I3-TD1
Introduction

Task 1: Galton Inheredity

For this task you need to use Galton heredity data. You can get the data by using the following R-code:

```
#install.packages("HistData") #for the first time you need to install the package
library(HistData)
data(Galton)
Galton<-data.frame(Galton)
```

a. Reconstruct the contingency table between the height of 928 adults children and the average height of their 205 set of parents.

Table 1: The contingency table between the height of 928 adults children and the average height of their 205 set of parents (columns)

	64	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	73
61.7	1	1	1	0	0	1	0	1	0	0	0
62.2	0	1	0	3	3	0	0	0	0	0	0
63.2	2	4	9	3	5	7	1	1	0	0	0
64.2	4	4	5	5	14	11	16	0	0	0	0
65.2	1	1	7	2	15	16	4	1	1	0	0
66.2	2	5	11	17	36	25	17	1	3	0	0
67.2	2	5	11	17	38	31	27	3	4	0	0
68.2	1	0	7	14	28	34	20	12	3	1	0
69.2	1	2	7	13	38	48	33	18	5	2	0
70.2	0	0	5	4	19	21	25	14	10	1	0
71.2	0	0	2	0	11	18	20	7	4	2	0
72.2	0	0	1	0	4	4	11	4	9	7	1
73.2	0	0	0	0	0	3	4	3	2	2	3
73.7	0	0	0	0	0	0	5	3	2	4	0

b. Reconstruct the scatter plot of and regression line between the height of children and average height of their parents.

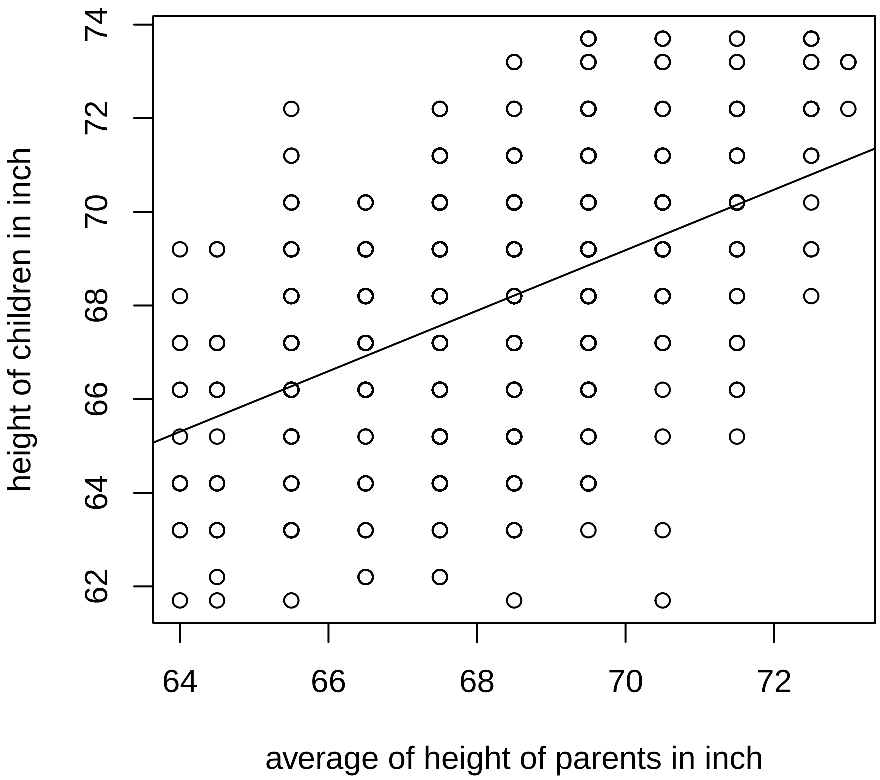


Figure 1: Scatter plot including a regression line between the height of children and the average height of their parents

Task 2: Munich Rent Index of 1999

For this task you need to use Munich rent index of 1999 data. You can get the data by using the following R-code:

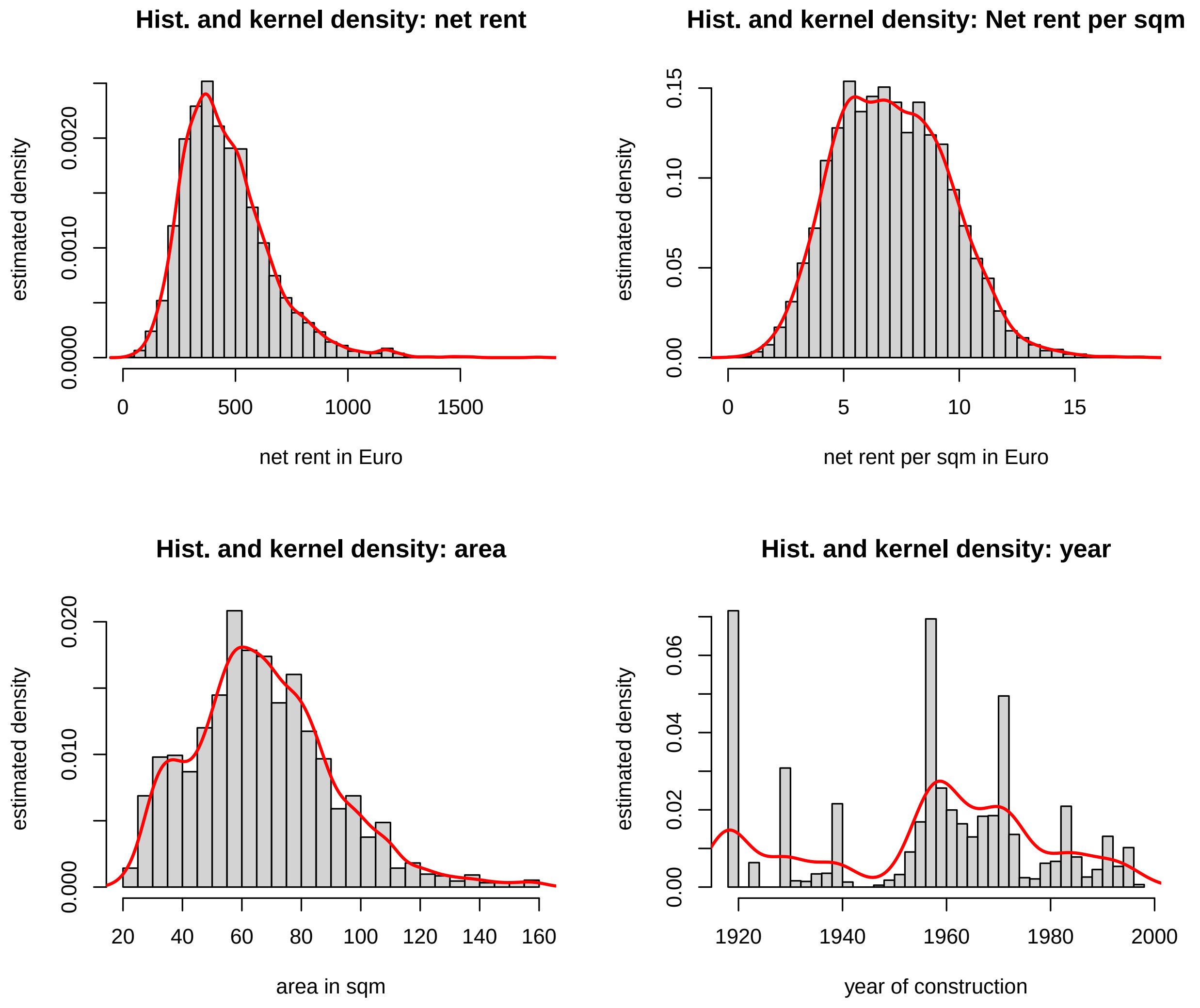
```r
#install.packages("gamlss.data") #for the first time you need to install the package
library(gamlss.data)
data(rent99)
rent99<-data.frame(rent99)
```

Structure of the data:

```r
library(dplyr)
glimpse(rent99)
```

```
## Rows: 3,082
## Columns: 9
## $ rent     <dbl> 109.94872, 243.28204, 261.64102, 106.41026, 133.38461, 339.02~
## $ rentsqm  <dbl> 4.228797, 8.688646, 8.721369, 3.547009, 4.446154, 11.300851, ~
## $ area     <int> 26, 28, 30, 30, 30, 30, 31, 31, 32, 33, 34, 35, 35, 36, 38, 3~
## $ yearc    <dbl> 1918, 1918, 1918, 1918, 1918, 1918, 1918, 1918, 1918, 1918, 1~
## $ location <fct> 2, 2, 1, 2, 2, 2, 1, 1, 1, 2, 2, 1, 2, 2, 1, 1, 2, 2, 1, 2, 2~
## $ bath     <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ kitchen  <fct> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ cheating <fct> 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 0, 1, 1, 1, 0, 0, 1, 0, 1, 0, 0~
## $ district <int> 916, 813, 611, 2025, 561, 541, 822, 1713, 1812, 152, 943, 171~
```

a. Reconstruct the histograms and kernel density estimates below.


Hist. and kernel density: net rent

Hist. and kernel density: Net rent per sqm

Hist. and kernel density: area

Hist. and kernel density: year

Figure 2: Histogram and kernel density estimators for the continuous variables *rent*, *rentsqm*, *area*, and *yearc*

b. Reconstruct the scatter plots below.

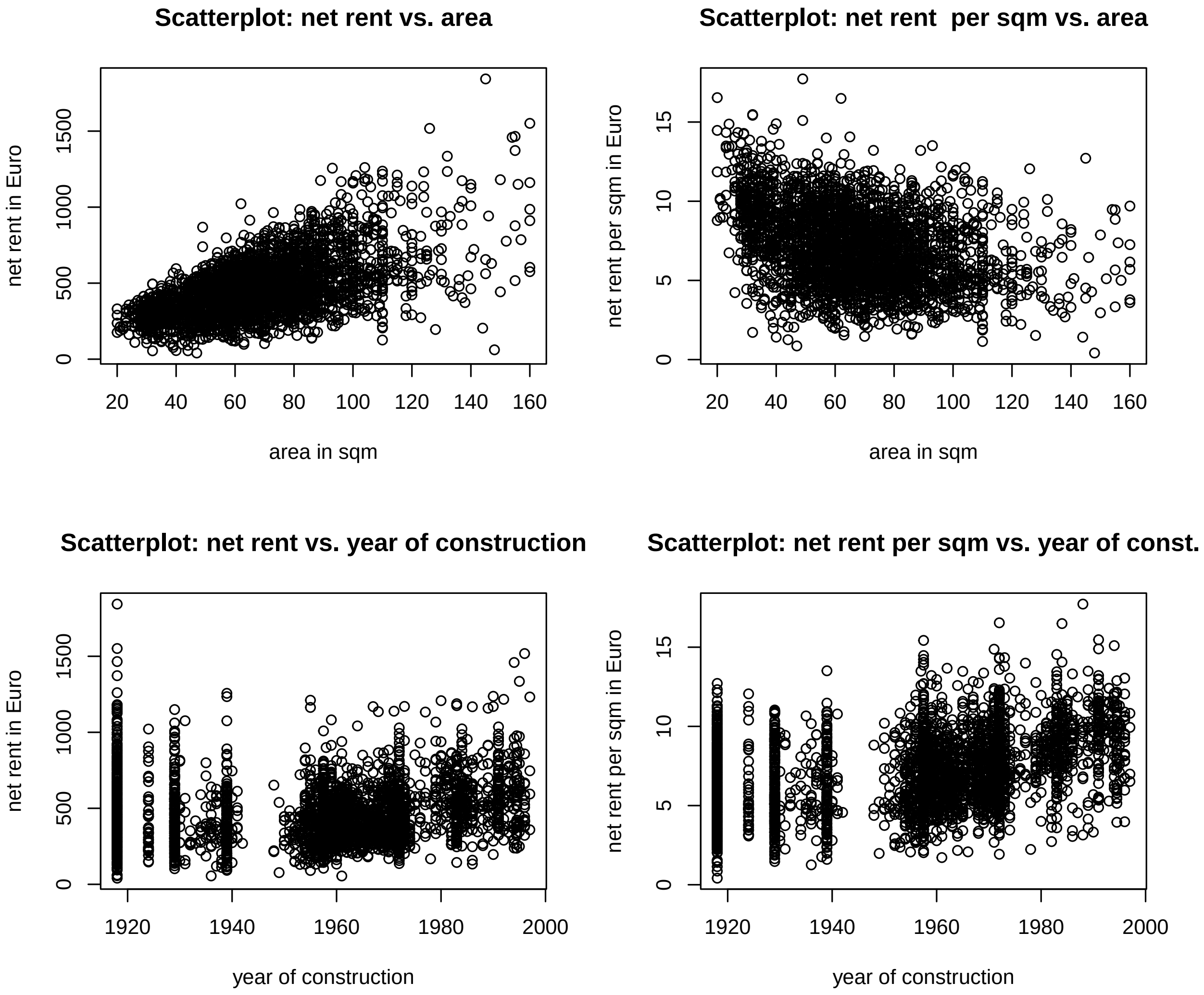


Figure 3: Scatter plots between net rent (left) / net rent per sqm (right) and the covariates area and year of construction

c. Reconstruct the cluster scatter plot below.

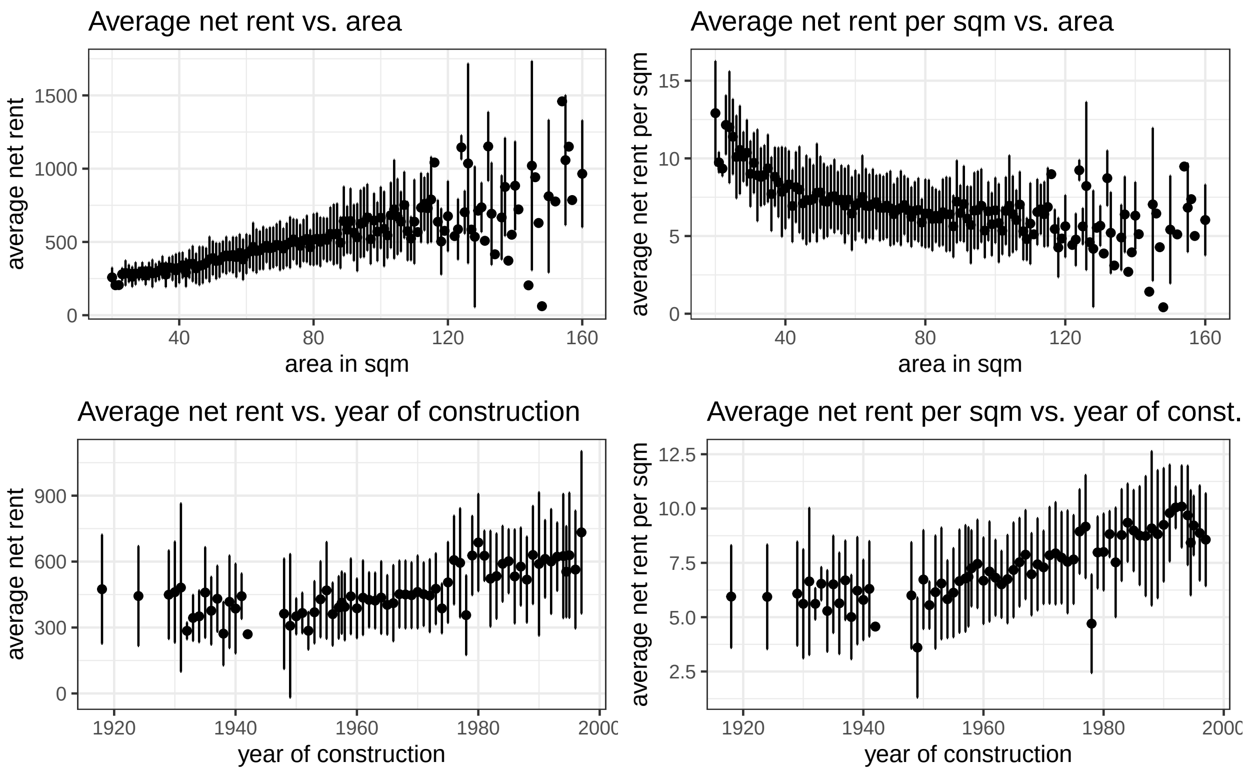


Figure 4: Average net rent (left) and net rent per sqm (right) plus/minus one standard deviation versus area and year of construction

d. Reconstruct the box plots and smooth density estimators below.

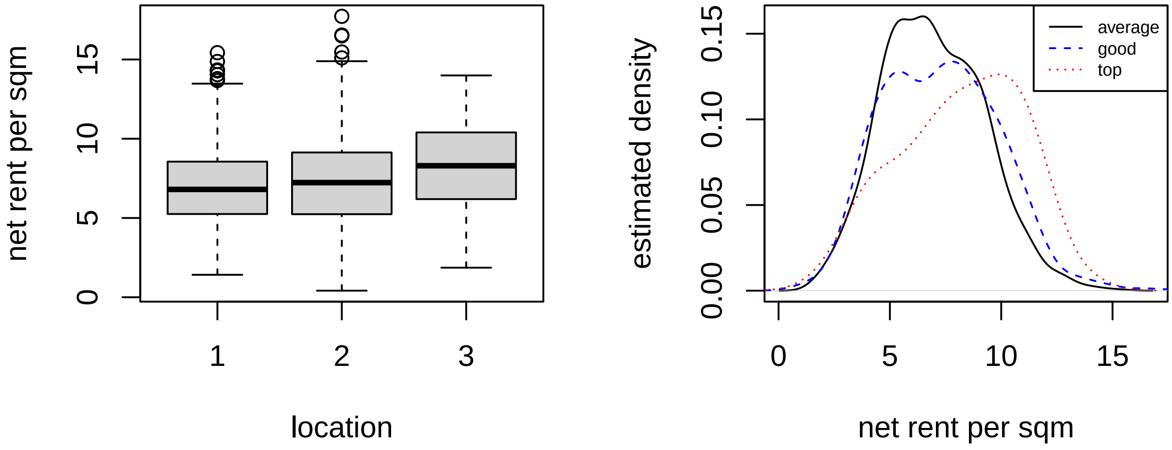


Figure 5: Distribution of net rent per sqm clustered according to location

Task 3: Fuel Consumption

The goal of this task is to understand how fuel consumption varies over the 50 United States and the District of Columbia (Federal Highway Administration, 2001).

Variables in the Fuel Consumption Data:

- `Drivers`: number of licensed drivers in the state
- `FuelC`: gasoline sold for road use, thousand of gallons
- `Income`: per person personal information for the year 2001, in thousands of dollars
- `Miles`: miles of Federal-aid highway miles in the state
- `Pop`: 2001 population age 16 and over
- `Tax`: gasoline state tax rate, cents per gallon

You can obtain the fuel consumption data by using the following `R-code`:

```
#install.packages("alr4") #for the first time you need to install the package
library(alr4)
data(fuel2001)
fuel2001<-data.frame(fuel2001)
```

a. Create 3 more following variables and add to the fuel data consumption.

- `Fuel`: 1000×`FuelC/Pop`
- `Dlic`: 1000×`Drivers/Pop`
- `log(Miles)`: natural logarithm of `Miles`

b. Based on the goal of the task

- Define response variable
- Study the overview of each variable by using initial descriptive and graphical univariate analysis
- Construct the correlation plots across the variables
- Visualize the relation between response variables and predictor variables.