Homework #2

1. Flowers

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

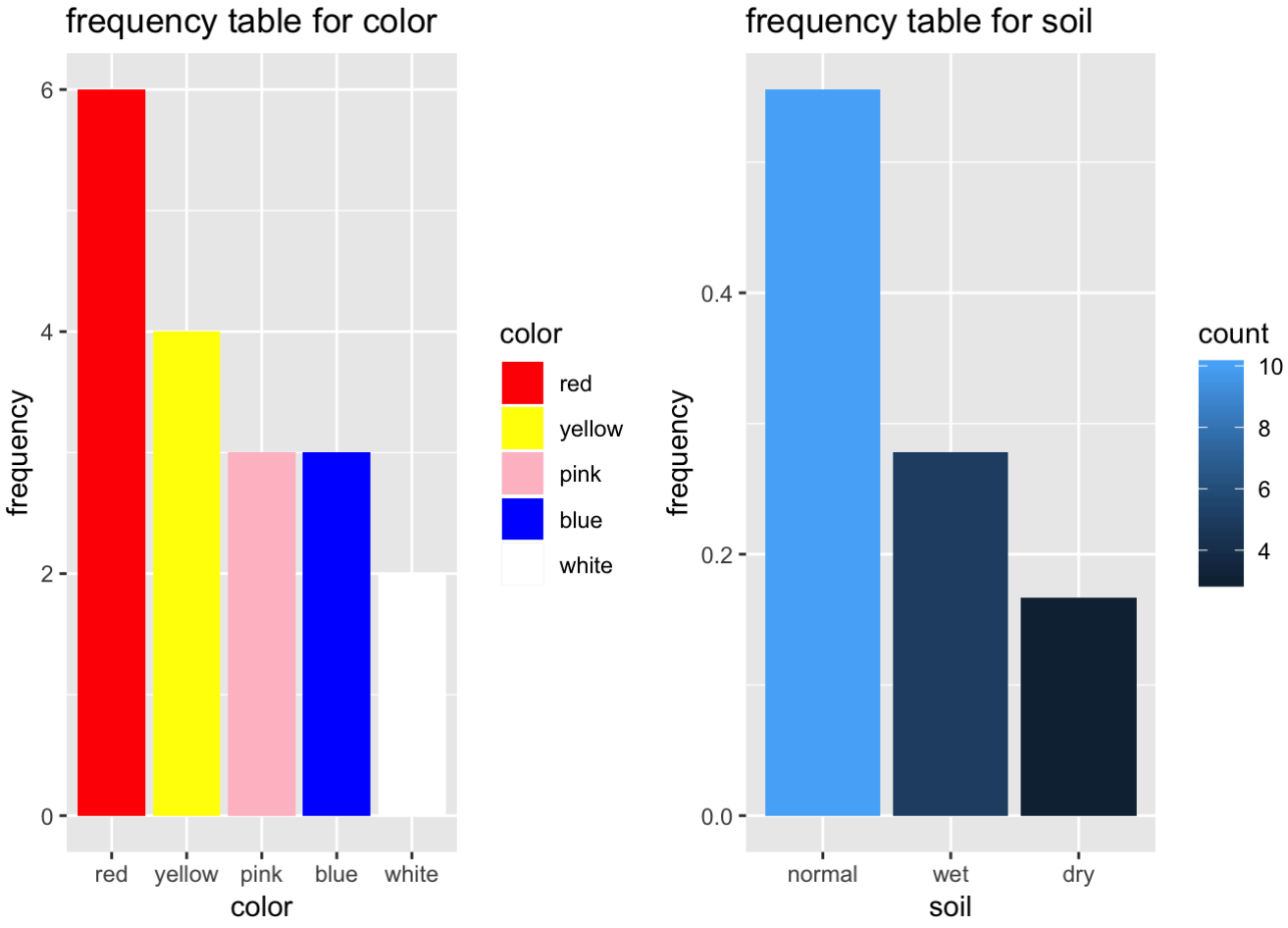
a. Rename the column names and recode the levels of categorical variables to descriptive names. For example, “V1” should be renamed “winters” and the levels to “no” or “yes”. Display the full dataset.

```
library(cluster)
library(dplyr)
library(plyr)
library(ggplot2)
library(gridExtra)
library(extracat)
names(flower) = c("winters","shadow","tuber","color","soil","preference","height","dista
nce")
flower$winters = mapvalues(flower$winters, from = c("0","1"),to = c("no","yes"))
flower$shadow = mapvalues(flower$shadow,from = c("0","1"), to = c("no","yes"))
flower$tuber = mapvalues(flower$tuber,from = c("0","1"), to = c("no","yes"))
flower$color = mapvalues(flower$color,from = c("1","2","3","4","5"), to =c("white","yell
ow","pink","red","blue"))
flower$soil = mapvalues(flower$soil,from = c("1","2","3"), to = c("dry","normal","wet"))
print.data.frame(flower)
```

```
##    winters shadow tuber  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.

```
flower1 <- within(flower, color<-factor(color, levels = names(sort(table(color),decreasi
ng = TRUE))))
g1 = ggplot(flower1, aes(x = flower1$color)) +
  geom_bar(aes(fill = flower1[,"color"])) +
  scale_fill_manual(values = levels(flower1[,"color"])) + labs(fill= "color") +
  xlab("color") +
  ylab("frequency") +
  ggtitle("frequency table for color")

flower2 <- within(flower, soil<-factor(soil, levels = names(sort(table(soil),decreasing
 = TRUE))))
g2 = ggplot(flower2, aes(x = flower2$soil, fill = ..count..)) +
  geom_bar(aes(y = (..count..)/sum(..count..))) +
  xlab("soil") +
  ylab("frequency") +
  ggtitle("frequency table for soil")
grid.arrange(g1,g2,ncol= 2)
```

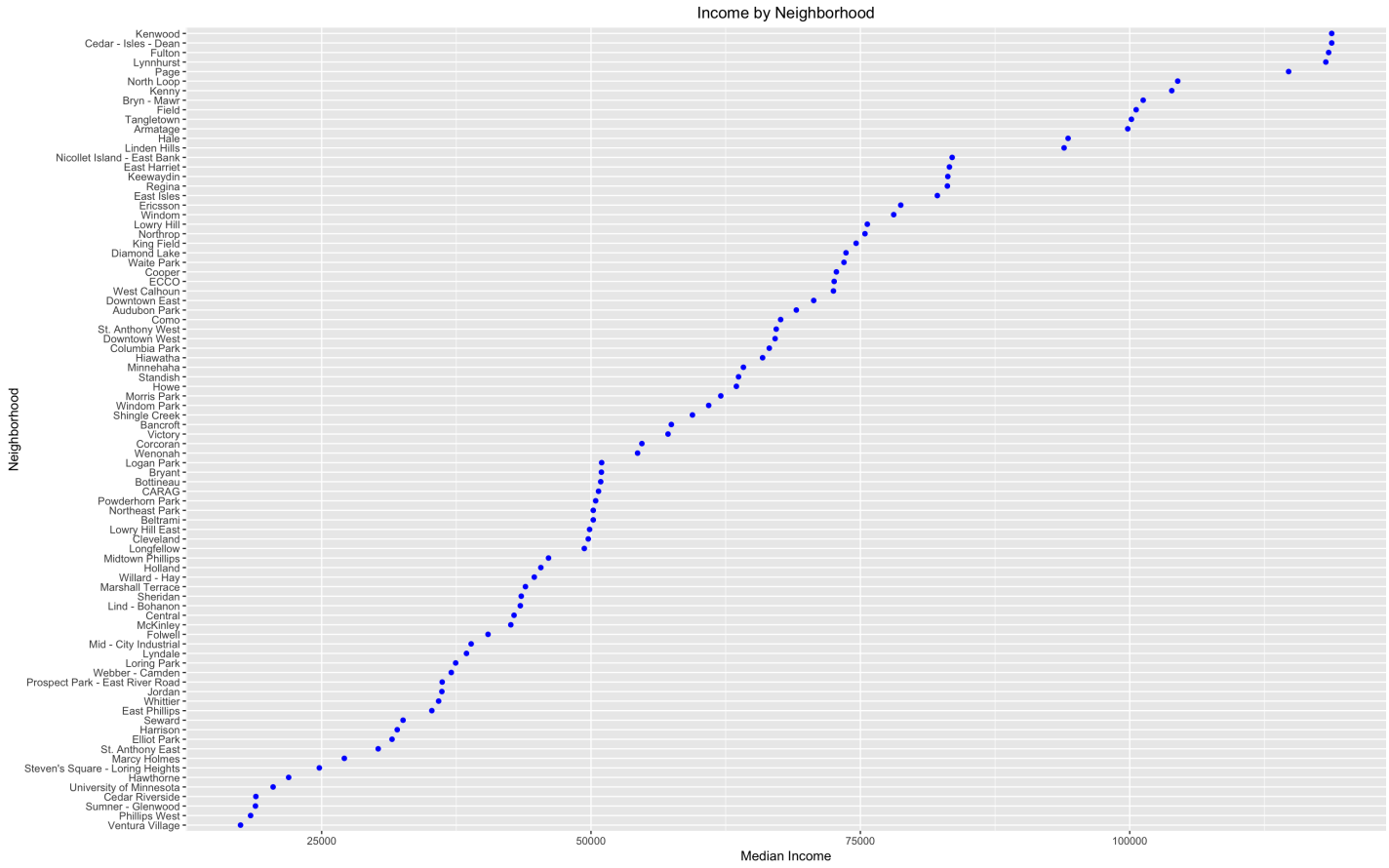


2. Minneapolis

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

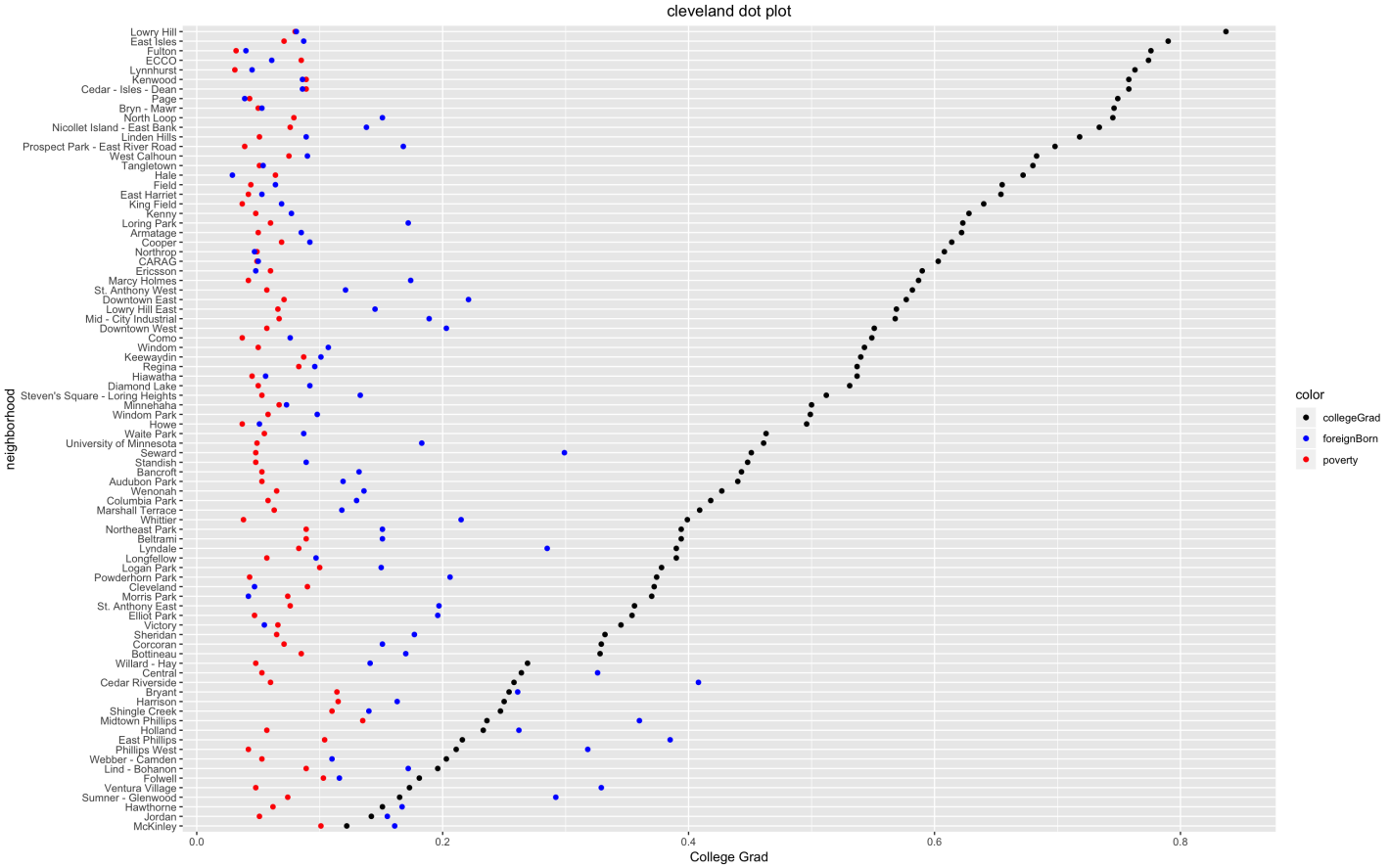
a. Create a Cleveland dot plot showing estimated median household income by neighborhood.

```
library(carData)
g3 = ggplot(MplsDemo, aes(x = hhIncome,y = reorder(neighborhood,hhIncome))) +
  geom_point(color = "blue") +
  xlab("Median Income") +
  ylab("Neighborhood") +
  theme(plot.title = element_text(hjust = 0.5)) +
  ggtitle("Income by Neighborhood")
g3
```



b. Create a Cleveland dot plot to show percentage of foreign born, earning less than twice the poverty level, and with a college degree in different colors. Data should be sorted by college degree.

```
MplsDemoV1 = MplsDemo %>%
  select(neighborhood,foreignBorn,poverty,collegeGrad)
g2b = ggplot(MplsDemoV1) +
  geom_point(aes(x = collegeGrad,y = reorder(neighborhood,collegeGrad, fun = median),gro
up = neighborhood, color = "black")) +
  geom_point(aes(x = poverty, y = neighborhood,color = "red")) +
  geom_point(aes(x = foreignBorn, y = neighborhood,color = "blue")) +
  xlab("College Grad") +
  ylab("neighborhood") +
  scale_colour_manual(name = 'color', values = c('black' = 'black','blue' = 'blue','red'
 = 'red'),labels = c('collegeGrad','foreignBorn','poverty')) +
  ggtitle("cleveland dot plot") +
  theme(plot.title = element_text(hjust =0.5))
g2b
```



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

For the neighborhood with low fraction of college degree, the fraction of foreignBorn is generally higher though the trend is fairly weak.There are several outliers such as Lyndale and Seward.Yet the poverty seems doesn't vary according to the fraction of foreignborn and fraction of those who have college degree.

3. Taxis

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

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

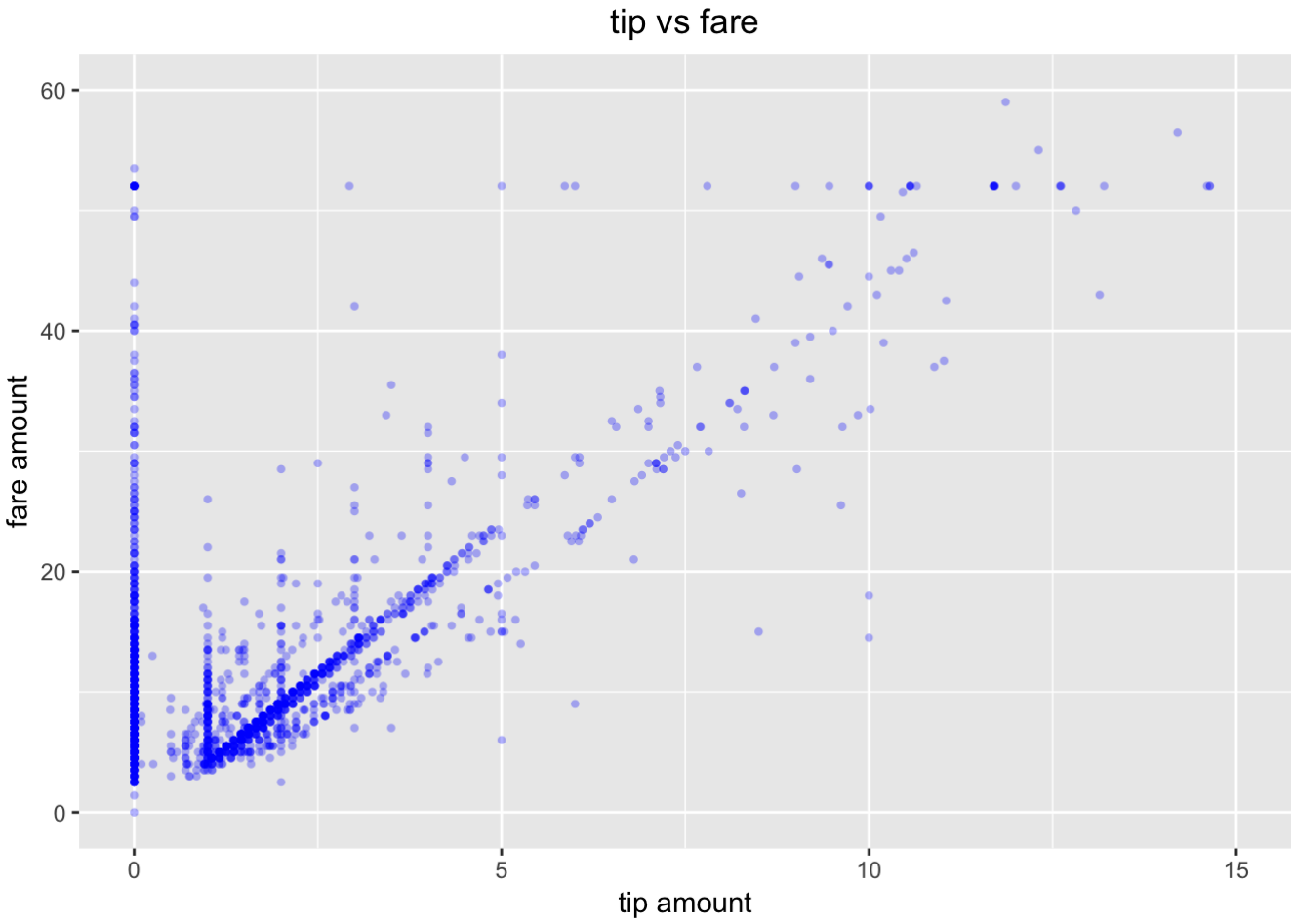
It's a large file so work with a reasonably-sized random subset of the data.

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

```
data = read.csv("yellow_tripdata_2018-06.csv")
df = data[sample(nrow(data),2000),]
```

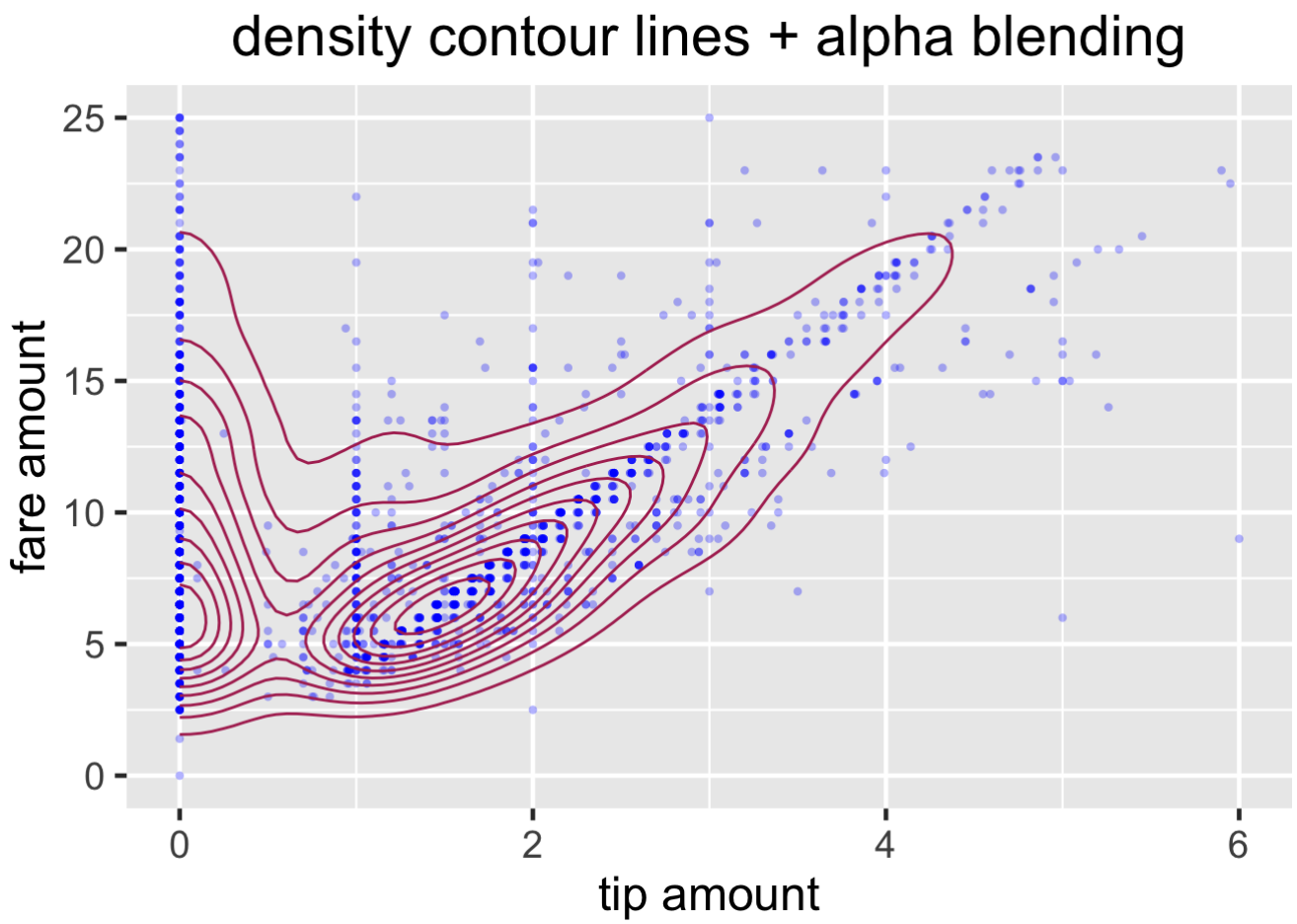
a. Points with alpha blending

```
ggplot(df,aes(x = df$tip_amount, y = df$fare_amount)) +
  geom_point(alpha = 0.3, color = "blue",stroke = 0) +
  xlab("tip amount") +
  ylab("fare amount") +
  ggtitle("tip vs fare") +
  xlim(0,15)+
  ylim(0,60)+
  theme(plot.title = element_text(hjust = 0.5))
```



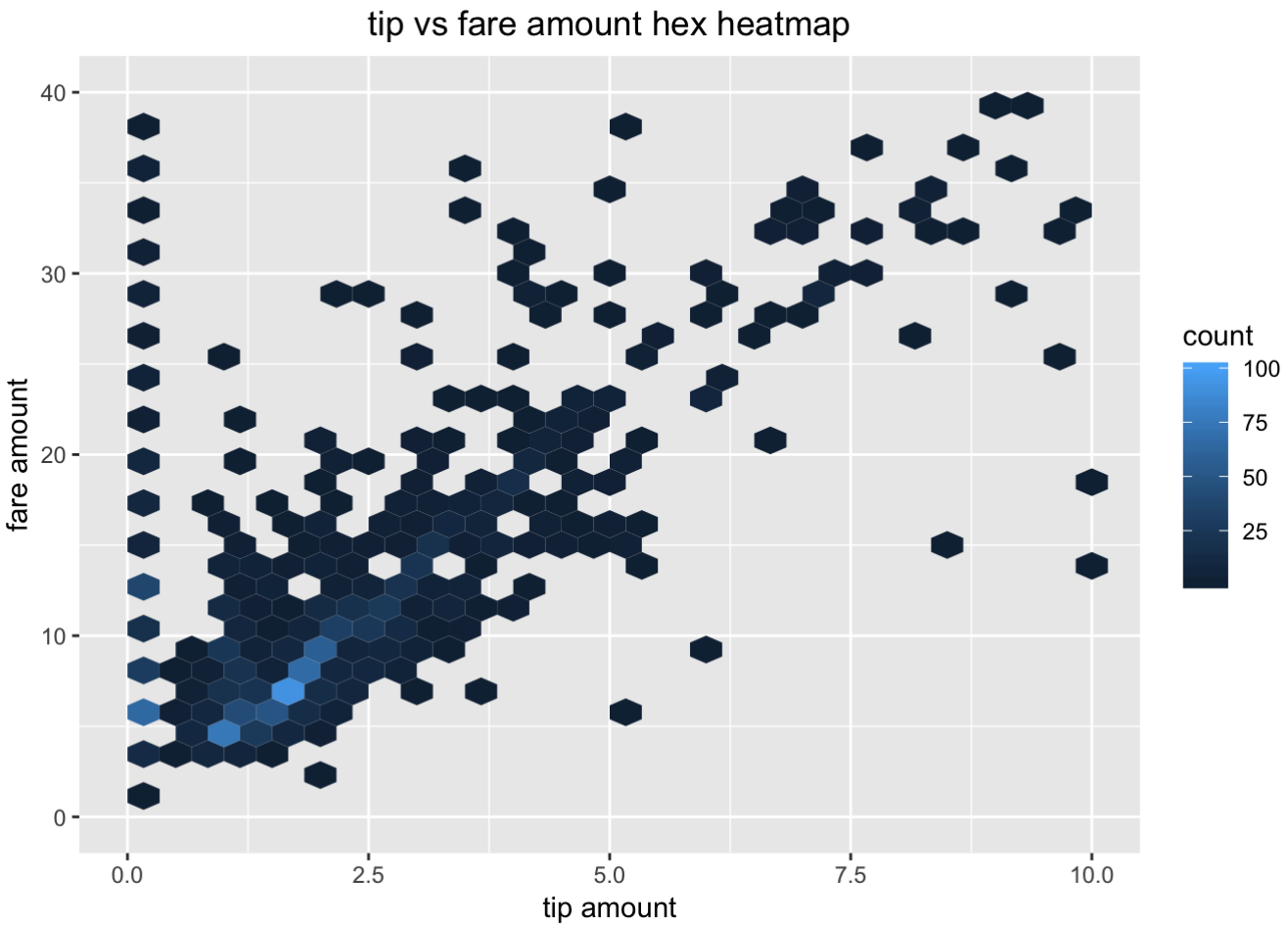
b. Points with alpha blending + density estimate contour lines

```
g3b = ggplot(df, aes(x = df$tip_amount,y=df$fare_amount)) +
  geom_point(alpha = 0.3, color = "blue",stroke = 0) +
  geom_density_2d(color = "maroon") +
  theme_grey(18) +
  xlim(0,6) +
  ylim(0,25) +
  xlab("tip amount") +
  ylab("fare amount") +
  ggtitle("density contour lines + alpha blending") +
  theme(plot.title = element_text(hjust = 0.5))

g3b
```



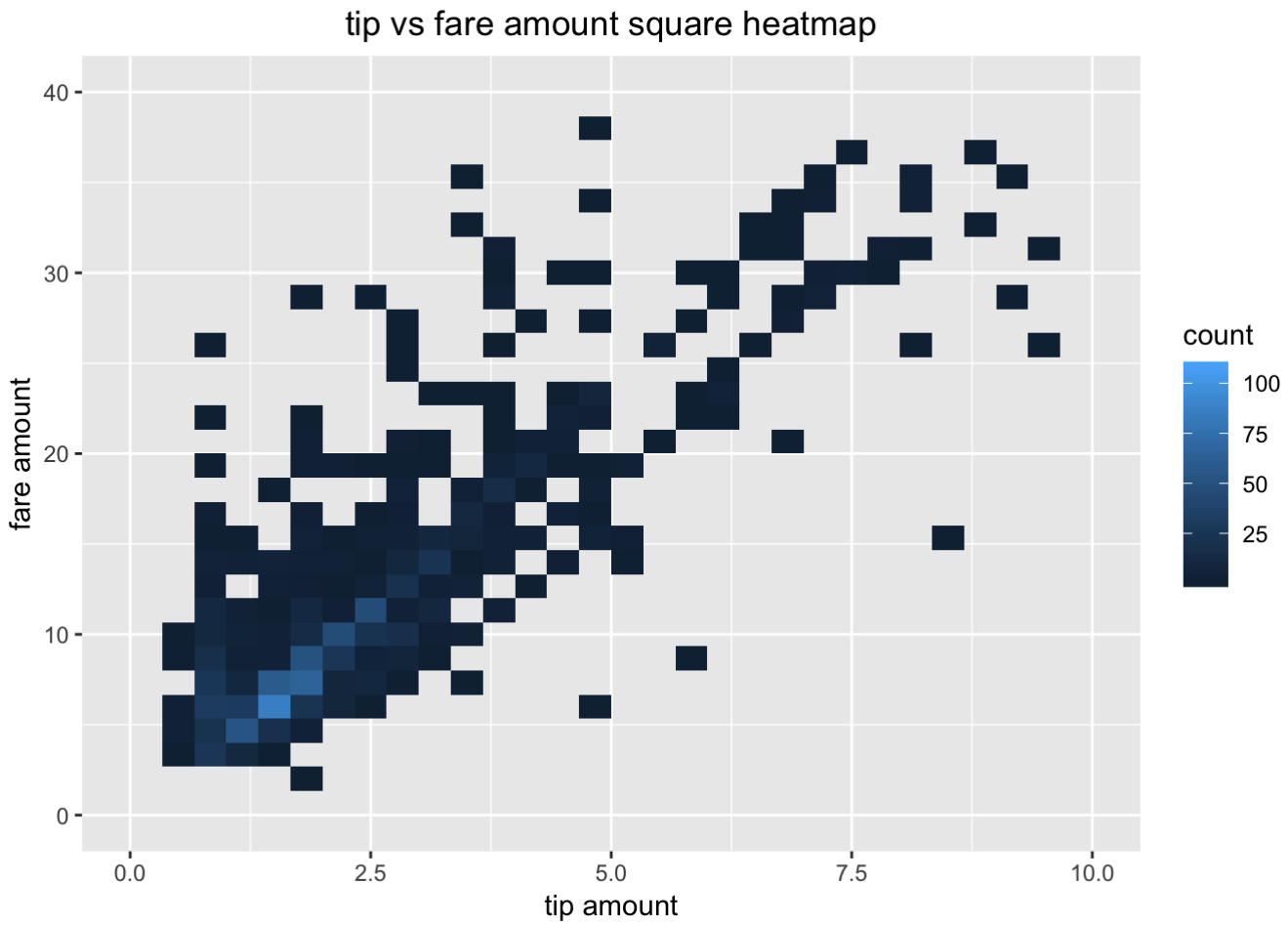
c. Hexagonal heatmap of bin counts

```
g3c <- ggplot(df, aes(x = df$tip_amount,y = df$fare_amount)) +
  geom_hex() +
  xlim(0,10) +
  ylim(0,40) +
  xlab("tip amount") +
  ylab("fare amount") +
  ggtitle("tip vs fare amount hex heatmap") +
  theme(plot.title = element_text(hjust = 0.5))

g3c
```



d. Square heatmap of bin counts

```
g3c <- ggplot(df, aes(x = df$tip_amount,y = df$fare_amount)) +
  geom_bin2d() +
  xlim(0,10) +
  ylim(0,40) +
  xlab("tip amount") +
  ylab("fare amount") +
  ggtitle("tip vs fare amount square heatmap") +
  theme(plot.title = element_text(hjust = 0.5))

g3c
```



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

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

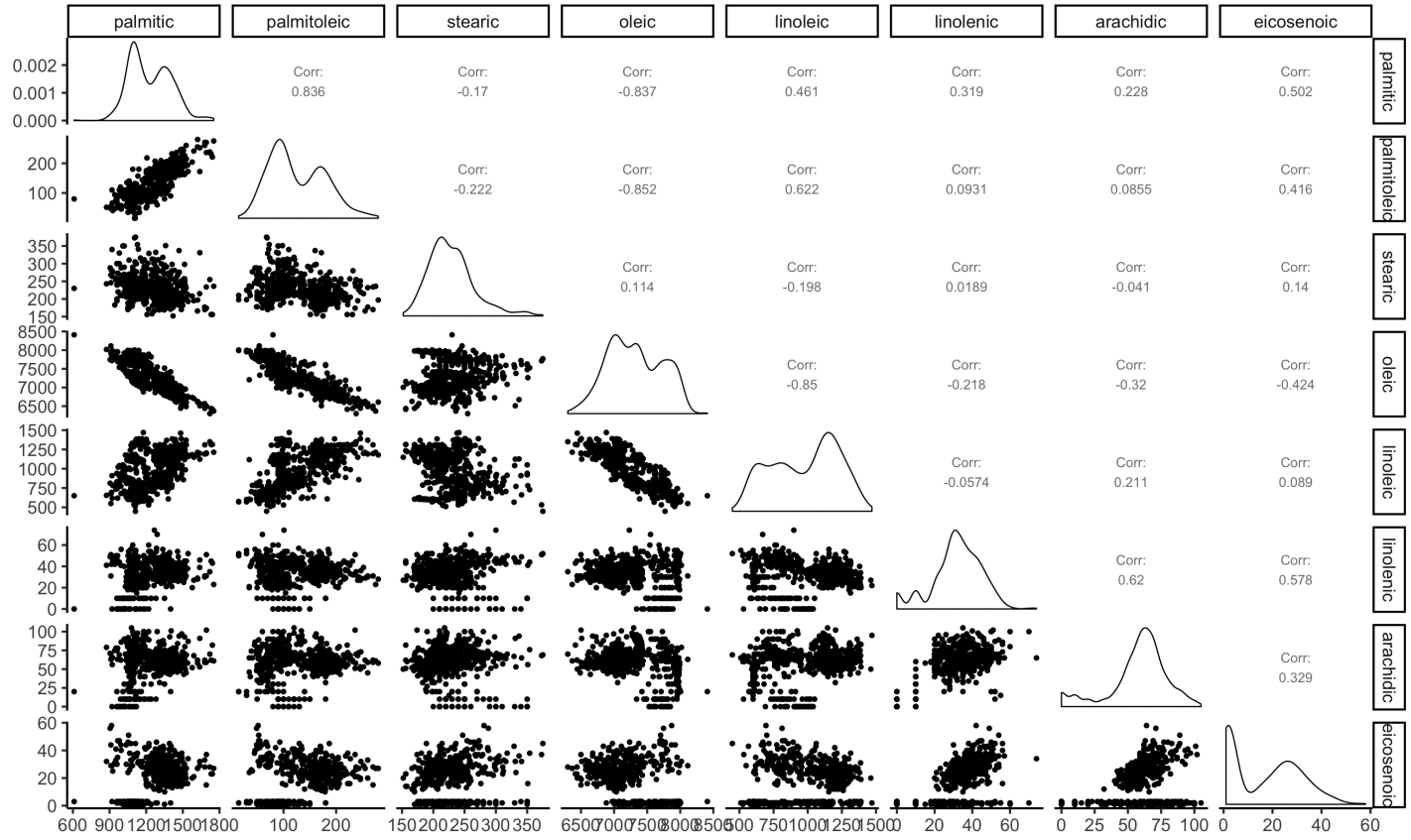
(a)There are many people doesn’t give tip as we can see a lots of data points line up at zero axis. and most people doesn’t give tip more than 5 dollars.(b)The distribution looks like a bimodel. There are two peaks one at zero tips and the other is about 1.7 dollars. From this graph, we can observe that most people don’t give tips more than 4 dollars. (c) The hex heat map further confirm that most people give tips at about 2 dollars because we can observe a cluster at tip amount 1.8 dollars and 5 dollar fare rate. However we can also observe a weak trend that higher fare rate might correspond to a higher tip amount. (d)In general, the majority of the data points fall with in about 6 dollar tip amount and 30 dollar fare amount. The light color indicate a cluster.For the fare amount that fall within 10 dollars, the tip amount is usually about 2 dollars.

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?

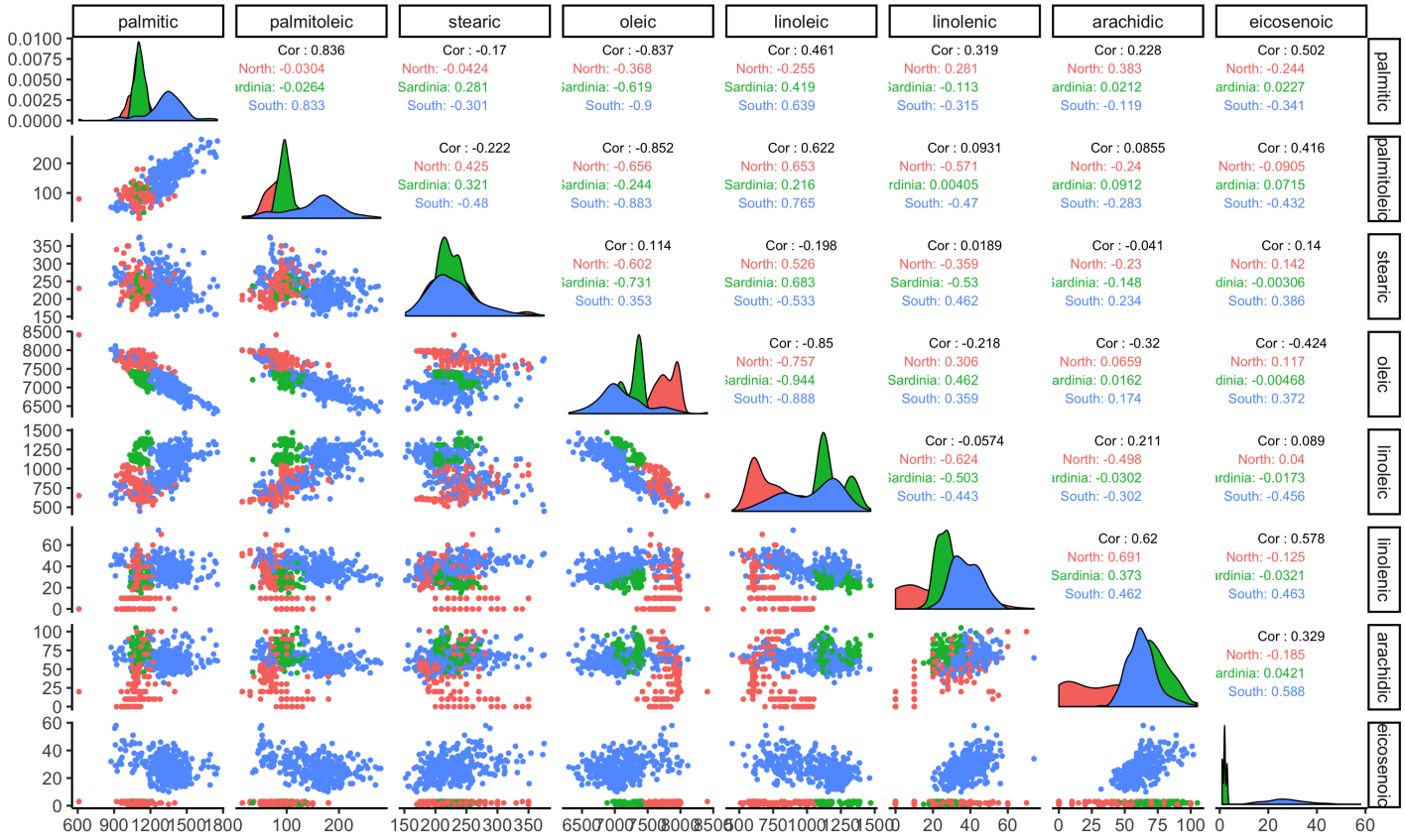
```
library(GGally)
columns = 3:(ncol(olives) - 1)
ggpairs(olives,columns = columns,title = '',axisLabels = "show",
        columnLabels = colnames(olives[, columns])) + theme_classic(18)
```



The top 3 pairs (palmitic,palmitoleic) ,(linoleic, palmitoleic),(arachidic,linolenic) have most positive colerration.The top 3 pairs (oleic,palmitic), (oleic, palmitoleic),(oleic,linoleic) have most negative colerration.

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

```
library(GGally)
columns = 3:(ncol(olives) - 1)
ggpairs(olives,columns = columns,title = '',axisLabels = "show",
        mapping=ggplot2::aes(colour =olives$Region),
        columnLabels = colnames(olives[, columns])) + theme_classic(18)
```



For olives that come from Sardinia and South, they contain almost zero eicosenoic fatty acid. For the strongly negatively correlated fatty acid pairs, olives from South contribute most to the negative correlation while the olives from Sardinia tend to cluster and the olives from North tend to be more spread. In general for the olives from Sardinia, the amount of their fatty acid is more tend to clustered from the figure showed above. The the fatty acid of olives from North are more likely to scatter in general from the graph above.

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)

fitb1 = lm(Type~., wine)
c1 = order(fitb1$coefficients)[1:4]
c2 =order(fitb1$coefficients, decreasing = TRUE)[1:3]
coll =c(c1,c2)
wine$Type <- mapvalues(wine$Type, from = c("1","2","3"), to = c("Barolo","Grignolino",
"Barvera"))
ggparcoord(wine, columns=c(17,18,21,7,16), groupColumn = 'Type') +
  ggtitle("Parallel coordinate for Wine type with 6 most prominent variable") +
  xlab("prominent variable") +
  theme(plot.title = element_text(hjust = 0.5))
```

b. Explain what you discovered.

From the graph, we can observe flavanoids for Barvera wine is much lower than that of Barolo and Grignolino while Barolo has the highest value.The value of non-flavanoid phenols are higher for barvera when compare of it against barolo while the value for grignolino spread sharply. The Hue value is lower fro barvera is lower yet hue value for grignolino spread widely again.The value of uronic acids is higher for barvera. For the value of total phenols, Baroloe has higher values than barvera and Grignolino. Throughout the five variables, grignolino are more spreadout than Barolo and Barvera which clearly clustered.