I3-TD1
Introduction

Section 1: Data Visualization

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:

```r
#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.

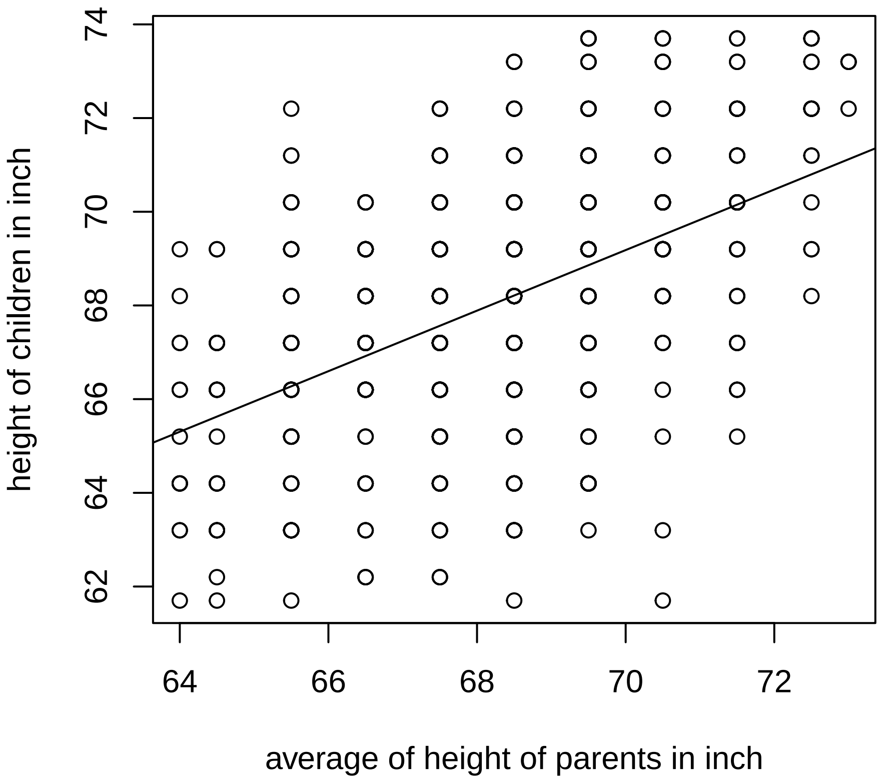


Figure 1: Scatter plot including a regression line between the height of children and the average height of their parents

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.

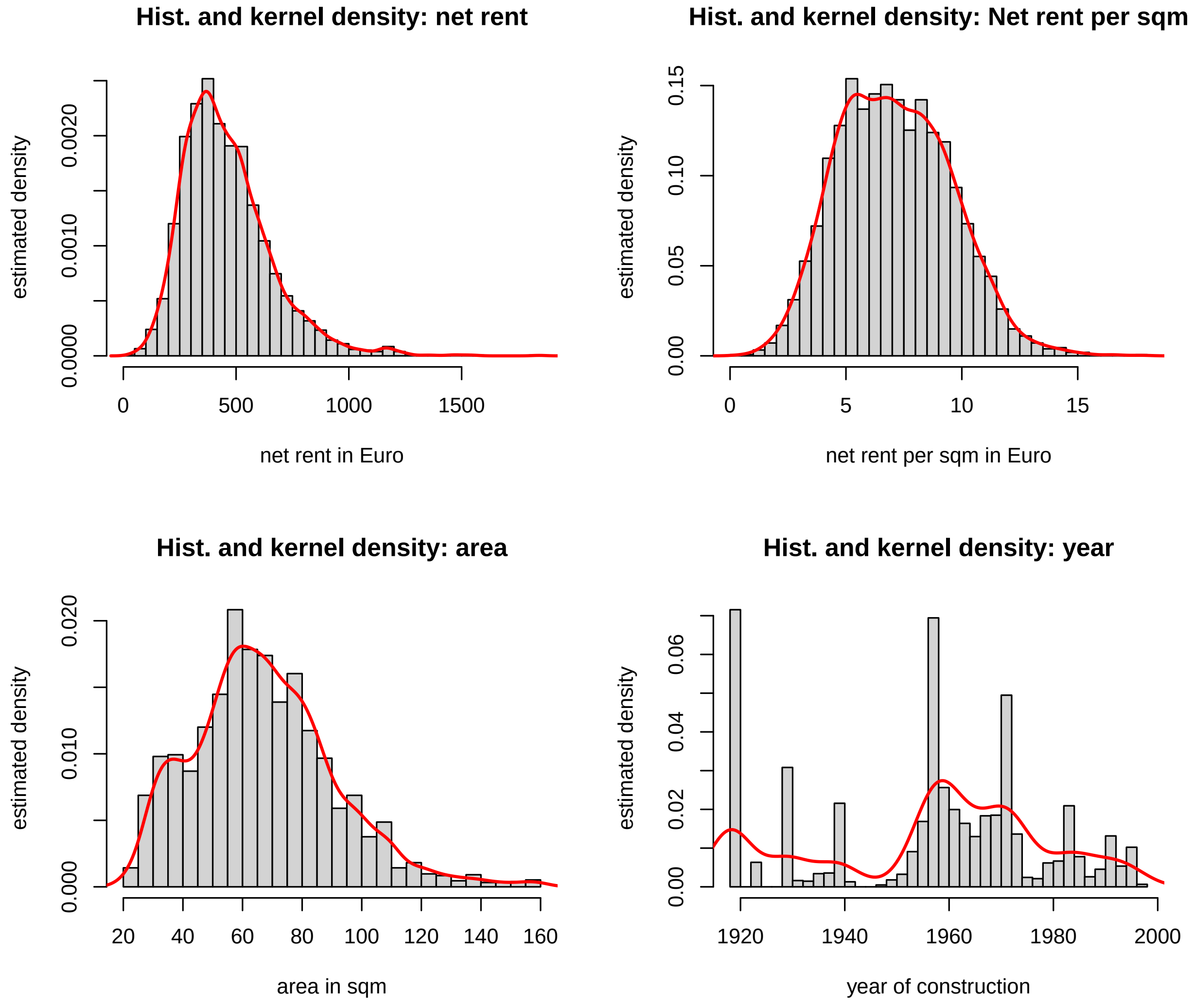


Figure 2: Histogram and kernel density estimators for the continuous variables *rent*, *rentsqm*, *area*, and *yearc*

b. Reconstruct the scatter plots below.

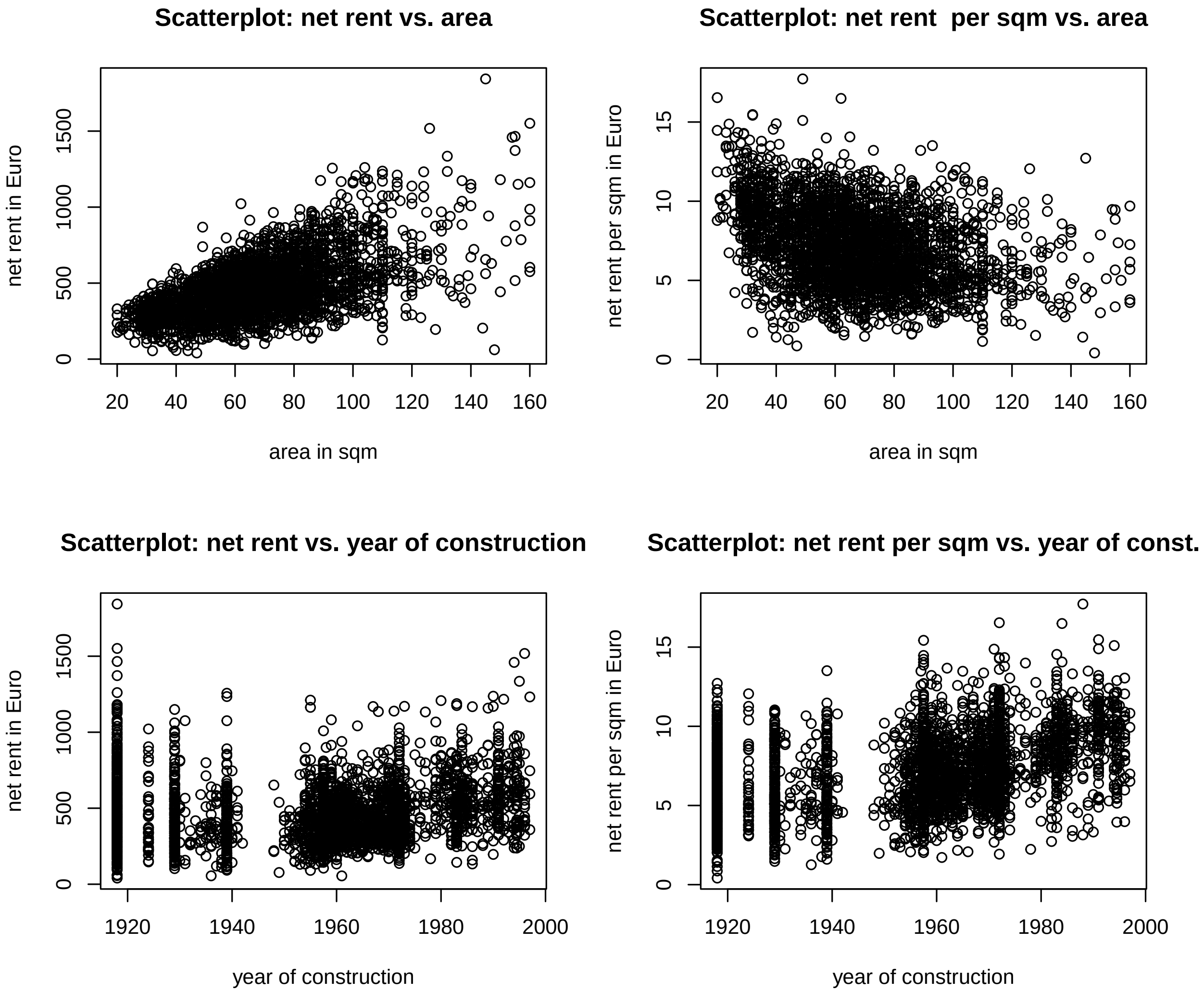


Figure 3: Scatter plots between net rent (left) / net rent per sqm (right) and the covariates area and year of construction

c. Reconstruct the cluster scatter plot below.

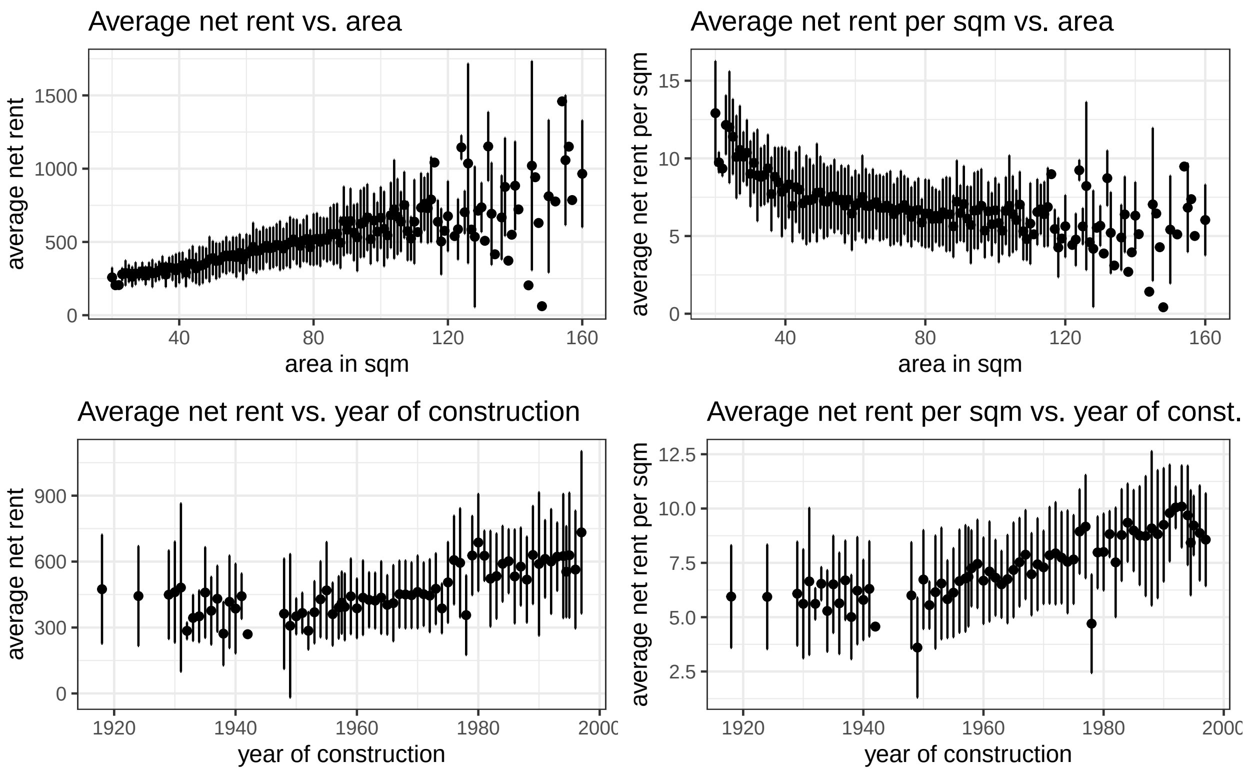


Figure 4: Average net rent (left) and net rent per sqm (right) plus/minus one standard deviation versus area and year of construction

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

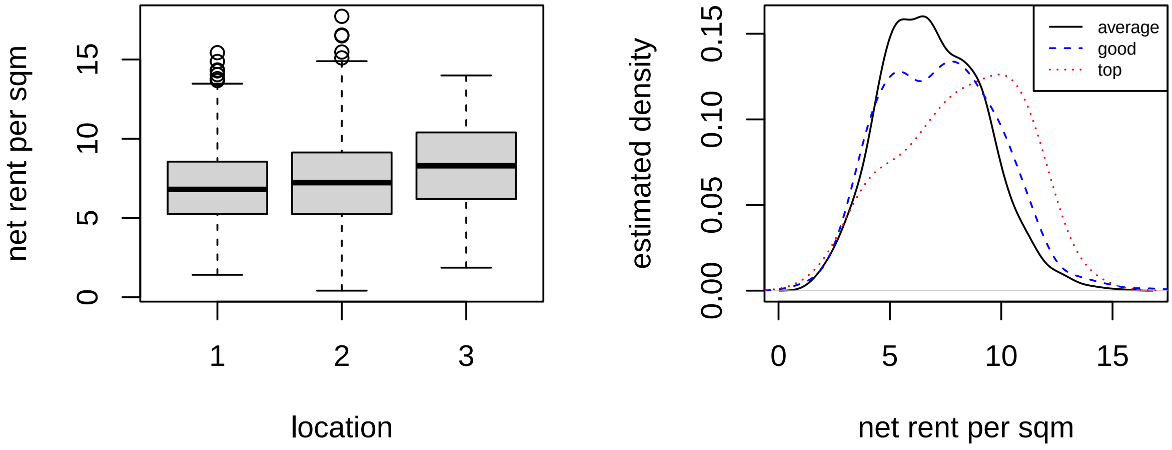


Figure 5: Distribution of net rent per sqm clustered according to location

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`:

```r
#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.

Section 2: the nassCDS data

In this section of the exam, we focus on the nassCDS data which is a US data from police-reported car crashes (1997-2002) in which there is a harmful event (people or property). Data are restricted to front-seat occupants, include only a subset of the variables recorded. More information about the dataset can be found using the following link: https://www.rdocumentation.org/packages/DAAG/versions/1.22/topics/nassCDS. The data is a part of the DAAG R package. To get an access to the data you first need to install the package.

```r
library("DAAG")
data(nassCDS)
#names(nassCDS)
#head(nassCDS)
#str(nassCDS)
```

Question 1

In this question we focus on the accident's outcome (the variable dead) and seatbelt usage (the variable seatbelt).

1. How many individuals used seatbelt?
2. What is the distribution of seatbelt usage across the accident's outcome factor ? Produce a 2X2 table that shows the number of seatbelt users (belted/none) and accident's outcome (alive/dead)?
3. Write a function that can be used to conduct inference for proportions in two independent populations. The null hypothesis is that there is no difference between the proportions in the two populations. Test the null hypothesis against a two sided alternative. The input of the function should be the 2X2 table in the previous item (Question 1.2) and the output should be the test statistic and the p value. Apply your function to test the null hypothesis that the proportion of deaths among individuals who used seatbelt is equal to the proportion of deaths among the individuals who did not use seatbelt.
4. Use a barplot to visualize the distribution of the seatbelt usage across the factor levels of the accident's outcome.

Solution for question 1.1

Solution for question 1.2

Solution for question 1.3

Solution for question 1.4

Question 2

In this question we focus on the outcome of the accident (dead/alive, the variable dead) and the
age of the occupant (the variable ageOFocc).

1. What is the mean and standard deviation of the age of occupant by accident outcome?
2. Use a boxplot to visualize the distribution of the occupants' age by accident outcome and add the data points on the boxplot.
3. Calculate a 95% confidence interval for the mean difference of the age of occupant using t distribution.

Solution for question 2.1

Solution for question 2.2

Solution for question 2.3

Question 3

1. Visualize the distribution of the occupant age by sex.
2. How many occupants over the age of 50 years old survived the accident?
3. Add a binary variable AgeOFocc_class that takes the value of 1 when the occupant age is over 50 years and 0 for when the occupant age is 50 years or less.
4. Create a data frame, nassCDS_o50, containing occupants older than 50 years old. This data frame should contain the variables dead, airbag, weight, and injSeverity. Remove the observations with missing values.
5. What is the dimension of the new data ?

6. Among the occupants who are older than 50 years old, use a barplot to visualize the distribution of airbag across the levels of the accident outcome (dead/alive). The variable dead should be on the x-axis.
7. Among the occupants who are older than 50 years old visualize the distribution of airbag across the level of injury sevirity (the variable injSeverity).

Solution for question 3.1

Solution for question 3.2

Solution for question 3.3

Solution for question 3.4

Solution for question 3.5

Solution for question 3.6

Solution for question 3.7

Question 4

Write a R function that receives as an input the nassCDS dataset. The function should conduct the following analysis:

1. Select only the observations for which the accident outcome is "dead".
2. Calculate percentage of deaths out of the overall number of observations.
3. Calculate the percentages of females and males among the occupants who died in the accident.
4. Show the most frequent severity of their injuries.
5. Calculate the minimum and maximum age of the occupant (the variable ageOFocc).
6. produce a histogram with the severity of injuries on the x axis and the occupant's age on the y axis
7. This **SINGLE** Function should return **two** outputs:
   - Numerical output: 4.2,4.3,4.4 and 4.5 as a table.
   - Graphical output: 4.6 as a plot.

Solution for question 4.1

Solution for question 4.2

Solution for question 4.3

Solution for question 4.4

Solution for question 4.5

Solution for question 4.6

Solution for question 4.7

Question 5

1. Create a new data frame which contains only occupants who used seatbelt.

2. How many occupants used seatbelt ?
3. Among the individuals who used seatbelt, how many died and how many survived the accident ?
4. Among the individuals who used seatbelt, how many were drivers among the individuals who died and how many were passengers among the individuals who survived the accident (use the variable occRole to identify drivers/passengers) ?
5. Sort the data frame according to the injury's severity and the occupant age.
6. Print the 25 occupants with the highest weight.

Solution for question 5.1

Solution for question 5.2

Solution for question 5.3

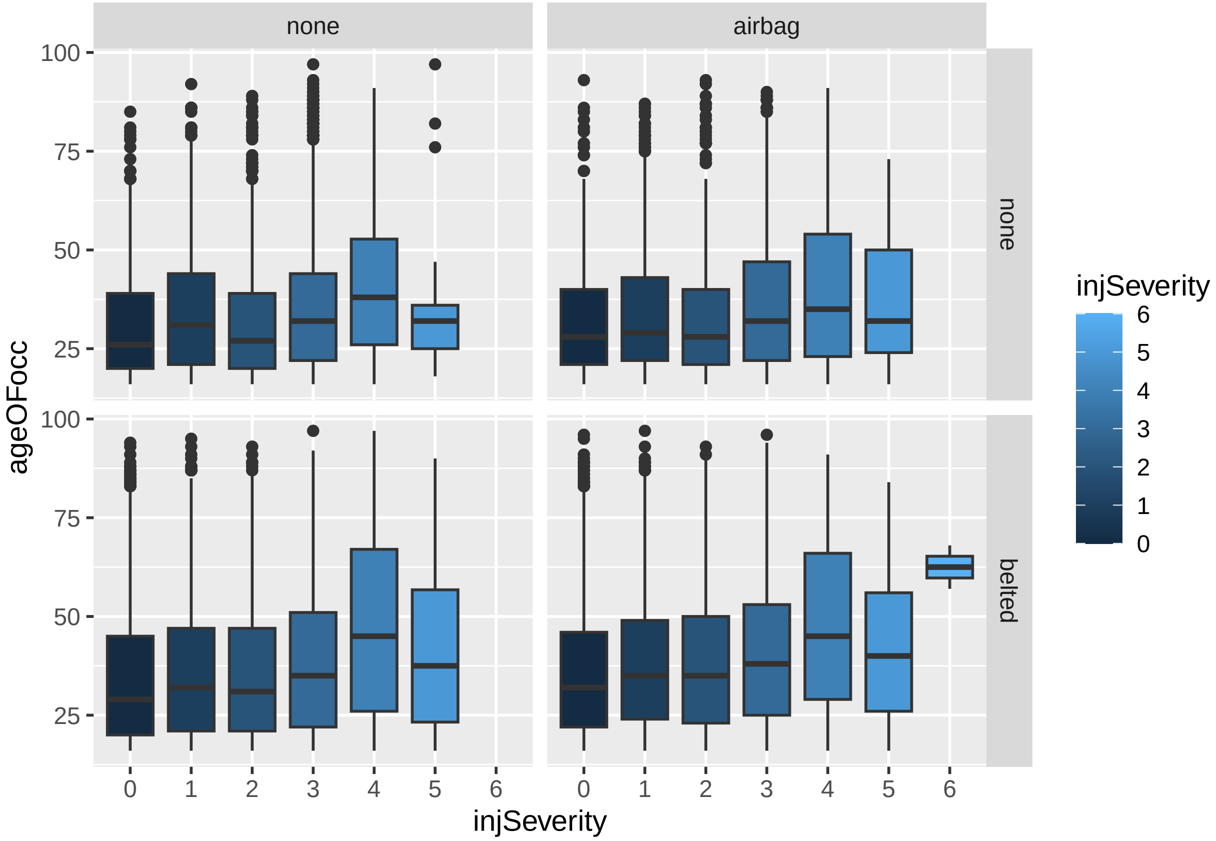
Solution for question 5.4

Solution for question 5.5

Solution for question 5.6

Question 6

1. Produce the figure below.



Solution for question 6.1