class10/Halloween Candy

Longmei Zhang A17012012

```r
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?

```r
nrow(candy)
```

```
[1] 85
```

There are 85 different candy types in this dataset

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

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

```
[1] 38
```

There are 38 fruity candy types in the dataset

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

```r
candy["Milky Way", "winpercent"]
```

```
[1] 73.09956
```

```r
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
```

```r
candy|>
  filter(rownames(candy) == "Haribo Happy Cola") |>
  select(winpercent)
```

```
                  winpercent
Haribo Happy Cola   34.15896
```

My favourite is MilkyWay, and its winpercent is 73.10%

> Q find candy with winpercent above 50%

```
candy |>
  filter(winpercent > 50) |>
  filter(fruity == 1)
```

```
                            chocolate fruity caramel peanutyalmondy nougat
Air Heads                           0      1       0             0      0
Haribo Gold Bears                   0      1       0             0      0
Haribo Sour Bears                   0      1       0             0      0
Lifesavers big ring gummies         0      1       0             0      0
Nerds                               0      1       0             0      0
Skittles original                   0      1       0             0      0
Skittles wildberry                  0      1       0             0      0
Sour Patch Kids                     0      1       0             0      0
Sour Patch Tricksters               0      1       0             0      0
Starburst                           0      1       0             0      0
Swedish Fish                        0      1       0             0      0
                            crispedricewafer hard bar pluribus sugarpercent
Air Heads                                  0    0   0        0        0.906
Haribo Gold Bears                          0    0   0        1        0.465
Haribo Sour Bears                          0    0   0        1        0.465
Lifesavers big ring gummies                0    0   0        0        0.267
Nerds                                      0    1   0        1        0.848
Skittles original                          0    0   0        1        0.941
Skittles wildberry                         0    0   0        1        0.941
Sour Patch Kids                            0    0   0        1        0.069
Sour Patch Tricksters                      0    0   0        1        0.069
Starburst                                  0    0   0        1        0.151
Swedish Fish                               0    0   0        1        0.604
                            pricepercent winpercent
Air Heads                          0.511   52.34146
Haribo Gold Bears                  0.465   57.11974
Haribo Sour Bears                  0.465   51.41243
Lifesavers big ring gummies        0.279   52.91139
Nerds                              0.325   55.35405
Skittles original                  0.220   63.08514
Skittles wildberry                 0.220   55.10370
Sour Patch Kids                    0.116   59.86400
Sour Patch Tricksters              0.116   52.82595
Starburst                          0.220   67.03763
Swedish Fish                       0.755   54.86111
```

```
#same results
#top.candy <- candy[candy$winpercent > 50]
#top.candy[candy$fruity == 1]
```

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

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

```
[1] 76.7686
```

The winepercent value of Kit kat is 76.78%

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

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

```
[1] 49.6535
```

The winpercent value of Tootsie Roll Snack Bars is 49.65

To get a quick insight into a new dataset some folks like using skimr package and function `skim()`

```r
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 value is measured on different scale than everything else. Need to scale the data before doing analysis like PCA.

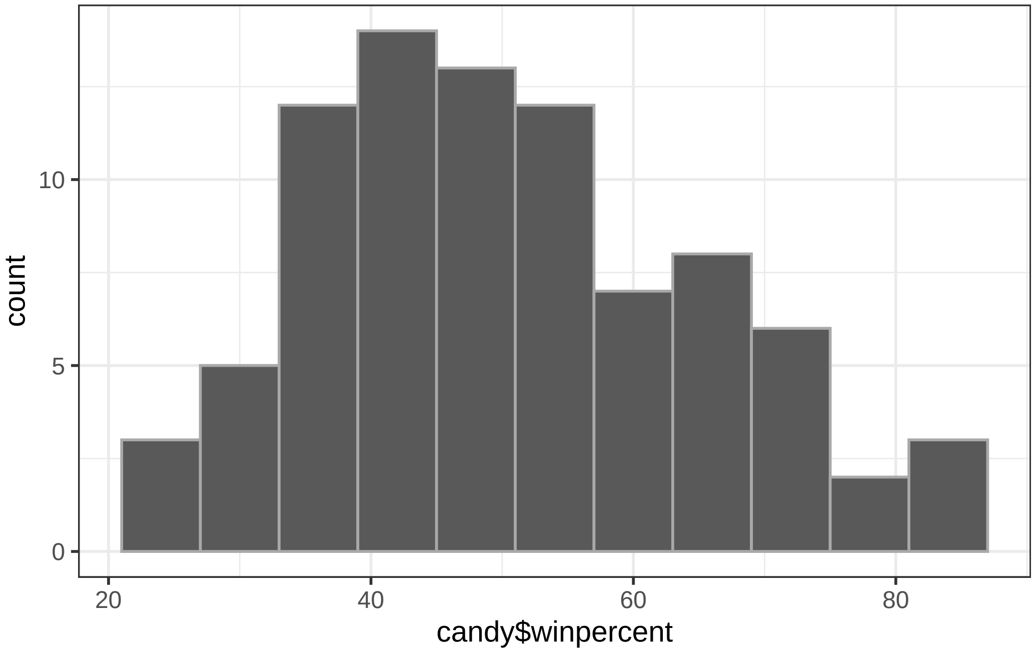
> Q7. What do you think a zero and one represent for the candy$chocolate column?

1 Means the candy is chocolate type while 0 means the candy is not chocolate type.

> Q8. Plot a histogram of winpercent values

```r
#hist(candy$winpercent, breaks = 10)

library("ggplot2")
ggplot(candy, aes(candy$winpercent)) +
  geom_histogram(binwidth = 6, color = "dark grey") +
  theme_bw()
```

```
Warning: Use of `candy$winpercent` is discouraged.
i Use `winpercent` instead.
```



Q9. Is the distribution of winpercent values symmetrical?

No, its skewed to the left

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

```
summary(candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.14   47.83   50.32   59.86   84.18 
```

Using median to represent the center, the center is blow 50%

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

```
fruit.candy <- candy |>
  filter(candy$fruity == 1)

summary(fruit.candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  22.45   39.04   42.97   44.12   52.11   67.04 
```

```
choco.candy <- candy |>
  filter(candy$chocolate == 1)

summary(choco.candy$winpercent)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  34.72   50.35   60.80   60.92   70.74   84.18 
```

Chocolate candy have a higher median and mean, meaning that chocolate candies are ranked higher on average.

> Q12. Is this difference statistically significant?

```
t.test(choco.candy$winpercent, fruit.candy$winpercent)
```

```

	Welch Two Sample t-test

data:  choco.candy$winpercent and fruit.candy$winpercent
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 
```

According to the p-value, which is very small, it is very unlikely to get this result by chance. The difference between winpercent of chocolate and fruity is significant.

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

```
#play <- c("d", "a", "c")
#sort(play)
#order(play)
head(candy[order(candy$winpercent), ], 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
```

the least favourite 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?

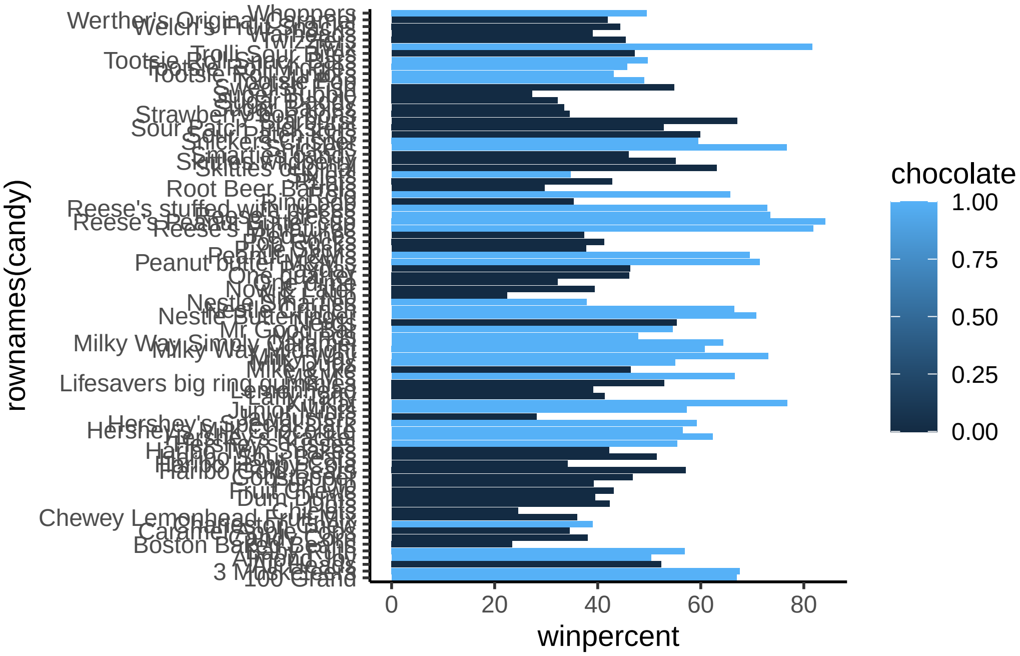
```r
head(candy[order(candy$winpercent, decreasing = T), ], 5)
```

```
                          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
```

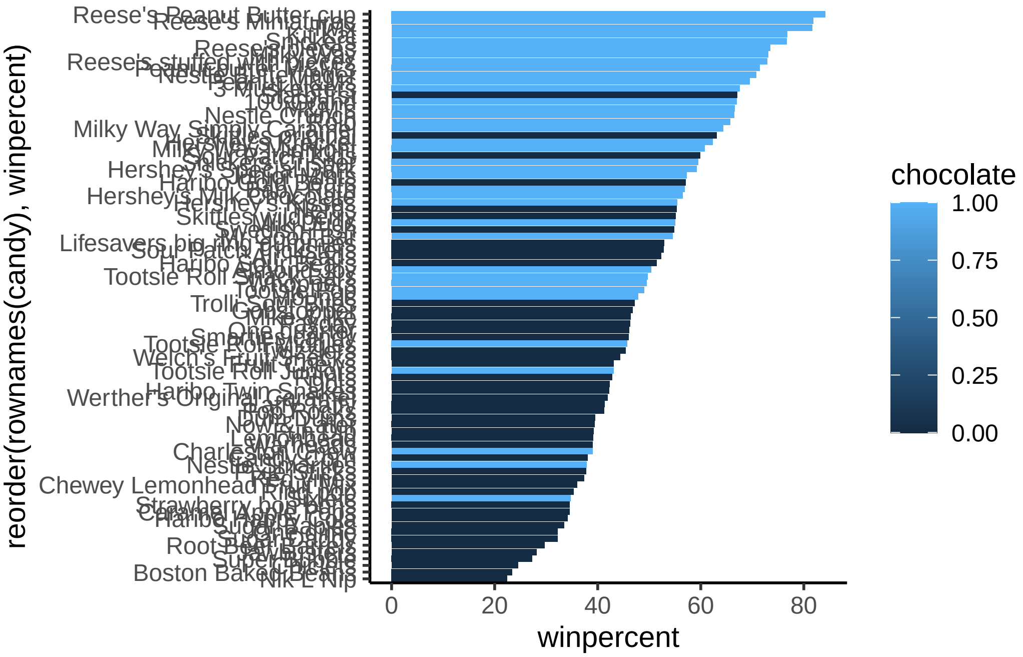
Q15. Make a first barplot of candy ranking based on winpercent values.

```
ggplot(candy)+
  aes(winpercent, rownames(candy), fill = chocolate) +
  geom_col() +
  theme_classic()
```



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

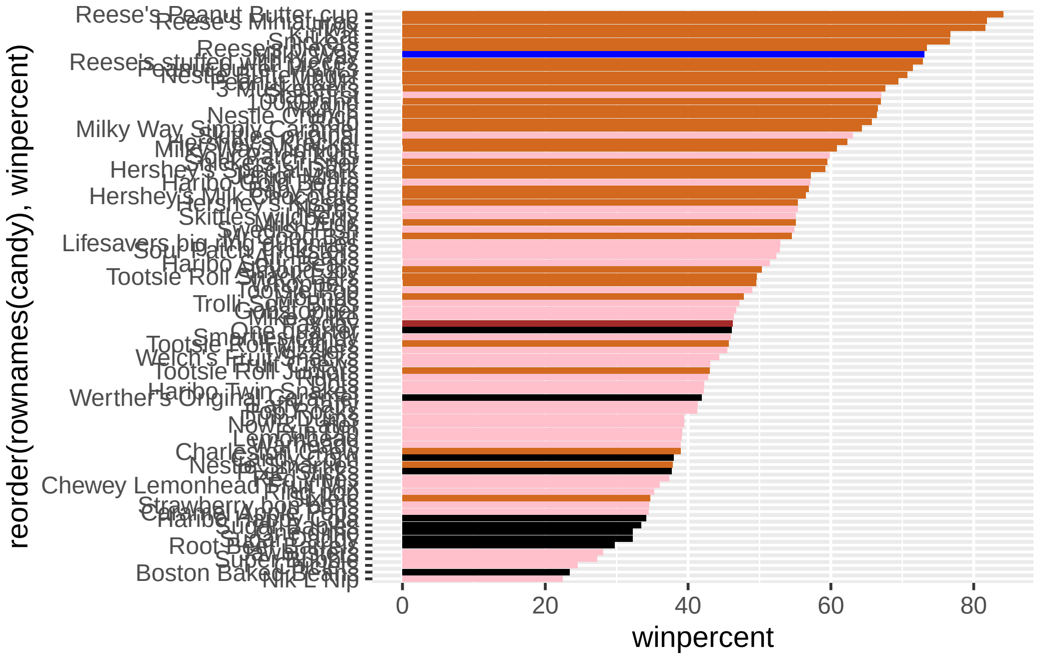
```
ggplot(candy)+
  aes(winpercent, reorder(rownames(candy), winpercent), fill = chocolate) +
  geom_col() +
  theme_classic()
```



Designing a more specialized colore scheme where we can see both chocolate and bar and fruity etc. all from the same plot. Change the color vector.

```r
mycols <- rep("black", nrow(candy))
mycols[as.logical(candy$bar)] = "brown"
mycols[as.logical(candy$chocolate)] = "chocolate"
mycols[as.logical(candy$fruity)] = "pink"

# Use blue for favorite candy
mycols[rownames(candy) == "Milky Way"] = "blue"
```

```r
ggplot(candy)+
  aes(winpercent, reorder(rownames(candy), winpercent)) +
  geom_col(fill = mycols)
```



Q17. What is the worst ranked chocolate candy?

The worst ranked chocolate candy is Sixlets

```
choco = candy[as.logical(candy$chocolate), ]
head(choco[order(choco$winpercent), ],5)
```

```
                    chocolate fruity caramel peanutyalmondy nougat
Sixlets                     1      0       0              0      0
Nestle Smarties             1      0       0              0      0
Charleston Chew             1      0       0              0      1
Tootsie Roll Juniors        1      0       0              0      0
Tootsie Roll Midgies        1      0       0              0      0
                    crispedricewafer hard bar pluribus sugarpercent
Sixlets                            0    0   0        1        0.220
Nestle Smarties                    0    0   0        1        0.267
Charleston Chew                    0    0   1        0        0.604
Tootsie Roll Juniors               0    0   0        0        0.313
Tootsie Roll Midgies               0    0   0        1        0.174
                    pricepercent winpercent
Sixlets                    0.081   34.72200
Nestle Smarties            0.976   37.88719
Charleston Chew            0.511   38.97504
```

```
Tootsie Roll Juniors        0.511   43.06890
Tootsie Roll Midgies        0.011   45.73675
```

Q18. What is the best ranked fruity candy?

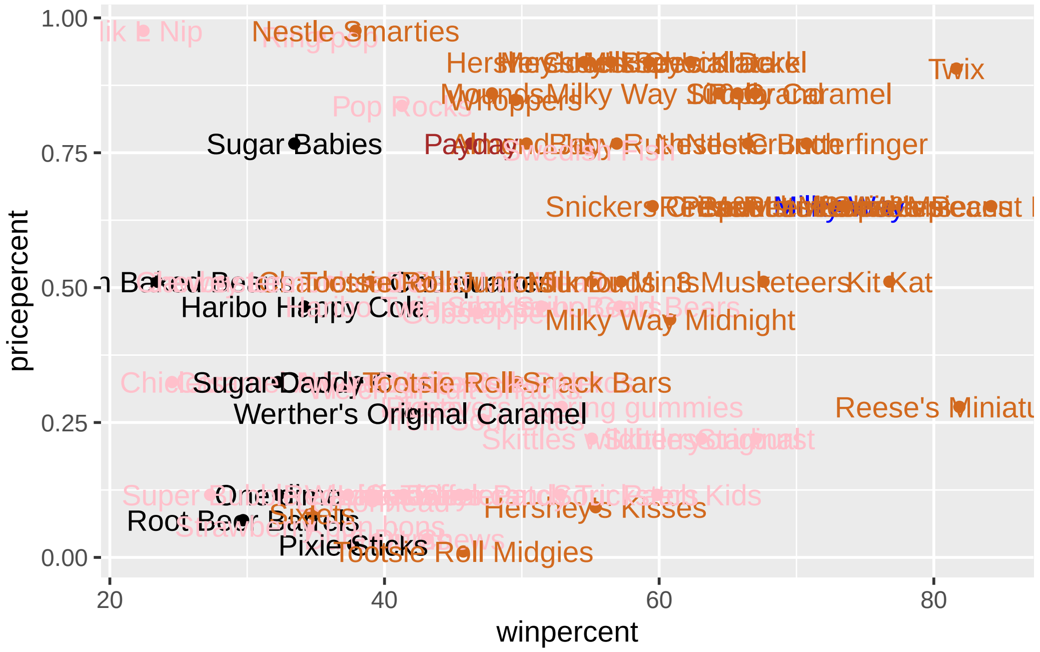
The best ranked fruity candy is Starburst.

```
fruity = candy[as.logical(candy$fruity), ]
head(fruity[order(fruity$winpercent, decreasing = T), ],5)
```

```
                  chocolate fruity caramel peanutyalmondy nougat
Starburst                 0      1       0              0      0
Skittles original         0      1       0              0      0
Sour Patch Kids           0      1       0              0      0
Haribo Gold Bears         0      1       0              0      0
Nerds                     0      1       0              0      0
                  crispedricewafer hard bar pluribus sugarpercent pricepercent
Starburst                        0    0   0        1        0.151        0.220
Skittles original                0    0   0        1        0.941        0.220
Sour Patch Kids                  0    0   0        1        0.069        0.116
Haribo Gold Bears                0    0   0        1        0.465        0.465
Nerds                            0    1   0        1        0.848        0.325
                  winpercent
Starburst           67.03763
Skittles original   63.08514
Sour Patch Kids     59.86400
Haribo Gold Bears   57.11974
Nerds               55.35405
```

##Taking a look at price percent

```
ggplot(candy) +
  aes(winpercent, pricepercent, label = rownames(candy)) +
  geom_point(col = mycols) +
  geom_text(col = mycols)
```



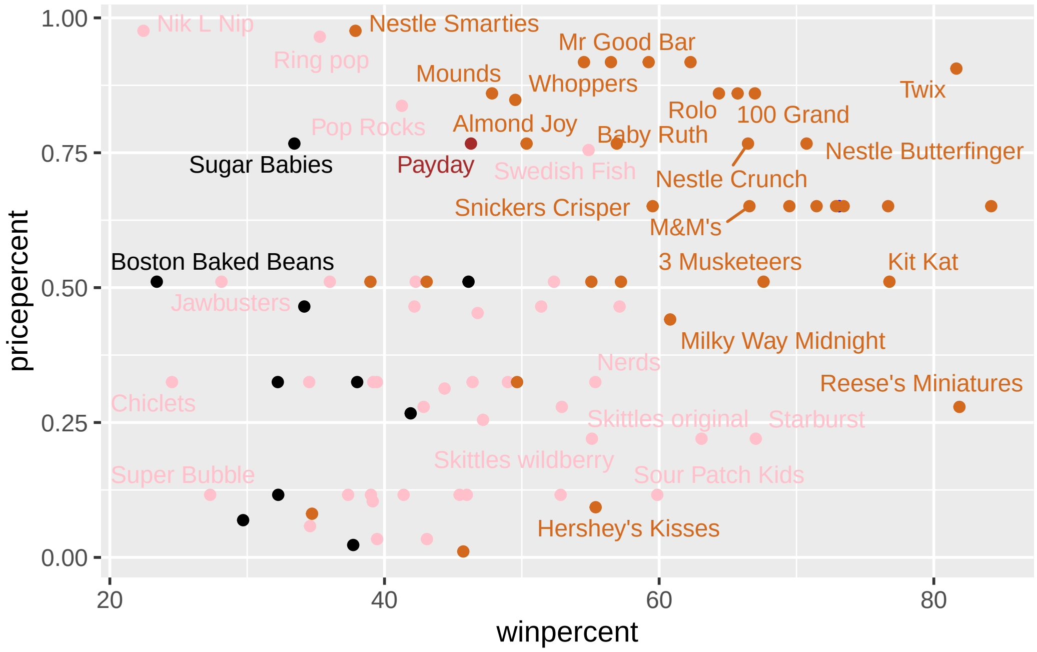
Making labels visible and not overlapping

```
library("ggrepel")
```

```
Warning: package 'ggrepel' was built under R version 4.3.3
```

```
ggplot(candy) +
  aes(winpercent, pricepercent, label = rownames(candy)) +
  geom_point(col = mycols) +
  geom_text_repel(col = mycols, max.overlaps = 8, size = 3.3)
```

```
Warning: ggrepel: 52 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?

Reese's Miniatures is ranked high in terms of winpercent for the least money.

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

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

```r
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
```

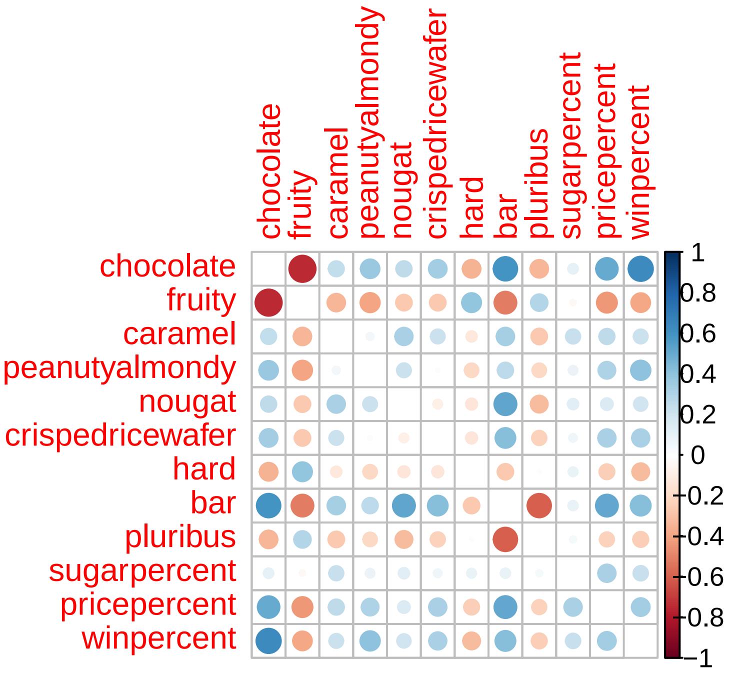
Exploring the correlation structure

```
library(corrplot)
```

```
Warning: package 'corrplot' was built under R version 4.3.3
```

```
corrplot 0.95 loaded
```

```
cij <- cor(candy)
corrplot(cij, diag = F)
```



> Q22. Examining this plot what two variables are anti-correlated (i.e. have minus values)?

Chocolate and Fruity are most anti-correlated

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

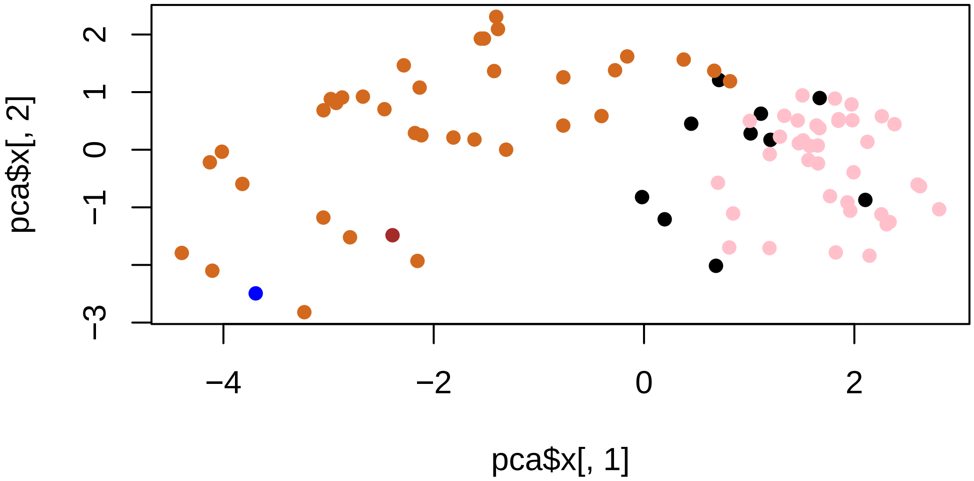
Chocolate and winpercent, and chocolate and bar are most positively correlated.

PCA

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

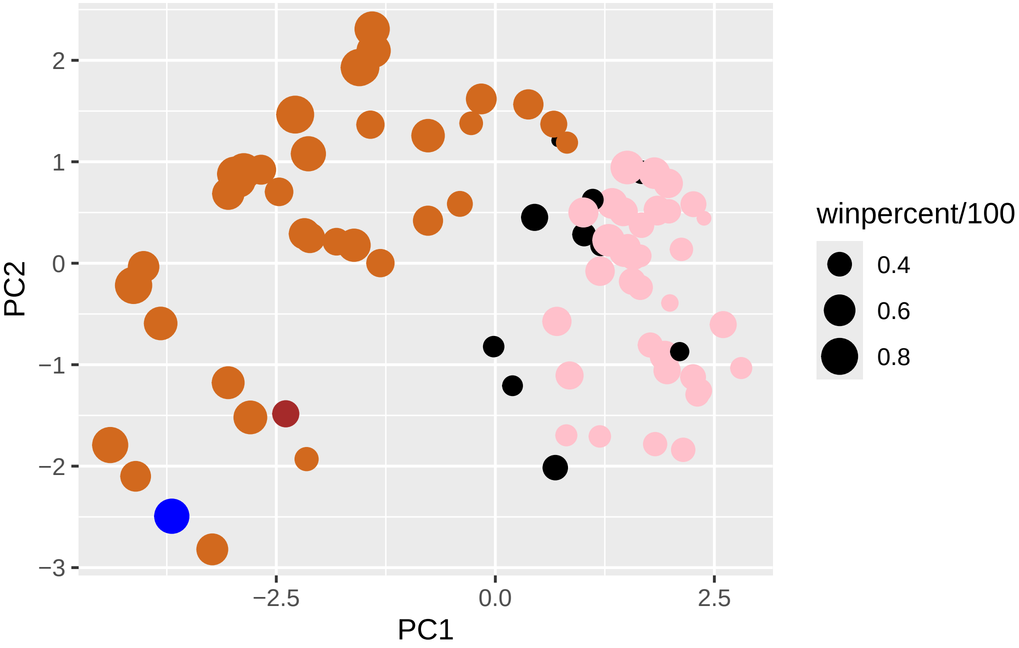
```
Importance of components:
                          PC1    PC2    PC3     PC4    PC5     PC6     PC7
Standard deviation     2.0788 1.1378 1.1092 1.07533 0.9518 0.81923 0.81530
Proportion of Variance 0.3601 0.1079 0.1025 0.09636 0.0755 0.05593 0.05539
Cumulative Proportion  0.3601 0.4680 0.5705 0.66688 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$x[,1], pca$x[,2], col = mycols, pch=16)
```



We can make a much nicer plot with the ggplot2 package. ggplot works best when we supply an input data.frame that includes a separate column for each of the aesthetics you would like displayed in your final plot. To accomplish this we make a new data.frame here that contains our PCA results with all the rest of our candy data. We will then use this for making plots below

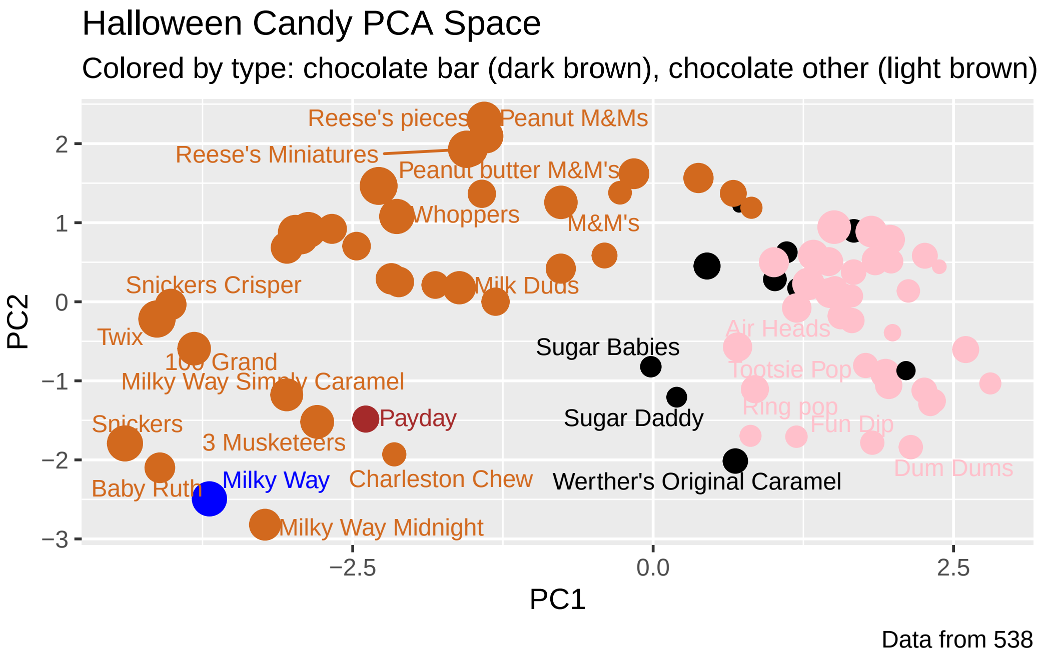
```
my_data <- cbind(candy, pca$x[,1:3])

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

p
```



We can use the `ggrepel` package and the function `ggrepel::geom_text_repel()` to label up the plot with non overlapping candy names like

```
p + geom_text_repel(size=3.3, col=mycols, 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
```

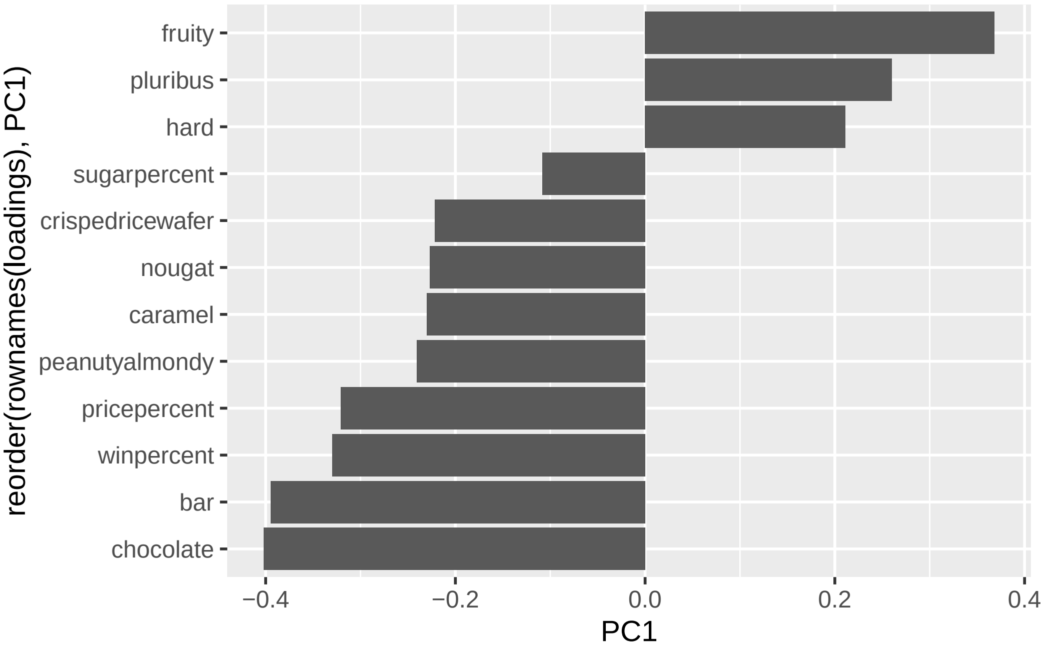


We can also generate interactive plot using `plotly`

```
#library(plotly)
#ggplotly(p)
```

How do the original variables (columns) contribute to the new PCs. Looking at PC1

```
loadings <- as.data.frame(pca$rotation)
ggplot(loadings) +
  aes(PC1, reorder(rownames(loadings), PC1)) +
  geom_col()
```



Q24. What original variables are picked up strongly by PC1 in the positive direction? Do these make sense to you?

Fruity strongly contributed to PC1 in the positive direction. It make sense since Fruity candy types are mainly located on the right side (positive direction of PC1) of the PC1 vs. PC2 graph.