Homework #2

Youki Cao

Oct 09 2018

1. Flowers

Data: `flowers` dataset in **cluster** package

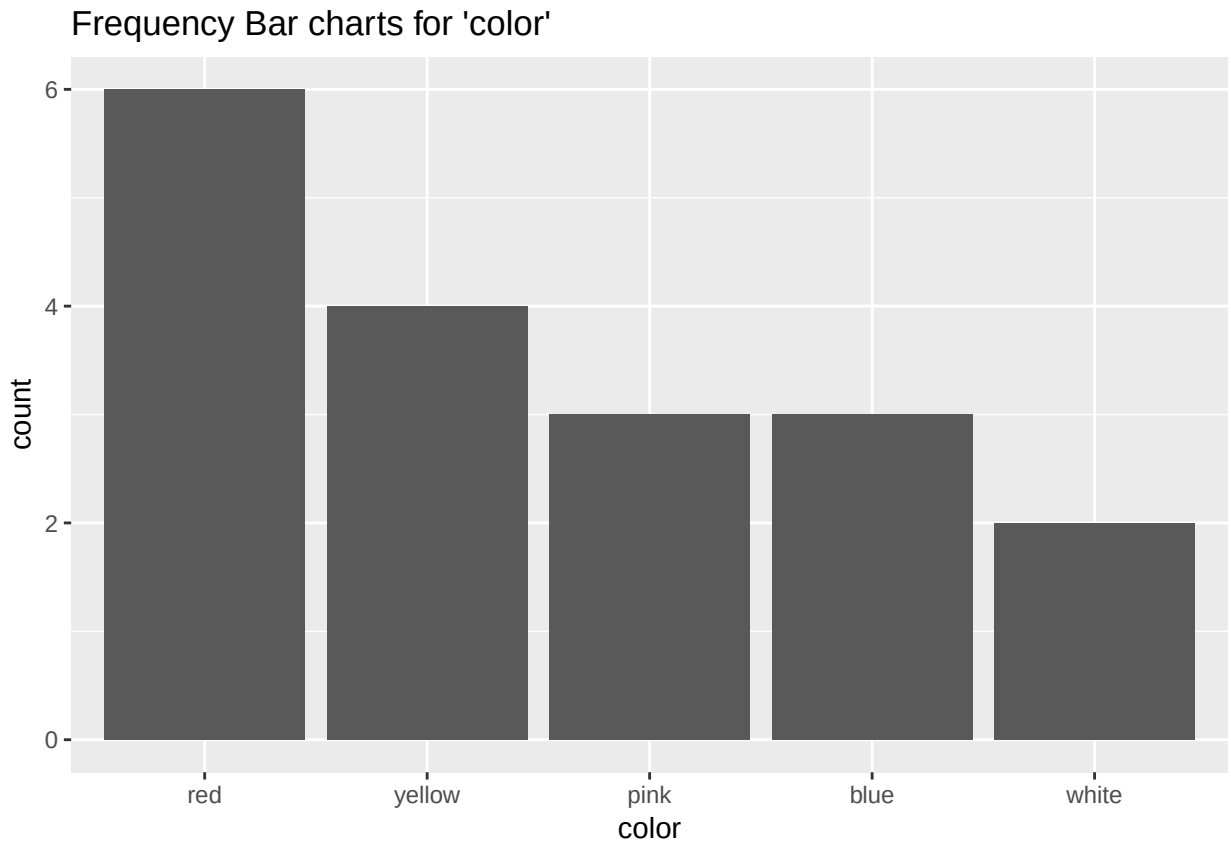
(a) Rename the column names and recode the levels of categorical variables to descriptive names. For example, "V1" should be renamed "winters" and the levels to "no" or "yes". Display the full dataset.

```
library(cluster)
df1 <- flower
colnames(df1) <- c('winters', 'shadow', 'tubers', 'color', 'soil',
                   'preference', 'height', 'distance')
levels(df1$winters) <- c('no', 'yes')
levels(df1$shadow) <- c('no', 'yes')
levels(df1$tubers) <- c('no', 'yes')
levels(df1$color) <- c('white', 'yellow', 'pink', 'red', 'blue')
levels(df1$soil) <- c('dry', 'normal', 'wet')
df1
```

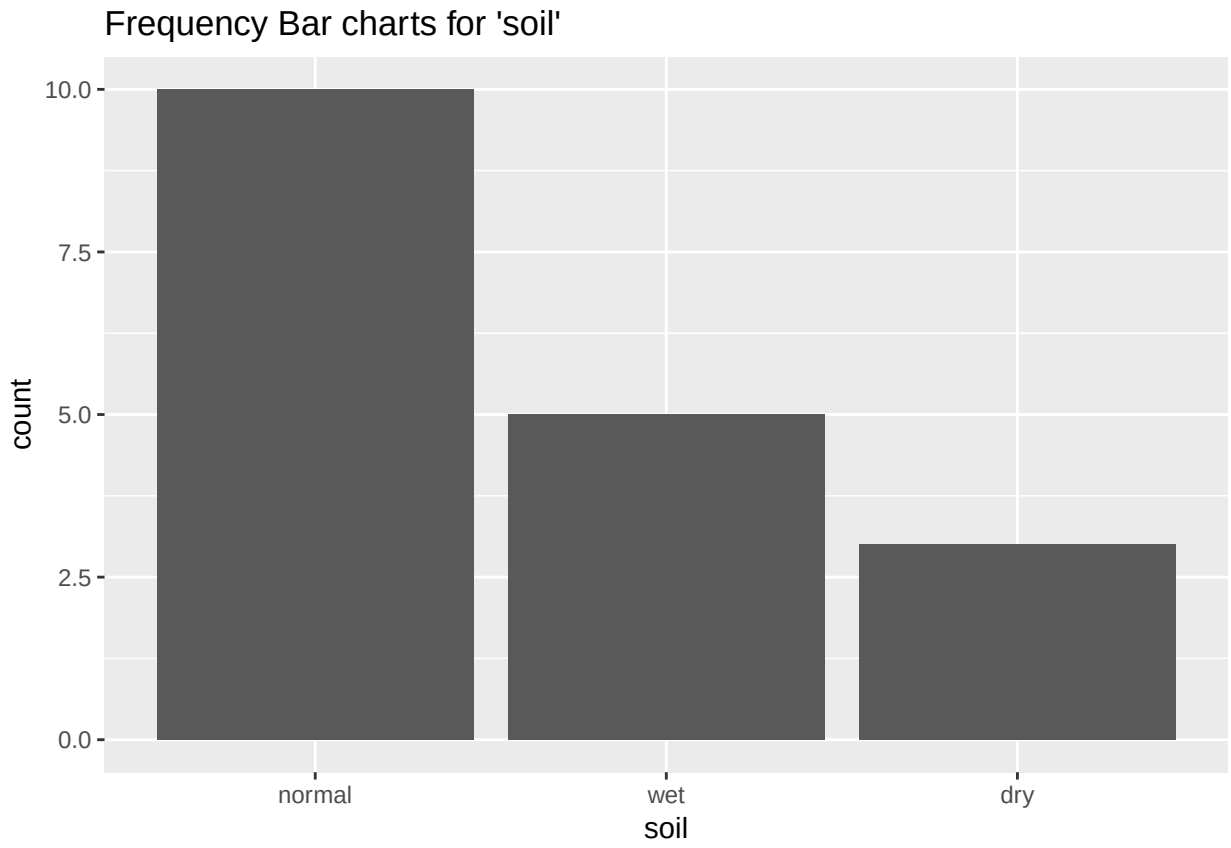
```
##    winters shadow tubers  color   soil preference height distance
## 1       no    yes    yes    red    wet         15     25       15
## 2      yes     no     no yellow    dry          3    150       50
## 3       no    yes     no   pink    wet          1    150       50
## 4       no     no    yes    red normal         16    125       50
## 5       no    yes     no   blue normal          2     20       15
## 6       no    yes     no    red    wet         12     50       40
## 7       no     no     no    red    wet         13     40       20
## 8       no     no    yes yellow normal          7    100       15
## 9      yes    yes     no   pink    dry          4     25       15
## 10     yes    yes     no   blue normal         14    100       60
## 11     yes    yes    yes   blue    wet          8     45       10
## 12     yes    yes    yes  white normal          9     90       25
## 13     yes    yes     no  white normal          6     20       10
## 14     yes    yes    yes    red normal         11     80       30
## 15     yes     no     no   pink normal         10     40       20
## 16     yes     no     no    red normal         18    200       60
## 17     yes     no     no yellow normal         17    150       60
## 18      no     no    yes yellow    dry          5     25       10
```

(b) Create frequency bar charts for the `color` and `soil` variables, using best practices for the order of the bars.

```
library(ggplot2)
df1 <- within(df1, color <- factor(color, levels = names(sort(table(color), decreasing = TRUE))))
p1b1 <- ggplot(df1, aes(color)) +
  geom_bar() +
  ggtitle("Frequency Bar charts for 'color'")
p1b1
```




```
library(ggplot2)
df1 <- within(df1, soil <- factor(soil, levels = names(sort(table(soil), decreasing = TRUE))))
p1b2 <- ggplot(df1, aes(soil)) +
  geom_bar() +
  ggtitle("Frequency Bar charts for 'soil'")
p1b2
```

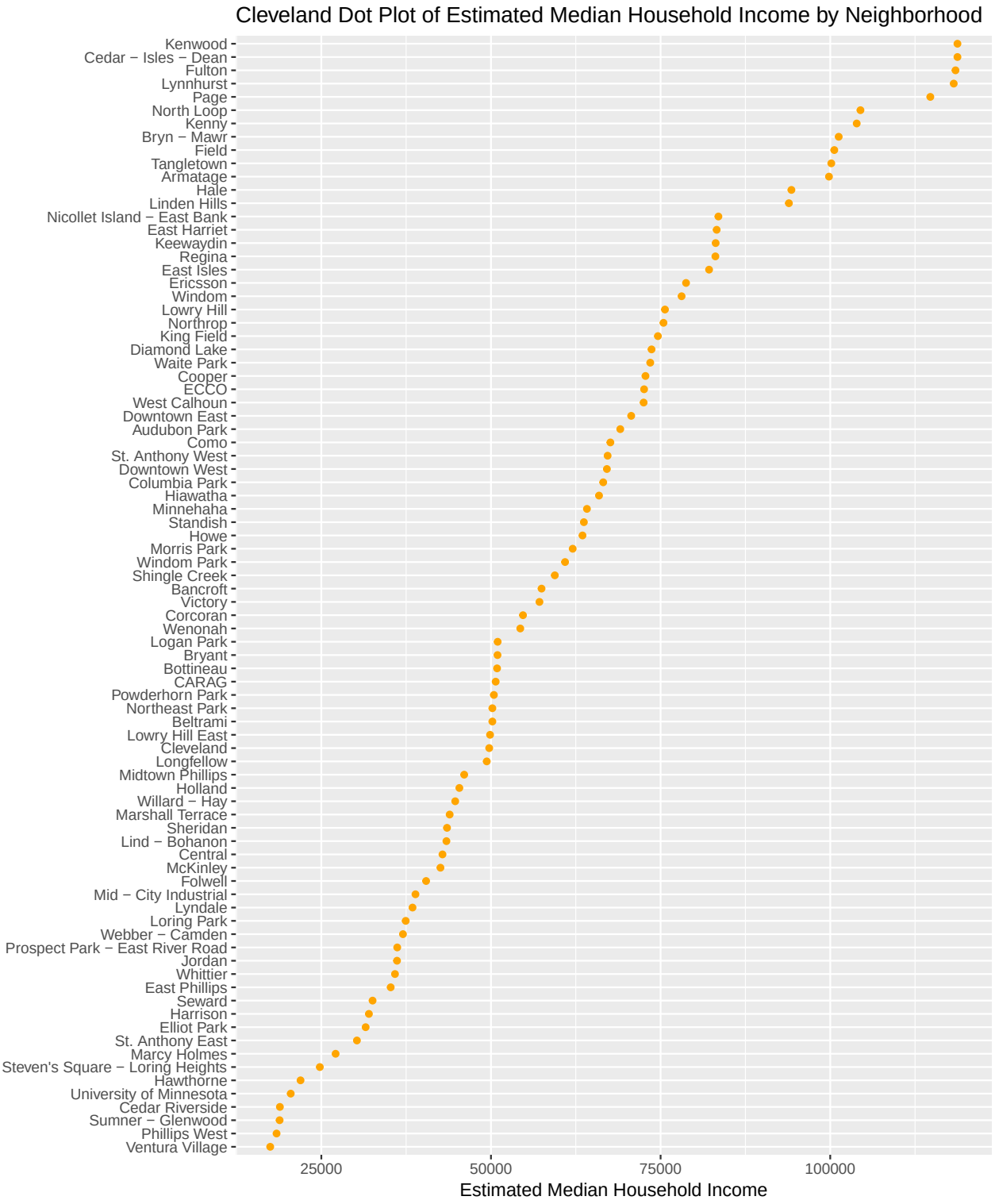


2. Minneapolis

Data: `MplsDemo` dataset in **carData** package

(a) Create a Cleveland dot plot showing estimated median household income by neighborhood.

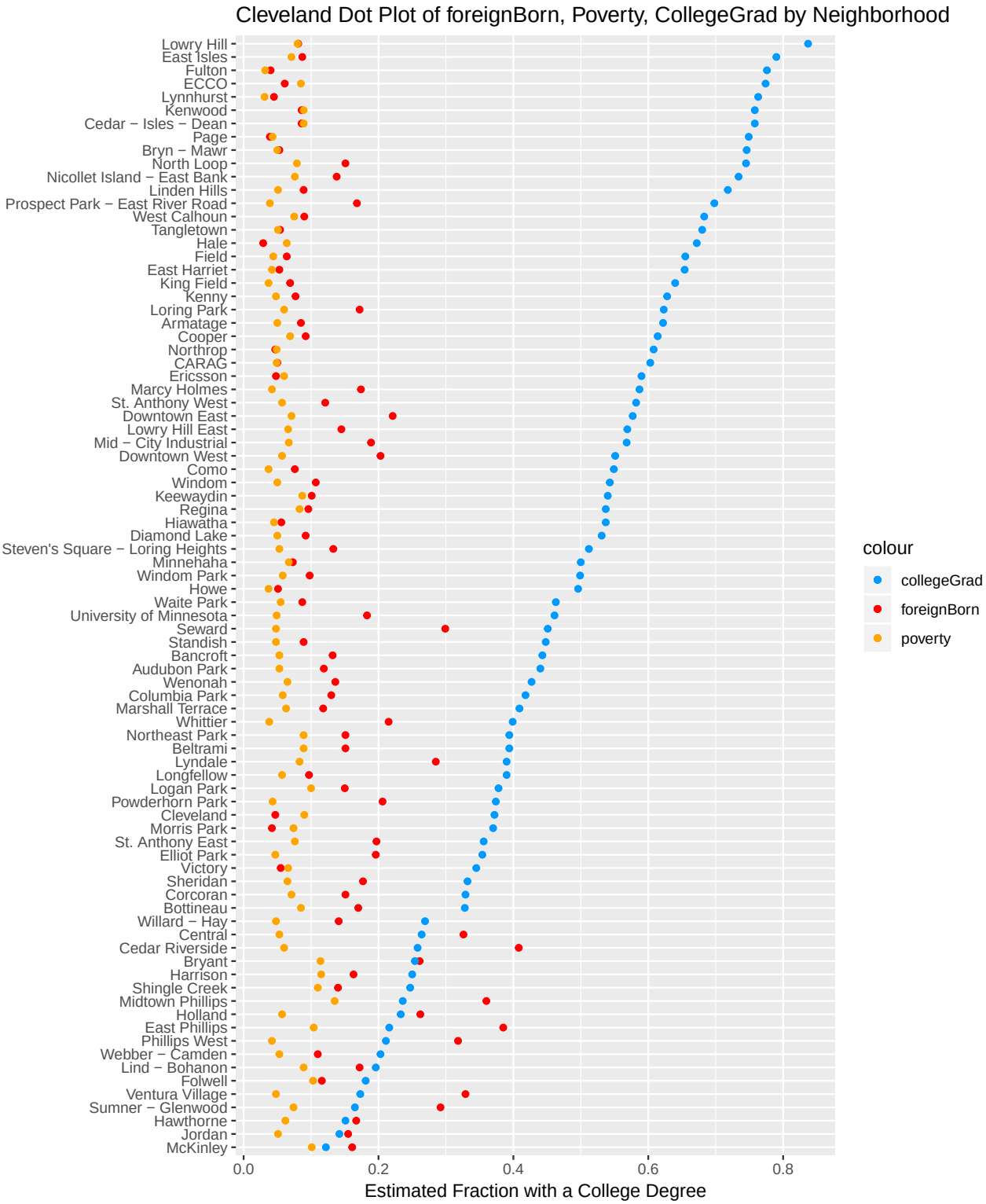
```
library(carData)
df2 <- MplsDemo
p2a <- ggplot(df2, aes(x = hhIncome, y = reorder(neighborhood, hhIncome))) +
  geom_point(color = "orange") +
  xlab("Estimated Median Household Income") + ylab("") +
  ggtitle("Cleveland Dot Plot of Estimated Median Household Income by Neighborhood")
p2a
```



(b) Create a Cleveland dot plot *with multiple dots* to show percentage of 1) foreign born, 2) earning less than twice the poverty level, and 3) with a college degree *by neighborhood.* Each of these three continuous variables should appear in a different color. Data should be sorted by college degree.

```
p2b <- ggplot(df2, aes(x = collegeGrad, y = reorder(neighborhood, collegeGrad))) +
  geom_point(aes(x = foreignBorn, color='foreignBorn')) +
```

```
  geom_point(aes(x = poverty, color='poverty')) +
  geom_point(aes(x = collegeGrad, color='collegeGrad')) +
  scale_colour_manual(values = c("#0099FF", "red","orange")) +
  xlab("Estimated Fraction with a College Degree") + ylab("") +
  ggtitle("Cleveland Dot Plot of foreignBorn, Poverty, CollegeGrad by Neighborhood")
p2b
```



(c) What patterns do you observe? What neighborhoods do not appear to follow these patterns?

- From the plot I observe that the estimated fraction with a college degree is somehow negatively related to the fraction of the population estimated to be foreign born. The exceptions of this pattern are Vleveland, Morris Park, Victory, Hawthorne, Jordan, McKinley.

- Also I think poverty does not have a clear relationship with collegeGrad. But we can indeed find a pattern that poverty fraction is almost less than foreignBorn fraction. The neighborhoods that do not follow this pattern are: ECCO, Kenwood, Cedar-Isles-Dean, Page, Hale, Northrop, Ericsson, Cleveland, Morris Park, Victory.

3. Taxis

Data: NYC yellow cab rides in June 2018, available here:

http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml

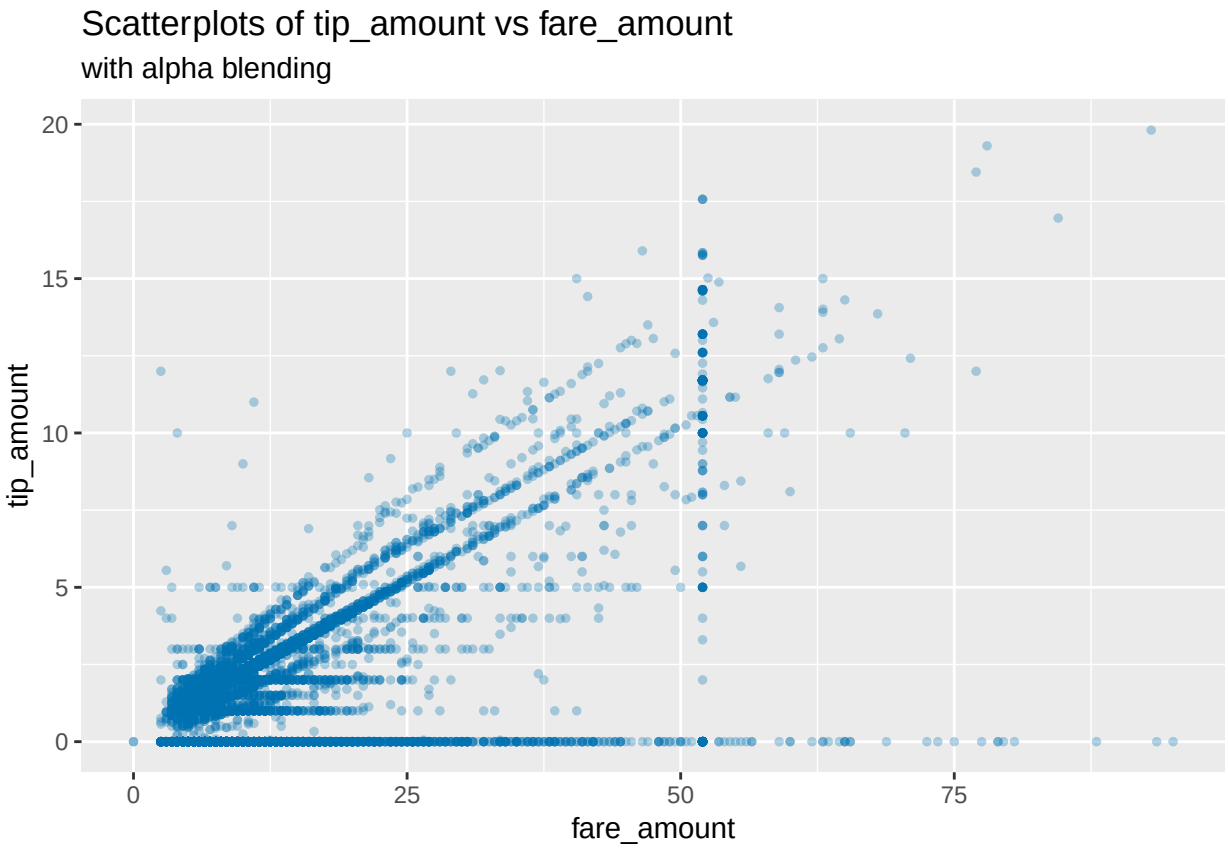
It's a large file so work with a reasonably-sized random subset of the data.

Draw four scatterplots of `tip_amount` vs. `far_amount` with the following variations:

(a) Points with alpha blending

```
# read data, and choose a random set of it
car.rides <- read.csv('yellow_tripdata_2018-06.csv')
sub1 <- subset(car.rides, fare_amount >= 0 & fare_amount < 100)
sub2 <- subset(sub1, tip_amount >= 0 & tip_amount < 20)
df3 <- sub2[sample(1: nrow(sub2), 10000), ]
```

```
p3a <- ggplot(df3, aes(x = fare_amount, y = tip_amount)) +
  geom_point(alpha=0.3, col = "#0072B2", size = 1) +
  ggtitle("Scatterplots of tip_amount vs fare_amount", subtitle = "with alpha blending")
p3a
```

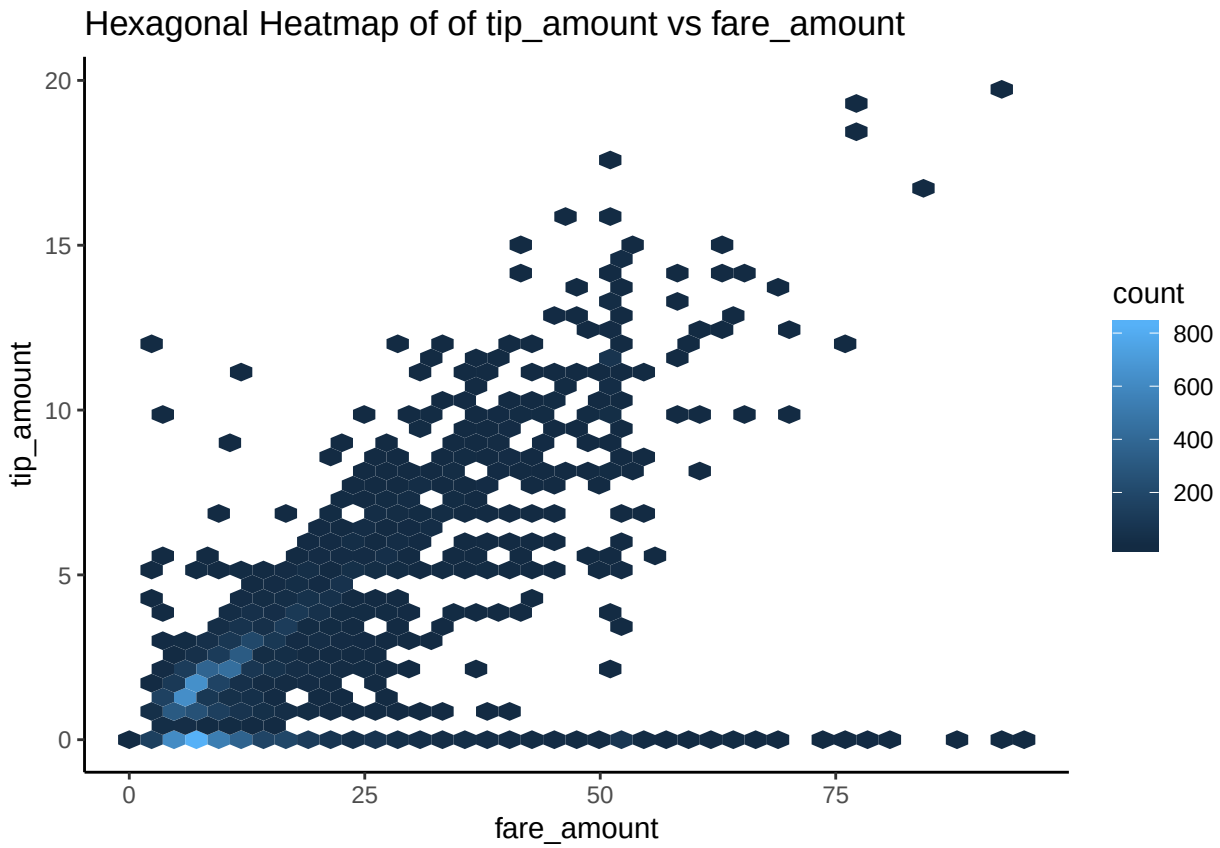


(b) Points with alpha blending + density estimate contour lines

```r
# We can contrain the range of fare_amount and tip_amount, and get a better view
p3b <- ggplot(df3, aes(x = fare_amount, y = tip_amount)) +
  geom_point(alpha=0.3, col = "#0072B2", size = 1) +
  geom_density_2d(col = "#CC79A7", size = 0.7) +
  coord_cartesian(xlim = c(0, 40), ylim = c(0, 7)) +
  ggtitle("Scatterplots of tip_amount vs fare_amount", subtitle = "with Alpha Blending and Density Esti
p3b
```

(c) Hexagonal heatmap of bin counts

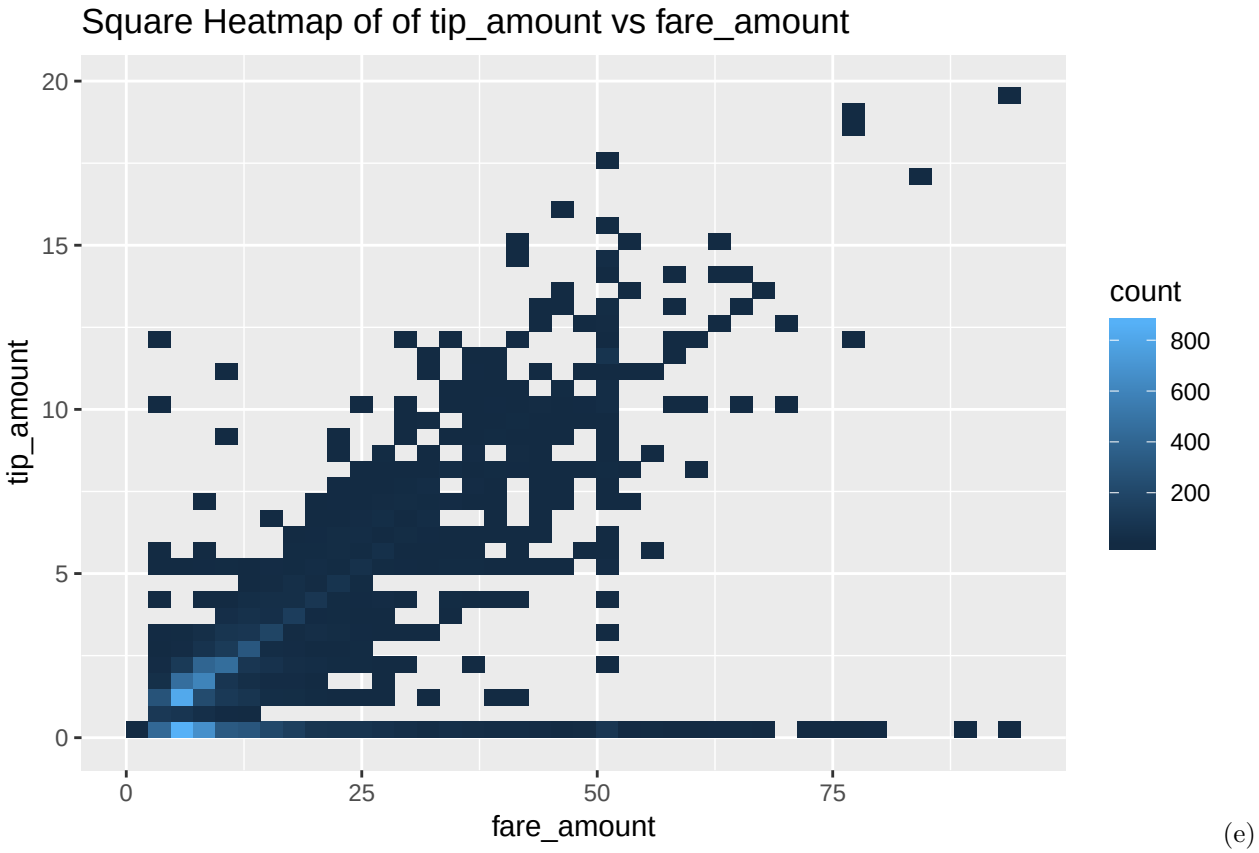
```r
p3c <- ggplot(df3, aes(x = fare_amount, y = tip_amount)) +
  geom_hex(bins = 40) +
  ggtitle("Hexagonal Heatmap of of tip_amount vs fare_amount") +
  theme_classic()
p3c
```



(d) Square heatmap of bin counts

For all, adjust parameters to the levels that provide the best views of the data.

```
p3d <- ggplot(df3, aes(x = fare_amount, y = tip_amount)) +
  geom_bin2d(bins = 40) +
  ggtitle("Square Heatmap of of tip_amount vs fare_amount") +
  theme_grey()
p3d
```



(e) Describe noteworthy features of the data, using the "Movie ratings" example on page 82 (last page of Section 5.3) as a guide.

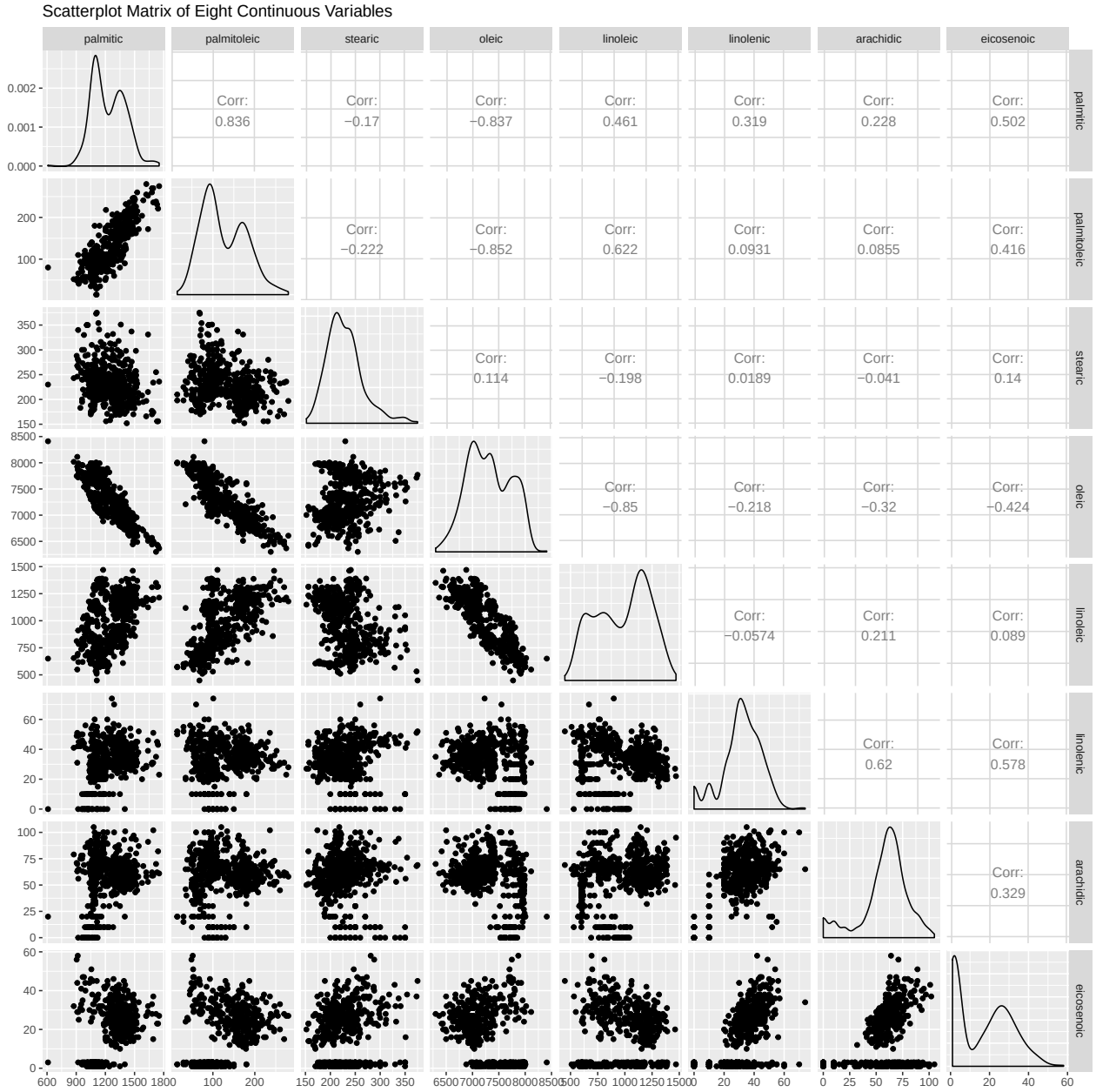
- From the scatterplots, we find many people pay extremely low tips or even no tips no matter what their fare amount is.
- Except those who pays no tips, we may find from the scatterplots that tips are generally in proportional to fares. It shows that most people tend to pay tips as one particular percentage of the fair amount.
- From the heatmap, we can easily notice that most of trips have very low tips and fares (less than 20). This shows that people's demand for taxi trips are mainly for short distance trips.
- From the scatterplots, we find that people love to pay tips of integers, especially \$5 tips. In the plot, we can find there are people paying \$5 tips whatever their fairs are.
- We also find for the fair amount around \$52, people gives a wide range of tips: from 0 to 20.
- We can also notice that for the trip with relatively low fare price (around \$3 - \$10), people pay tips at will, that is not quite in proportion to fare price.
- Finally we may find that for trips with quite high fares, there are two situations: either pay tips in proportion to fares, nor pay almost no tips. There are seldom people pay tips betweeen them.

4. Olive Oil

Data: `olives` dataset in **extracat** package

(a) Draw a scatterplot matrix of the eight continuous variables. Which pairs of variables are strongly positively associated and which are strongly negatively associated?

```
library(extracat)
library(GGally)
df4 <- olives[3:10]
p4a <- ggpairs(df4, title = "Scatterplot Matrix of Eight Continuous Variables")
p4a
```

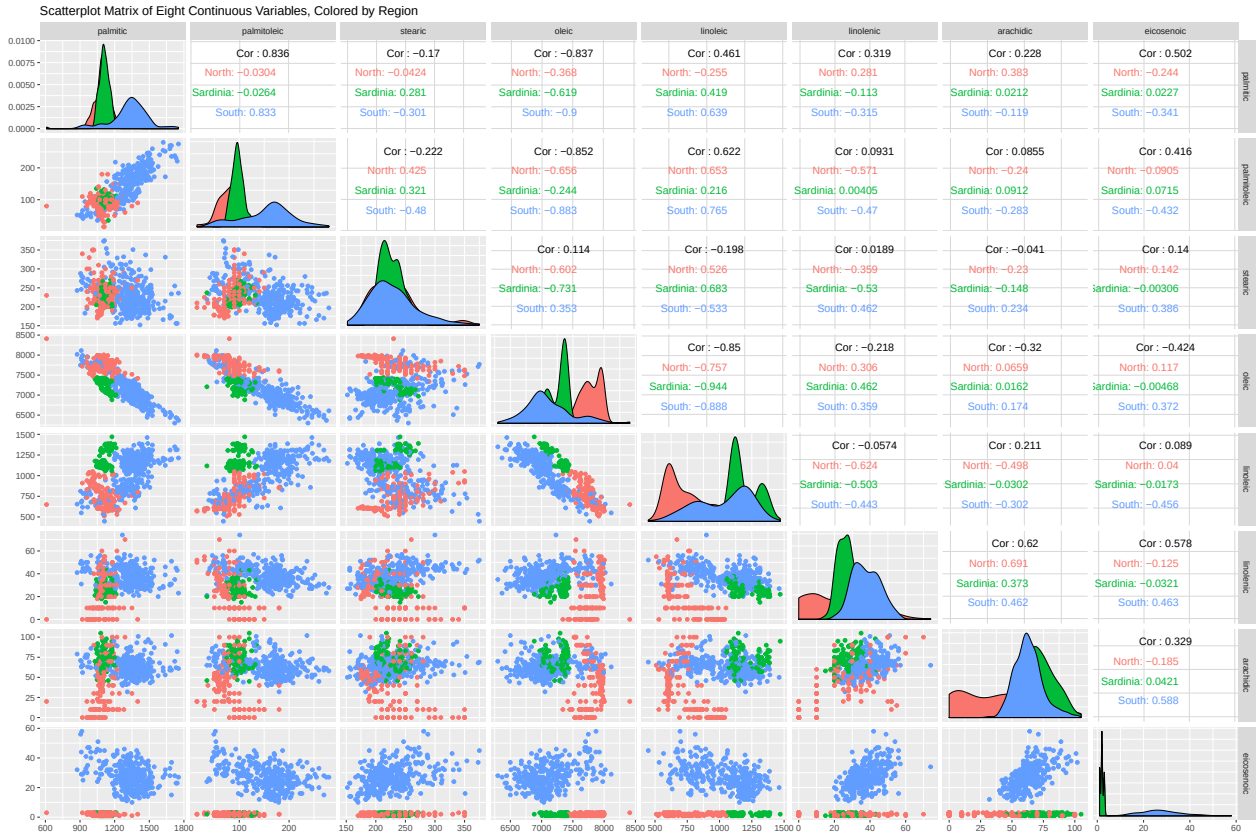


- According to the scatterplot matrix, if one variable systematically incremented by one variable , that is the plot is close to a line with positive gradient, then we may see them strongly positively linearly related. If the plot of two variables are similar to a line with negative gradient, then we may see them strongly negatively linearly related. Otherwise, there is not obvious linear relationshp between them.
- From the scatterplot matrix, it is easy to find that palmitic and palmitoletic are strongly positively associated. This can also be verified by the correlation coefficient, which is 0.836.
- We also notice that, palmitic and oleic, palmitoletic and oleic, oleic and linoleic are strongly netatively

associated, these facts can also be verified because their correlation coefficients are less than -0.8.

(b) Color the points by region. What do you observe?

```
library(extracat)
library(GGally)
df4 <- olives[3:10]
p4b <- ggpairs(olives,
               column = c(3:10),
               mapping = ggplot2::aes(colour = Region),
               title = "Scatterplot Matrix of Eight Continuous Variables, Colored by Region"
               )
p4b
```



- From the scatterplot matrix colored by region, we find that the strong positive/negative relationships we find in question 4(a) are mainly contributed by South region. Except that the relationships between oleic and lioleic are all significantly negatively linear in all the three areas.
- Also, in Sardina and North, eicoseoic is very close to 0, with no relationshp to any other variables.
- Overall, we can find the variables in South has much more association with each other than the variables in the other two regions. Especially for North region, there is almost no linear relationship between two variables.

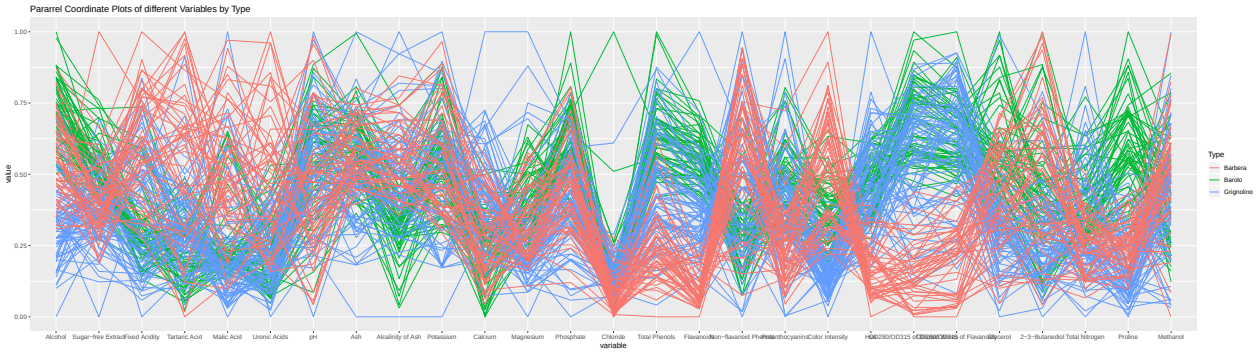
5. Wine

Data: `wine` dataset in **pgmm** package

(Recode the `Type` variable to descriptive names.)

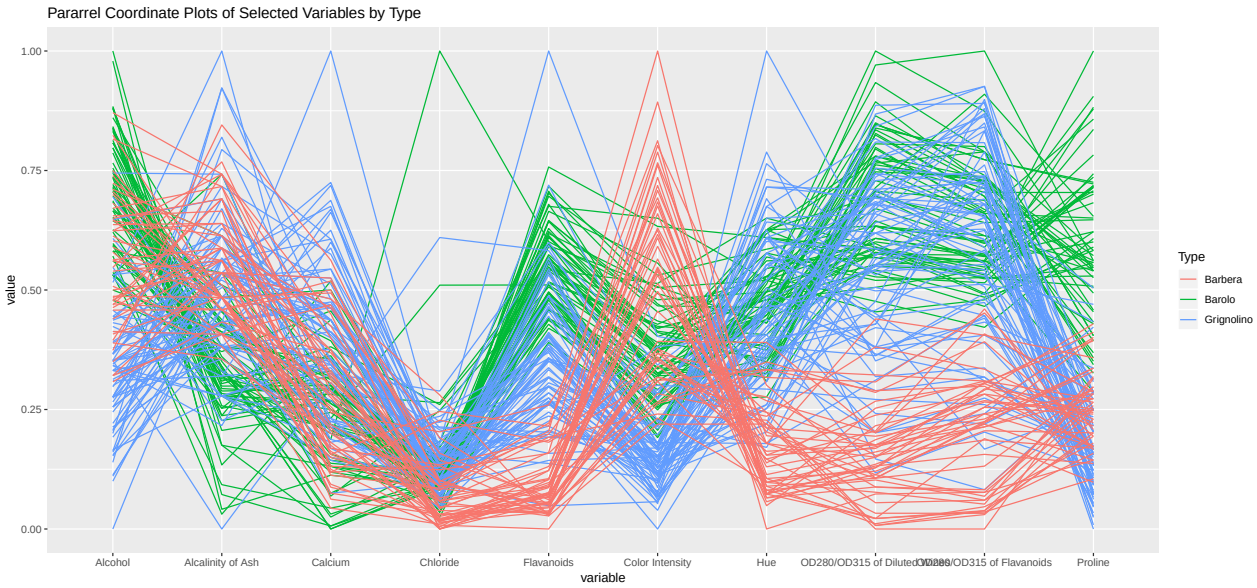
(a) Use parallel coordinate plots to explore how the variables separate the wines by `Type`. Present the version that you find to be most informative. You do not need to include all of the variables.

```
library(pgmm)
data(wine)
df5 <- wine
df5[df5$Type == 1, 'Type'] <- 'Barolo'
df5[df5$Type == 2, 'Type'] <- 'Grignolino'
df5[df5$Type == 3, 'Type'] <- 'Barbera'
p5a <- ggparcoord(df5,
                  groupColumn = 'Type',
                  columns = c(2: 28),
                  title = 'Pararrel Coordinate Plots of different Variables by Type',
                  scale = 'uniminmax')
p5a
```



- To further have a better view, I choose some of the variables that can better separte these three types of wine. And plot the plots as below:

```
p5a2 <- ggparcoord(df5,
                  groupColumn = 'Type',
                  columns = c(2, 10, 12, 15, 17, 20, 21, 22, 23, 27),
                  title = 'Pararrel Coordinate Plots of Selected Variables by Type',
                  scale = 'uniminmax')
p5a2
```



(b) Explain what you discovered.

- Based on the parallel coordinate plots, we dicover Calcium, Chloride and Flavanoids can separate three types of wine well. Because the lines of different types in these variables are gathered together, and separate from other types. For example, for Flavanoids, values of Barbera is the lowest, Grignolino ranks next, and Barolo is the highest.
- For variables Alcohol, Alcalinity and Proline, we can separate Barolo apart from the other two types because the green line is away from the other two kinds of lines.
- For variables Intensity, we can separate Grignolino from the other two types, because blue line is the lowest, can we can find a boundary to separate it to the others.
- For Hue, OD280/OD315 of Diluted Wines,OD280/OD315 of Flavanoids, we can separate Barbera easily from the other two types also by finding a boundary to tell them apart.