
## [1] 44.11974
```

Chocolate on average is higher ranked than fruity candy.

Q12. Is this difference statistically significant?

```
t.test(choc_per,fruit_per)
```

```
##
##  Welch Two Sample t-test
##
## data:  choc_per and fruit_per
## 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
```

It is statistacally significant since the p value is 2.9e-08, which is much lower than 0.05, which is usually the accepted value of when data is statistcally significant.

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

```
head(candy[order(candy$winpercent),], n=15)
```

```
##                            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
## Sugar Daddy                        0      0       1               0      0
## One dime                           0      0       0               0      0
## Sugar Babies                       0      0       1               0      0
## Haribo Happy Cola                  0      0       0               0      0
## Caramel Apple Pops                 0      1       1               0      0
## Strawberry bon bons                0      1       0               0      0
## Sixlets                            1      0       0               0      0
## Ring pop                           0      1       0               0      0
## Chewey Lemonhead Fruit Mix         0      1       0               0      0
##                            crispedricewafer hard bar pluribus sugarpercent
## Nik L Nip                                 0    0   0        1        0.197
## Boston Baked Beans                        0    0   0        1        0.313
## Chiclets                                  0    0   0        1        0.046
## Super Bubble                              0    0   0        0        0.162
## Jawbusters                                0    1   0        1        0.093
## Root Beer Barrels                         0    1   0        1        0.732
## Sugar Daddy                               0    0   0        0        0.418
## One dime                                  0    0   0        0        0.011
## Sugar Babies                              0    0   0        1        0.965
## Haribo Happy Cola                         0    0   0        1        0.465
## Caramel Apple Pops                        0    0   0        0        0.604
## Strawberry bon bons                       0    1   0        1        0.569
## Sixlets                                   0    0   0        1        0.220
## Ring pop                                  0    1   0        0        0.732
## Chewey Lemonhead Fruit Mix                0    0   0        1        0.732
##                            pricepercent winpercent
## Nik L Nip                         0.976   22.44534
## Boston Baked Beans                0.511   23.41782
## Chiclets                          0.325   24.52499
## Super Bubble                      0.116   27.30386
## Jawbusters                        0.511   28.12744
## Root Beer Barrels                 0.069   29.70369
## Sugar Daddy                       0.325   32.23100
## One dime                          0.116   32.26109
## Sugar Babies                      0.767   33.43755
## Haribo Happy Cola                 0.465   34.15896
## Caramel Apple Pops                0.325   34.51768
## Strawberry bon bons               0.058   34.57899
## Sixlets                           0.081   34.72200
## Ring pop                          0.965   35.29076
## Chewey Lemonhead Fruit Mix        0.511   36.01763
```

The five least liked cany types in the set 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=15)
```

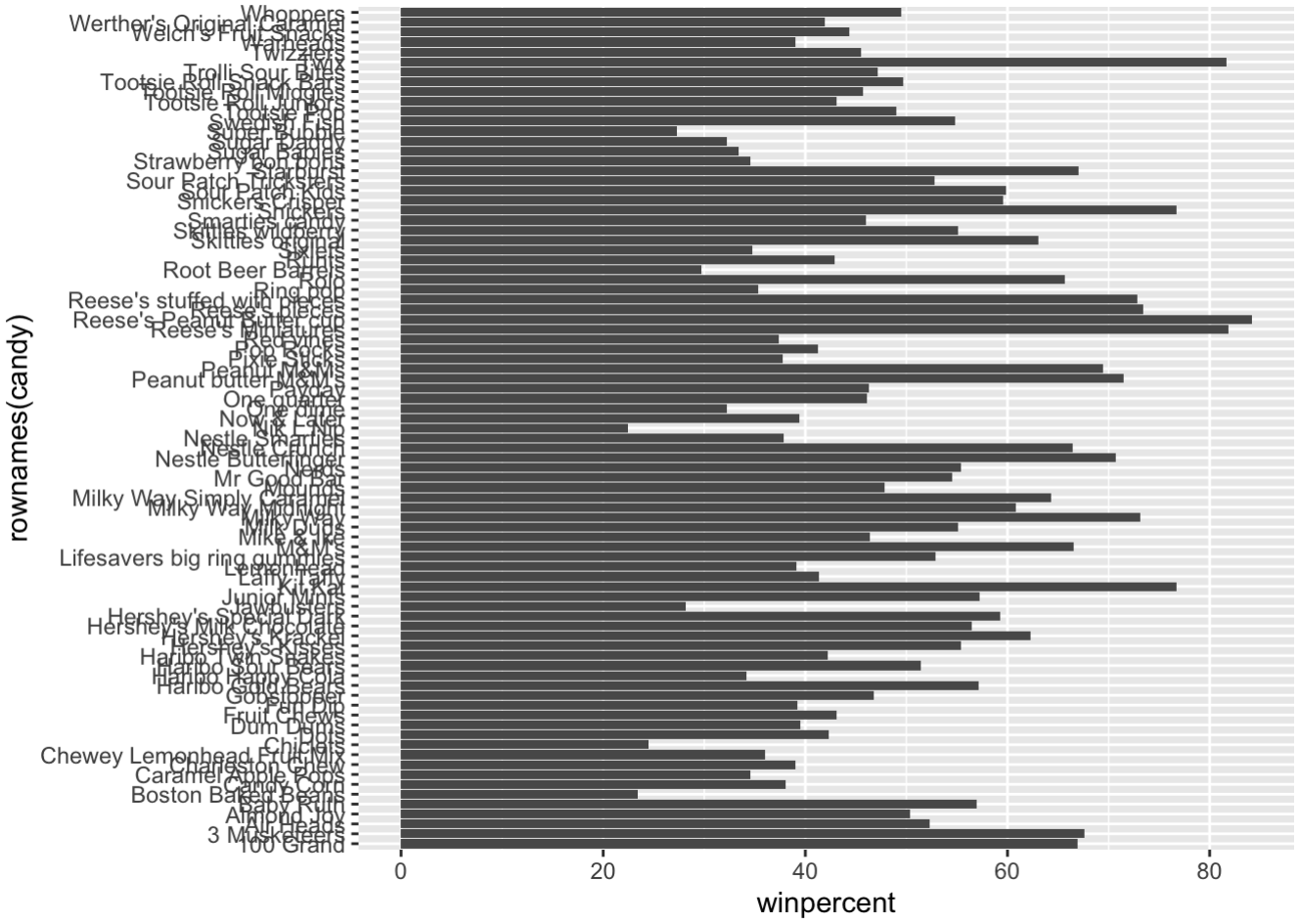
```
##                             chocolate fruity caramel peanutyalmondy nougat
## M&M's                               1      0       0               0      0
## 100 Grand                           1      0       1               0      0
## Starburst                           0      1       0               0      0
## 3 Musketeers                        1      0       0               0      1
## Peanut M&Ms                         1      0       0               1      0
## Nestle Butterfinger                 1      0       0               1      0
## Peanut butter M&M's                 1      0       0               1      0
## Reese's stuffed with pieces         1      0       0               1      0
## Milky Way                           1      0       1               0      1
## Reese's pieces                      1      0       0               1      0
## Snickers                            1      0       1               1      1
## Kit Kat                             1      0       0               0      0
## Twix                                1      0       1               0      0
## Reese's Miniatures                  1      0       0               1      0
## Reese's Peanut Butter cup           1      0       0               1      0
##                             crispedricewafer hard bar pluribus sugarpercent
## M&M's                                      0    0   0        1        0.825
## 100 Grand                                  1    0   1        0        0.732
## Starburst                                  0    0   0        1        0.151
## 3 Musketeers                               0    0   1        0        0.604
## Peanut M&Ms                                0    0   0        1        0.593
## Nestle Butterfinger                        0    0   1        0        0.604
## Peanut butter M&M's                        0    0   0        1        0.825
## Reese's stuffed with pieces                0    0   0        0        0.988
## Milky Way                                  0    0   1        0        0.604
## Reese's pieces                             0    0   0        1        0.406
## Snickers                                   0    0   1        0        0.546
## Kit Kat                                    1    0   1        0        0.313
## Twix                                       1    0   1        0        0.546
## Reese's Miniatures                         0    0   0        0        0.034
## Reese's Peanut Butter cup                  0    0   0        0        0.720
##                             pricepercent winpercent
## M&M's                              0.651   66.57458
## 100 Grand                          0.860   66.97173
## Starburst                          0.220   67.03763
## 3 Musketeers                       0.511   67.60294
## Peanut M&Ms                        0.651   69.48379
## Nestle Butterfinger                0.767   70.73564
## Peanut butter M&M's                0.651   71.46505
## Reese's stuffed with pieces        0.651   72.88790
## Milky Way                          0.651   73.09956
## Reese's pieces                     0.651   73.43499
## Snickers                           0.651   76.67378
## Kit Kat                            0.511   76.76860
## Twix                               0.906   81.64291
## Reese's Miniatures                 0.279   81.86626
## Reese's Peanut Butter cup          0.651   84.18029
```

The top 5 all time favorite candy types are Snickers, Kit Kat, Twix, Reese’s Minatures, and Reese’s Peanut Butter cup

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

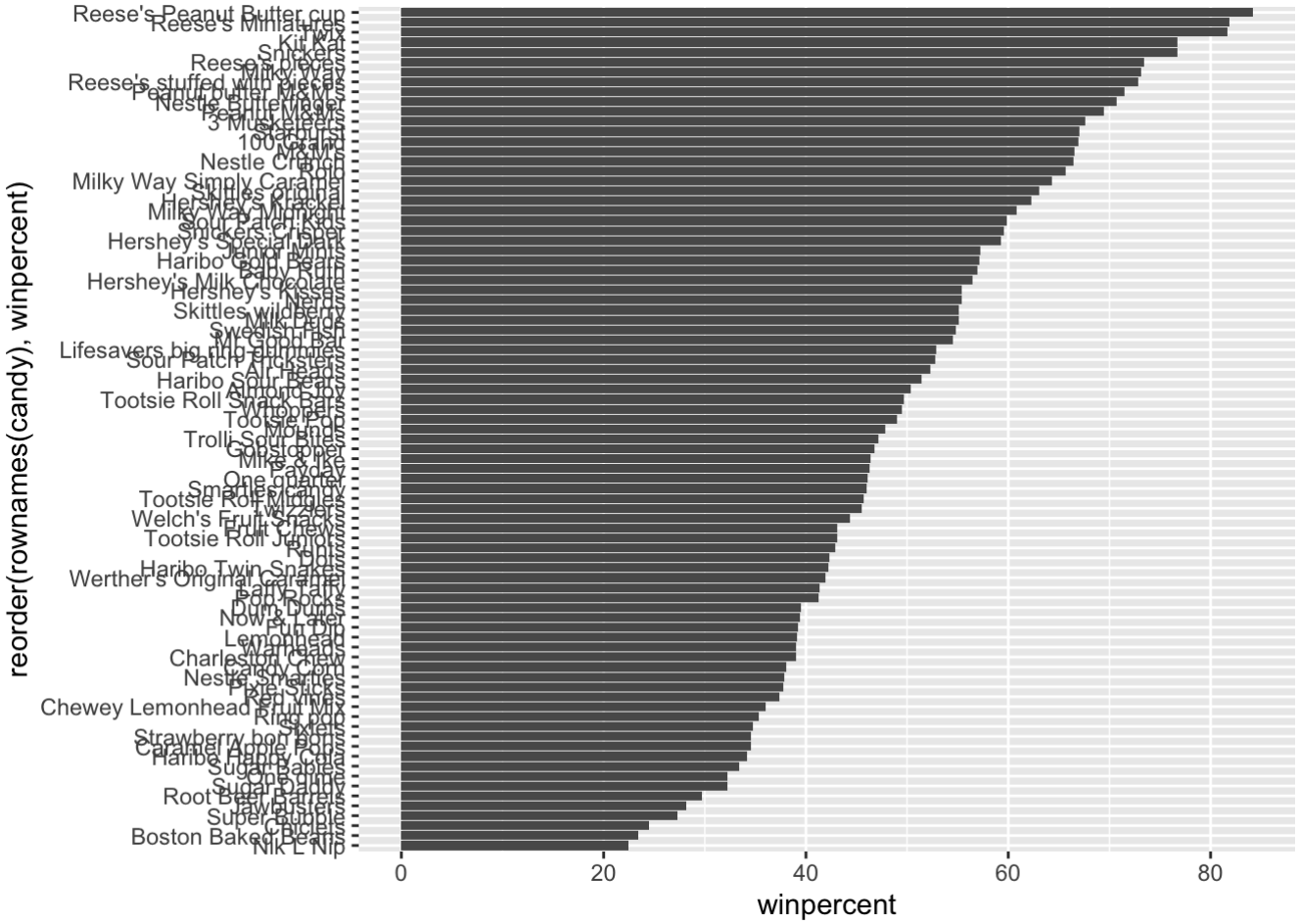
```
library(ggplot2)
```

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



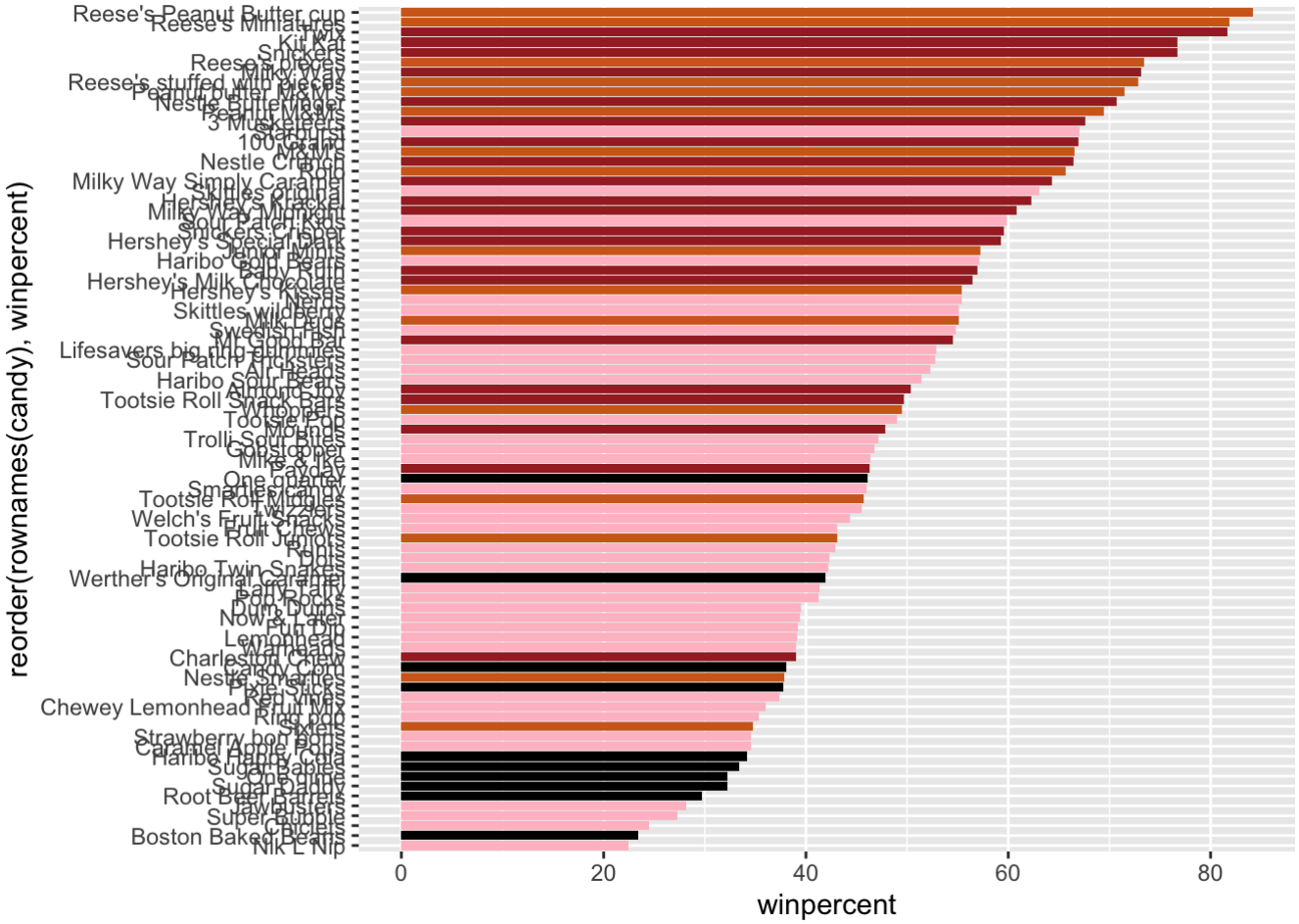
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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




```
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)
```



Q17. What is the worst ranked chocolate candy?

Sixlets

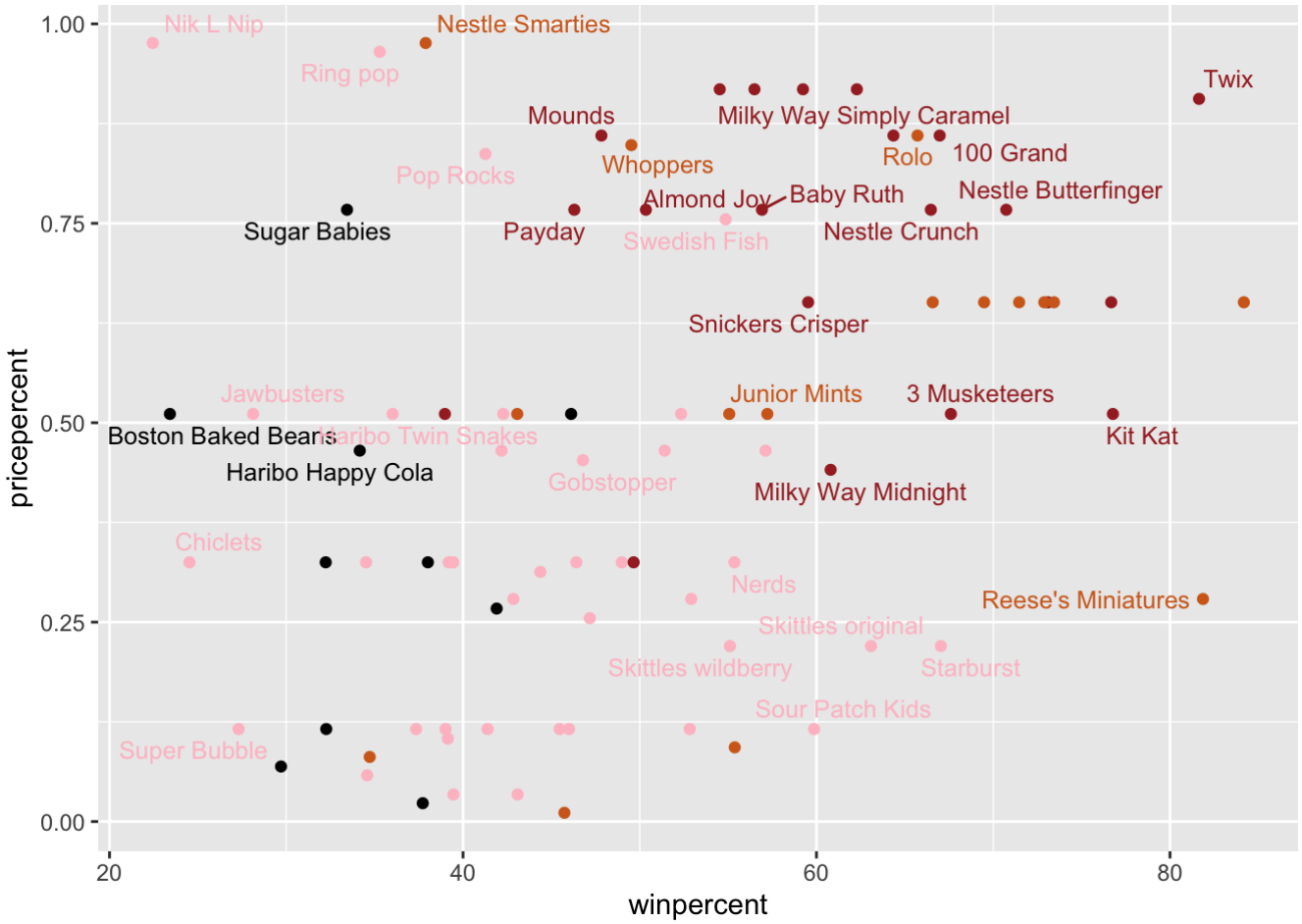
Q18. What is the best ranked fruity candy?

Starburst

```
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: 50 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?

Reeses Minature since it has over 80% winpercent with a little over 25% pricepercent.

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

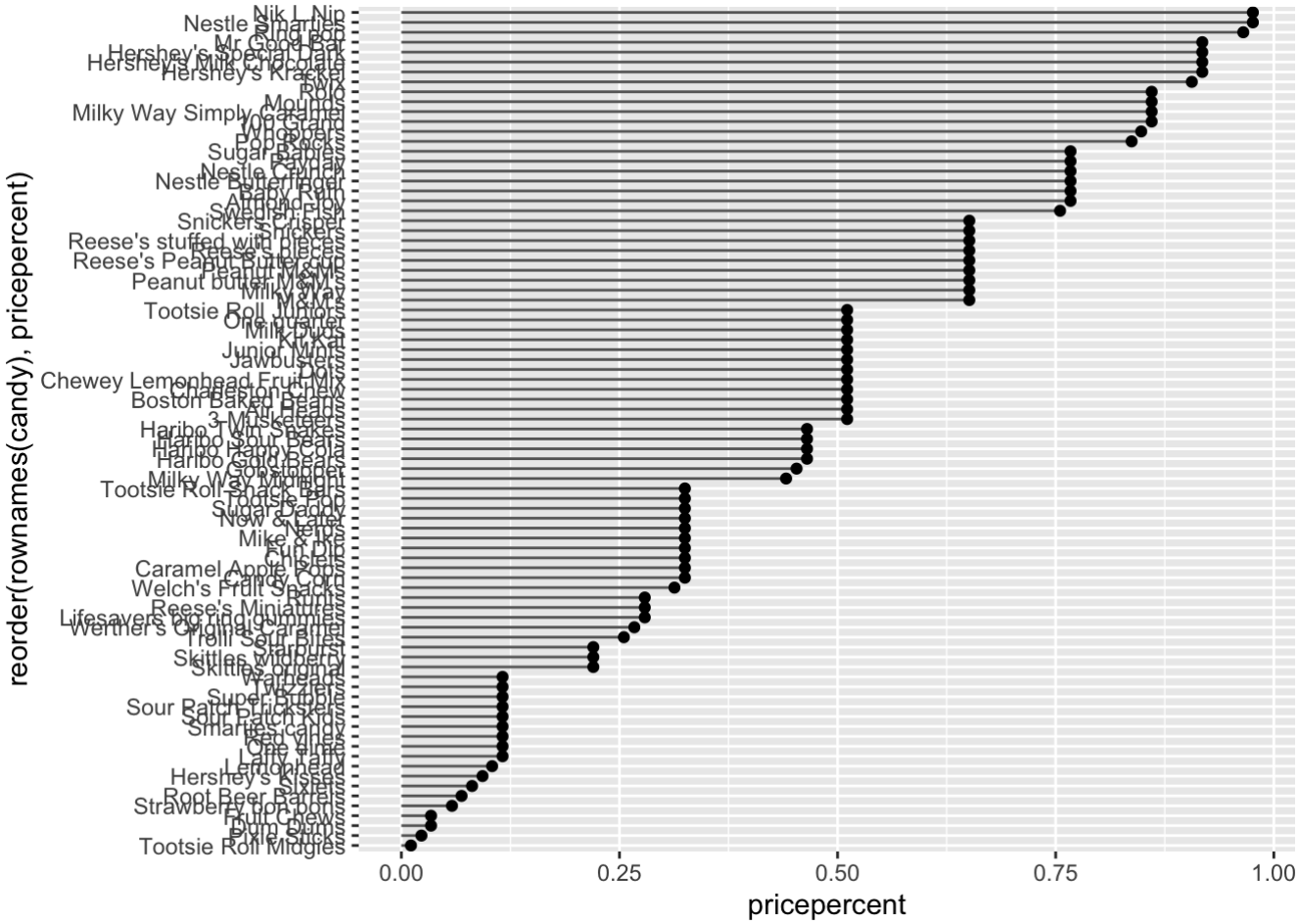
```
ord <- order(candy$pricepercent, decreasing = TRUE)
head( candy[ord,c(11,12)], n=5 )
```

```
##                           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
```

The top 5 most expensive candies are Nik L Lip, Ring pop, Nestl Smarties, Hershey Krackel, and Hersheys Milk Chocolate. The most unpopular candy of these are 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().

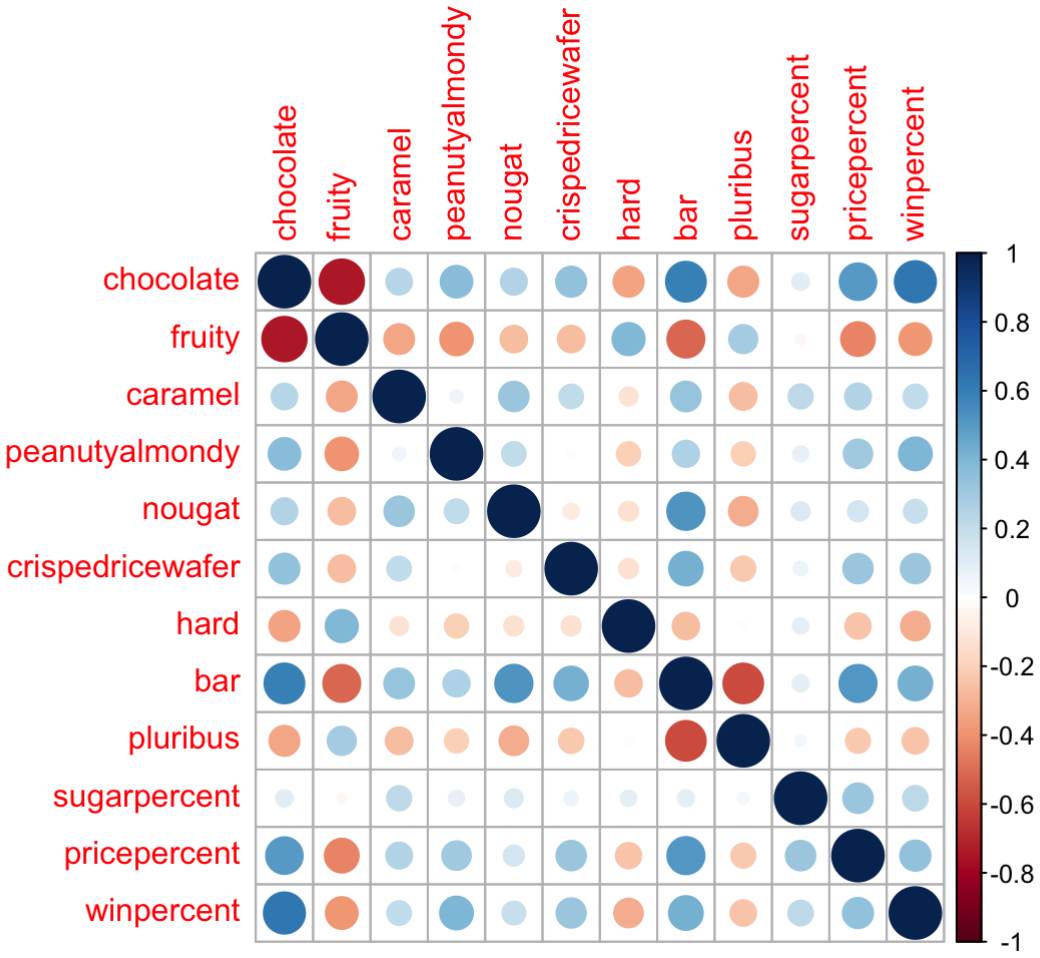
```
# Make a lollipop chart of pricepercent
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```




```
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 and Chocolate

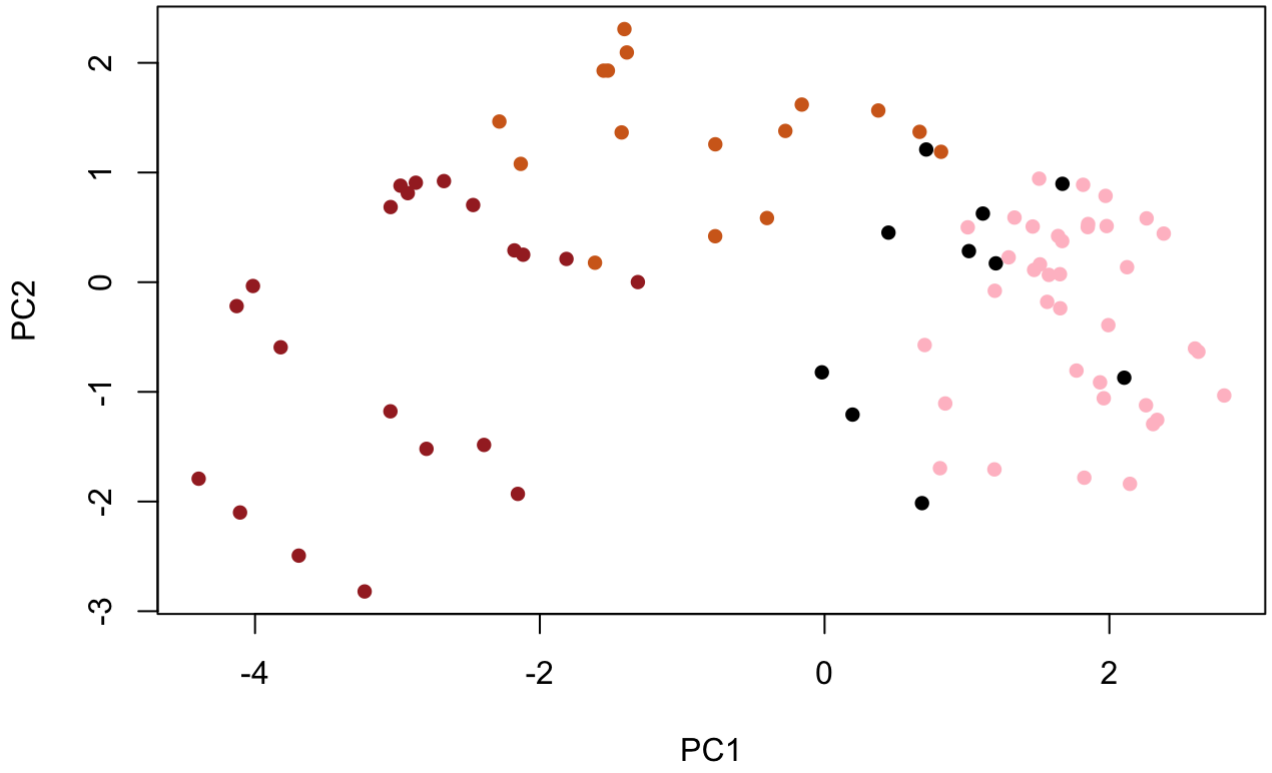
Q23. Similarly, what two variables are most positively correlated?

Chocolate and bar or chocolate and winpercent.

```
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
```

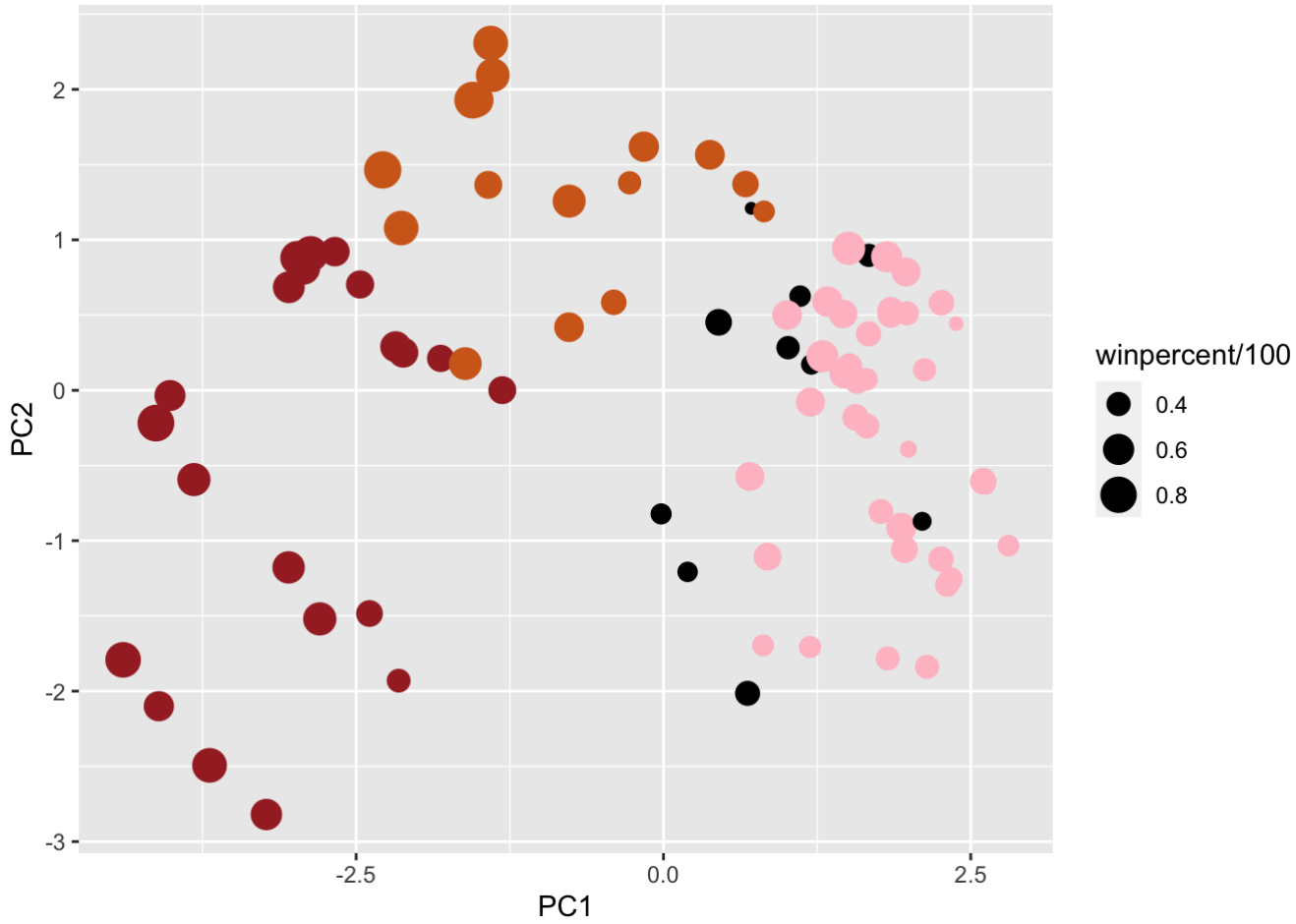
```
plot(pca$x[,1:2],col=my_cols,pch=16)
```




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

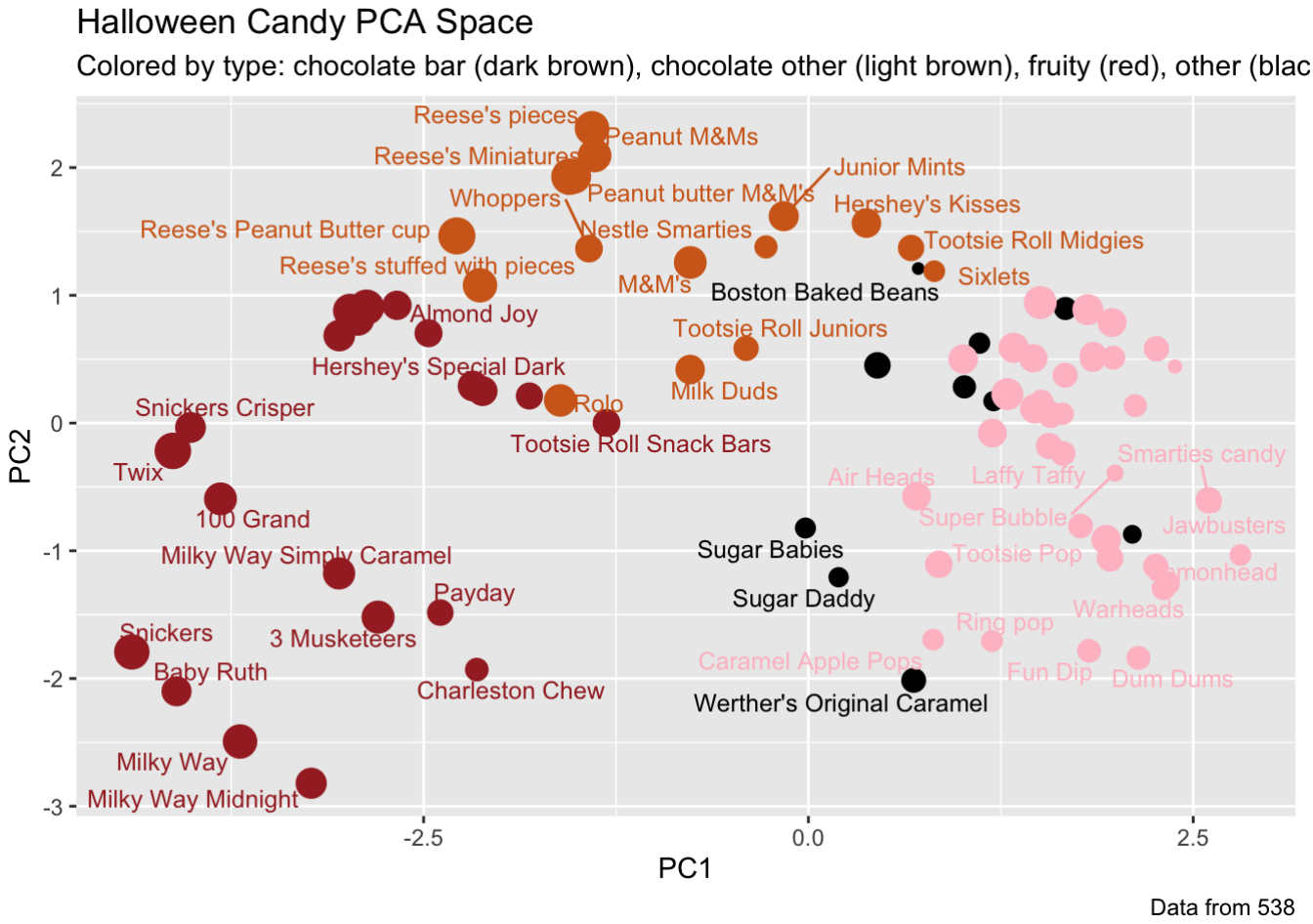
```
p <- ggplot(my_data) +
        aes(x=PC1, y=PC2,
            size=winpercent/100,
            text=rownames(my_data),
            label=rownames(my_data)) +
        geom_point(col=my_cols)

p
```




```
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 bro
wn), fruity (red), other (black)",
       caption="Data from 538")
```

```
## Warning: ggrepel: 39 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```




```
library(plotly)
```

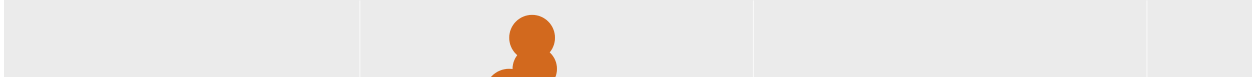
```
##
## Attaching package: 'plotly'
```

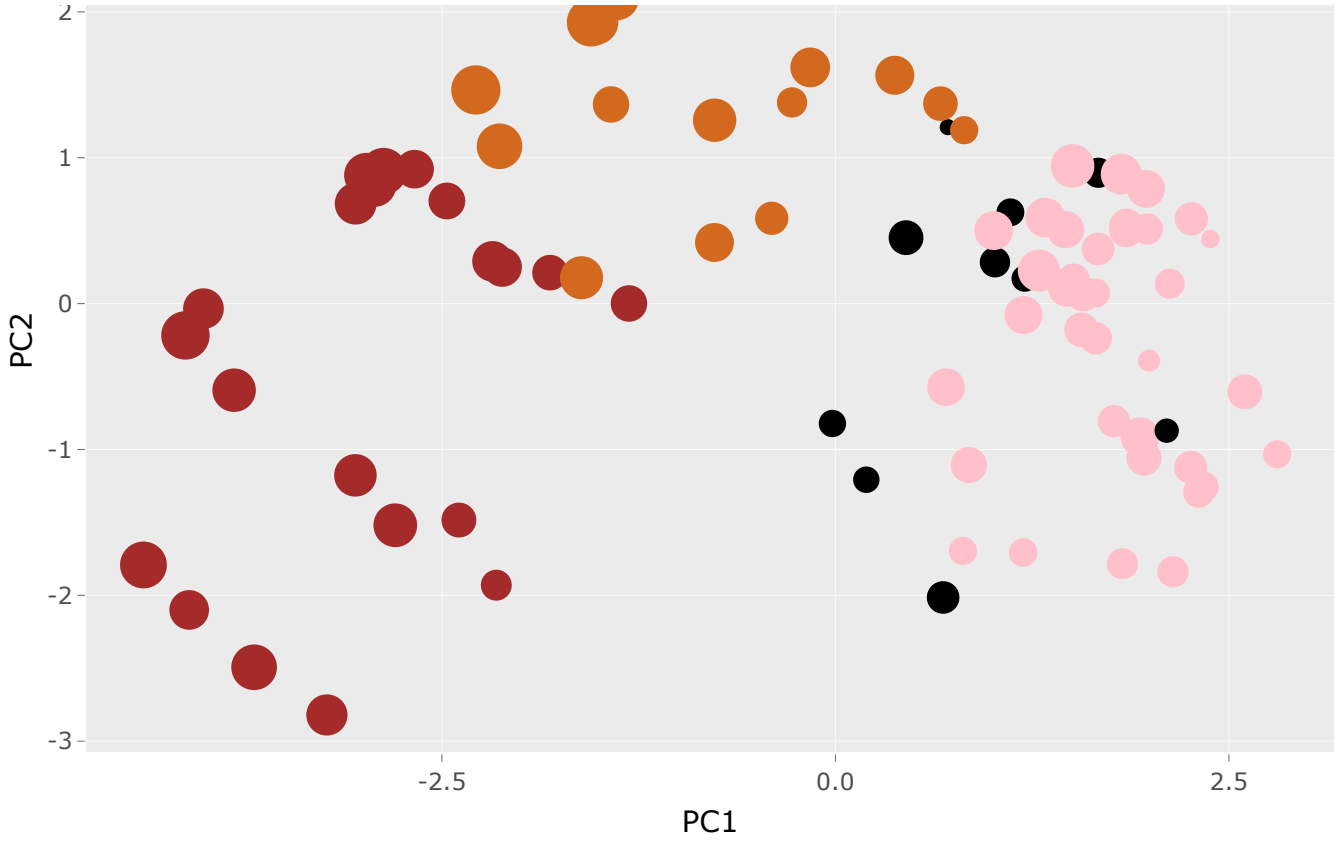
```
## The following object is masked from 'package:ggplot2':
##
##     last_plot
```

```
## The following object is masked from 'package:stats':
##
##     filter
```

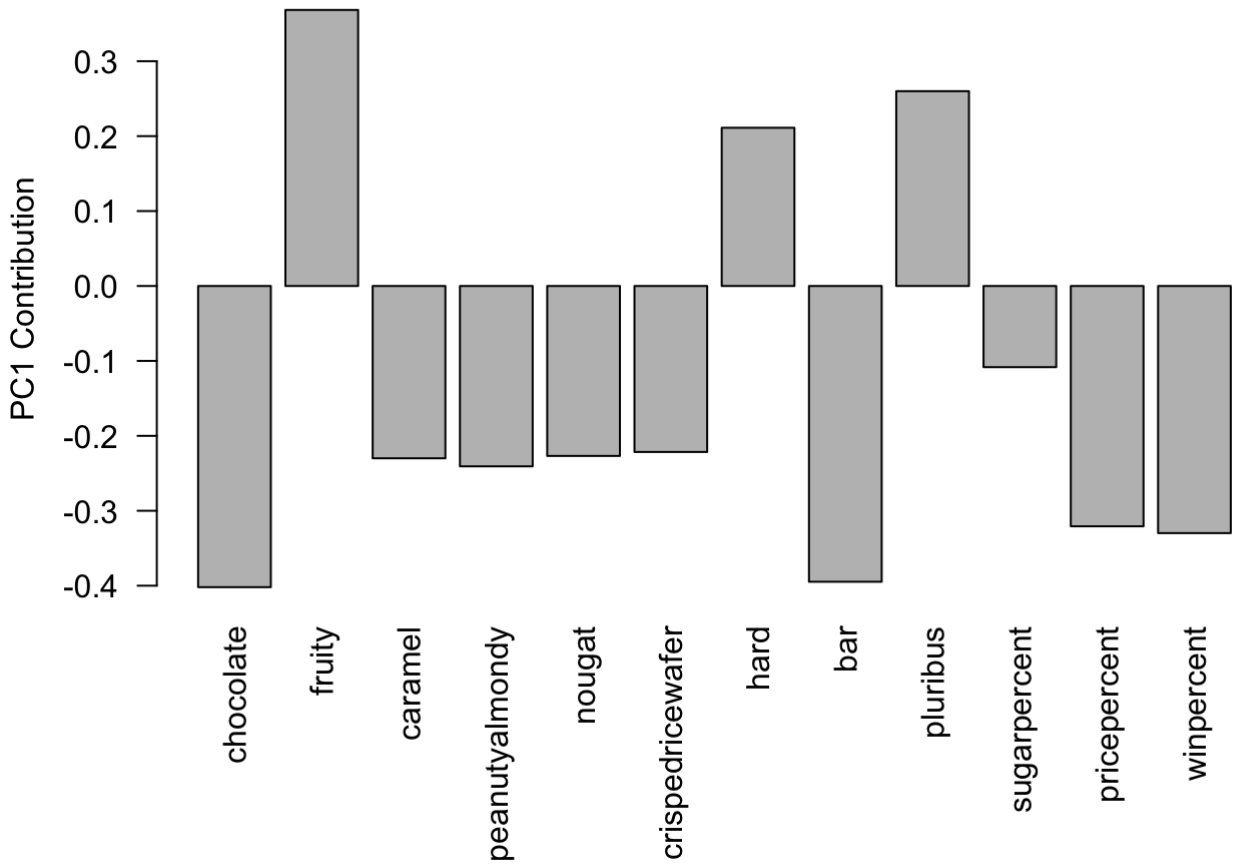
```
## The following object is masked from 'package:graphics':
##
##     layout
```

```
ggplotly(p)
```






```
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?

The original variables that are picked up strongly by PC1 in positive direction are fruity, hard, and pluribus.