Class 10

Arturo Agramont

Importing candy data

```
candy = read.csv("candy-data.csv", 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
```

There are 85 kinds of candy in this dataset.

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

```
table(candy$fruity)
```

```
 0  1
47 38
```

There are 38 fruity candy types in the data set.

Viewing win percent

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

```
[1] 81.64291
```

Twix in this case has an 81.64% win rate in match-ups against other candies.

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

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

```
[1] 59.864
```

The win percent value for my favorite candy, Sour Patch Kids, is 59.86%

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

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

```
[1] 76.7686
```

Kit Kat has a 76.76% win rate

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

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

```
[1] 49.6535
```

Tootsie rolls have a 49.65% win percent value.

We are going to use skim to look at the data

```
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?

It appears that win percent is on a different scale than the rest of the columns.

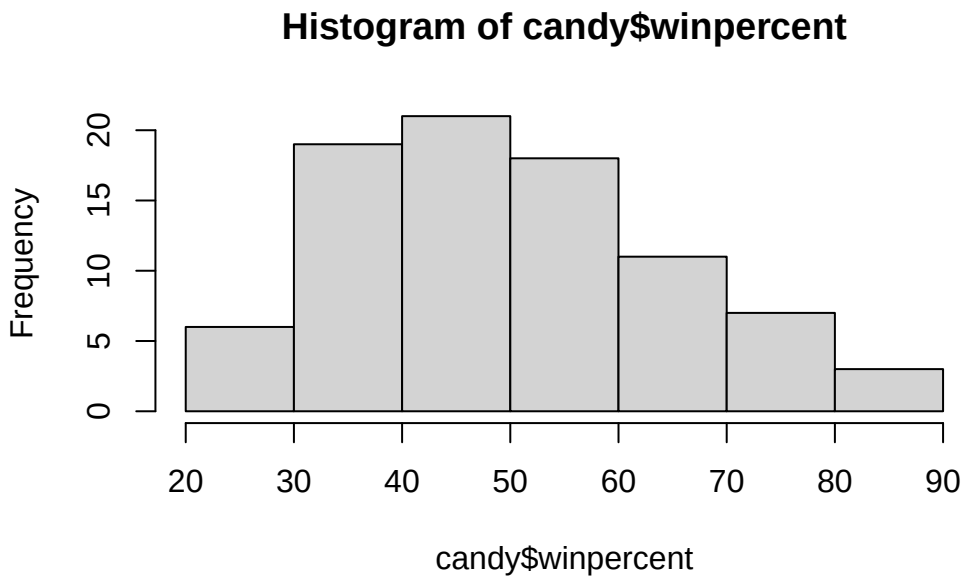
- **Q7**. What do you think a zero and one represent for the `candy$chocolate` column?

A zero represents that the candy has no chocolate and a 1 is that the candy contains chocolate.

We will now start making plots

- **Q8**. Plot a histogram of `winpercent` values

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



- **Q9**. Is the distribution of `winpercent` values symmetrical?

The distribution is not symmetrical but is shifted to the left.

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

Below 50%

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

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

```
[1] 44.11974
```

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

```
[1] 60.92153
```

On average chocolate is chosen more and higher ranked.

- **Q12**. Is this difference statistically significant?

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

```
    Welch Two Sample t-test

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

This is a significant difference.

Overall candy rankings

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

  We will use order() to arrange the data set by win percent

```
head(candy[order(candy$winpercent),])
```

```
                  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
                   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
Root Beer Barrels                 0    1   0        1        0.732        0.069
                   winpercent
Nik L Nip            22.44534
Boston Baked Beans   23.41782
Chiclets             24.52499
Super Bubble         27.30386
Jawbusters           28.12744
Root Beer Barrels    29.70369
```

The 5 least likes are Nik L Nip, Boston Baked Beans, Chiclets, Super Bubble, Jawbusters

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

```
tail(candy[order(candy$winpercent),])
```

```
                         chocolate fruity caramel peanutyalmondy nougat
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
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
Reese's pieces                0.651   73.43499
Snickers                      0.651   76.67378
Kit Kat                       0.511   76.76860
Twix                          0.906   81.64291
```

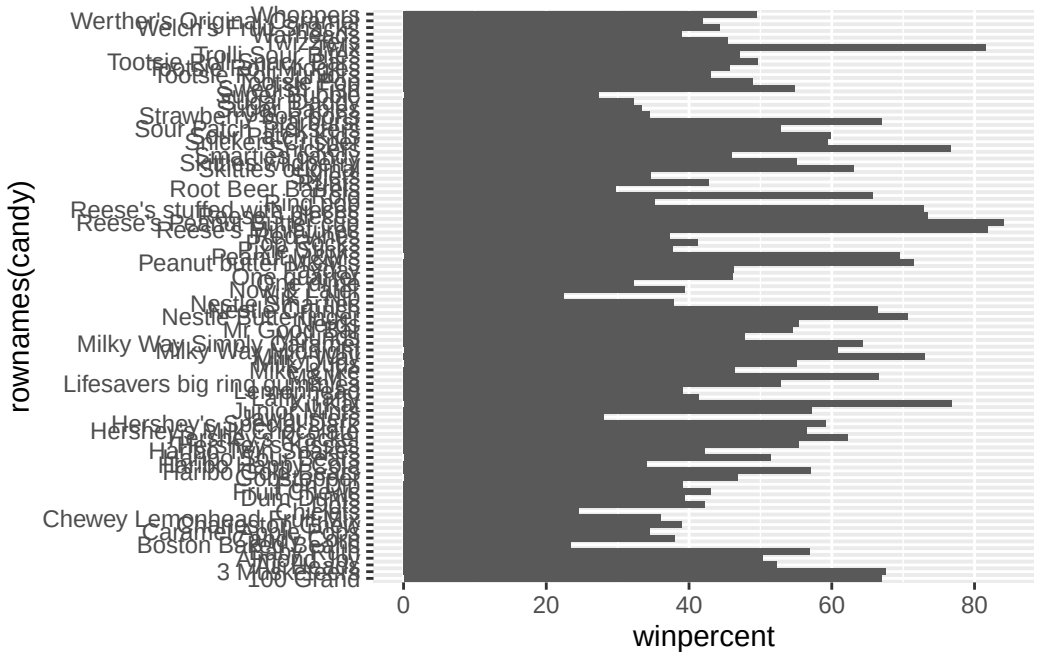
```
Reese's Miniatures                0.279   81.86626
Reese's Peanut Butter cup         0.651   84.18029
```

Reese's Peanut Butter Cup, Reese's Miniatures, Twix, Kit Kat, and Snickers have the highest win percent.

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

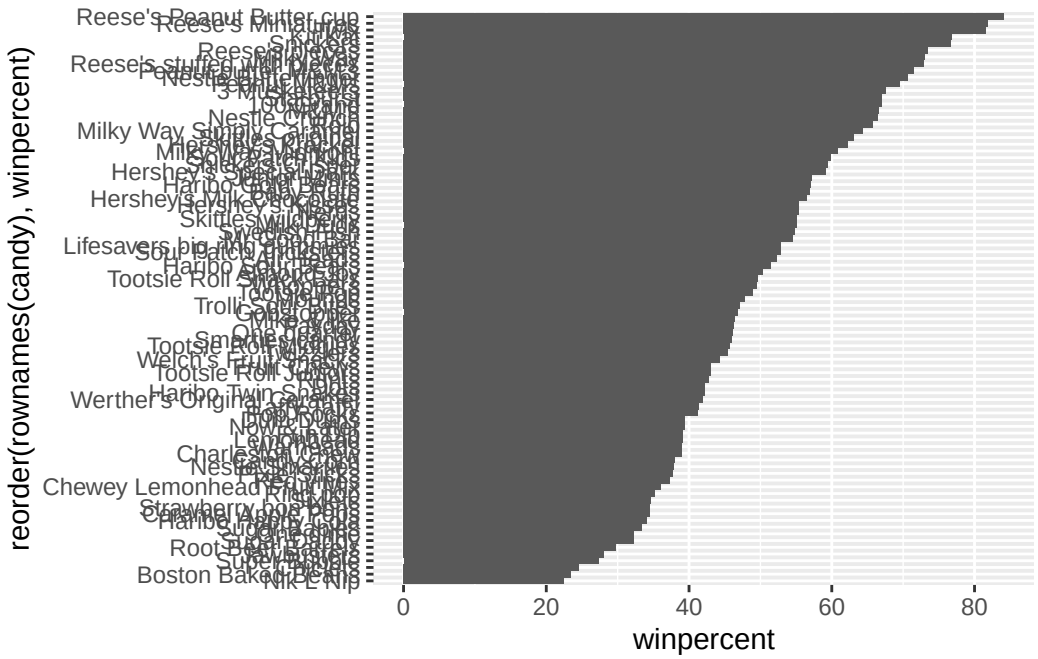
```
library(ggplot2)

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

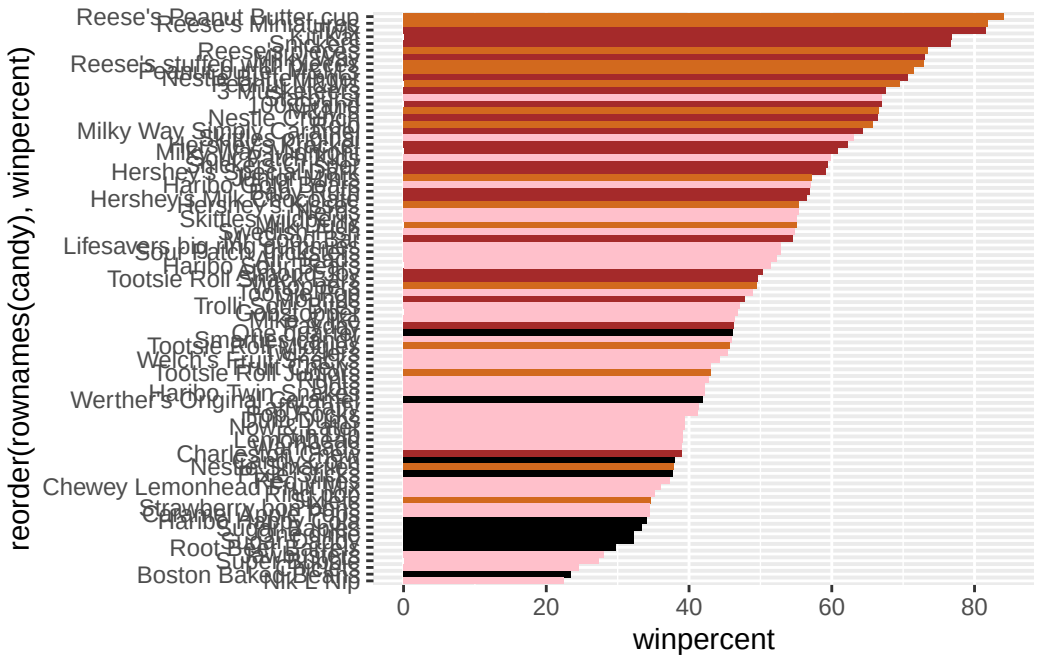


Now we are going to add color to the graph

We will assign colors to a variable to apply to the bar graph

```
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 are the worst ranked chocolate candy.

- **Q18**. What is the best ranked fruity candy?

Starburst is the best ranked fruity candy

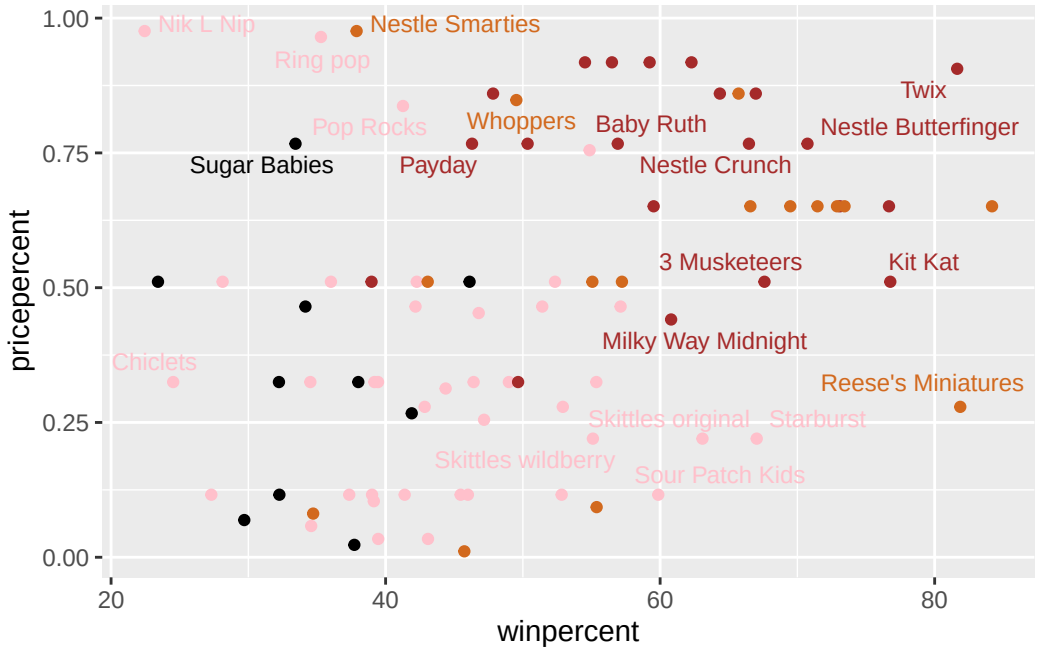
Price percent

We are going to make a plot of win percent vs price percent, we will use ggplot and ggrepel

```
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: 65 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, starburst, kit kat are the highest ranked in win percent for the least amount of money

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

We are going to change the order of the data in order to see the most expensive candy and the least popular of the 5 most expensive.

```
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 least popular expensive candy is Nik L Nip.

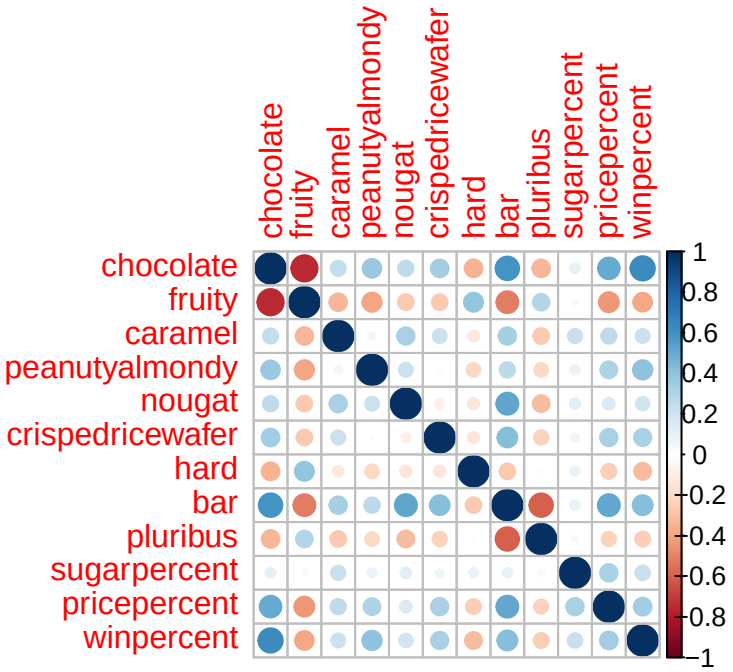
Correlation Structure

We are going to use corrplot to explore correlation

```
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)?

Win percent and price percent are the most anti-correlated

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

Chocolate and Fruity are the most positive correlated

Using PCA

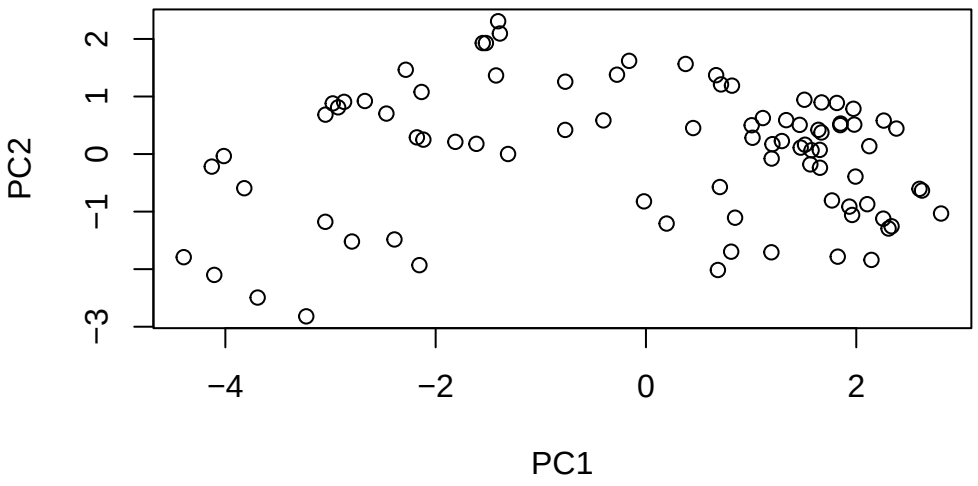
We will use PCA analysis on the data

```
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
```

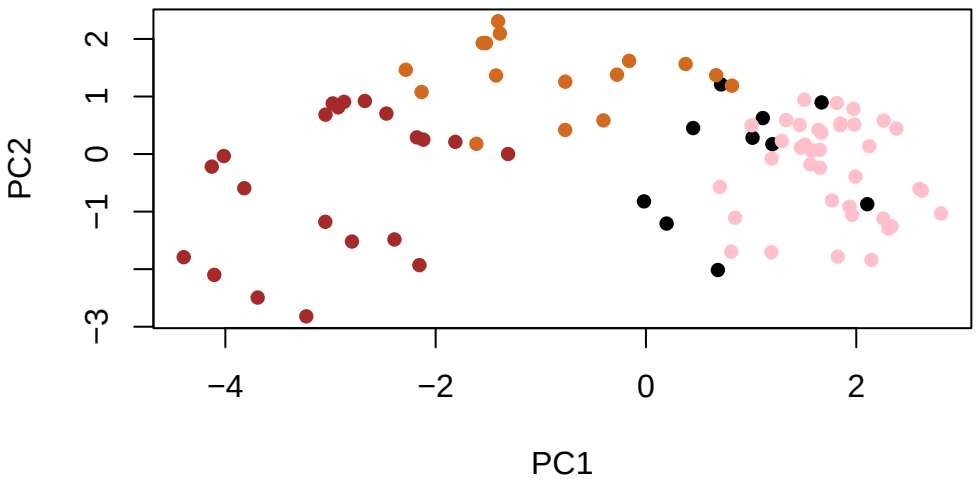
Now we will plot PC1 vs PC2

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



Now adding color

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


PC2
2
1
0
−1
−3
−4
−2
0
2
PC1