Homework #2

Hrishikesh Telang (hnt2107)

10/09/2018

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(ggplot2)
library(dplyr)
library(cluster)
data(flower)
names(flower)
```

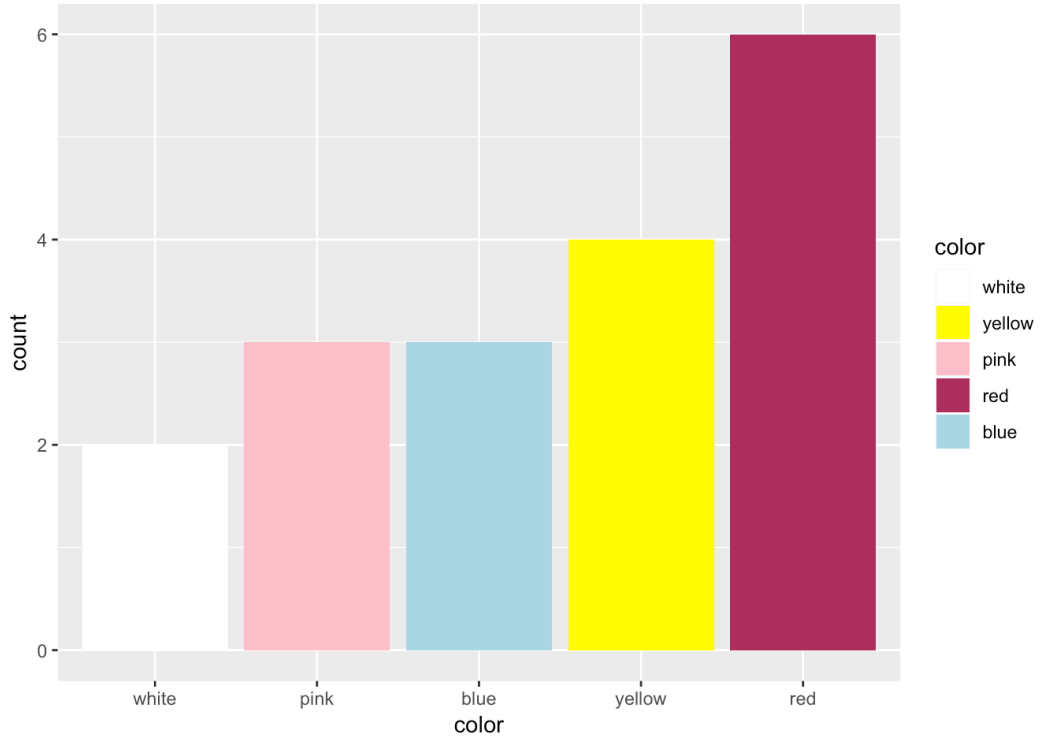
```
## [1] "V1" "V2" "V3" "V4" "V5" "V6" "V7" "V8"
```

```
name<-c("winters","shadow","tubers","color","soil","preference","height","distance")
colnames(flower)<-name

flower[,1]<-factor(flower[,1],levels=c(0,1),labels=c("no","yes"))
flower[,2]<-factor(flower[,2],levels=c(0,1),labels=c("no","yes"))
flower[,3]<-factor(flower[,3],levels=c(0,1),labels=c("no","yes"))
flower[,4]<-factor(flower[,4],levels=c(1,2,3,4,5),labels=c("white","yellow","pink","red","blue"))
flower[,5]<-factor(flower[,5],levels=c(1,2,3),labels=c("dry","normal","wet"))

flower
```

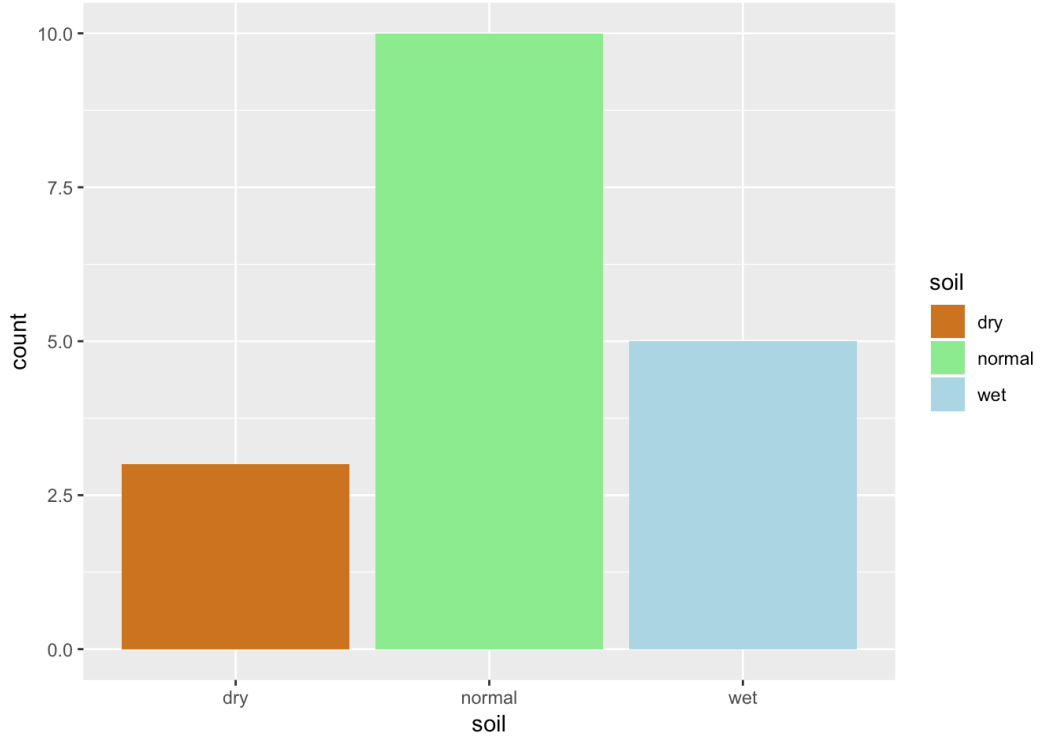
```
##    winters shadow tubers  color   soil preference height distance
## 1       no    yes    yes    red    wet         15     25       15
## 2      yes     no     no yellow    dry          3    150       50
## 3       no    yes     no   pink    wet          1    150       50
## 4       no     no    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.

```
myColors <- c("white", "yellow", "pink","maroon","lightblue")
p<-ggplot(flower,aes(x=reorder(color, table(color)[color])),xlab("colors"))+geom_bar(aes(fill=factor(color)))

p+scale_fill_manual(values=myColors)+ labs(x="color",fill = "color")
```




```
ggplot(flower,aes(x=soil))+geom_bar(aes(fill=factor(soil)))+scale_fill_manual(values=c("#cc7722","lightgreen","lightblue"))+labs(x="soil",fill = "soil")
```

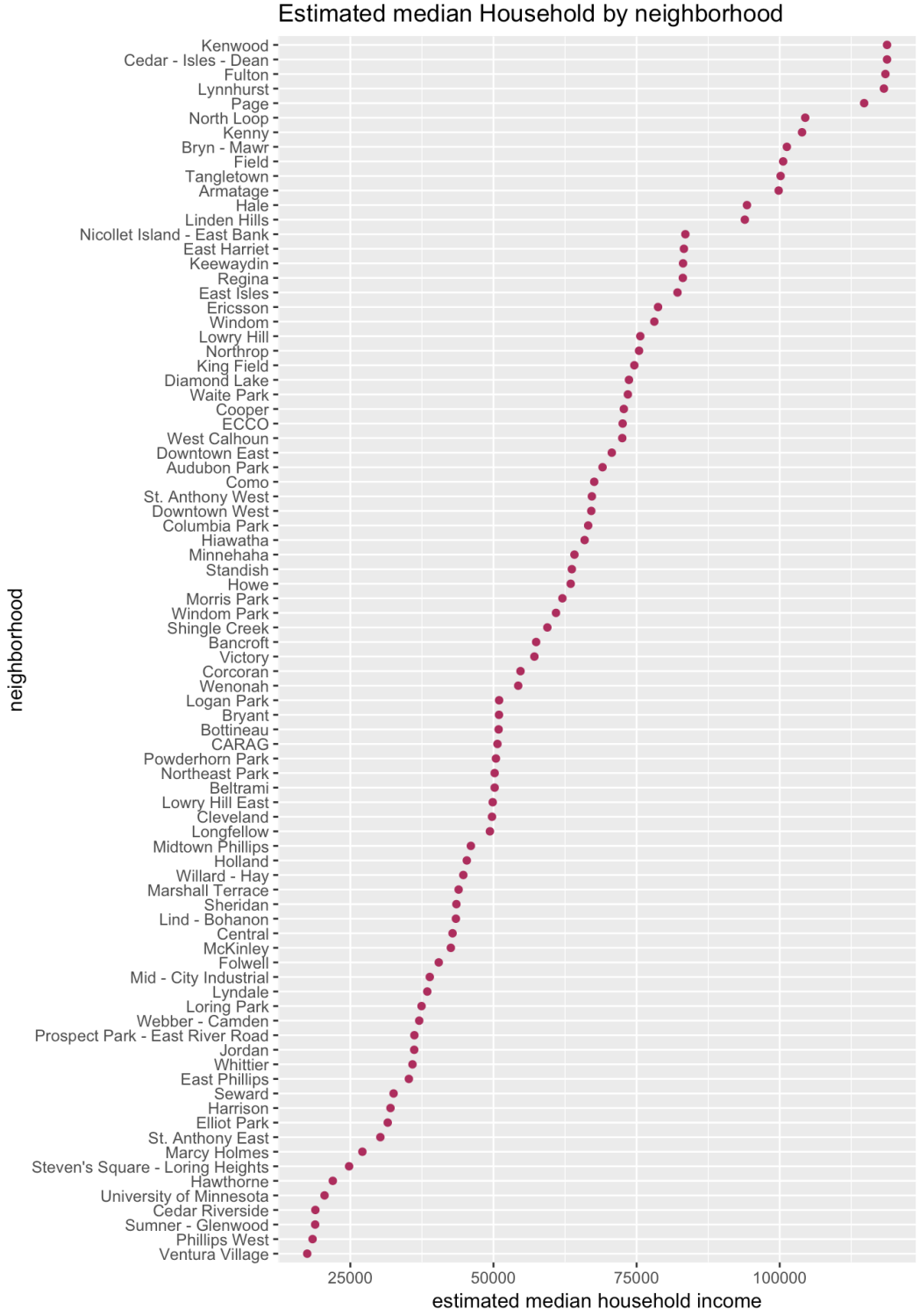


2. Minneapolis

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

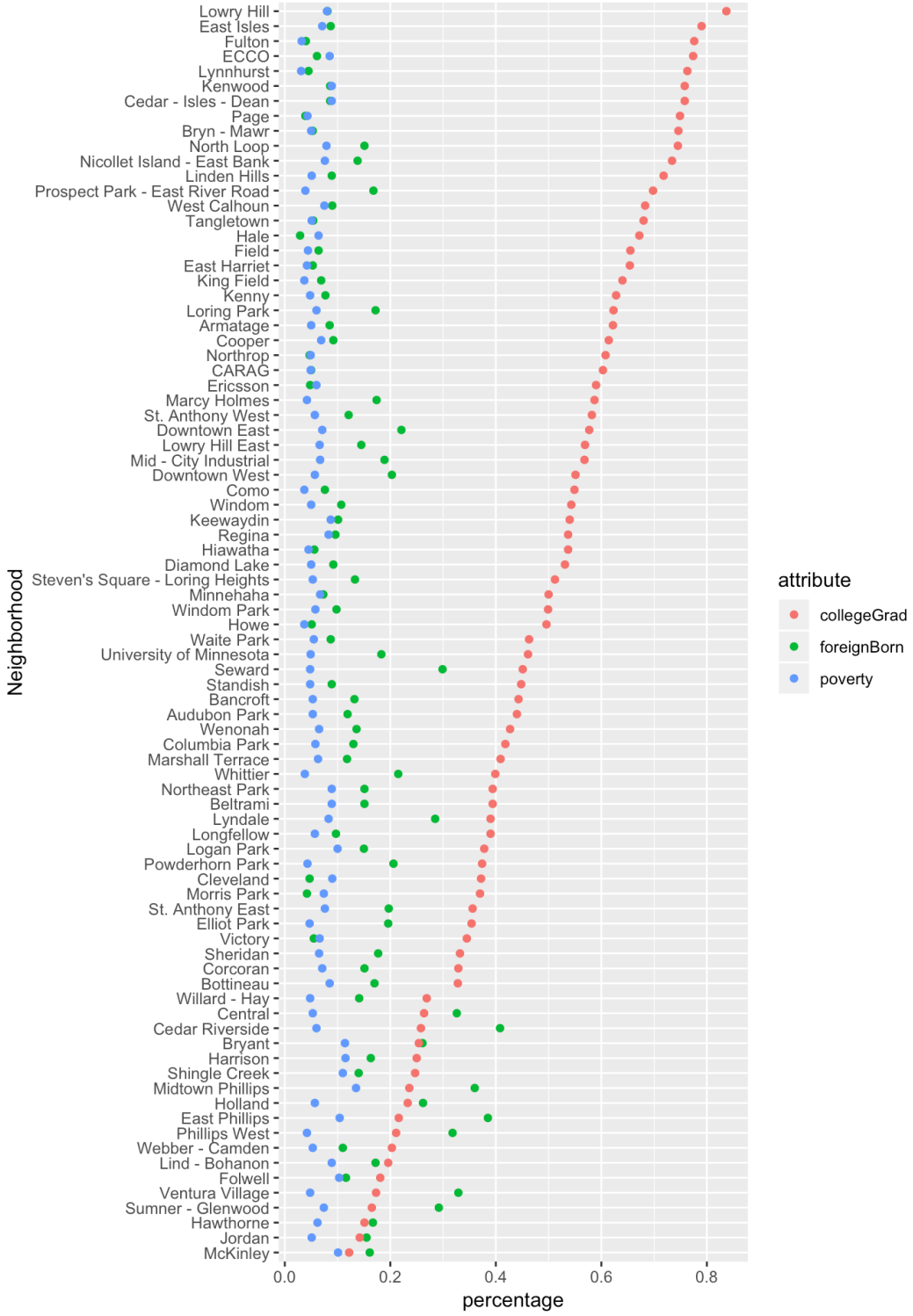
a. Create a Cleveland dot plot showing estimated median household income by neighborhood.

```
library(carData)
data(MplsDemo)
ggplot(MplsDemo,aes(x=hhIncome,y=reorder(neighborhood,hhIncome))) +geom_point(color="maroon")+ylab("neighborhood"
)+xlab("estimated median household income")+ggtitle("Estimated median Household by neighborhood")
```



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

```
library(tidyr)
Mp<-MplsDemo %>%gather(attribute, percentage, -neighborhood,-population,-white,-black,-hhIncome)

ggplot(Mp, aes(x=percentage,y=reorder(neighborhood,percentage*(attribute=='collegeGrad'),max))) +
  geom_point(aes(color = attribute))+ylab("Neighborhood")
```



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

Generally, it can be observed that most neighborhoods have similar proportions of foreign born residents and people living below twice the poverty level, regardless of the number of people with college degrees.

However,neighborhoods that have the lowest proportion (about 25%) of residents whith college degree are seen to have a higher proportion of foreign born residents compared to other neighborhoods.

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.

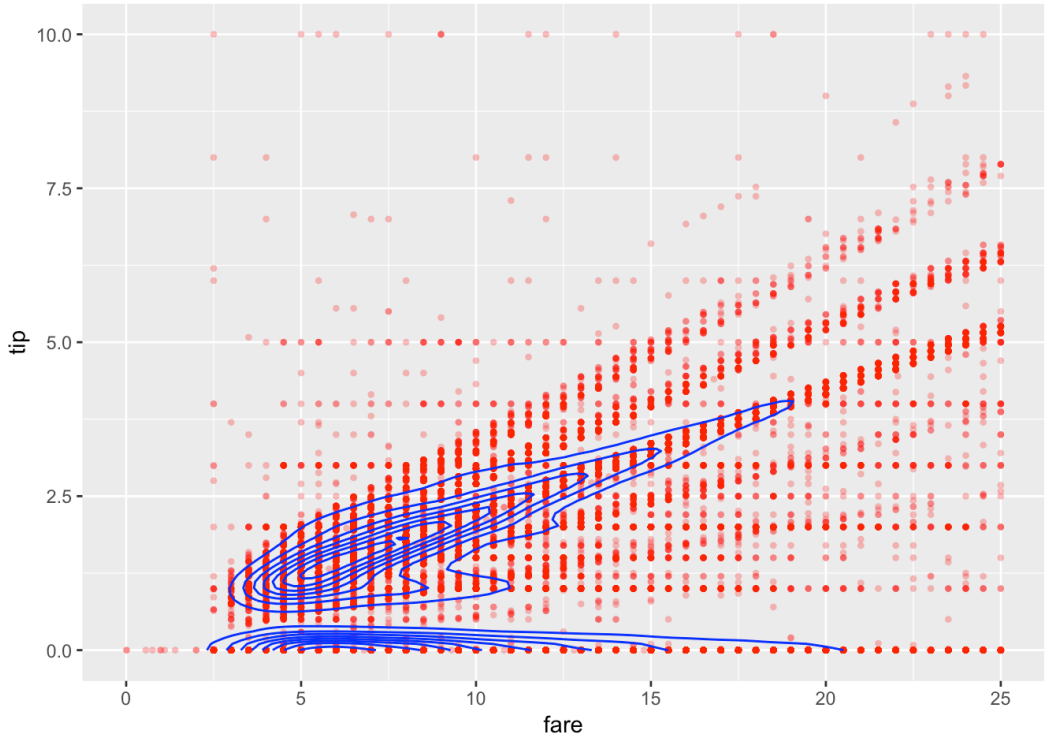
Draw four scatterplots of `tip_amount` vs. `far_amount` with the following variations:

a. Points with alpha blending

```
library(tidyverse)
library(dplyr)
library(plotly)
library(viridis)
df <- read_csv('/Users/hrishikeshtelang/Downloads/yellow_tripdata_2018-06.csv')
x<-sample_n(df, 50000)
x<-x[x$fare_amount>=0,]

g1<-ggplot(filter(x,fare_amount>0 & fare_amount<100 & tip_amount <50), aes(tip_amount, fare_amount)) + geom_point
(alpha = .3, color = "red",stroke=0)+xlab("tip")+ylab("fare")
g1<-g1+coord_flip()
g1
```

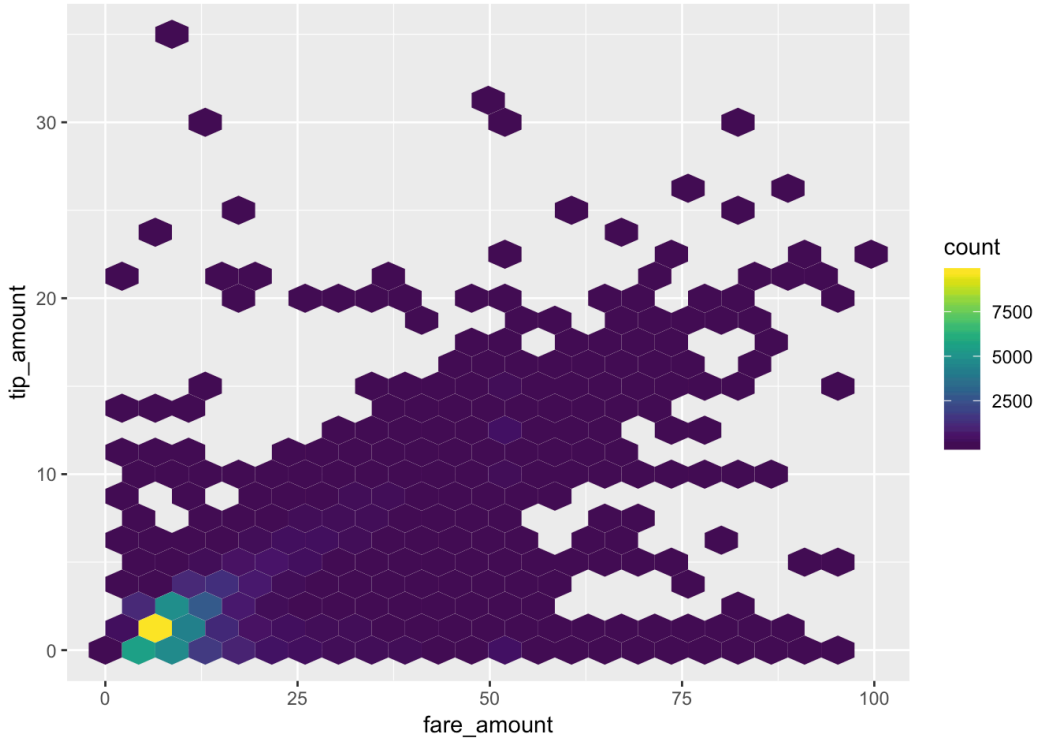
b. Points with alpha blending + density estimate contour lines

```
g1+geom_density_2d(color="blue")+xlim(0,10)+ylim(0,25)
```



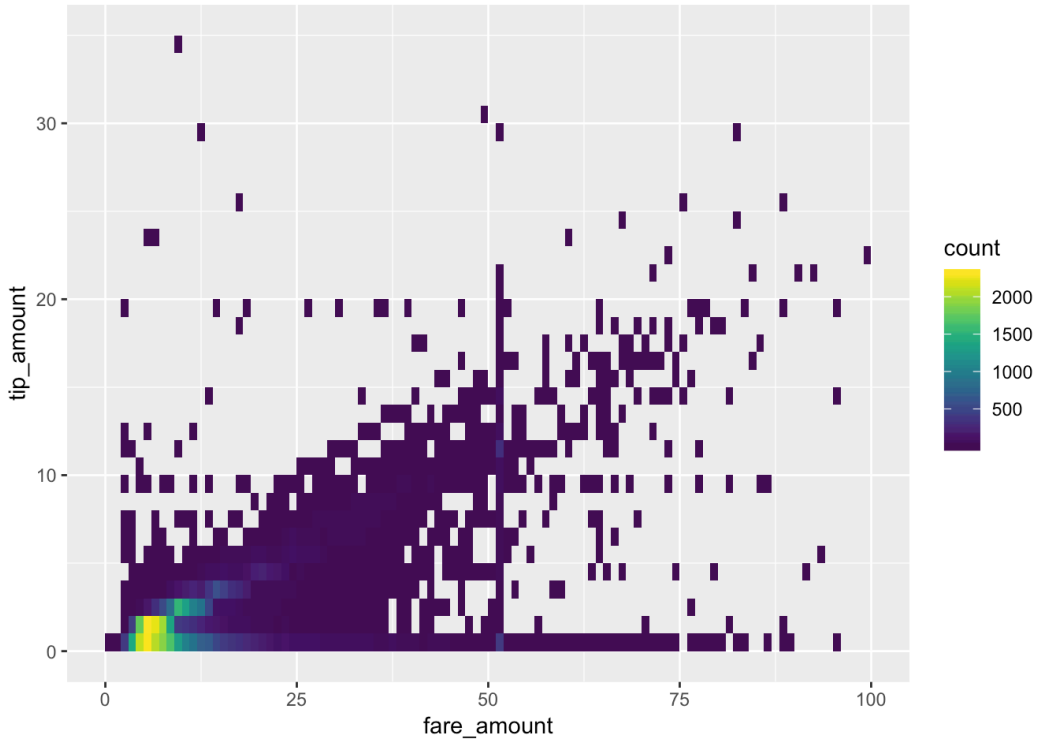
(c) Hexagonal heatmap of bin counts

```
g2<-ggplot(filter(x,fare_amount>0 & fare_amount<100 & tip_amount <50), aes(tip_amount, fare_amount)) + geom_hex(b
inwidth = c(2.5,2.5))+scale_fill_viridis()+coord_flip()
g2
```



d. Square heatmap of bin counts

```
g3<-ggplot(filter(x,fare_amount>0 & fare_amount<100 & tip_amount <50), aes(tip_amount, fare_amount))+scale_fill_v
iridis()+coord_flip() + geom_bin2d(binwidth = c(1,1))
g3
```



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.

Outliers- In the original scatterplot of the subset, we can see that there are some disproportionate tips paid for smaller fares (eg. tip of almost 50 for zero fare and tips of almost five times the fare for low fare values).

Zero tips- Judging by the concentration of red dots on the X- axis, we can see that a lot of people do not tip at all regardless what the fare amount is.

Flat fare- There are multiple tips being paid for the same amount of fare which results in a vertical line of sorts for a fare of close to 50$. This suggests that there is a flat fare in existence for some rides.

Linear pattern- The tips and fares are generally proportional to each other and grow in a linear fashion. It can also be noticed that certain tip amounts (eg.2,5 and 10) are very common which implies that people tend to favour round numbers for tips, especially when the fares aren’t high.

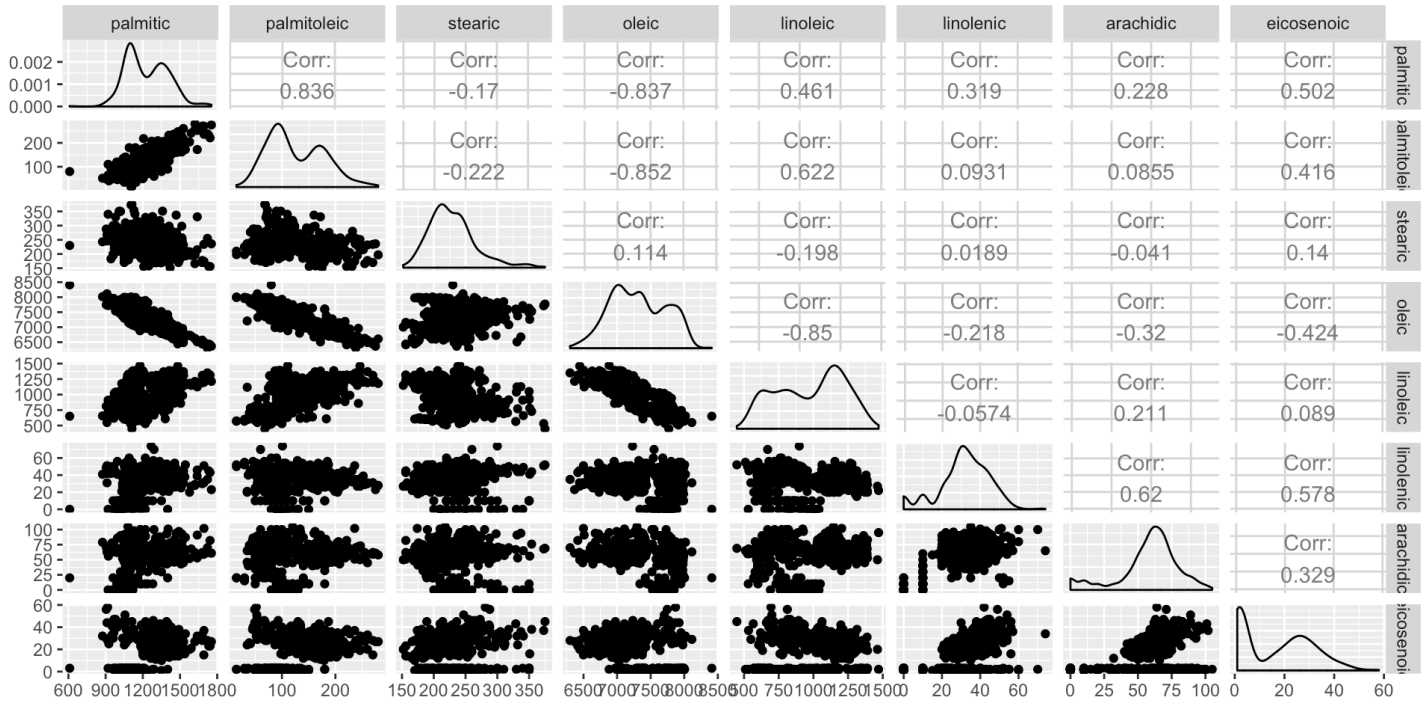
Density- It can be observed from the Hexagonal and Square heat maps that maximum density is observed in the lower left corner, which implies that this specific region of the graph contains most frequently occurring values of fares and tips for a sample of 50,000

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)
library(extracat)
data(olives)
library(dplyr)
mat<-olives%>% dplyr::select(palmitic, palmitoleic, stearic,oleic,linoleic,linolenic,arachidic,eicosenoic)
ggpairs(mat,echo=FALSE)
```



It can be seen from the scatterplot matrix that strong posotive correlation os obtained for:

1.Palmitic-Palmitoleic (0.836)

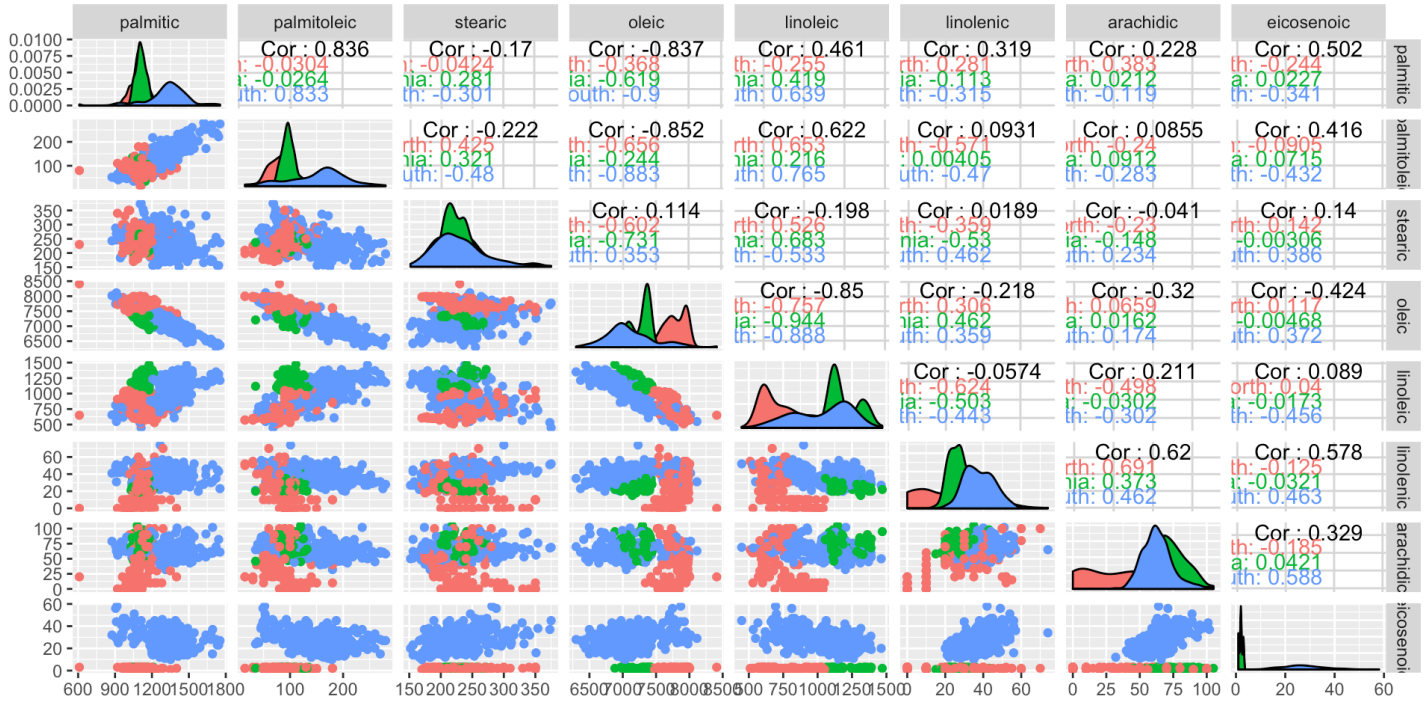
Strong Negative Correlation: 1.Palmitic-Oleic (-0.837)

2.Palmitoleic-Oleic (-0.852)

3.Oleic-Linoleic (-0.85)

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

```
ggpairs(mat,aes(color=olives$'Region'),echo=FALSE)
```



We can observe that the wine from the South Region is the one that most contributes to the correlation values obtained.

The North Region gives a lot of outliers and sometimes, the North region just has a value close zero especially for eicosenioic acid.

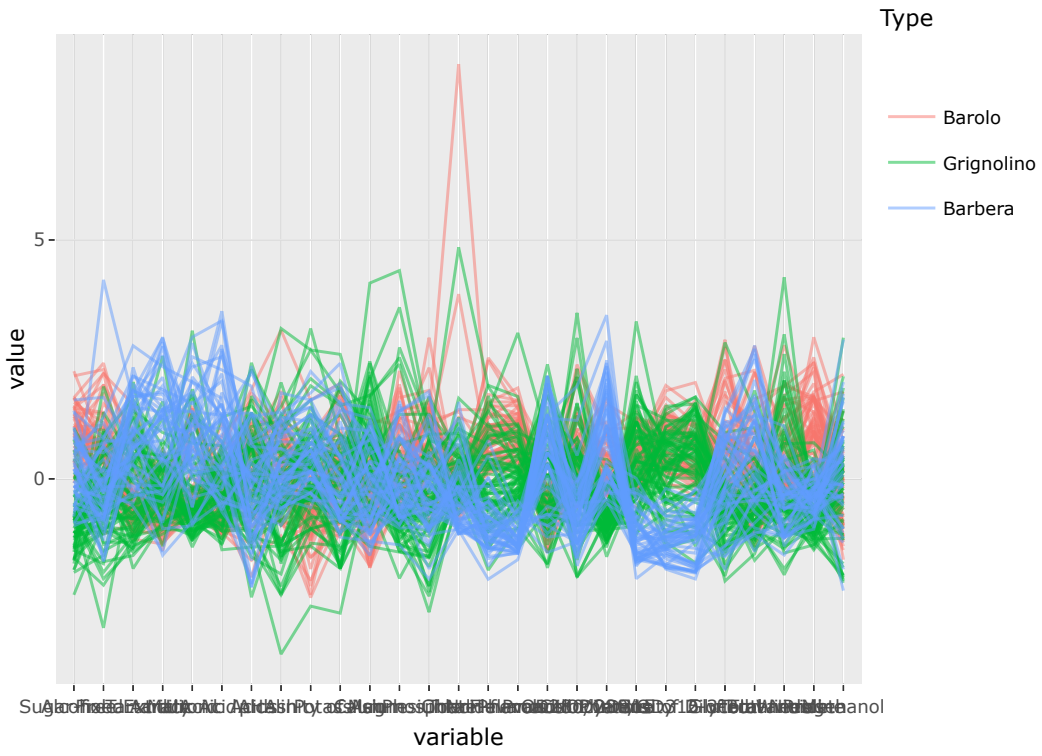
5. Wine

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

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

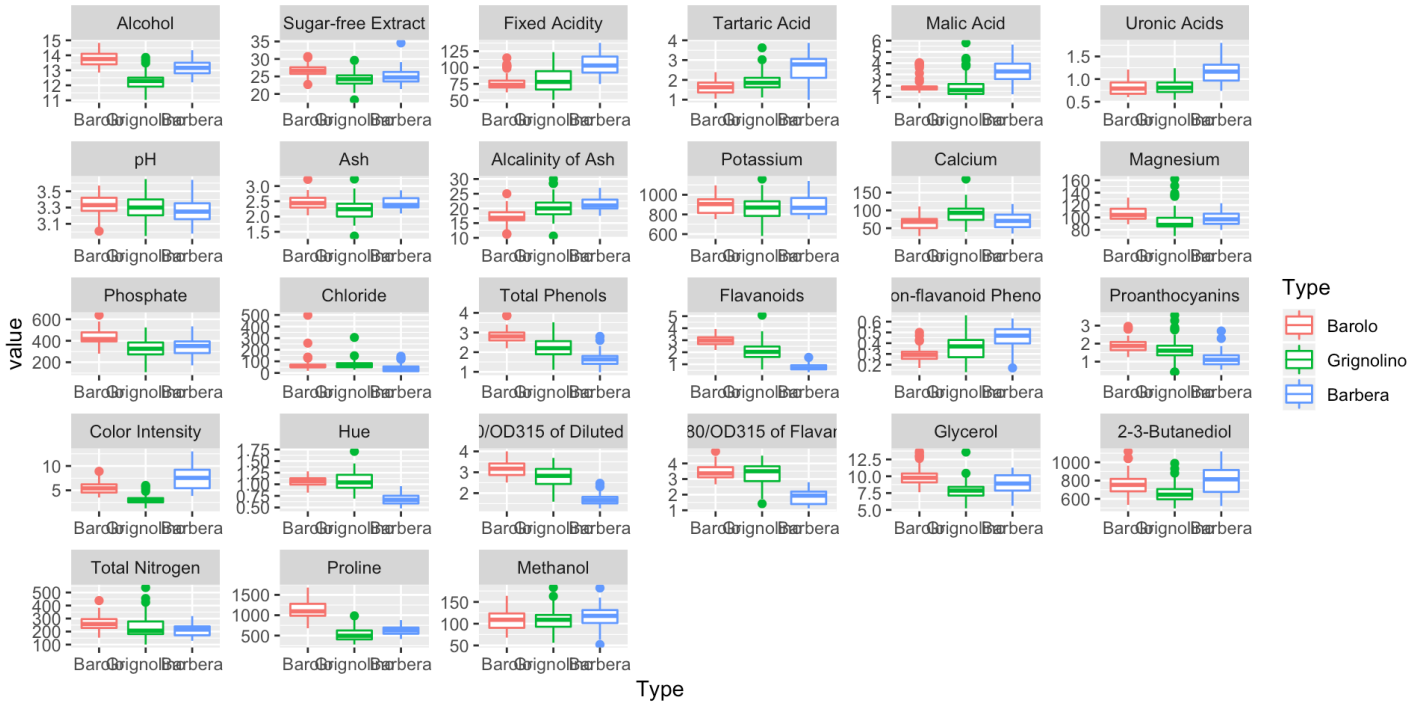
a. Use parallel coordinate plots to explore how the variables separate the wines by `Type` . Present the version that you find to be most informative. You do not need to include all of the variables.

```
library(GGally)
library(pgmm)
data(wine)
wine[,1]<-factor(wine[,1],levels=c(1,2,3),labels=c("Barolo","Grignolino","Barbera"))

a1<-ggparcoord(wine, columns = 2:28, alphaLines = .5, groupColumn=1)

ggplotly(a1)
```



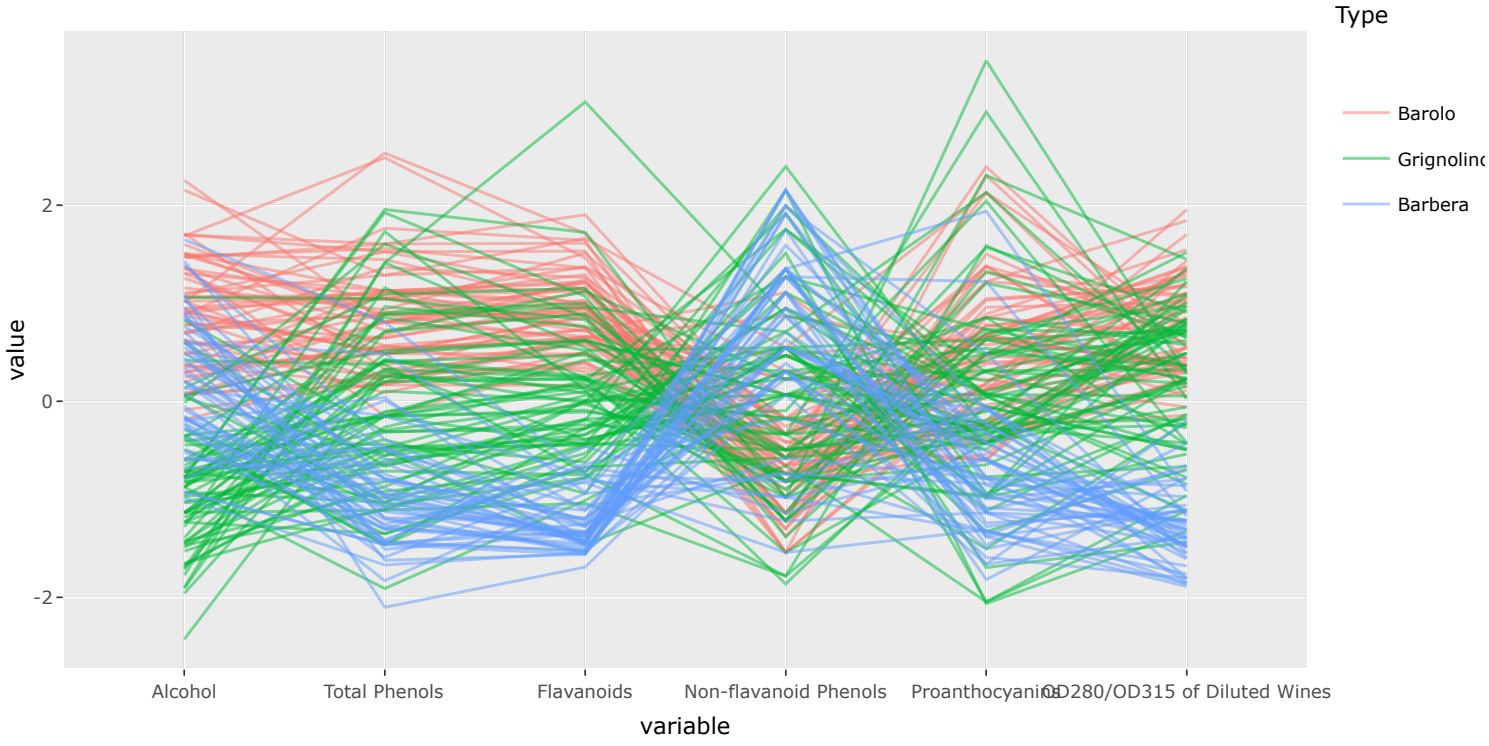
We can see that plotting a parallel co-ordinate plot for Type according to all colums gets very confusing. It is thus better to concentrate only on a few variables.

```
library(reshape2)
sample=melt(wine,id=c("Type"))

ggplot(subset(sample),aes(x=Type,y=value,color=Type))+geom_boxplot()+facet_wrap(~variable,scales="free")
```



It can be seen from the boxplots that columns 10, 16,17,18,19,22 have relatively different distributions compared to the other variables which have similar median values and distributions

```
a2<-ggparcoord(wine, columns = c(2,16:19,22), alphaLines = .5, groupColumn=1)

ggplotly(a2)
```



b. Explain what you discovered.

The second graph makes it easier to derive insights from the data based on the parallel co-ordinate plot.

We can see that alcohol content is high and Total-phenols are maximum in Barolo wines. Barolo wines also have the maximum OD315 of diluted wines.

Barbera wines have moderate alcohol content, the lowest amount of pheols and flavenoids amongst the three, and the highest content of non-flavenoid phenols. The OD315 of diluted wines is also the lowest for Barbera.

Grignolino wines have the lowest alcohol content and moderate flavenoids and OD315, but it can be observed that they have a lot of outlier values.