Class08

Alexander Liu (PID: 69026918)

```
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?

```
nrow(candy)
```

```
[1] 85
```

85 types.

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

```
sum(candy$fruity == 1)
```

```
[1] 38
```

38 types.

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

```
candy["100 Grand",]$winpercent
```

```
[1] 66.97173
```

My favorite is 100 Grand. Its `winpercent` is 66.97173

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

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

```
[1] 76.7686
```

Kit Kat's winpercent is 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's winpercent is 49.6535.

```
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?

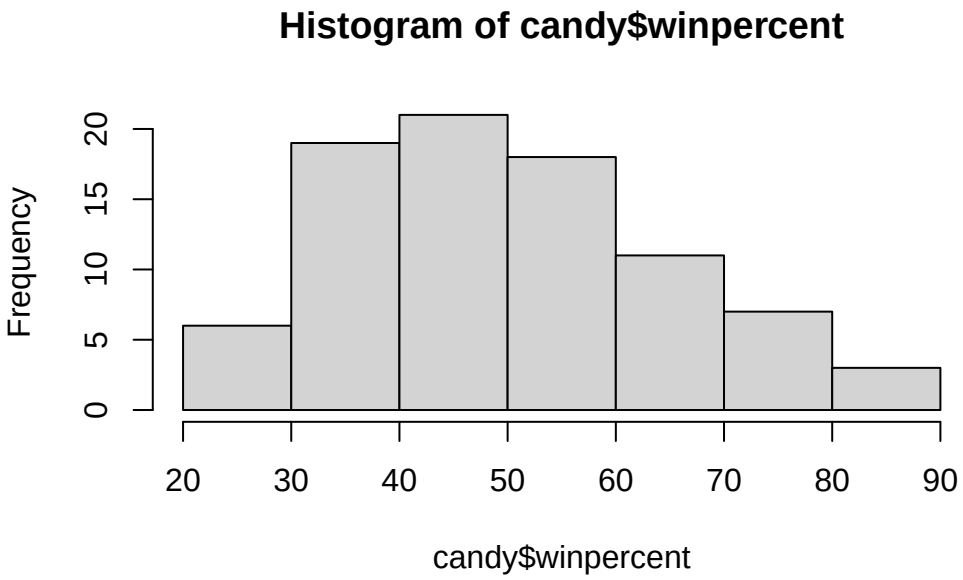
winpercent.

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

"1" indicates that the variable is chocolate, whereas "0" indicates not.

Q8. Plot a histogram of winpercent values

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



Q9. Is the distribution of winpercent values symmetrical?

According to the histogram, not.

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

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

```
[1] 47.82975
```

Below 50%.

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

```
chocolate <- candy[candy$chocolate == 1, "winpercent"]
fruity <- candy[candy$fruity == 1, "winpercent"]
mean(chocolate)
```

```
[1] 60.92153
```

```
mean(fruity)
```

```
[1] 44.11974
```

Higher.

Q12. Is this difference statistically significant?

```
  t.test(chocolate, fruity)
```

```
    Welch Two Sample t-test

data:  chocolate and 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
```

Yes.

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
```

```
  row.names.data.frame(head(candy[order(candy$winpercent),], n=5))
```

```
[1] "Nik L Nip"          "Boston Baked Beans" "Chiclets"
[4] "Super Bubble"       "Jawbusters"
```

```
row.names.data.frame(candy %>% arrange(winpercent) %>% head(5))
```

```
[1] "Nik L Nip"           "Boston Baked Beans" "Chiclets"
[4] "Super Bubble"        "Jawbusters"
```

"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?

```
row.names.data.frame(tail(candy[order(candy$winpercent),], n=5))
```

```
[1] "Snickers"                  "Kit Kat"
[3] "Twix"                      "Reese's Miniatures"
[5] "Reese's Peanut Butter cup"
```

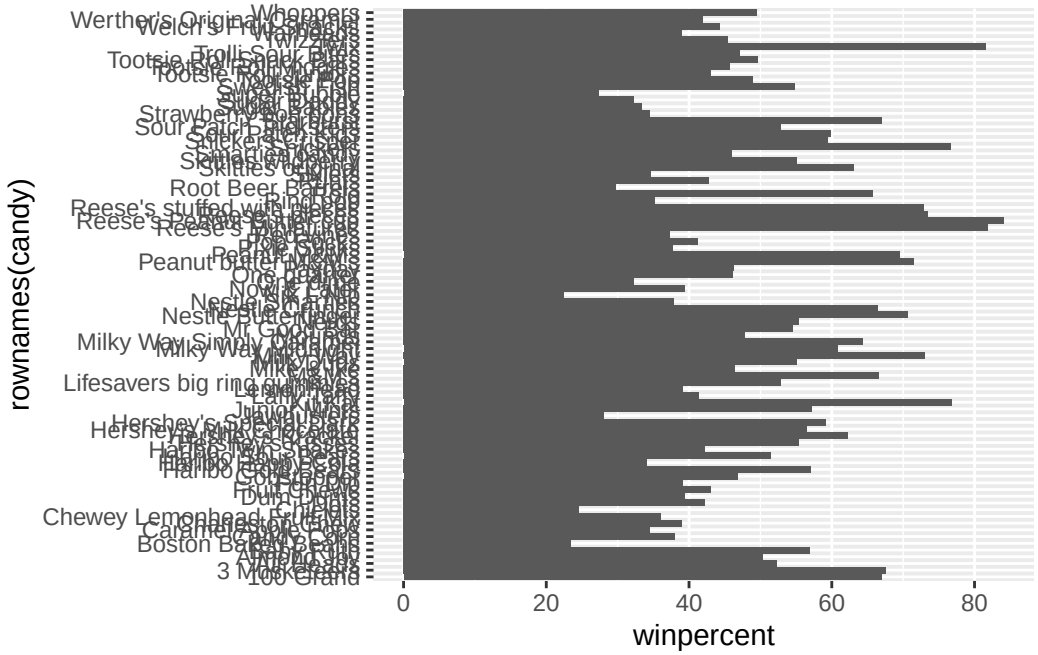
```
row.names.data.frame(candy %>% arrange(winpercent) %>% tail(5))
```

```
[1] "Snickers"                  "Kit Kat"
[3] "Twix"                      "Reese's Miniatures"
[5] "Reese's Peanut Butter cup"
```

"Snickers", "Kit Kat", "Twix", "Reese's Miniatures", and "Reese's Peanut Butter cup".

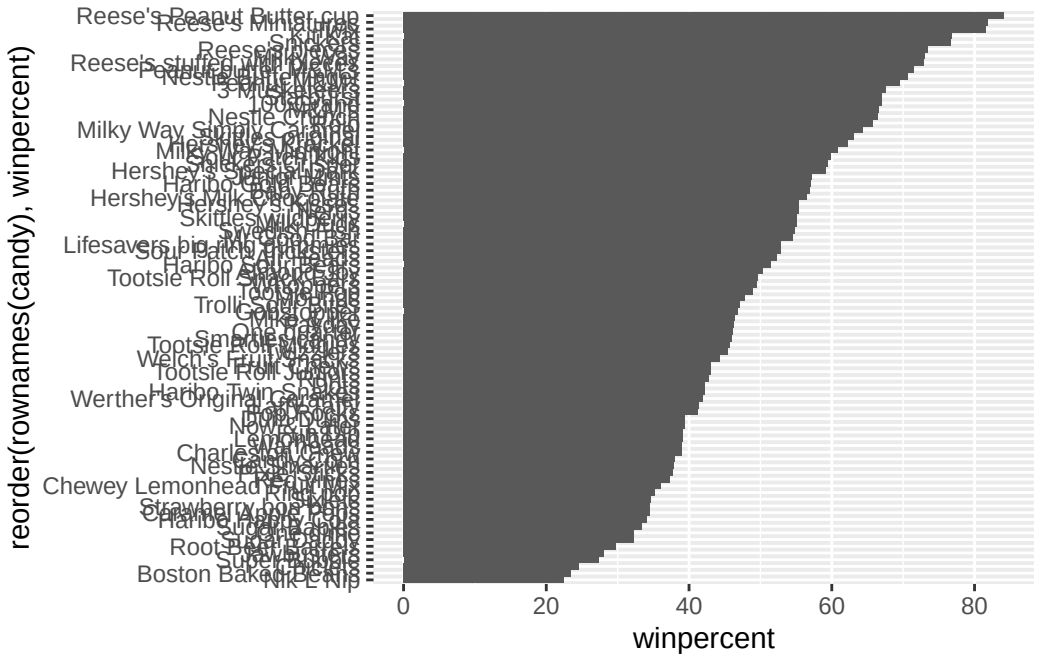
Q15. Make a first barplot of candy ranking based on winpercent values.

```
library(ggplot2)
ggplot(candy) +
  aes(winpercent, rownames(candy)) +
  geom_bar(stat = "identity")
```



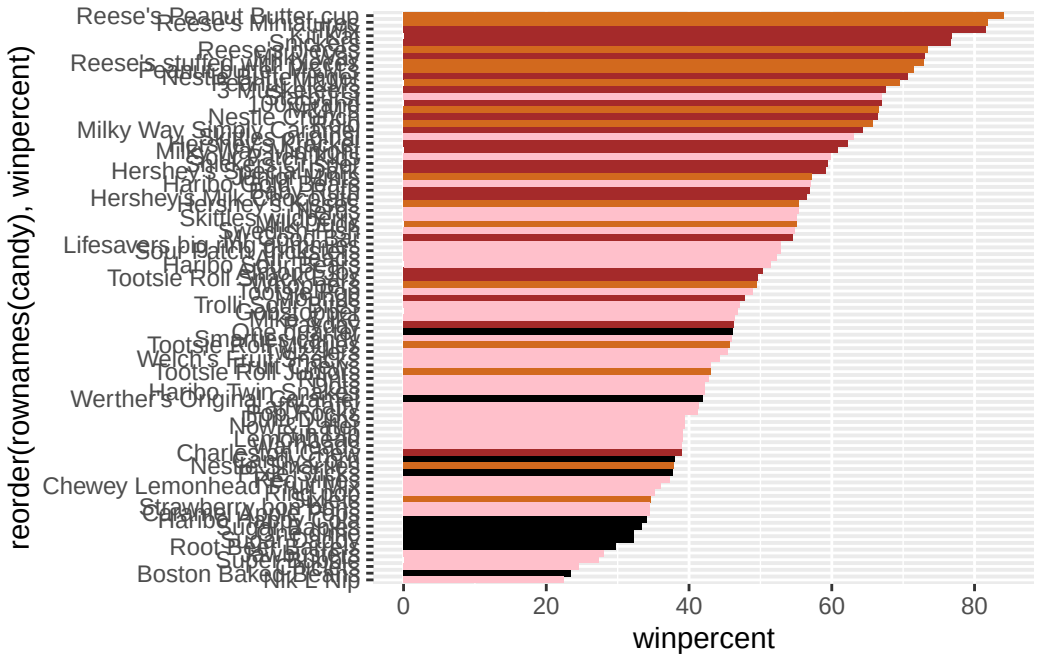
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

```
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)
```



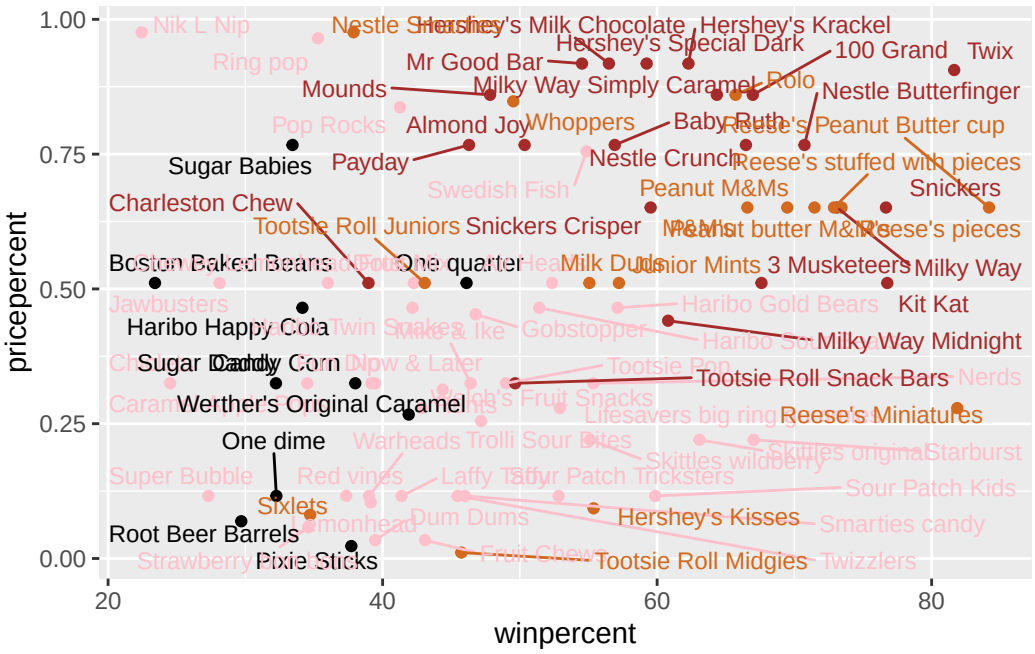
- Q17. What is the worst ranked chocolate candy?

"Nik L Nip"

- Q18. What is the best ranked fruity candy?

"Reese's Peanut Butter cup"

```
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 = 50)
```



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?

I calculated the cost efficiency (winpercent/pricepercent)

```
candy$bang <- candy$winpercent/candy$pricepercent
ord <- order(candy$bang, decreasing = TRUE)
head( candy[ord,c(11,12,13)], n=5 )
```

```
                    pricepercent winpercent      bang
Tootsie Roll Midgies        0.011   45.73675 4157.8862
Pixie Sticks                0.023   37.72234 1640.1016
Fruit Chews                 0.034   43.08892 1267.3212
Dum Dums                    0.034   39.46056 1160.6045
Strawberry bon bons         0.058   34.57899  596.1895
```

"Tootsie Roll Midgies".

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

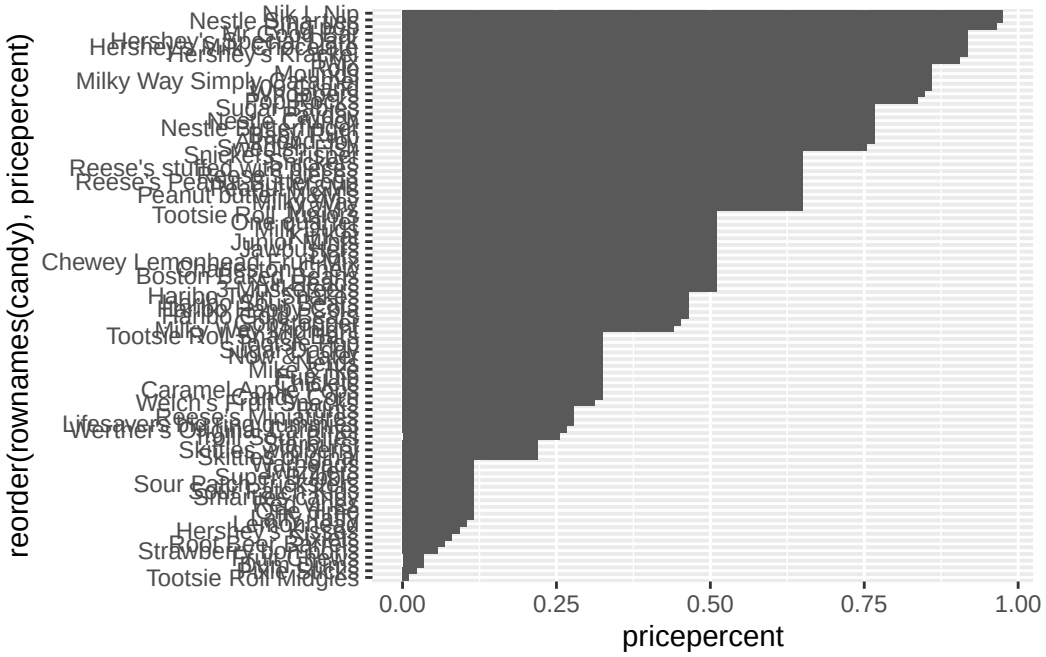
```
ord <- order(candy$pricepercent, decreasing = TRUE)
tail( candy[ord,c(11,12,13)], n=5 )
```

```
                     pricepercent winpercent      bang
Strawberry bon bons         0.058   34.57899  596.1895
Dum Dums                    0.034   39.46056 1160.6045
Fruit Chews                 0.034   43.08892 1267.3212
Pixie Sticks                0.023   37.72234 1640.1016
Tootsie Roll Midgies        0.011   45.73675 4157.8862
```

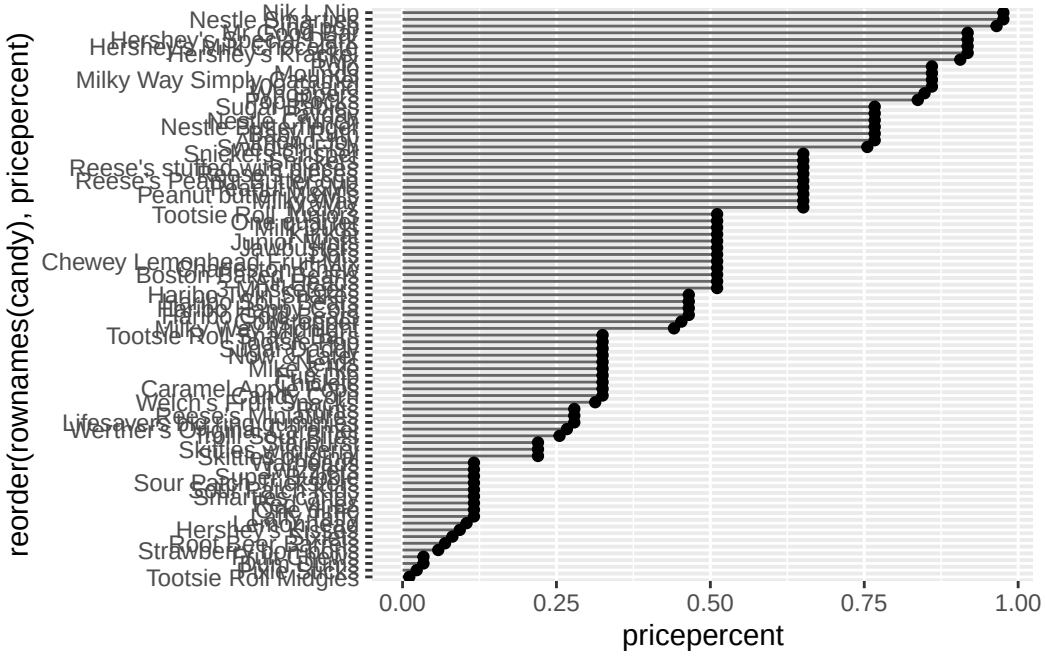
"Nik L Nip".

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_col()
```



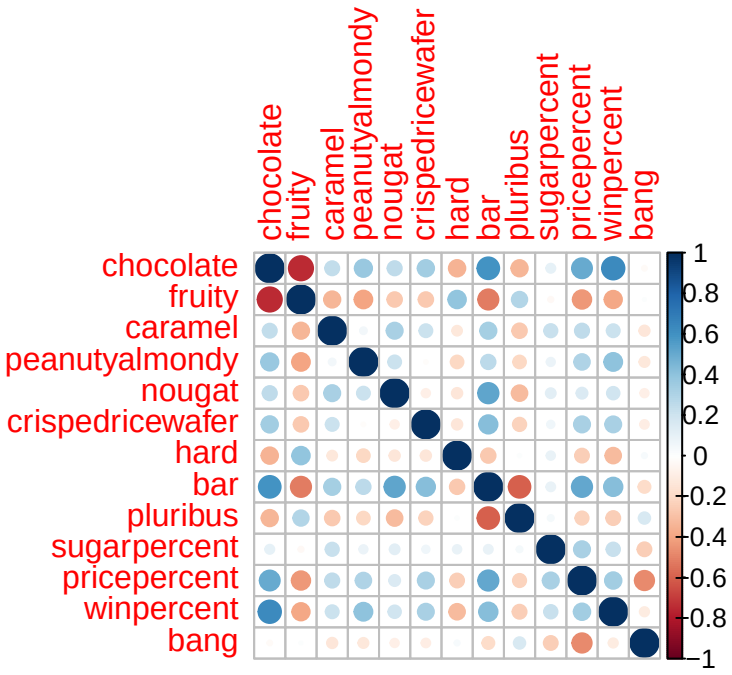

```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```




```
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)?

chocolate vs fruity, and pluribus vs bar

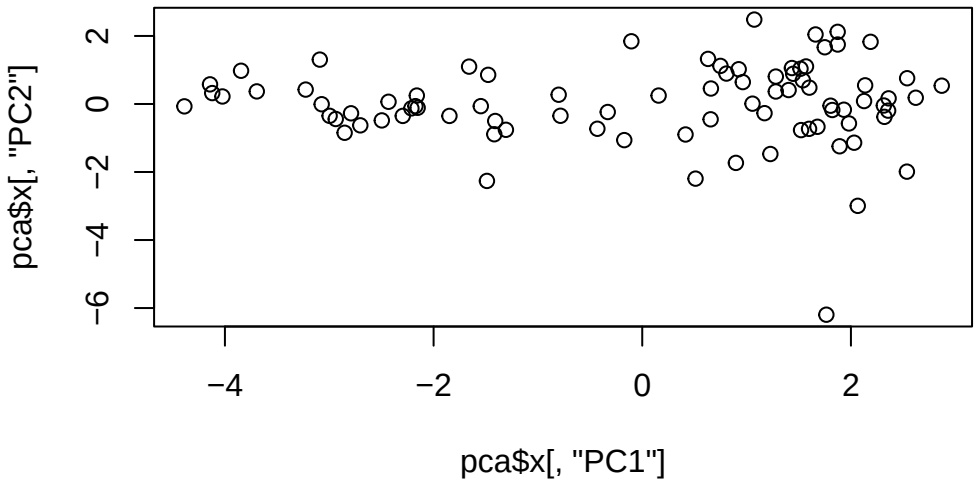
Q23. Similarly, what two variables are most positively correlated?

chocolate vs bar, and chocolate vs winpercent

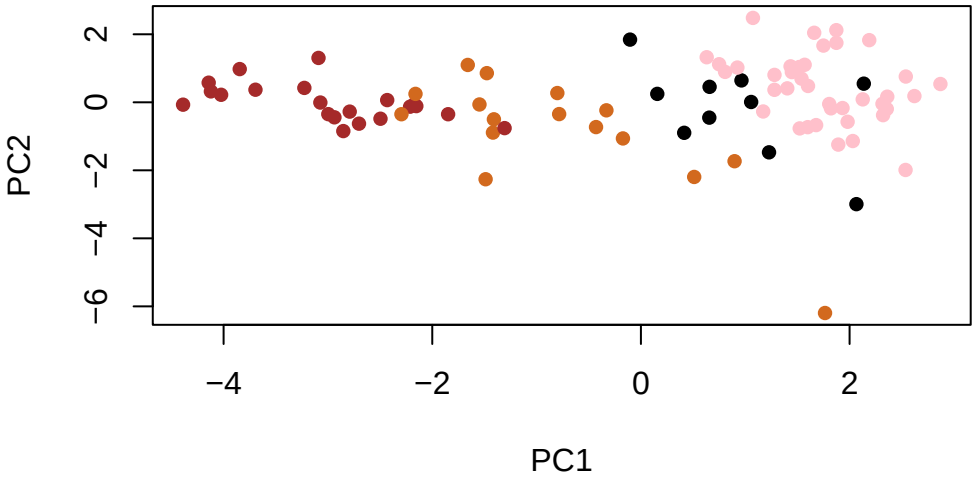
```
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
Importance of components:
                          PC1    PC2     PC3    PC4     PC5     PC6     PC7
Standard deviation     2.0938 1.2127 1.13054 1.0787 0.98027 0.93656 0.81530
Proportion of Variance 0.3372 0.1131 0.09832 0.0895 0.07392 0.06747 0.05113
Cumulative Proportion  0.3372 0.4503 0.54866 0.6382 0.71208 0.77956 0.83069
                           PC8     PC9    PC10    PC11    PC12    PC13
Standard deviation     0.78462 0.68466 0.66328 0.57829 0.43128 0.39534
Proportion of Variance 0.04736 0.03606 0.03384 0.02572 0.01431 0.01202
Cumulative Proportion  0.87804 0.91410 0.94794 0.97367 0.98798 1.00000
```

```
plot(pca$x[,"PC1"], pca$x[,"PC2"])
```



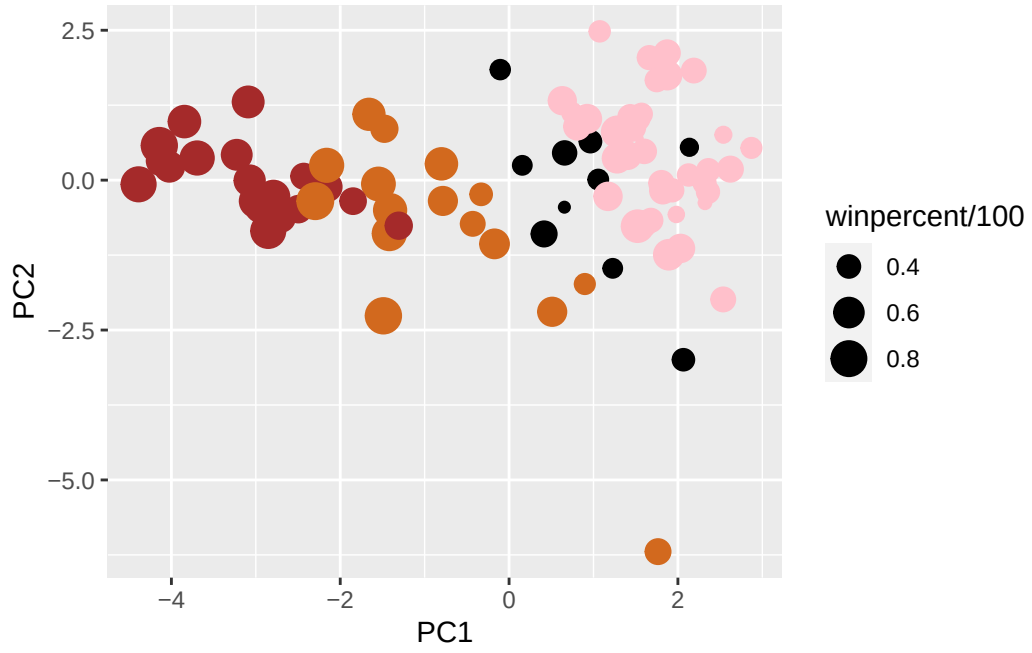

```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



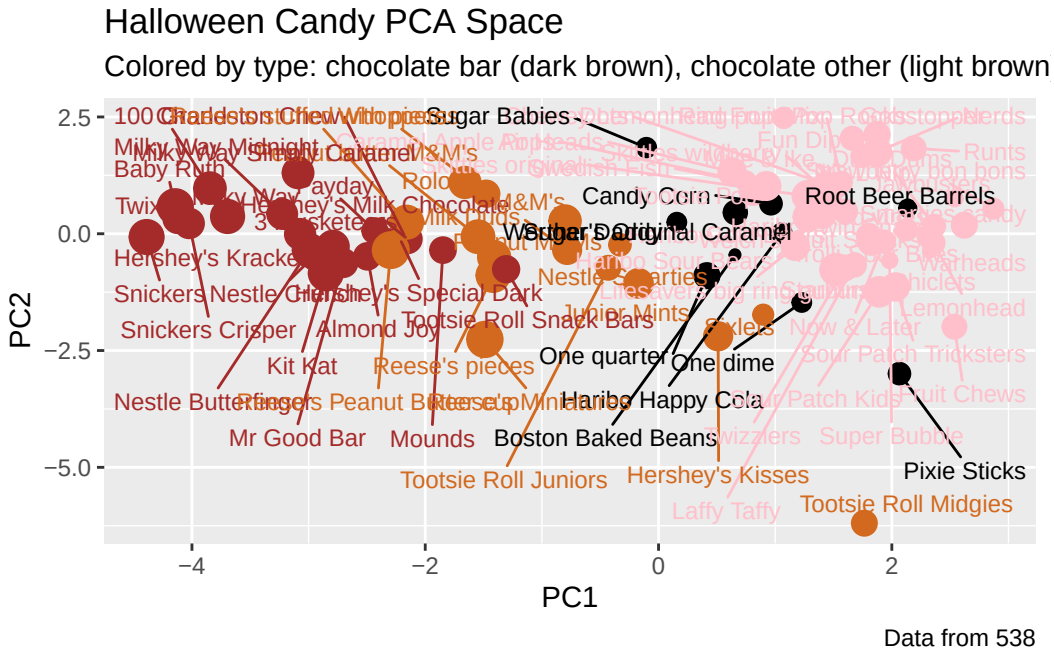

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

```
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 = 50)  +
  theme(legend.position = "none") +
  labs(title="Halloween Candy PCA Space",
       subtitle="Colored by type: chocolate bar (dark brown), chocolate other (light brown
       caption="Data from 538")
```



Data from 538

```
library(plotly)
```

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

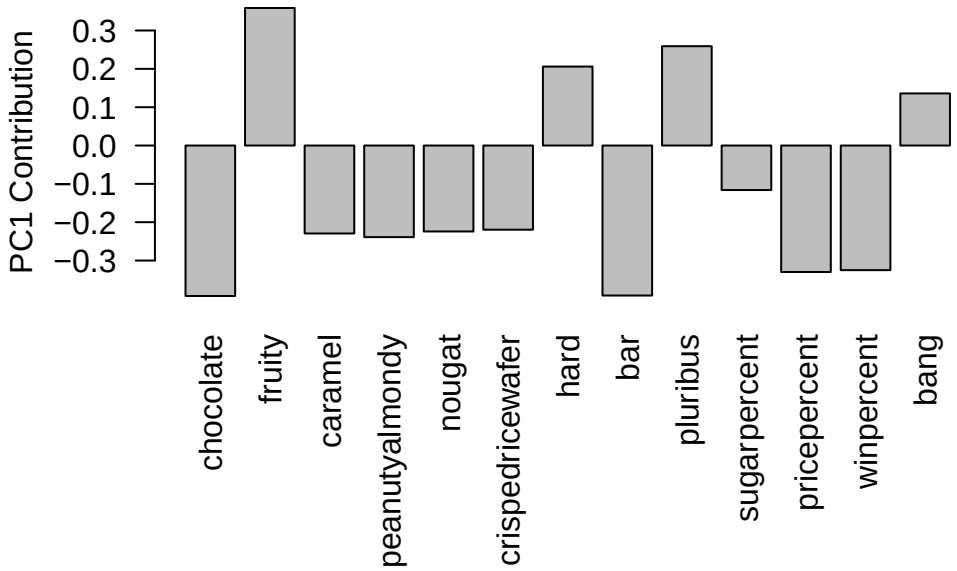
```
The following object is masked from 'package:ggplot2':

    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?

fruity, hard, and pluribus. Many candies which have these characteristics can be found at the positive side of the PC1-PC2 plot.