HW02

Yawen Han(yh3069)

10/7/2018

Problem 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.

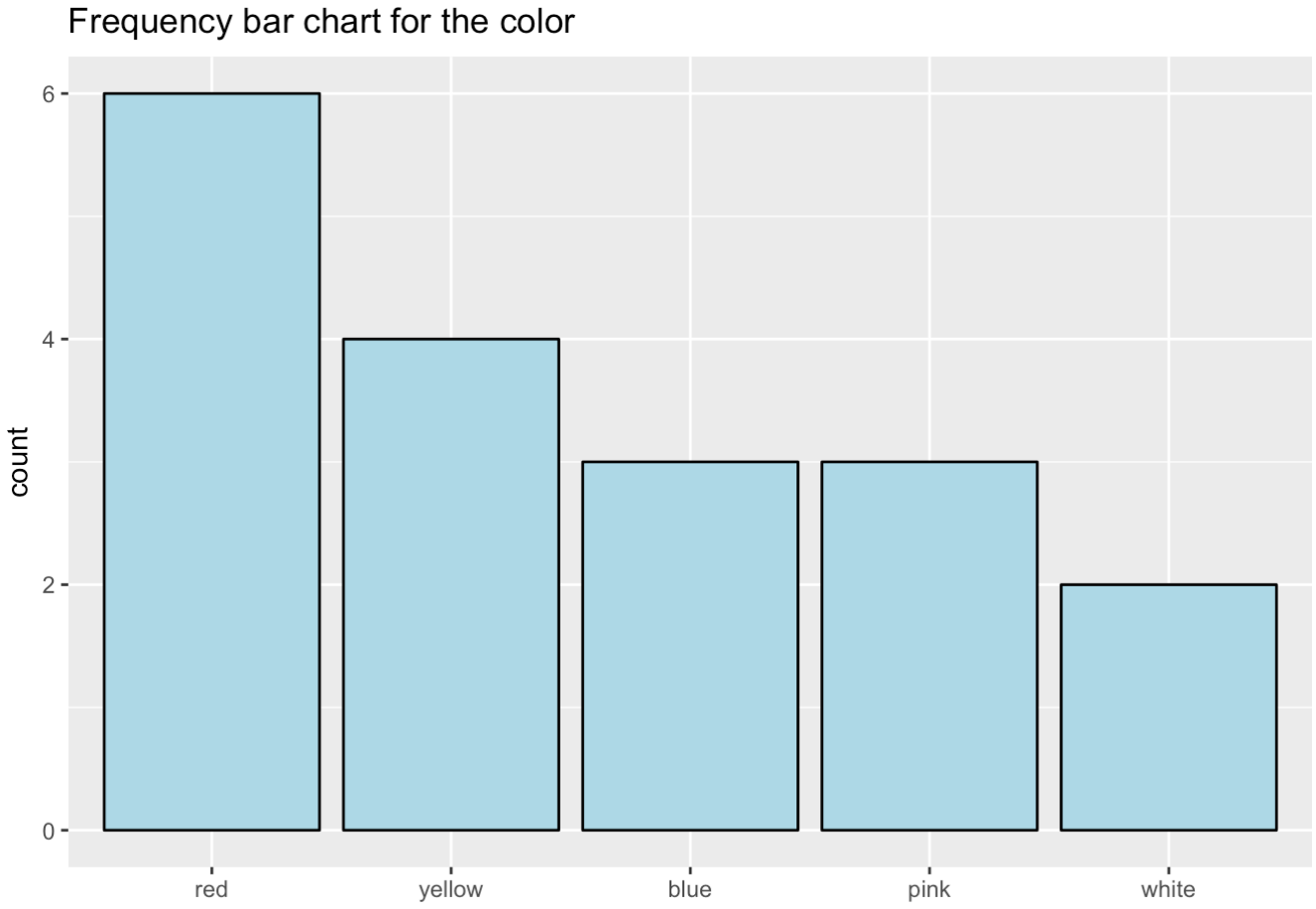
```
library(cluster)
library(ggplot2)
library("dplyr")
data(flower)
fa <- flower
# Rename the column names
fa <- rename(fa, winters=V1, shadow=V2, tubers=V3, color=V4, soil=V5, preference=V6, hei
ght=V7, distance=V8)
# recode the levels of categorical variables to descriptive names
fa$winters <- ifelse(fa$winters==0, "no", "yes")
fa$shadow <- ifelse(fa$shadow==0, "no", "yes")
fa$tubers <- ifelse(fa$tubers==0, "no", "yes")
fa$color <- ifelse(fa$color==1,"white",ifelse(fa$color==2,"yellow",ifelse(fa$color==3,"p
ink",ifelse(fa$color==4,"red","blue"))))
fa$soil <- ifelse(fa$soil==1,"dry",ifelse(fa$soil==2,"normal","wet"))
fa
```

```
##    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
```

After the renaming and recoding operations, the dataset is shown as above. As “preference”, “height” and “distance” are not categorical variables, their variables are not recoded.

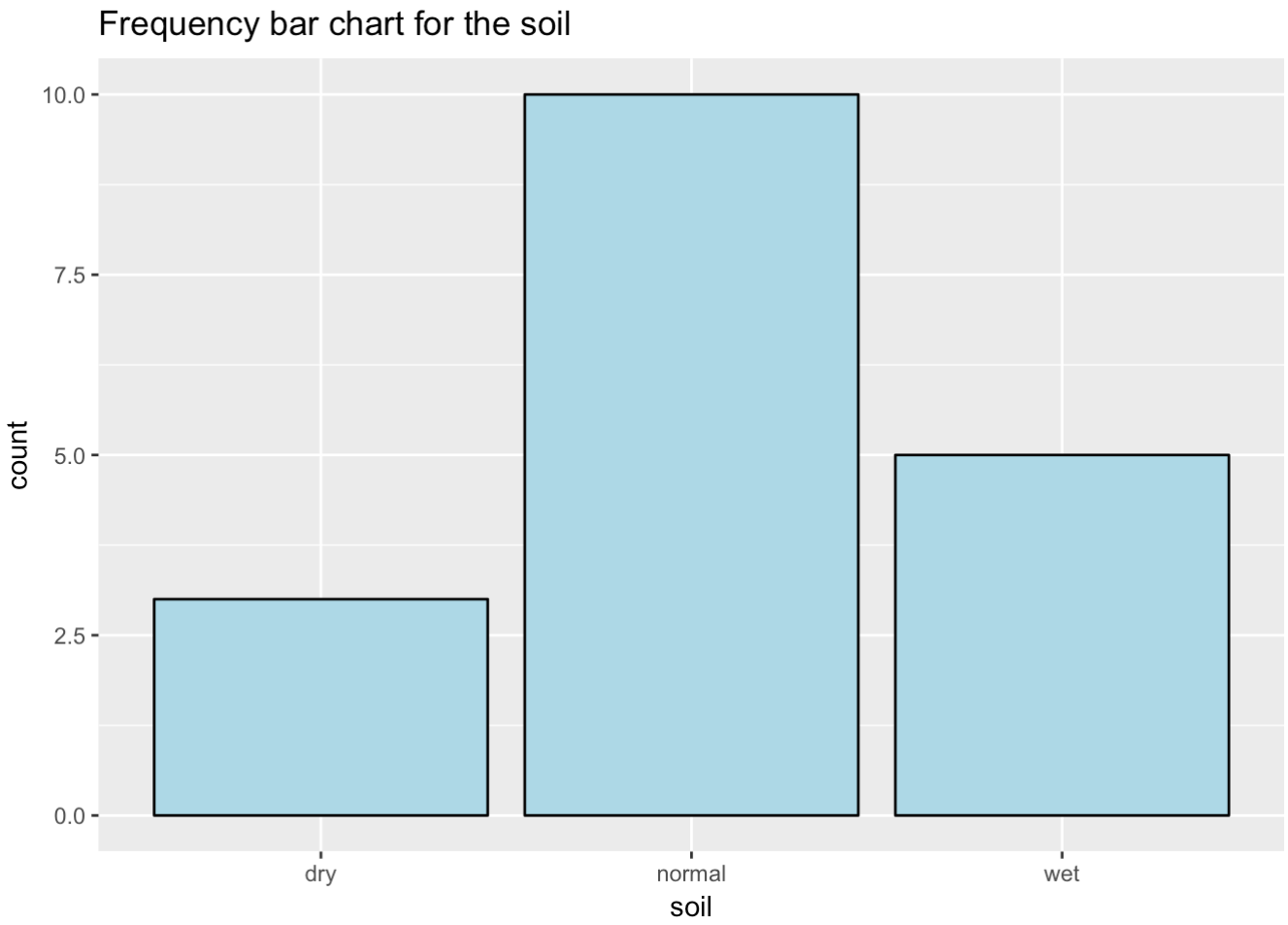
(b) Create frequency bar charts for the color and soil variables, using best practices for the order of the bars.

```
ggplot(fa, aes(x=reorder(color,color,function(x)-length(x))))+geom_bar(color="black",fil
l="lightblue")+ggtitle("Frequency bar chart for the color")+xlab("")
```



The frequency bar chart for the color is shown above. As “color” is a nominal variable with no fixed category order, the bar chart displays the bars with a decreasing count order for better observation.

```
ggplot(fa, aes(soil)) + geom_bar(color="black",fill="lightblue")+ggtitle("Frequency bar
 chart for the soil")
```



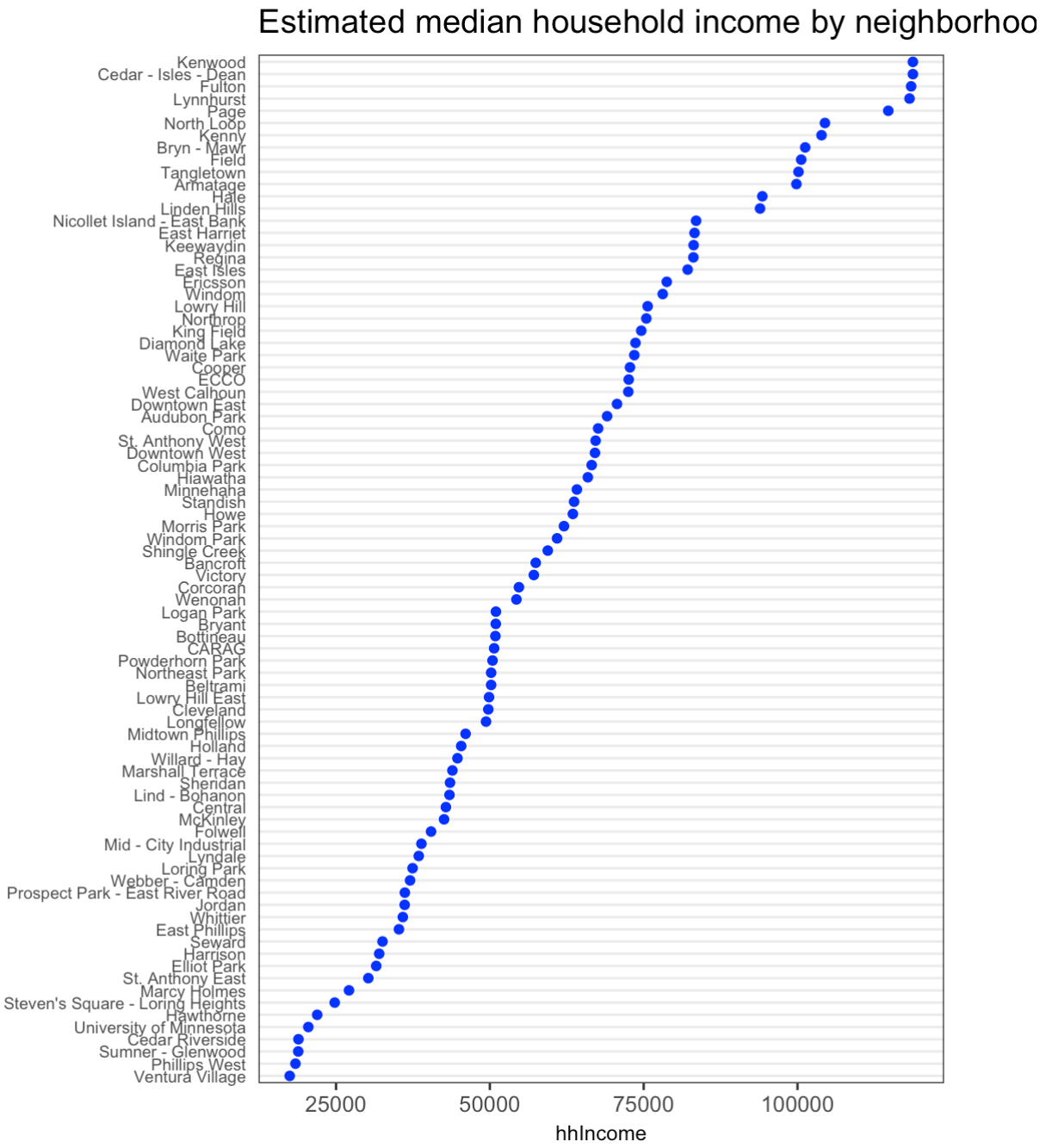
The frequency bar chart for the soil is shown above. As “soil” is an ordinal variable with a fixed category order from “day-normal-wet”, the bar chart displays the bars with its fixed order instead of the total count for better observation.

Problem 2 - Minneapolis

Data: MplsDemo dataset in carData package

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

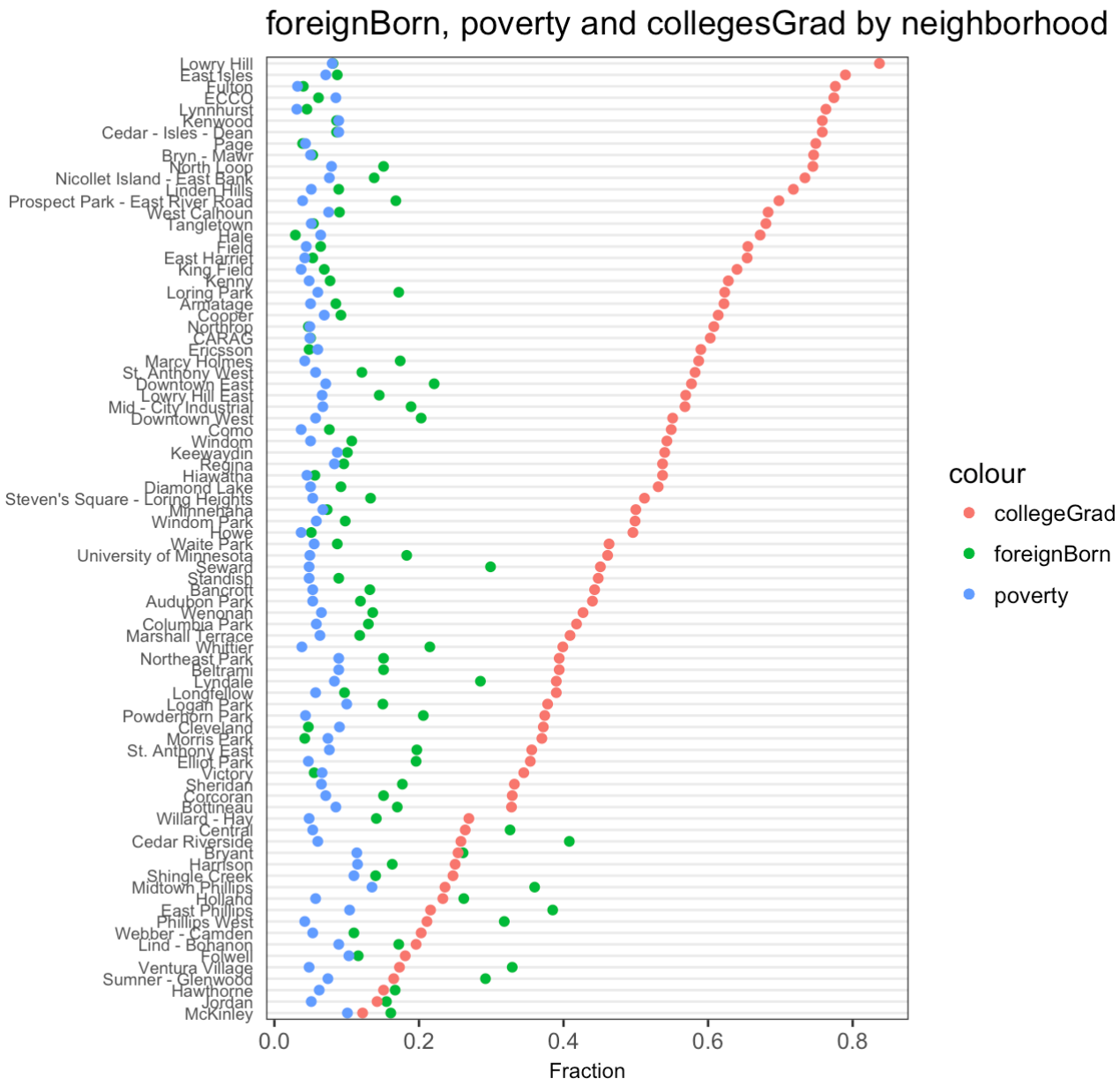
```
library(carData)
data("MplsDemo")
ma<-MplsDemo
## Cleveland Dot Plot theme
library(tidyverse)
theme_dotplot <- theme_bw(12) +
    theme(axis.text.y = element_text(size = rel(0.75)),
          axis.ticks.y = element_blank(),
          axis.title.x = element_text(size = rel(0.75)),
          panel.grid.major.x = element_blank(),
          panel.grid.major.y = element_line(size = 0.5),
          panel.grid.minor.x = element_blank(),
          aspect.ratio=1.5)
ma_cdp<-ggplot(ma, aes(x = hhIncome, y = fct_reorder(neighborhood, hhIncome))) +geom_poi
nt(color = "blue") + ylab("") +theme_dotplot + ggtitle("Estimated median household incom
e by neighborhood")
ma_cdp
```



The graph above is a Cleveland dot plot showing estimated median household income by neighborhood. The incomes of different neighborhood are shown with a decreasing order from top to bottom.

(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.

```
ggplot(ma,aes(y=fct_reorder(neighborhood,collegeGrad)))+geom_point(aes(x=foreignBorn,col
='foreignBorn')) +geom_point(aes(x=poverty,col="poverty"))+geom_point(aes(x=collegeGrad,
col="collegeGrad"))+xlab("Fraction") +ylab("")+ theme_dotplot + ggtitle("foreignBorn, po
verty and collegesGrad by neighborhood")
```



The graph above is a multiple dots of Cleveland dot plot showing the percentage of foreign-born, earning less than twice the poverty level, and the fraction with a college degree by neighborhood. Data are sorted by the college degree.

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

According to the Cleveland dot plot with multiple dots above, some patterns are observed below:

1. the “poverty fraction (estimated fraction earning less than twice the poverty level)” between different neighborhoods does not have an obvious difference, and stays at a relatively low level below 0.2;
2. the neighborhoods with a high “collegeGrad (estimated fraction with a college degree)” tends to have a low “foreignBorn rate (fraction of the population estimated to be foreign born)”, and vice versa.

Some neighborhoods appear do not follow the pattern(2), are the ones still keep a low “foreignBorn rate” with a relatively low “collegeGrad fraction”. For example: “Cleveland”, “Morris Park”, “Phillips West”, “Folwell”, etc.

Problem 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.

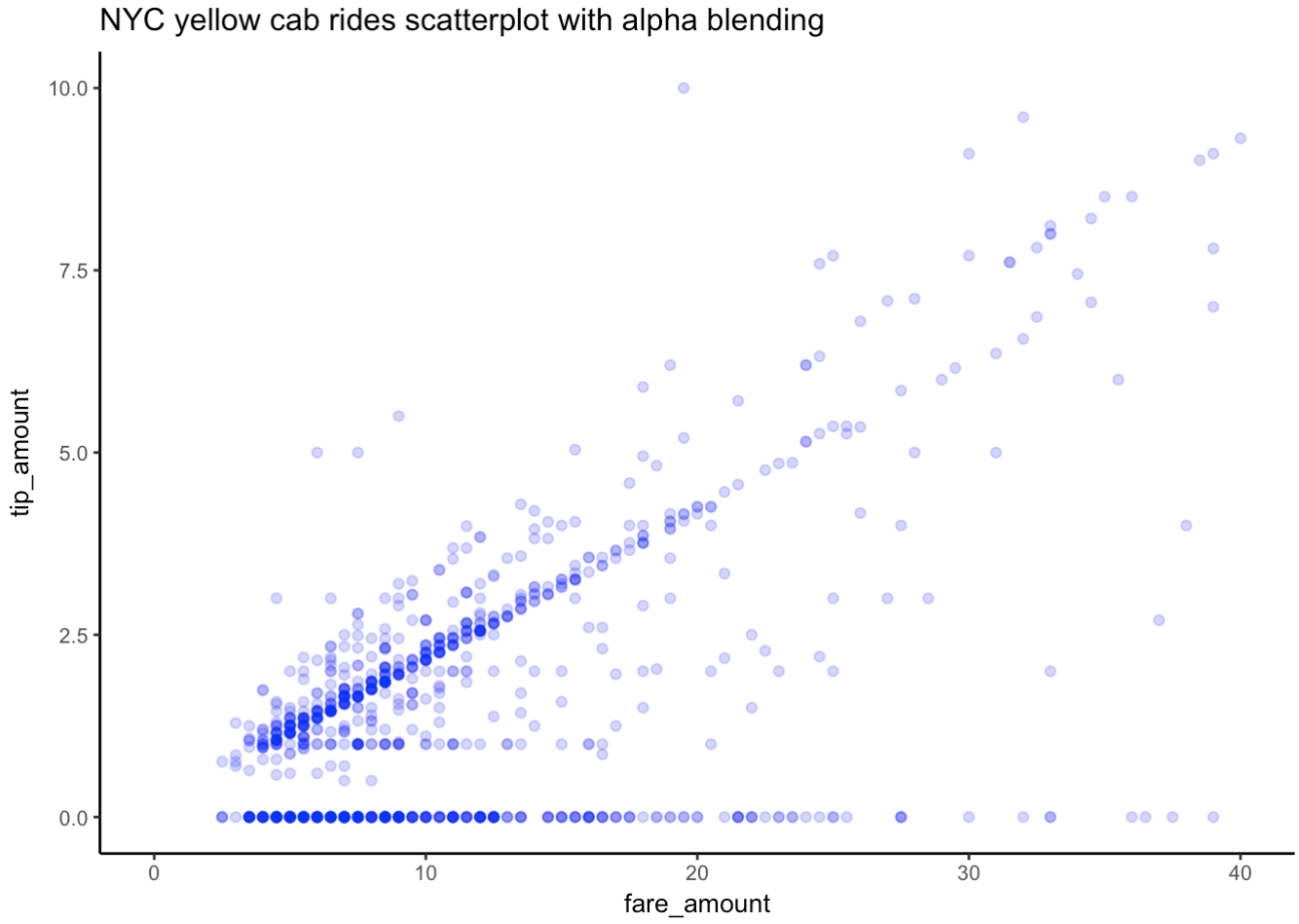
Draw four scatterplots of tip_amount vs. fare_amount with the following variations:

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

```
taxi <- read_csv("/Users/yawenhan/Desktop/Autumn2018/5702 EDAV/R Code/HW02/yellow_tripda
ta_2018-06.csv")
```

(a) Points with alpha blending

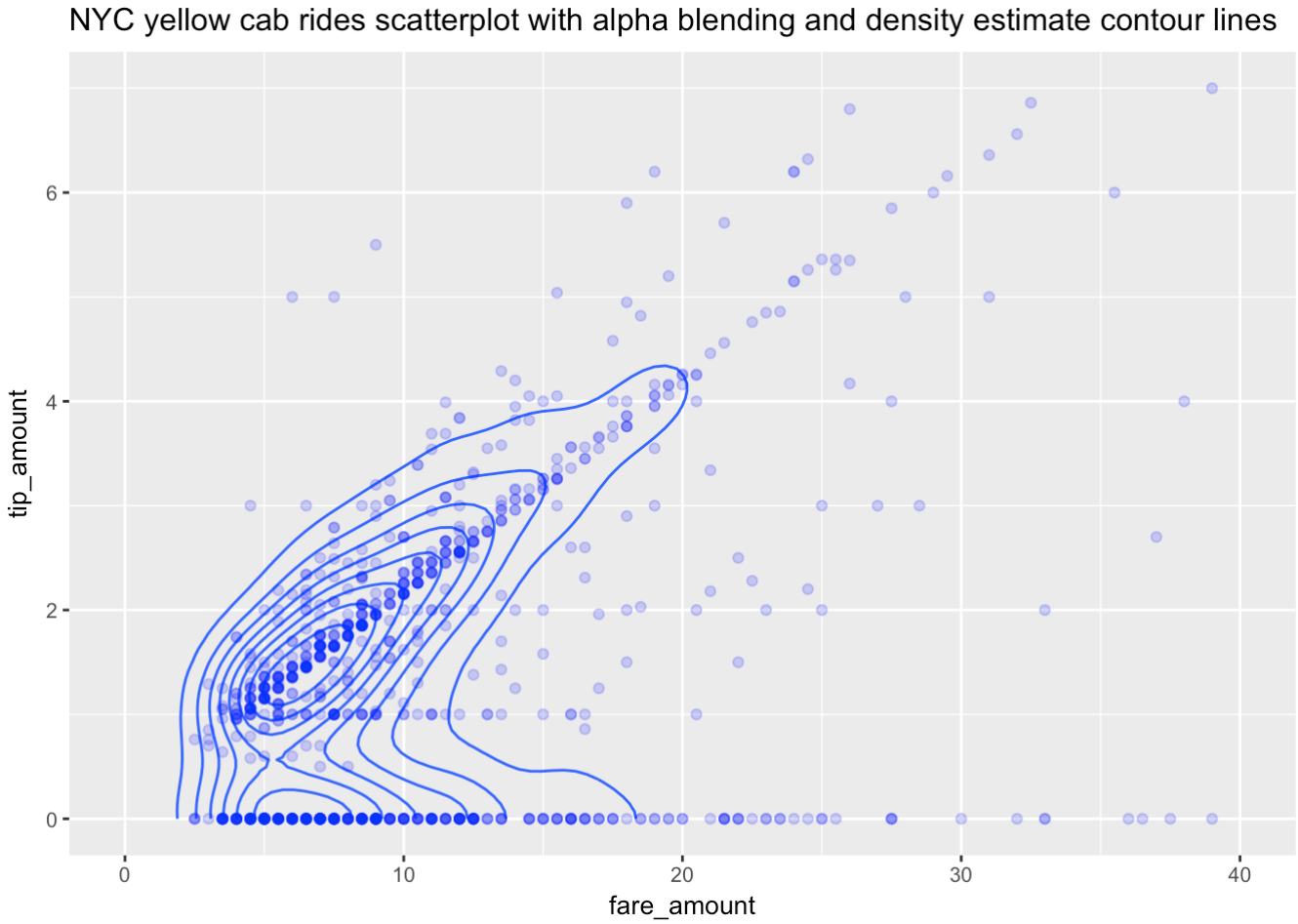
```
ta <- ggplot(taxi_sub, aes(fare_amount, tip_amount)) + geom_point(alpha = .2, color = "b
lue") + theme_classic(10) + xlim(0,40) + ylim(NA,10) + ggtitle("NYC yellow cab rides sca
tterplot with alpha blending")
ta
```



The graph above is scatterplot with alpha blending. To have a better view of the scatterplot, the data points that far away from the main cluster (points at upper right corner), and points that make no sense (value is negative) are excluded from the above plot. The scale limit for x-axis is (0,40), and y-axis is (Na,10). Moreover, the alpha transparency scale is set to be 0.2 to display the density of data points better.

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

```
tb <- ggplot(taxi_sub, aes(fare_amount, tip_amount)) + geom_density_2d() + theme_grey(10
) + geom_point(alpha = .2, color = "blue")+ xlim(0,40) + ylim(NA,7)+ ggtitle("NYC yellow
 cab rides scatterplot with alpha blending and density estimate contour lines")
tb
```



The graph above is scatterplot with alpha blending and density estimate contour lines. To have a better view of the scatterplot, the data points that far away from the main cluster (points at upper right corner), and points that make no sense (value is negative) are excluded from the above plot. The scale limit for x-axis is (0,40), and y-axis is (Na,7). Moreover, the alpha transparency scale is set to be 0.2 to display the density of data points better.

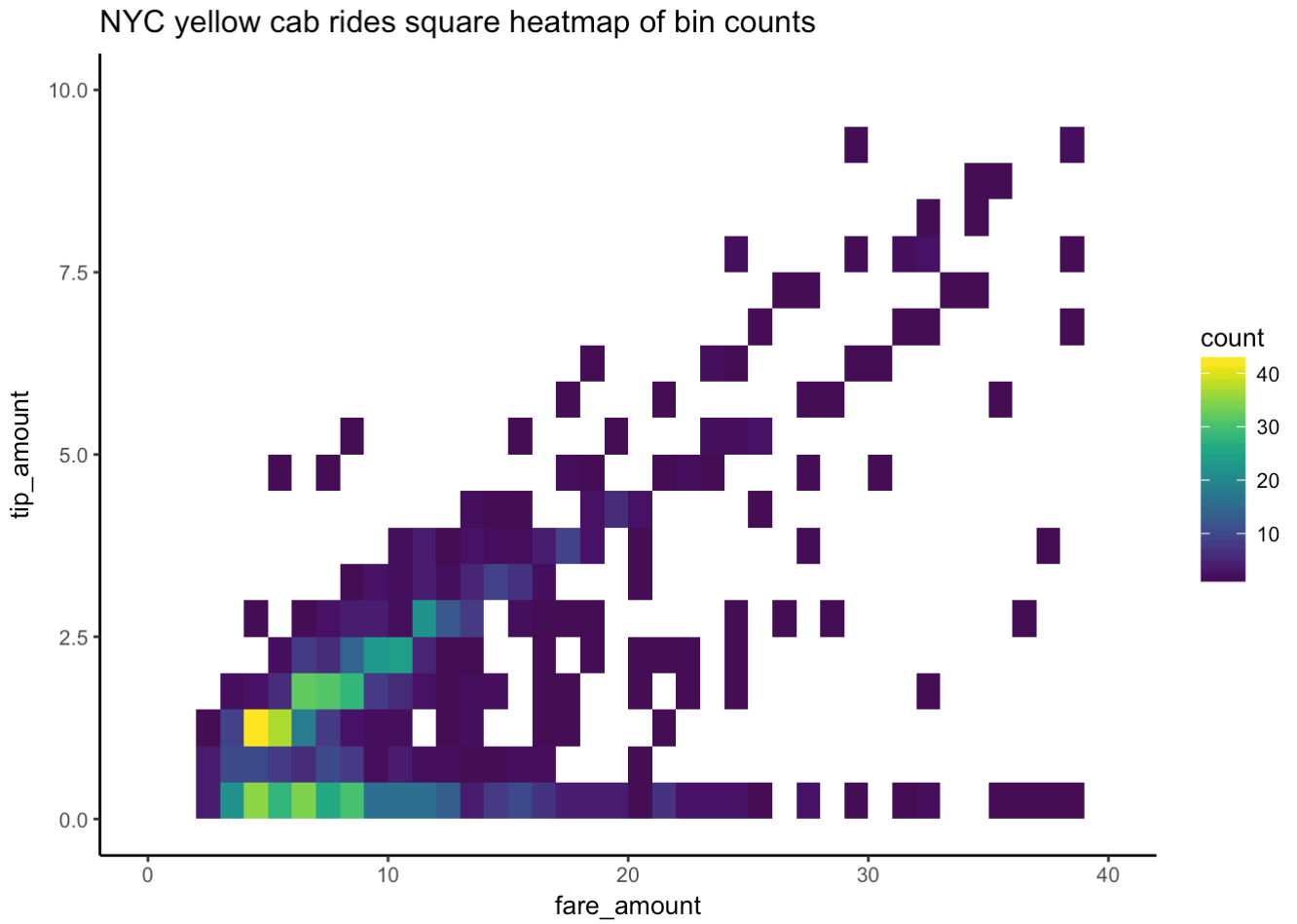
(c) Hexagonal heatmap of bin counts

```
library(viridis)
tc <- ggplot(taxi_sub, aes(fare_amount, tip_amount)) + geom_hex(binwidth = c(1, 0.5)) +
 scale_fill_viridis() + theme_classic(12)+ xlim(0,40) + ylim(NA,10)+ theme_classic(10)+
 ggtitle("NYC yellow cab rides hexagonal heatmap of bin counts")
tc
```

The graph above is hexagonal heatmap of bin counts. To have a better view of the heatmap, the data points that far away from the main cluster (points at upper right corner), and points that make no sense (value is negative) are excluded from the above plot. The scale limit for x-axis is (0,40), and y-axis is (Na,10). After the tradeoff between “displaying the difference of counts value for each region clearly” and “not too many details that hurts the observation”, the binwidth of the heatmap is set to be (1, 0.5) to display the bin counts better.

(d) Square heatmap of bin counts

```
td <- ggplot(taxi_sub, aes(fare_amount, tip_amount)) + geom_bin2d(binwidth = c(1, 0.5))
 + scale_fill_viridis() + theme_classic(12)+ xlim(0,40) + ylim(NA,10)+ theme_classic(10)
+ ggtitle("NYC yellow cab rides square heatmap of bin counts")
td
```



The graph above is square heatmap of bin counts. To have a better view of the heatmap, the data points that far away from the main cluster (points at upper right corner), and points that make no sense (value is negative) are excluded from the above plot. The scale limit for x-axis is (0,40), and y-axis is (Na,10). After the tradeoff between "displaying the difference of counts value for each region clearly" and "not too many details that hurts the observation", the binwidth of the heatmap is set to be (1, 0.5) to display the bin counts better.

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

Features of the data:

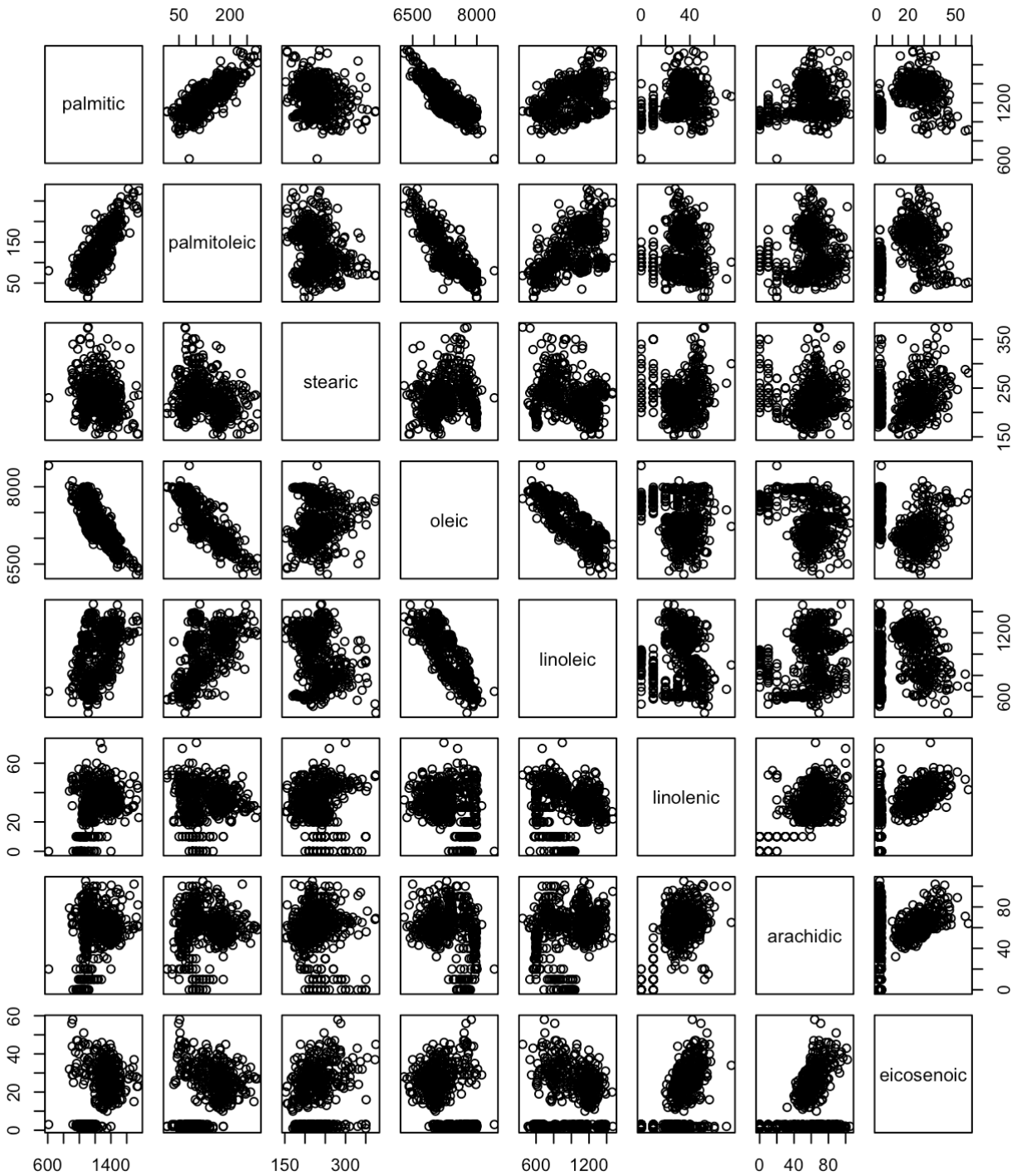
(1)There are no rides with a low fare amount and a high tip amount.
(2)No matter what the fare amount is, there will always be some rides has no tip.
(3)For rides that tips are given, the tip amount increases with the increasing of fare amount.
(4)For rides that tips are given, a few rides with a relatively high fare amount, over 50, look like outliers. They have a distinctly lower tip amount than other rides with similar fare amount.
(5)For rides that tips are given, there is almost no trip with a low fare amount and a high tip amount. Only one ride get the highest tip amount with a low fare amount, looks like a outlier.

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)
data(olives)
oldf <- olives
olvar <- oldf %>% dplyr::select(palmitic, palmitoleic, stearic, oleic, linoleic, linolen
ic, arachidic, eicosenoic)
plot(olvar)
```



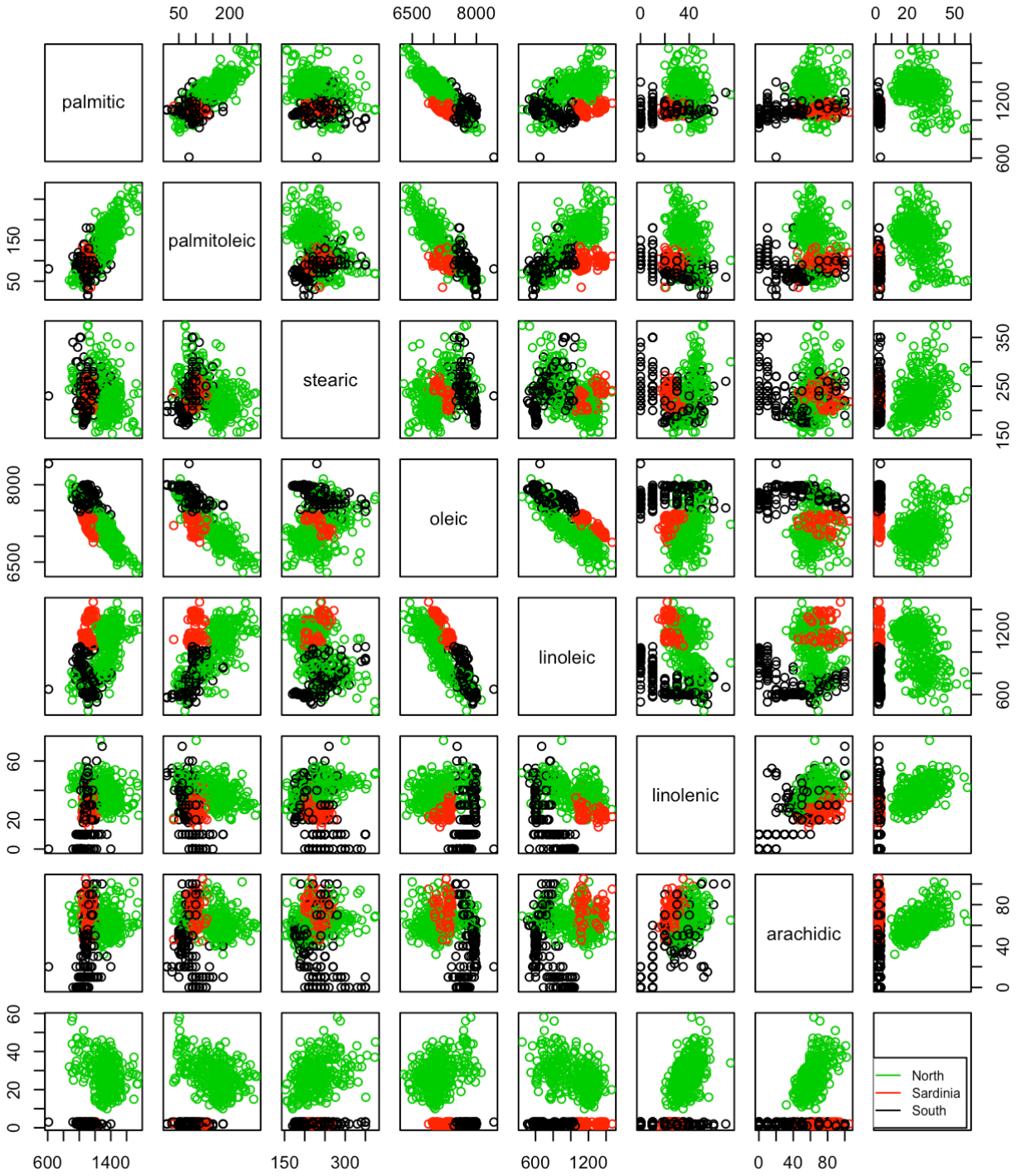
The scatterplot matrix of the eight continuous variables from olives dataset is shown above.

(1)The pairs of variables (palmitic, palmitoleic) is strongly positively associated, because as palmitoleic increases, so does palmitic.
(2)The pairs of variables (palmitic, oleic), (palmitoleic, oleic) is strongly negatively, associated, because as one variable increases, the other decreases in general.

For other pairs of variables, there is no stronly associated relationship been observed from the scatterplot matrix.

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

```
plot(olvar, col=oldf$Regio)
legend("bottomright", legend=levels(oldf$Regio),col=unique(oldf$Regio), cex=0.5,text.fon
t=0,lty=1:1)
```



Observations:

1. The data from North region can be perfectly separated from others by feature “eicosenoic”, because only North has a positive “eicosenoic” value.
2. The data from North region dominates the trend of each scatterplot as a result of its largest amount. In some scatterplot, the data from other two regions shows the similar correlation of two variables as North, such as (palmitic, oleic). While in other scatterplots, the data from other two regions do not share the same correlation as North, even exactly opposite, such as (stearic, oleic).

5. Wine

Data: wine dataset in pgmm package

(Recode the Type variable to descriptive names.)

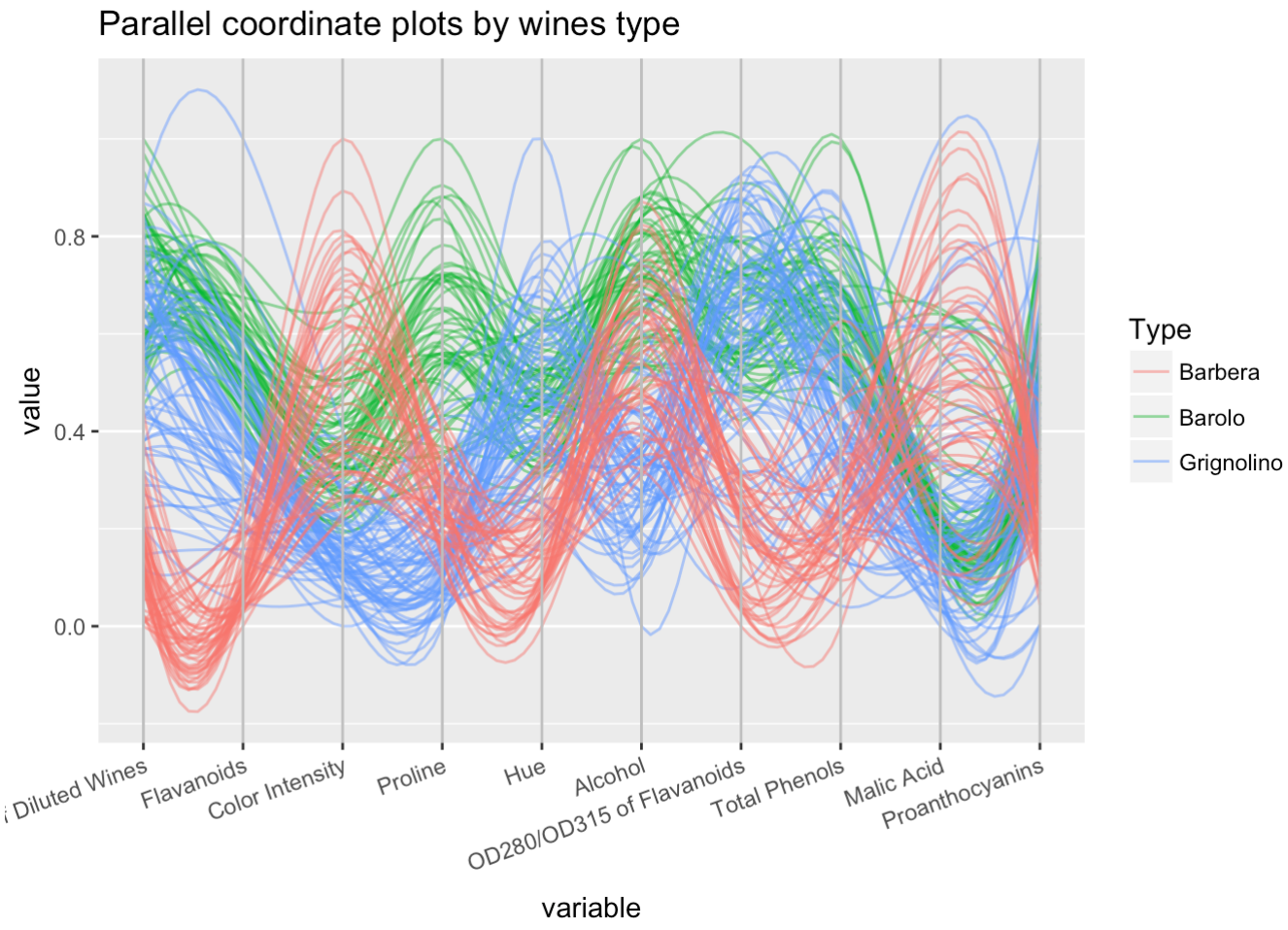
```
library(pgmm)
data(wine)
wa <- wine
# recode the type variable to descriptive names
wa$Type <- ifelse(wa$Type == 1, "Barolo", ifelse(wa$Type == 2, "Grignolino", "Barbera"))
```

The above table shows the recoded dataset, in which the “Type” variable is recoded 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(e1071)
library(caret)
library("dplyr")
library(rpart)
#control <- trainControl(method="repeatedcv", number=10, repeats=3)
# train the model
modelf <- rpart(Type~., data=wa)
# estimate variable importance
imp<-varImp(modelf)
selected <- c(rownames(imp)[order(imp$Overall, decreasing=TRUE)])
m<-wa[,c("Type",selected[1:10])]
```

```
# scale = std (default)
library(GGally)
ggparcoord(m, columns =2:11 , alphaLines = .5,scale = "uniminmax", splineFactor = 10, gr
oupColumn = 1)+ geom_vline(xintercept = 1:10, color = "grey")+theme(axis.text.x=element_
text(angle=20,hjust=1)) + ggtitle("Parallel coordinate plots by wines type")
```



Using the decision tree classifier, 10 features with greatest importance variance are selected to help separated the three types of wines. To have a better view of the plot, the plot is added with alpha transparency, rescale, splines and vertical lines. The above plot is very informative and can help classify different wine types efficiently.

(b) Explain what you discovered.

Observations:

In the descriptions below, the relative position is been described as "high", "low" or "medium", which is determined by the intersection of most lines with the vertical lines.

(1)For "Barbera" type wine - it usually has a low "Diluted Wines", a low "Flavanoids", a high "Color Intensity", a low "Proline", a low "Hue", a high "Alcohol", a low "OD280/OD315 of Flavanoids", a low "Total Phenols", a high "Malic Acid" and a low "Proanthocyanins".
(2)For "Barolo" type wine - it usually has a high "Diluted Wines", a high "Flavanoids", a medium "Color Intensity", a high "Proline", a medium "Hue", a high "Alcohol", a medium "OD280/OD315 of Flavanoids", a high "Total Phenols", a low "Malic Acid" and a high "Proanthocyanins".
(3)For "Grignolino" type wine - it usually has a medium "Diluted Wines", a medium "Flavanoids", a low "Color Intensity", a low "Proline", a high "Hue", a low "Alcohol", a high "OD280/OD315 of Flavanoids", a medium "Total Phenols", a low "Malic Acid" and a medium "Proanthocyanins".

In conclusion, the three wine types can be easily classified through the above procedure, according to the distinctive difference from the combination of these 10 features.