Class08_Halloween Candy Mini-Project

Changcheng Li (PID: A69027828)

```
candy_file <- "https://raw.githubusercontent.com/fivethirtyeight/data/master/candy-power-r
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 candy types.

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

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

```
[1] 38
```

There are 38 fruity candy types.

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

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

```
[1] 39.46056
```

Dum Dums, 39.46%

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

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

```
[1] 76.7686
```

76.77%

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

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

```
[1] 49.6535
```

49.65%

```
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?

The 'winpercent' column seems to be on a different scale (0~100) while others are probably 0~1

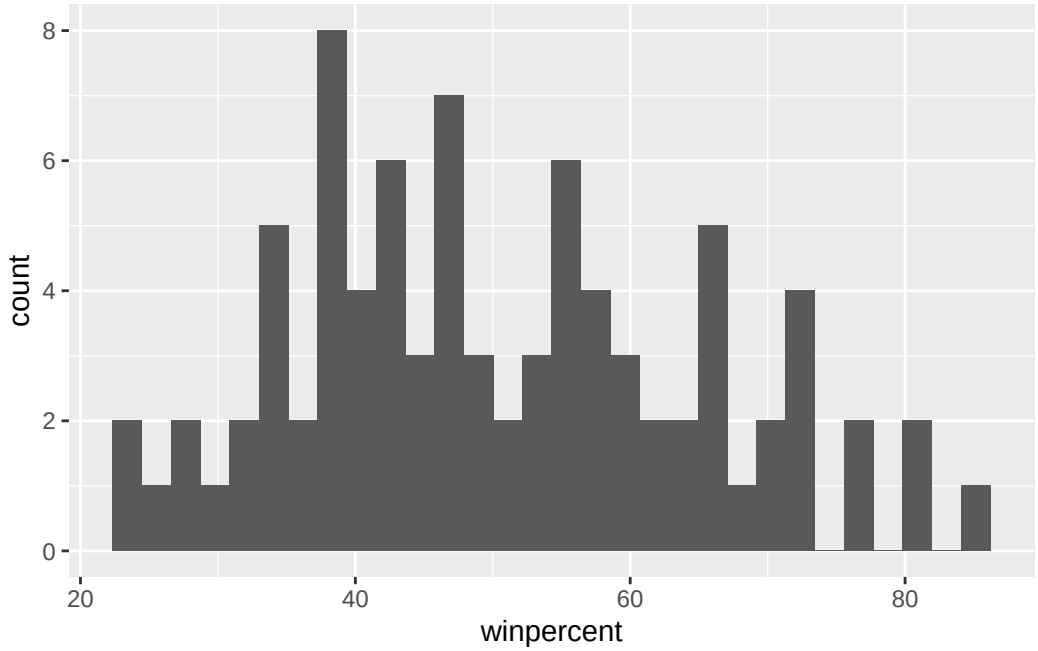
Q7. What do you think a zero and one represent for the candy$chocolate column?

It means 'chocolate' variable is binary: One candy type can be either chocolate or non-chocolate and there are no intermediates. '1' represent the candy is chocolate while '0' means it is not.

Q8. Plot a histogram of winpercent values

```
library(ggplot2)
ggplot(candy, aes(x=winpercent)) +
  geom_histogram()
```

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



Q9. Is the distribution of winpercent values symmetrical?

It seems that this distribution is not symmetrical: There is an enrichment on the lower side.

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

It is below 50%.

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

```
mean(candy$winpercent[as.logical(candy$chocolate)])
```

```
[1] 60.92153
```

```
mean(candy$winpercent[as.logical(candy$fruity)])
```

```
[1] 44.11974
```

Obviously chocolate candy is higher ranked than fruity candy.

Q12. Is this difference statistically significant?

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

```
    Welch Two Sample t-test

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

p-value = 2.871e-08 « 0.05, so this difference is significant.

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
```

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

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

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

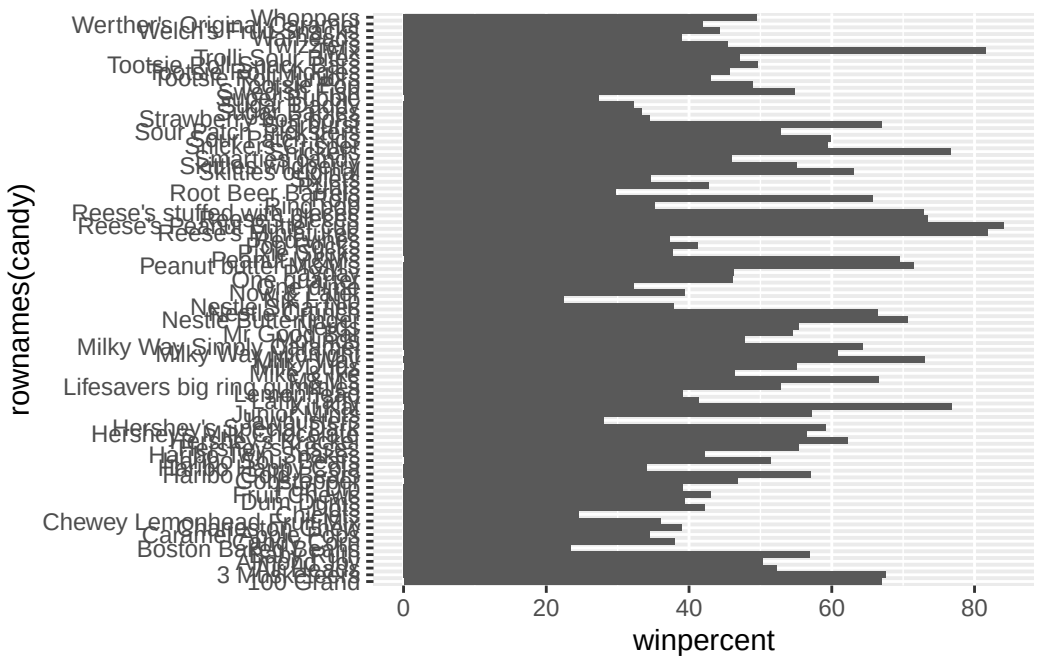
```
                        chocolate fruity caramel peanutyalmondy nougat
Reese's Peanut Butter cup         1      0       0              1      0
Reese's Miniatures                1      0       0              1      0
Twix                              1      0       1              0      0
Kit Kat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

Reese's Peanut Butter cup, Reese's Miniatures, Twix, Kit Kat and Snickers are the top5 favorite.

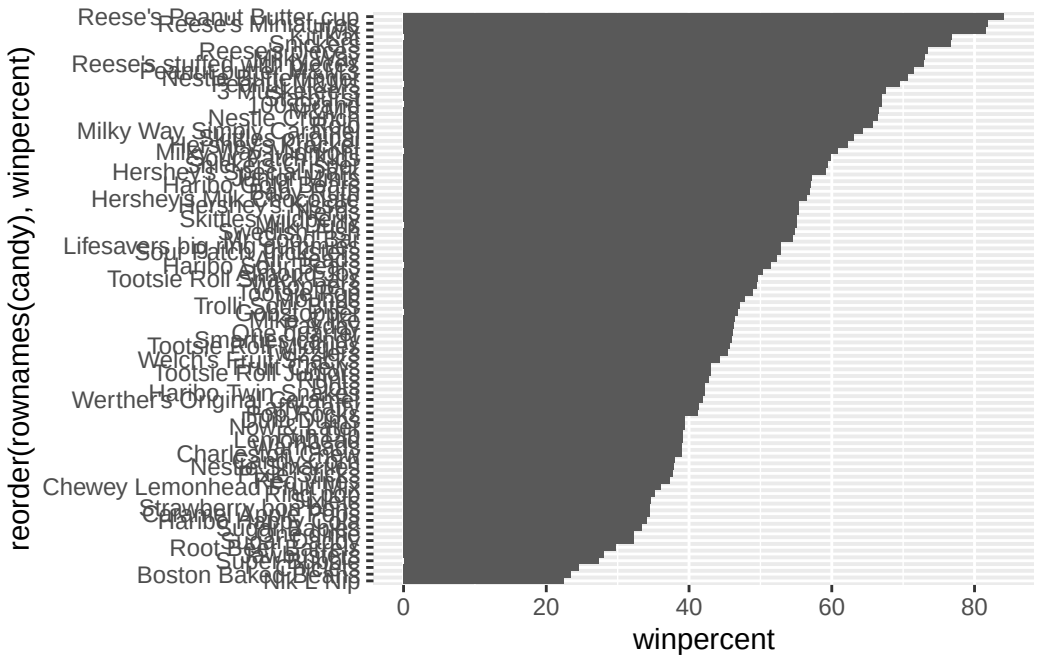
Q15. Make a first barplot of candy ranking based on winpercent values.

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



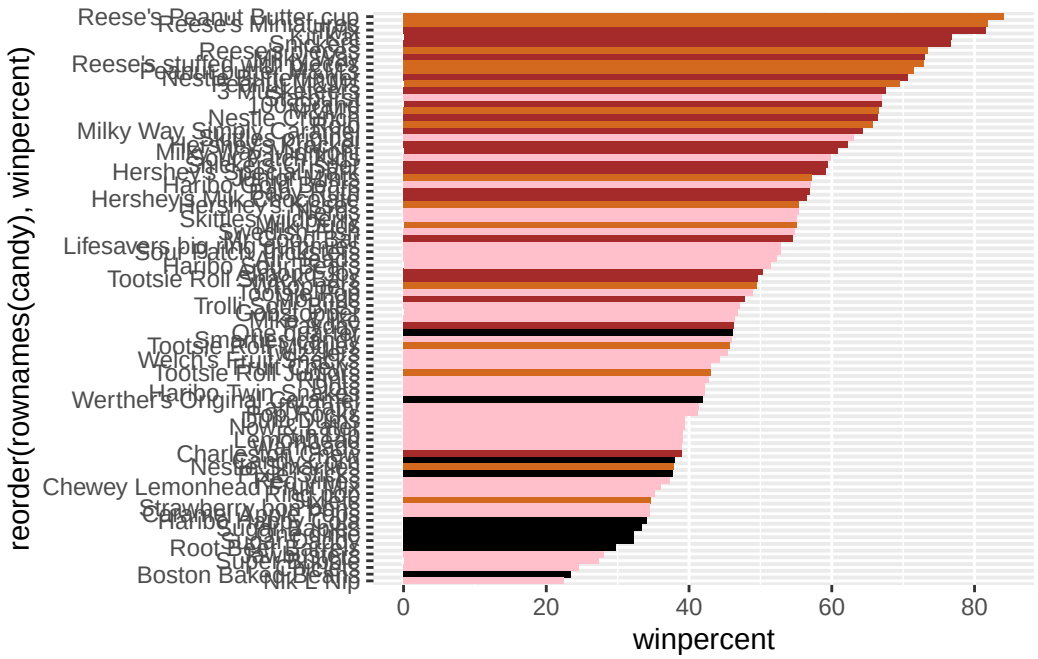
Q16. This is quite ugly, use the reorder() function to get the bars sorted by winpercent?

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




```
#Setting up color vectors for future plots
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"
```

```
#Bar plot with colors
ggplot(candy) +
  aes(winpercent, reorder(rownames(candy),winpercent)) +
  geom_col(fill=my_cols)
```



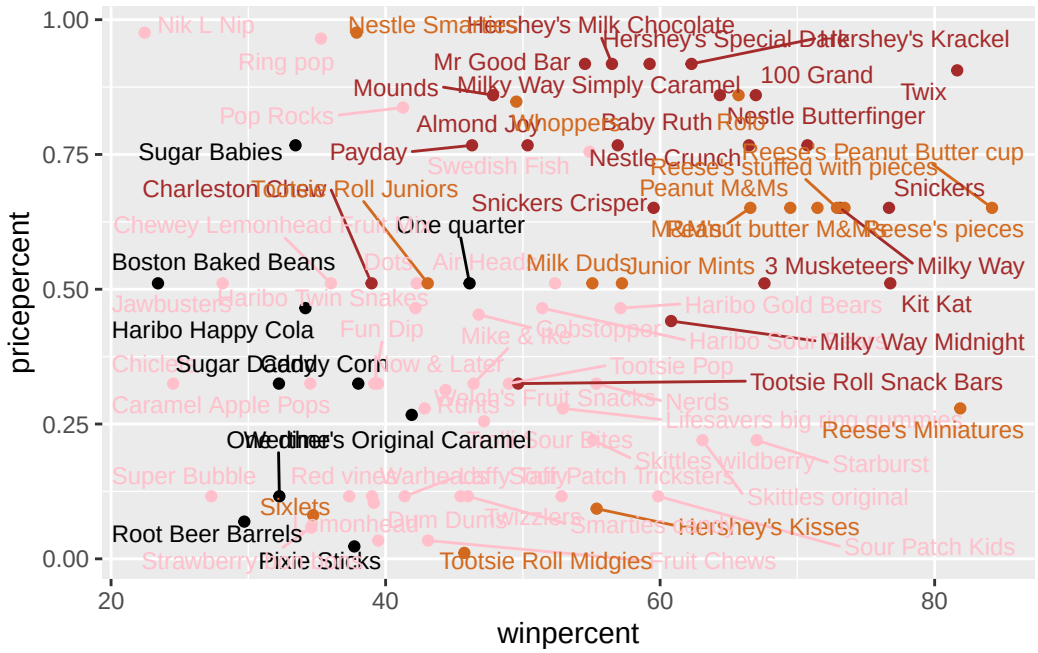
Q17. What is the worst ranked chocolate candy?

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 = 100)
```



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?

```
candy.bang <- candy
candy.bang$bang <- candy$winpercent / candy$pricepercent
ord <- order(candy.bang$bang, decreasing = TRUE)
head( candy.bang[ord,c(11,12,13)], n=100 )
```

	pricepercent	winpercent	bang
Tootsie Roll Midgies	0.011	45.73675	4157.88618
Pixie Sticks	0.023	37.72234	1640.10157
Fruit Chews	0.034	43.08892	1267.32122
Dum Dums	0.034	39.46056	1160.60452
Strawberry bon bons	0.058	34.57899	596.18952
Hershey's Kisses	0.093	55.37545	595.43498
Sour Patch Kids	0.116	59.86400	516.06895
Sour Patch Tricksters	0.116	52.82595	455.39609
Root Beer Barrels	0.069	29.70369	430.48829
Sixlets	0.081	34.72200	428.66667
Smarties candy	0.116	45.99583	396.51575
Twizzlers	0.116	45.46628	391.95071
Lemonhead	0.104	39.14106	376.35631

Laffy Taffy	0.116	41.38956	356.80653
Warheads	0.116	39.01190	336.30947
Red vines	0.116	37.34852	321.97002
Starburst	0.220	67.03763	304.71649
Reese's Miniatures	0.279	81.86626	293.42743
Skittles original	0.220	63.08514	286.75064
One dime	0.116	32.26109	278.11281
Skittles wildberry	0.220	55.10370	250.47134
Super Bubble	0.116	27.30386	235.37815
Lifesavers big ring gummies	0.279	52.91139	189.64656
Trolli Sour Bites	0.255	47.17323	184.99305
Nerds	0.325	55.35405	170.32015
Werther's Original Caramel	0.267	41.90431	156.94498
Runts	0.279	42.84914	153.58116
Tootsie Roll Snack Bars	0.325	49.65350	152.78001
Tootsie Pop	0.325	48.98265	150.71585
Kit Kat	0.511	76.76860	150.23210
Mike & Ike	0.325	46.41172	142.80528
Welch's Fruit Snacks	0.313	44.37552	141.77483
Milky Way Midnight	0.441	60.80070	137.87007
3 Musketeers	0.511	67.60294	132.29538
Reese's Peanut Butter cup	0.651	84.18029	129.30920
Haribo Gold Bears	0.465	57.11974	122.83815
Now & Later	0.325	39.44680	121.37477
Fun Dip	0.325	39.18550	120.57079
Snickers	0.651	76.67378	117.77846
Candy Corn	0.325	38.01096	116.95681
Reese's pieces	0.651	73.43499	112.80336
Milky Way	0.651	73.09956	112.28810
Junior Mints	0.511	57.21925	111.97505
Reese's stuffed with pieces	0.651	72.88790	111.96298
Haribo Sour Bears	0.465	51.41243	110.56437
Peanut butter M&M's	0.651	71.46505	109.77734
Milk Duds	0.511	55.06407	107.75748
Peanut M&Ms	0.651	69.48379	106.73393
Caramel Apple Pops	0.325	34.51768	106.20825
Gobstopper	0.453	46.78335	103.27450
Air Heads	0.511	52.34146	102.42949
M&M's	0.651	66.57458	102.26510
Sugar Daddy	0.325	32.23100	99.17230
Nestle Butterfinger	0.767	70.73564	92.22378
Snickers Crisper	0.651	59.52925	91.44278
Haribo Twin Snakes	0.465	42.17877	90.70704

One quarter	0.511	46.11650	90.24757
Twix	0.906	81.64291	90.11359
Nestle Crunch	0.767	66.47068	86.66321
Tootsie Roll Juniors	0.511	43.06890	84.28356
Dots	0.511	42.27208	82.72422
100 Grand	0.860	66.97173	77.87410
Rolo	0.860	65.71629	76.41429
Charleston Chew	0.511	38.97504	76.27209
Chiclets	0.325	24.52499	75.46150
Milky Way Simply Caramel	0.860	64.35334	74.82946
Baby Ruth	0.767	56.91455	74.20410
Haribo Happy Cola	0.465	34.15896	73.46012
Swedish Fish	0.755	54.86111	72.66372
Chewey Lemonhead Fruit Mix	0.511	36.01763	70.48460
Hershey's Krackel	0.918	62.28448	67.84802
Almond Joy	0.767	50.34755	65.64217
Hershey's Special Dark	0.918	59.23612	64.52737
Hershey's Milk Chocolate	0.918	56.49050	61.53649
Payday	0.767	46.29660	60.36062
Mr Good Bar	0.918	54.52645	59.39701
Whoppers	0.848	49.52411	58.40108
Mounds	0.860	47.82975	55.61599
Jawbusters	0.511	28.12744	55.04391
Pop Rocks	0.837	41.26551	49.30169
Boston Baked Beans	0.511	23.41782	45.82745
Sugar Babies	0.767	33.43755	43.59524
Nestle Smarties	0.976	37.88719	38.81884
Ring pop	0.965	35.29076	36.57073
Nik L Nip	0.976	22.44534	22.99728

Intuitively, it is Reese's Miniatures shown in the figure above. However, if you calculate the ratio of winpercent/pricepercent, Tootsie Roll Midgies offers the most bang for my buck.

> 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

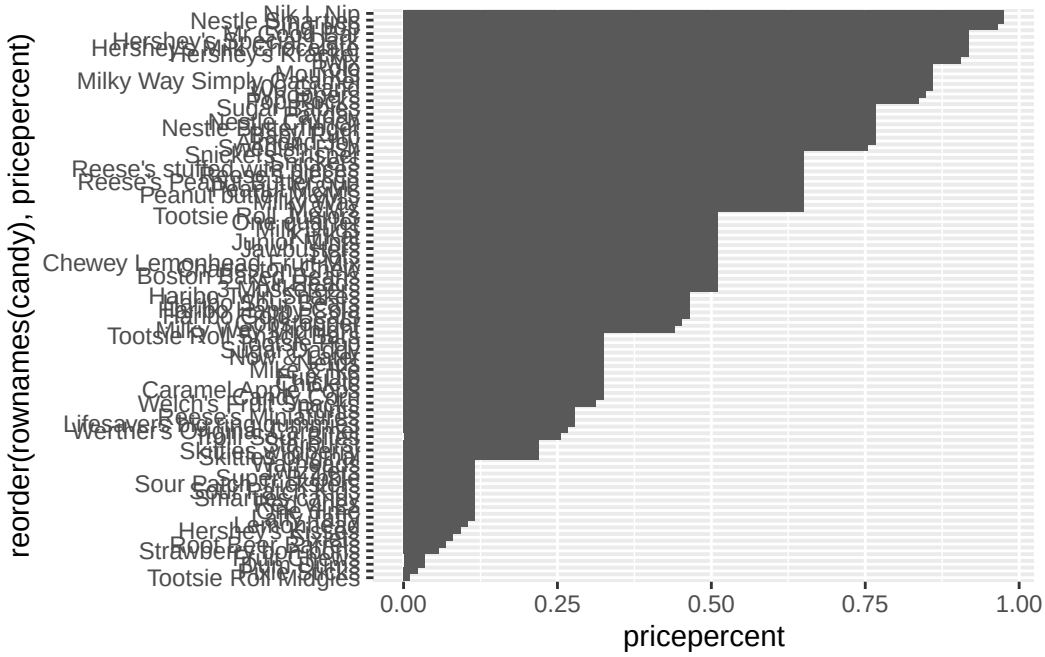
```
Ring pop                      0.965   35.29076
Hershey's Krackel             0.918   62.28448
Hershey's Milk Chocolate      0.918   56.49050
```

Nik L Nip, Nestle Smarties, Ring pop, Hershey's Krackel and Hershey's Milk Chocolate are the top 5 most expensive. Nik L Nip is the least popular

> 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().

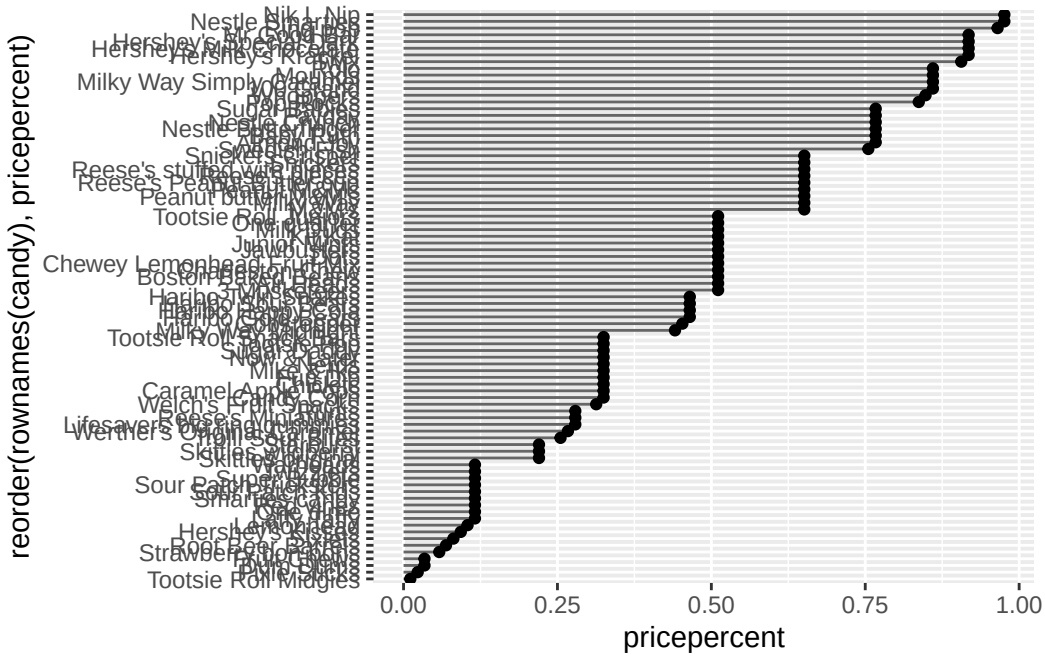
```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy),pricepercent)) +
  geom_col(stat="identity")
```

```
Warning in geom_col(stat = "identity"): Ignoring unknown parameters: `stat`
```




```
ggplot(candy) +
  aes(pricepercent, reorder(rownames(candy), pricepercent)) +
```

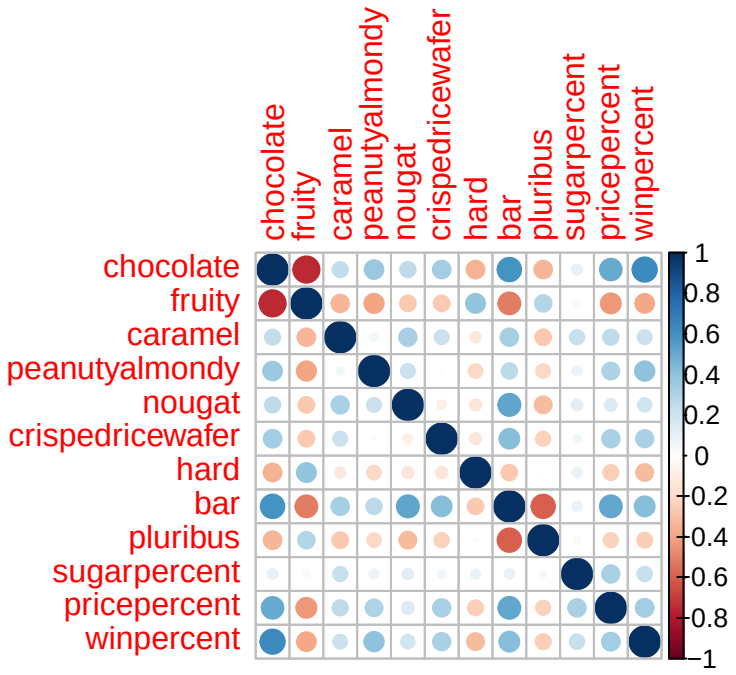
```
  geom_segment(aes(yend = reorder(rownames(candy), pricepercent),
                   xend = 0), col="gray40") +
    geom_point()
```




```
  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, pluribus and bar are the top 2 anti-correlated variable pairs.

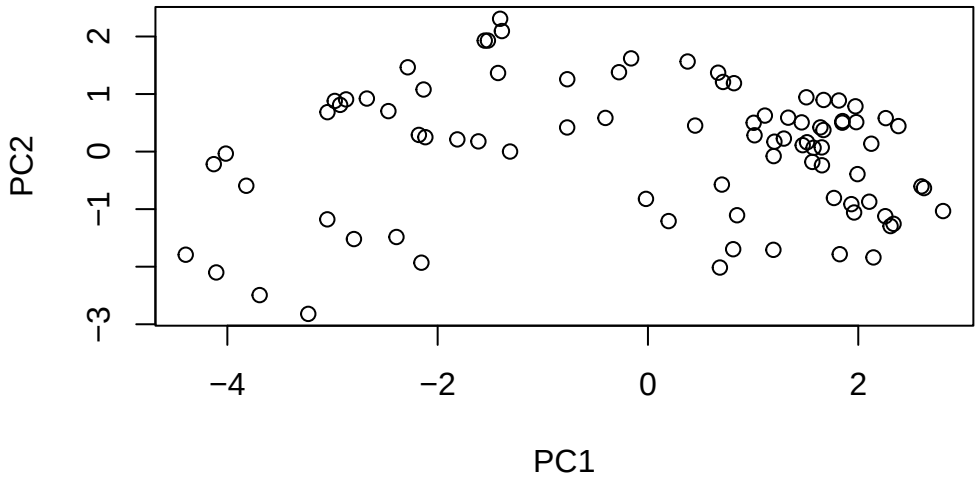
Q23. Similarly, what two variables are most positively correlated?

winpercent and chocolate, bar and chocolate are the top 2 positively correlated variable pairs.

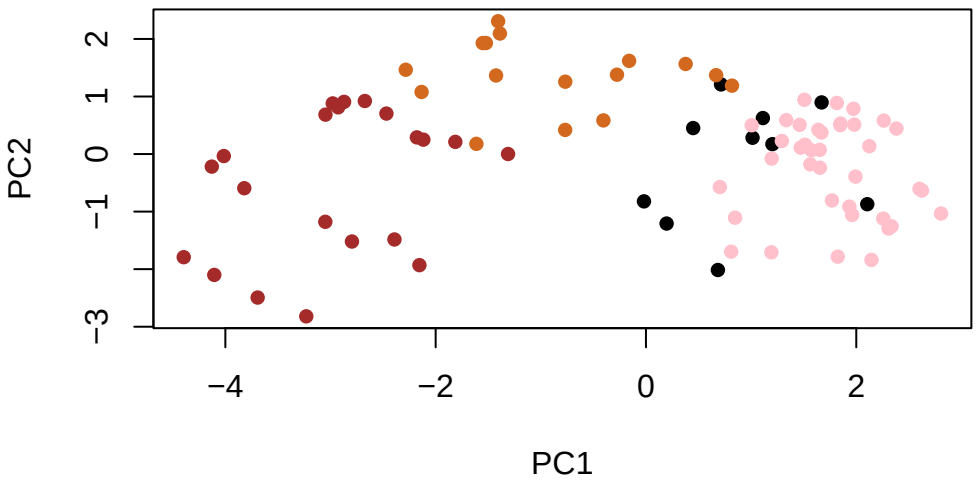
```
#Scale is used to uniform 0~1 and 0~100
pca <- prcomp(candy, scale=TRUE)
summary(pca)
```

```
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[,c(1,2)])
```

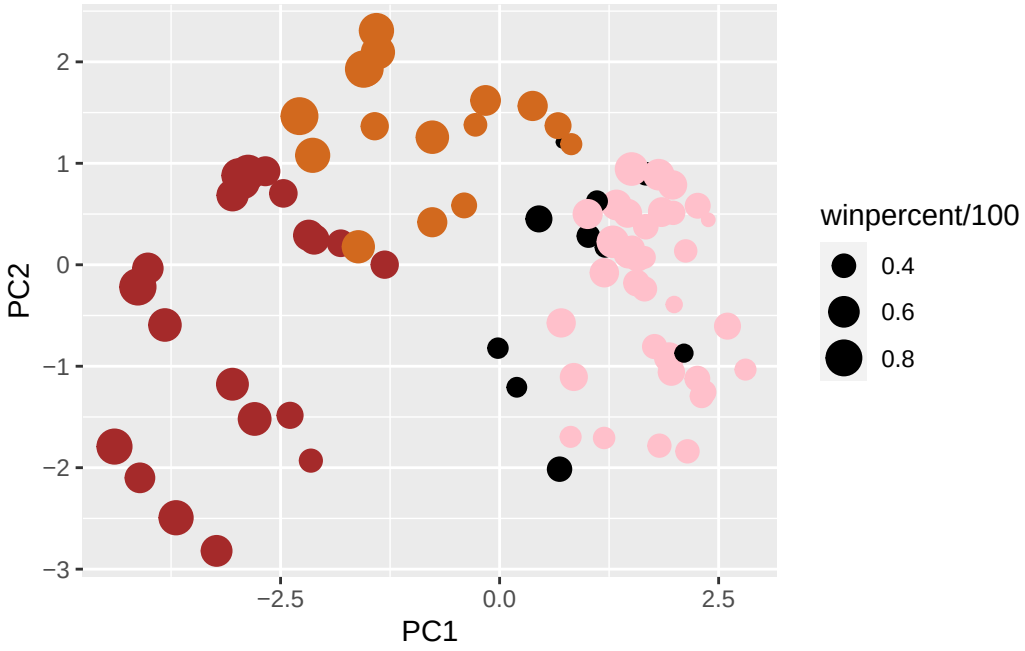



```
plot(pca$x[,1:2], col=my_cols, pch=16)
```



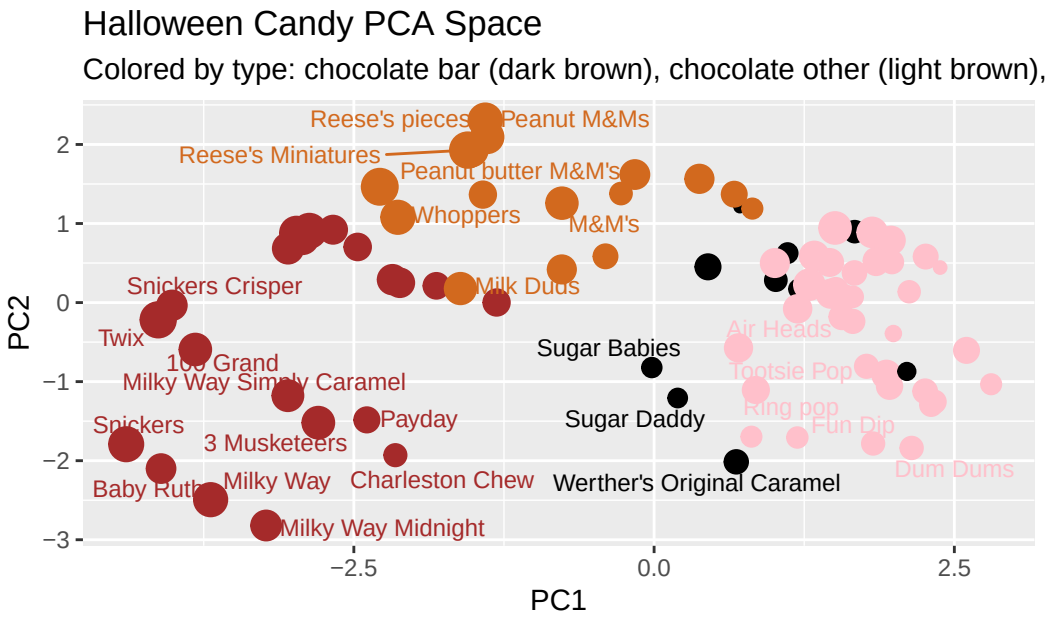

```
# Make a new data-frame with our PCA results and candy data
my_data <- cbind(candy, pca$x[,1:3])
```

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

```
Warning: ggrepel: 59 unlabeled data points (too many overlaps). Consider
increasing max.overlaps
```



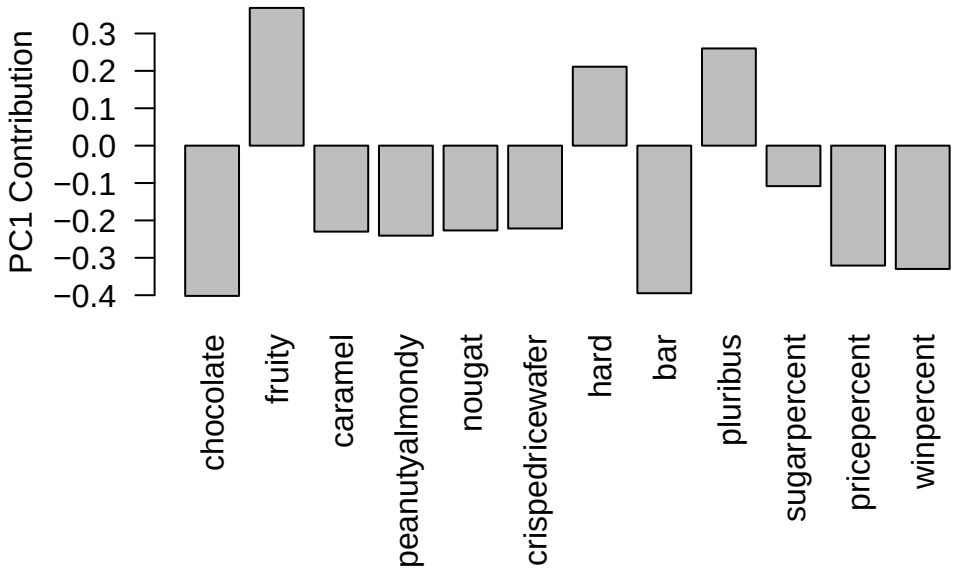

```
library(plotly)
```

```
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)
#This figure cannot be rendered to pdf
library(webshot)
```

```
webshot(url = "file:///E:/%E5%AD%A6%E4%B9%A0/UCSD/Course/BGGN213%20Found%20Bioinfor/Week4/
```

```
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 picked up strongly by PC1 in the positive direction. I think this tells us that fruity candies are often produced as hard candies. Also, as they are hard and small, they can be easily packaged in bags or boxes, so they are in the same direction with pluribus. In the negative direction, similarly, chocolate candies are commonly produced as bar candies.