Exercise 4

for Data Analysis

Valentino Lazarevic - 1223211

16.04.2024

Setup

```
setwd("/Users/valentino/Documents/Studium/Semester 4/Datenanalyse/Exercises")

ChildSmokers = read.csv("ChildSmokers.csv")
PISA = read.csv("PISA2018.csv")
Diamonds = read.csv("Diamonds.csv")

library(MASS)
```

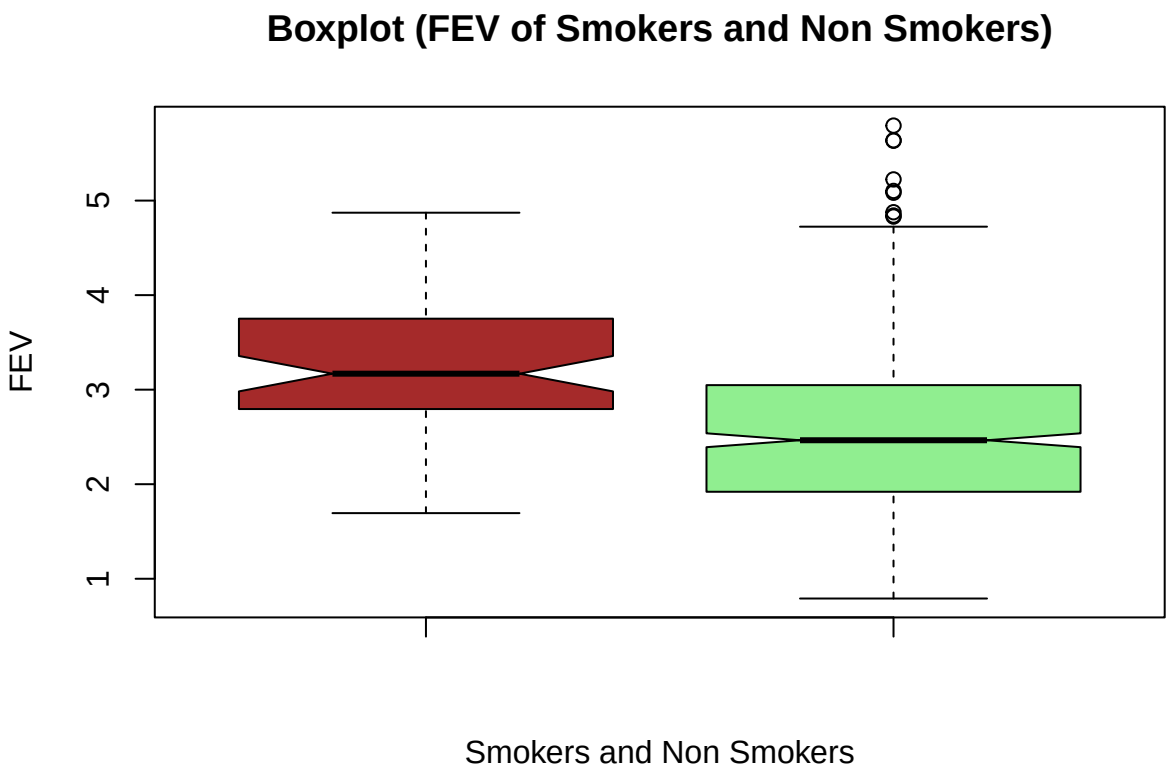
Exercise 1

Compare by parallel boxplots the FEV values for the non-smokers and smokers, and interpret the outcome (which is a bit surprising).

The two parallel boxplots illustrate the forced expiratory volume (FEV), a parameter predominantly affected by smoking and often indicative of heightened pulmonary disease risk when declining. The boxplot representing smokers exhibits a notably higher median FEV value (above 3) and also higher first and third quartiles compared to non-smokers. Interestingly, despite the expected association between reduced FEV and smoking, non-smokers display lower median and quartile values, initially surprising observers. This unexpected outcome warrants further exploration, as detailed in section 2.

Additionally, non-smokers include outliers with FEV values surpassing even the maximum observed among smokers. Moreover, the notches in the smokers' boxplot appear slightly larger, suggesting a smaller sample size and consequently a wider 95%-confidence interval for its median.

It's evident that smokers exhibit a higher median FEV score compared to non-smokers (around 2.50 versus 3.25), contrary to conventional expectations. Notably, some outliers in the non-smoker group demonstrate markedly superior results compared to smokers. Surprisingly, the lower whiskers of non-smokers extend to 0.75, possibly attributable to the smaller lung volume typically observed in younger children (5-13 years) compared to older children (13-18 years). This age-related dynamic might contribute to the denser distribution of smokers in the 13-18 age range.

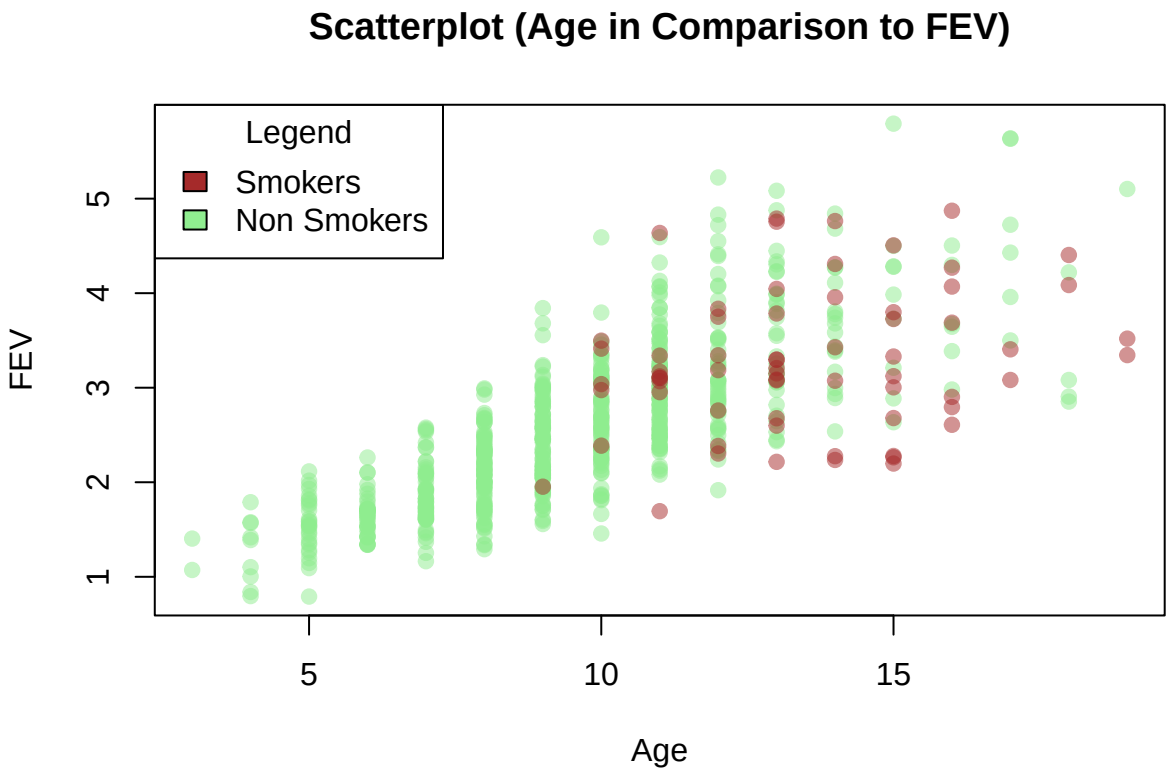


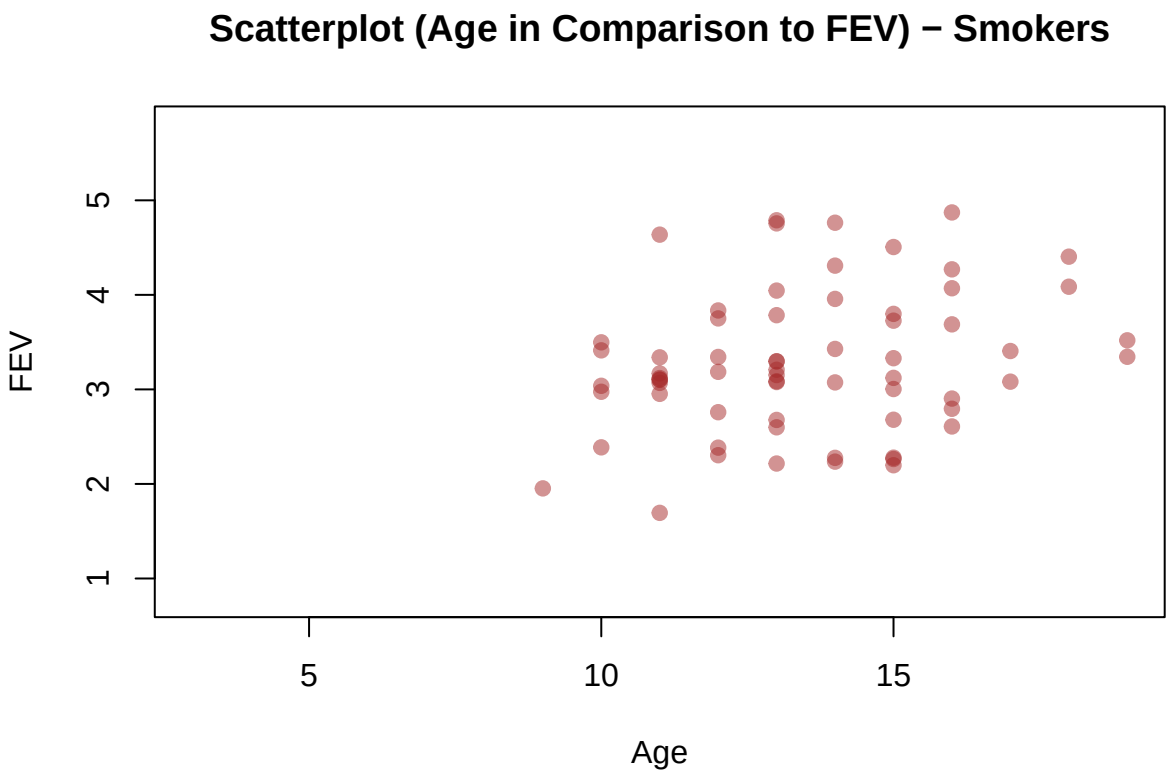
Exercise 2

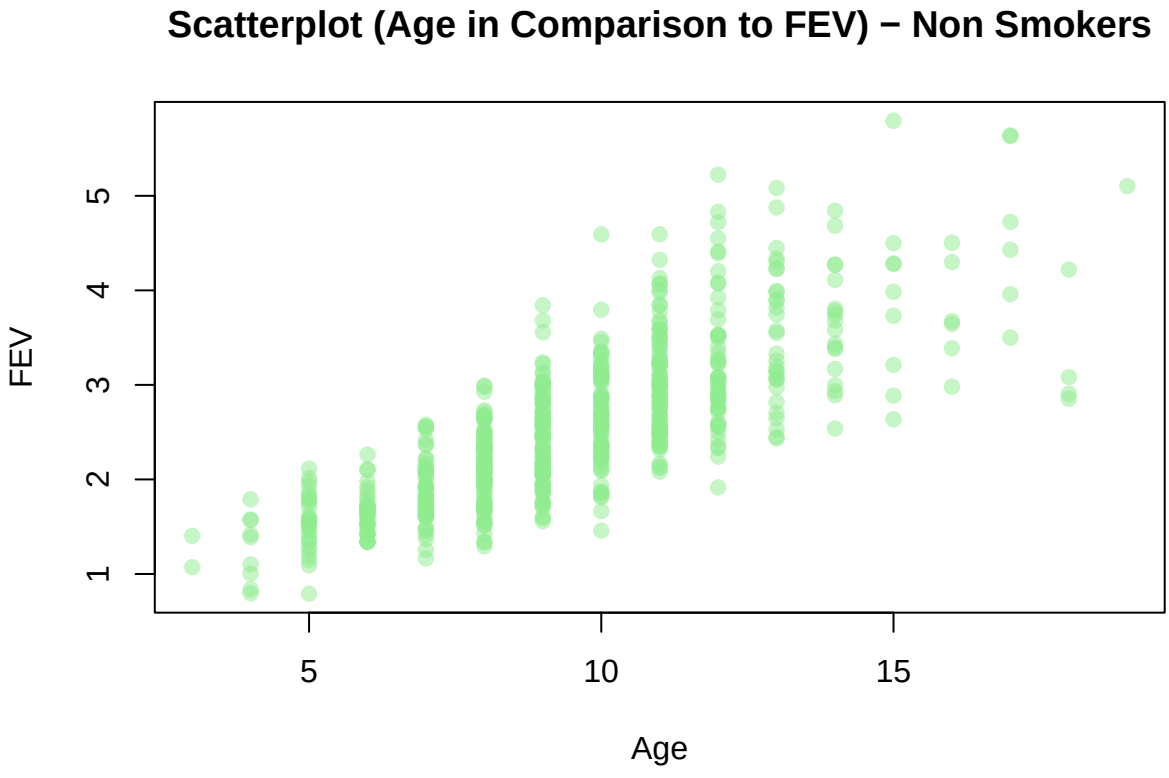
As the result in 1. is the opposite of what we would expect, we need to go deeper. Try to use the remaining variables within scatterplots to find an answer for the contradiction.

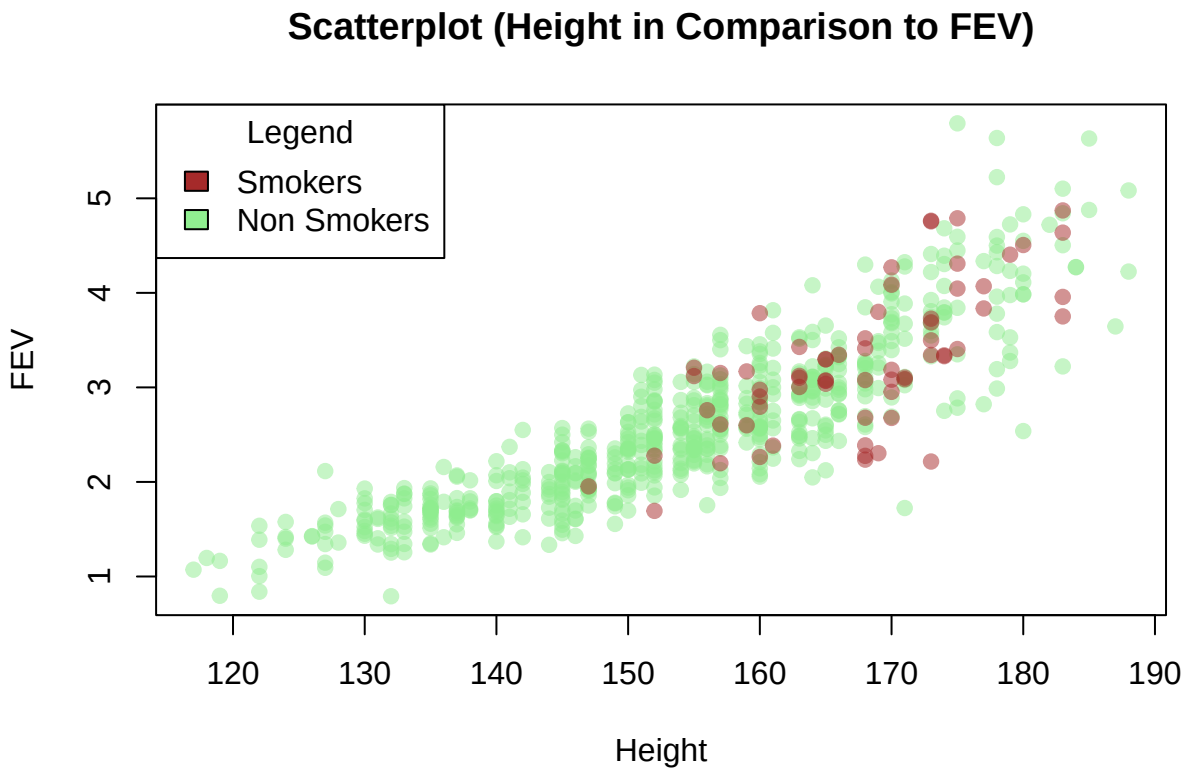
To untangle the unexpected discrepancy in FEV values between smokers and non-smokers, we need to explore additional variables in the dataset. Age and height show a clear link to increased FEV values, with older and taller participants generally exhibiting higher FEV. However, a closer look reveals that smokers are predominantly represented by certain age and height groups, while non-smokers span across various age ranges and heights, including younger and smaller individuals who typically have lower FEV.

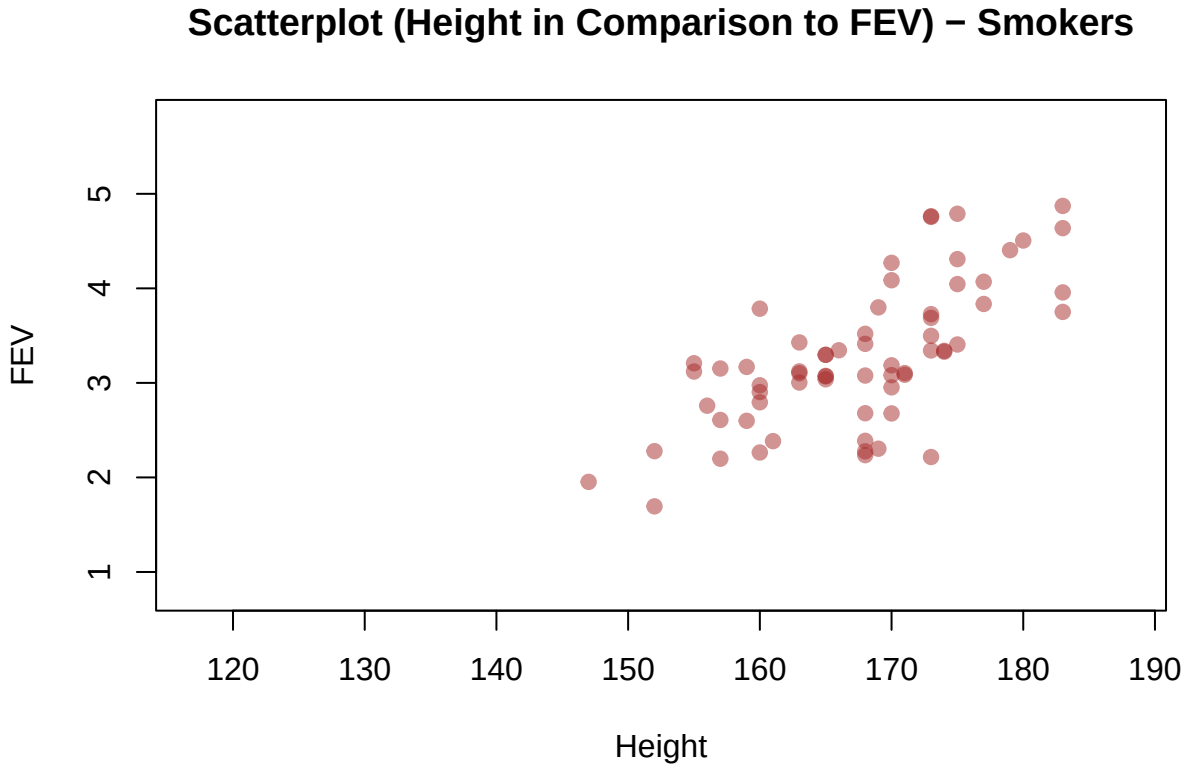
This disparity becomes clearer when examining scatterplots, where FEV values are consistently higher in children aged 10-20 compared to those aged 0-10, reflecting the trend seen in height and FEV values. However, there are notably few smokers under 10 years old, contributing to the surprising findings in the boxplots.

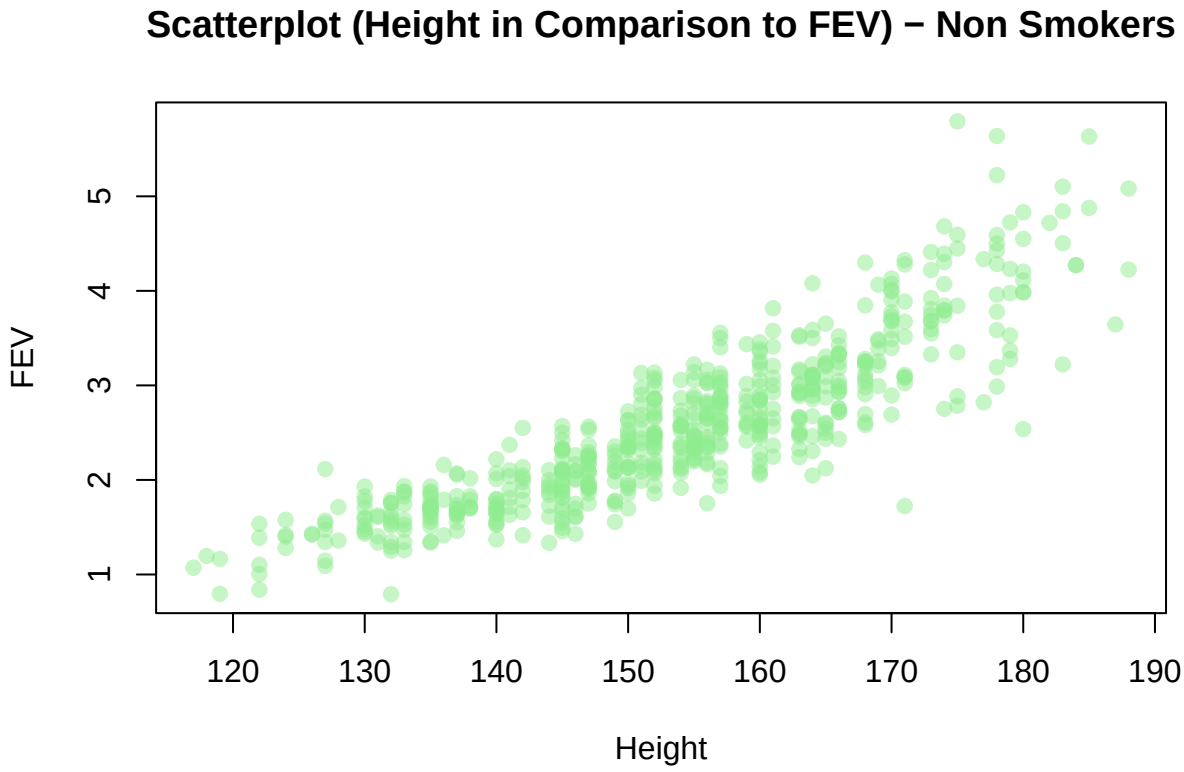



Scatterplot (Age in Comparison to FEV) − Smokers
FEV
1
2
3
4
5
5
10
15
Age


Scatterplot (Age in Comparison to FEV) − Non Smokers
FEV
1
2
3
4
5
5
10
15
Age


Scatterplot (Height in Comparison to FEV)
Legend
Smokers
Non Smokers
FEV
1
2
3
4
5
120
130
140
150
160
170
180
190
Height

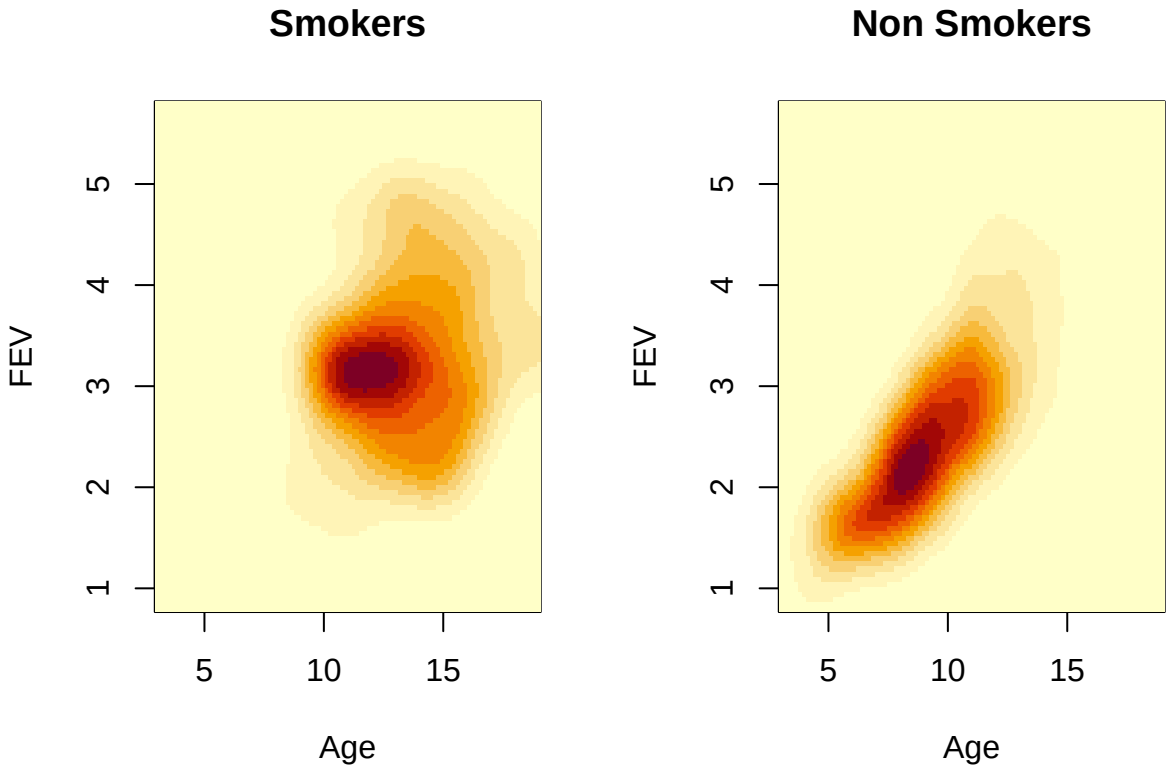

Scatterplot (Height in Comparison to FEV) − Smokers
FEV
1
2
3
4
5
120
130
140
150
160
170
180
190
Height

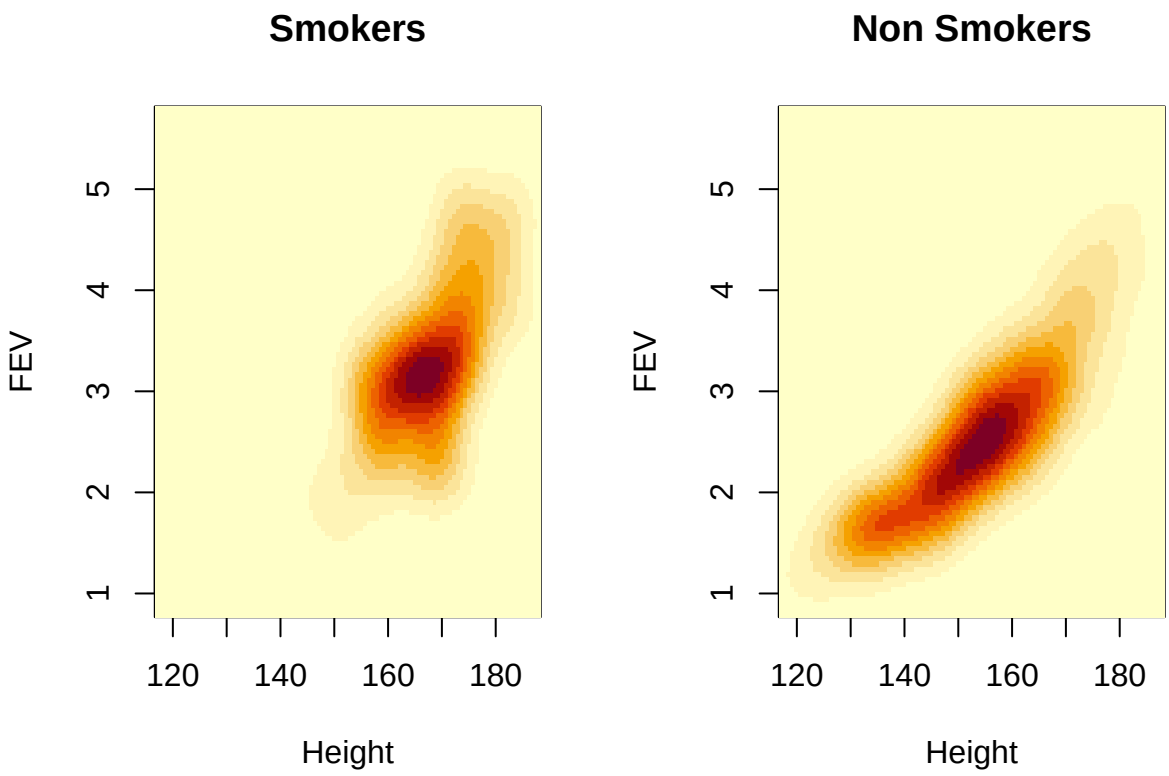

Scatterplot (Height in Comparison to FEV) − Non Smokers
FEV
1
2
3
4
5
120
130
140
150
160
170
180
190
Height

Exercise 3

Use bivariate density estimation, and visualize the results appropriately in order to explain the contradictory phenomenon seen in 1.

The higher FEV values seen in smokers result from the absence of younger and smaller participants in their dataset, as revealed by density estimations for Age and Height against FEV. Older and taller smokers show higher density at higher FEV values, while non-smokers peak at smaller age and height values, corresponding to smaller FEV values. The density shapes vary between groups, with non-smokers showing a clear trend of increasing density for age or height leading to higher FEV, while smokers exhibit a less clear correlation, particularly in Age. Additionally, FEV values concentrate between 7 and 10 years of age, reflecting similar lung volumes in children regardless of smoking. This concentration is more pronounced among smoking children, as shown in height and FEV density estimates, with a notable 20 cm difference between smokers and non-smokers.

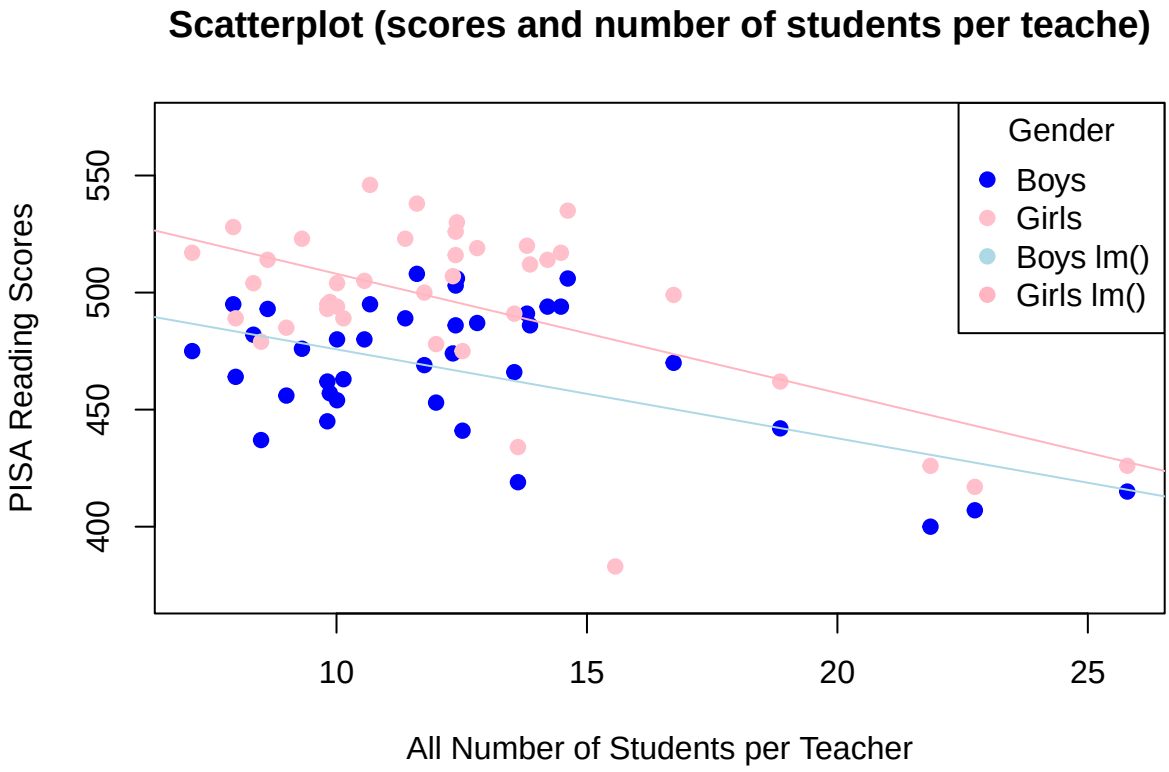



Smokers
Non Smokers
FEV
Height
1
2
3
4
5
120
140
160
180

Exercise 4

Inspect graphically in one scatterplot the relationship between the reading scores of boys and girls with the overall number of students per teacher. Fit a linear model of reading score versus number of students per teacher, using lm(), separately for the data of the boys and the girls, and show the resulting regression lines. What do you conclude?

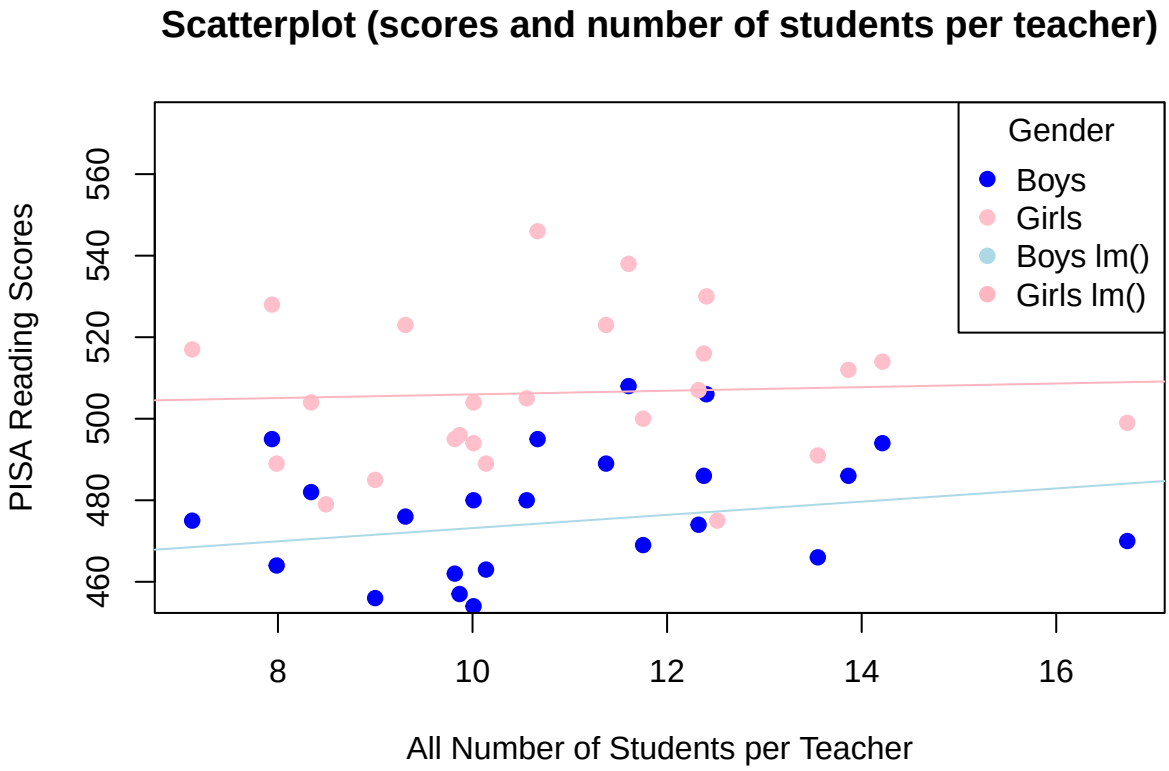
In both the comparison of reading scores across genders and scatterplot analysis, it's evident that girls generally outperform boys, especially in smaller class sizes. However, as class sizes increase, this difference diminishes. Regardless of gender, reading scores decline with higher student-to-teacher ratios, indicating a negative correlation.



Exercise 5

Do the same as in 4., but limited just to the European countries. Are the regression slopes significantly different from zero? If not, what does that mean?

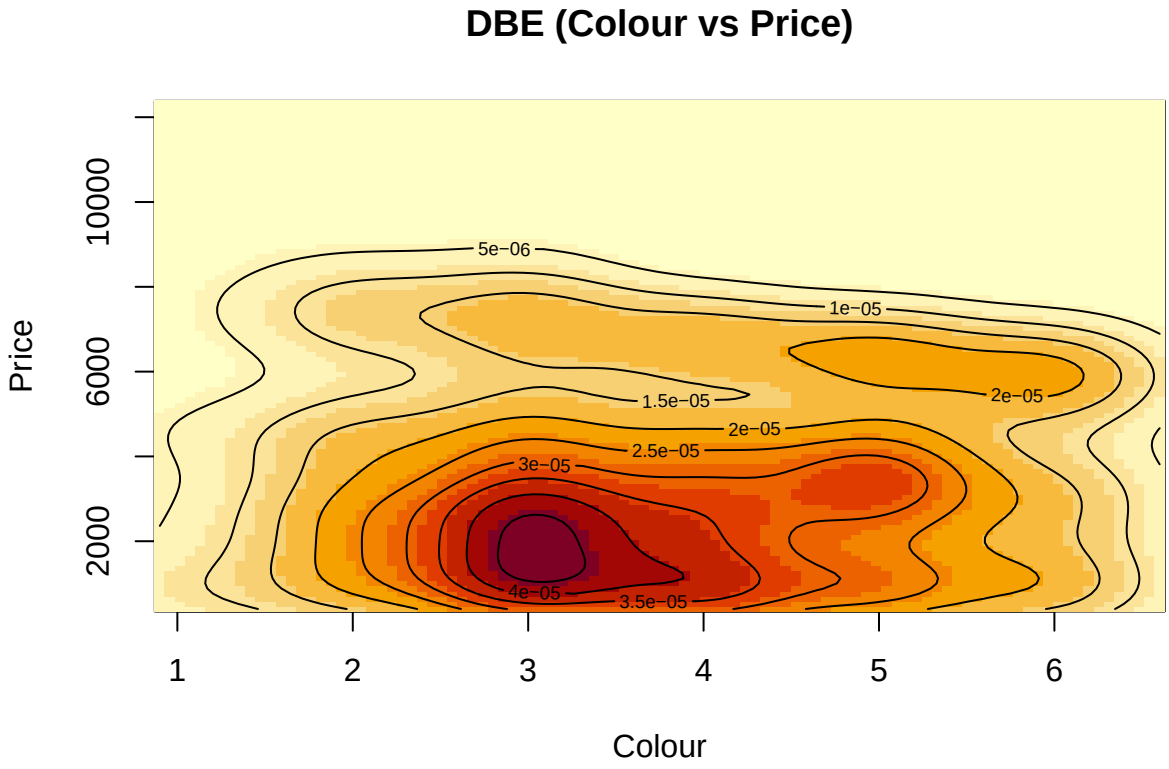
In both the global and European datasets comparing reading scores, girls consistently outperform boys. However, the influence of student-to-teacher ratios on these scores appears less pronounced in Europe. While globally, a negative correlation between student-to-teacher ratio and reading scores is observed, European data presents a different picture. In Europe, the regression lines for boys show no clear correlation with student-to-teacher ratios—neither negative nor positive. This discrepancy may stem from the significantly lower maximum student-to-teacher ratios in Europe compared to global averages, suggesting that a negative correlation may only manifest beyond a certain capacity, not reached in Europe. Additionally, the overall low number of students per teacher in European countries contributes to this difference, with girls consistently exhibiting higher reading proficiency regardless of class size. However, due to the small sample size and already low student-to-teacher ratios, the precision of the linear model may be diminished in the European dataset.



Exercise 6

Use bivariate density estimation to visualize the relationship between Price and Colour. Interpret the result.

The density plot illustrates that while lower color rankings typically lead to higher prices, there are unexpected patterns. Diamonds with higher color purity (ranked 3) are found predominantly in lower price ranges, while those with greater impurity (ranked 6) command the highest prices, challenging the notion of a clear correlation between color impurity and price. Additionally, the densest region corresponds to color values 3 and 4, with minimal density observed in the "colorless" category (ranked 1). Moderately dense regions for color values 5-6 are found at higher prices, likely due to the prevalence of yellow and brown shades.

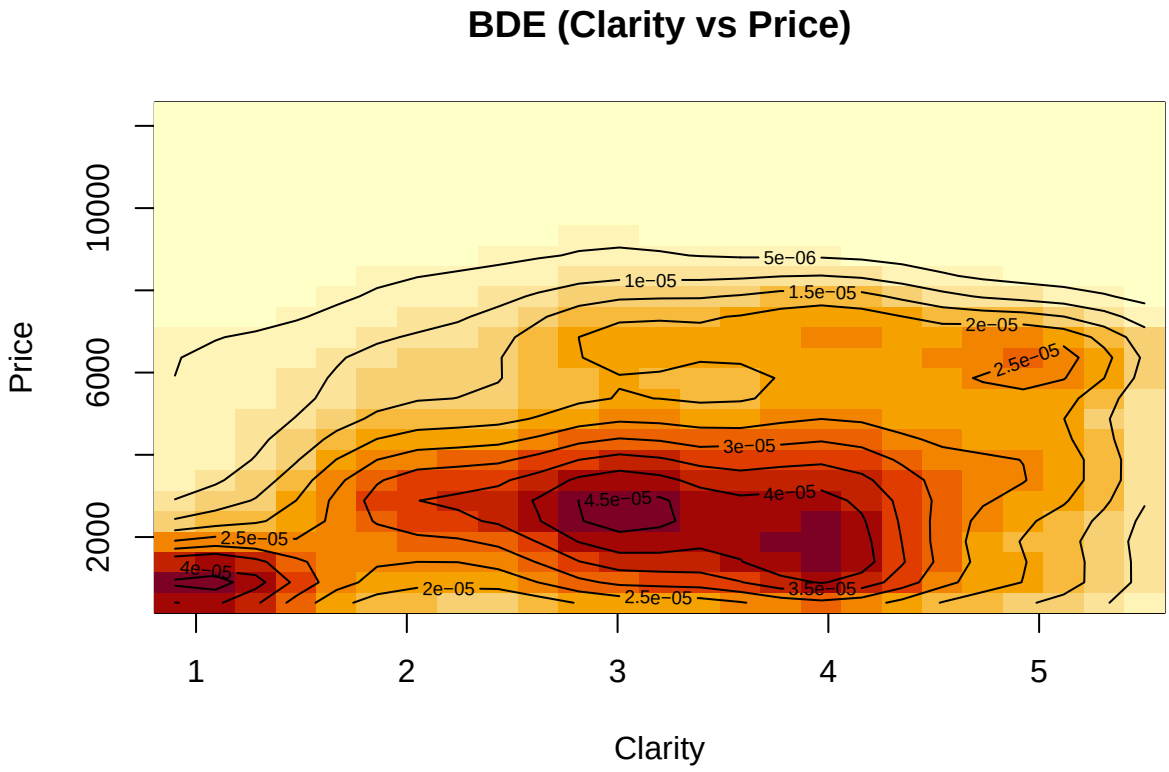


Exercise 7

Do the same as in 6. with the variable Price and Clarity.

Both color and clarity contribute to a diamond's purity ranking, ranging from flawless to very small inclusions. While one might expect higher clarity rankings to correlate with higher prices, the plot reveals unexpected patterns. All clarity ranges are present to an equal degree, with two distinct price ranges visible. Surprisingly, flawless diamonds have a high density at very low prices, while diamonds with lower clarity rankings exhibit their highest density at higher prices. This challenges the reliability of color and clarity alone in estimating diamond prices. Carat value, which considers weight and impurities, may offer a more successful approach.

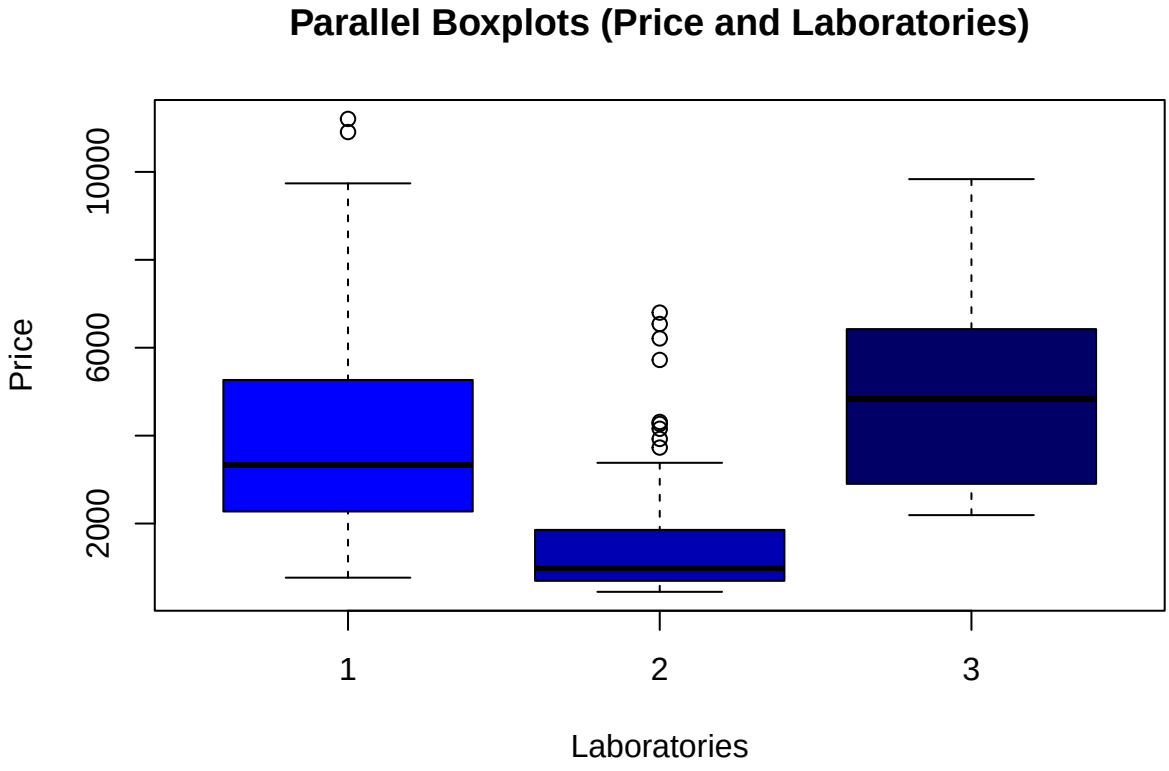
In the bivariate density estimation, flawless and low-priced diamonds are in high demand, while the highest concentration overall lies in the clarity range of 3-4. In the high-priced range, there is also a higher concentration of diamonds with certain clarity disturbances.

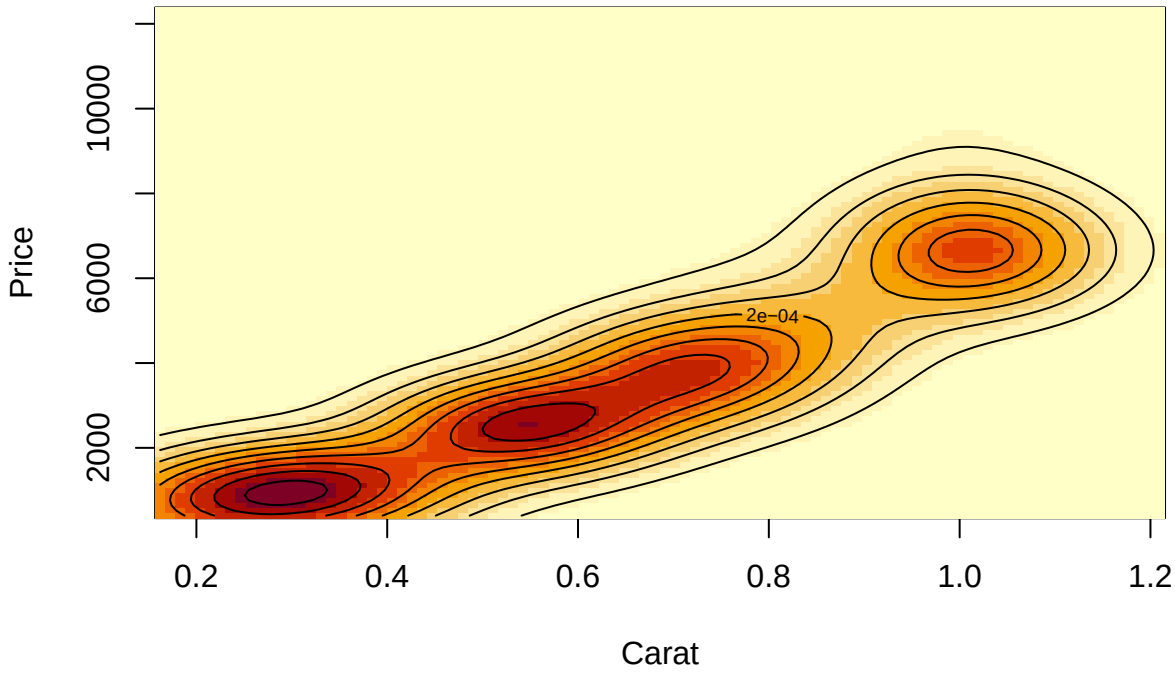


Exercise 8

Show parallel boxplots for the Price, split up into the three different laboratories. You will see a strong difference, which gives the impression that the gradings of the laboratories are unfair. Investigate and argue why this is not the case.

The boxplots reveal significant price differences among diamond laboratories, with laboratory 3 generally commanding higher prices. Laboratory 2 stands out for its notably lower prices, likely due to testing diamonds with very low carat values. This clarifies the price differences observed among the laboratories and prompts investigation into whether these differences stem from unfair pricing practices or variations in diamond quality. By establishing a clear link between carats and price, it may be possible to visualize which diamonds are tested in each laboratory.




BDE (Carat vs Price)
Price
10000
6000
2000
2e−04
0.2
0.4
0.6
0.8
1.0
1.2
Carat

Exercise 9

Try to indentify a functional relationship between Price and Carat. We shall look for a linear function (lm()), but might have to use an appropriate transformation first. Compare the predicted values $\hat{y}$ with the measured values y of Price. What do you conclude?

There is a high positive correlation between Carat and the Price.