Programming in R: Exam part 1 (13/12/2024)

Ziv Shkedy, Rahmasari Nur Azizah, Thi Huyen Nguyen, Bernard Isekah Osangir, Rudradev Sengupta, Ewoud De Troyer, and Marijke Van Moerbeke. (2024/2025)

Introduction

General information

The exam consists of 4 parts in which you are asked to conduct analysis of different datasets. The datasets are included in different R packages and you need to install the packages to access the data. Your analysis should be done using R and your answers should be given in R code. For example, if the question is

Question 0 (Example)

1. Draw a random sample of size 100 from N(0,1).
2. Produce a histogram for the sample.

Your solution should be

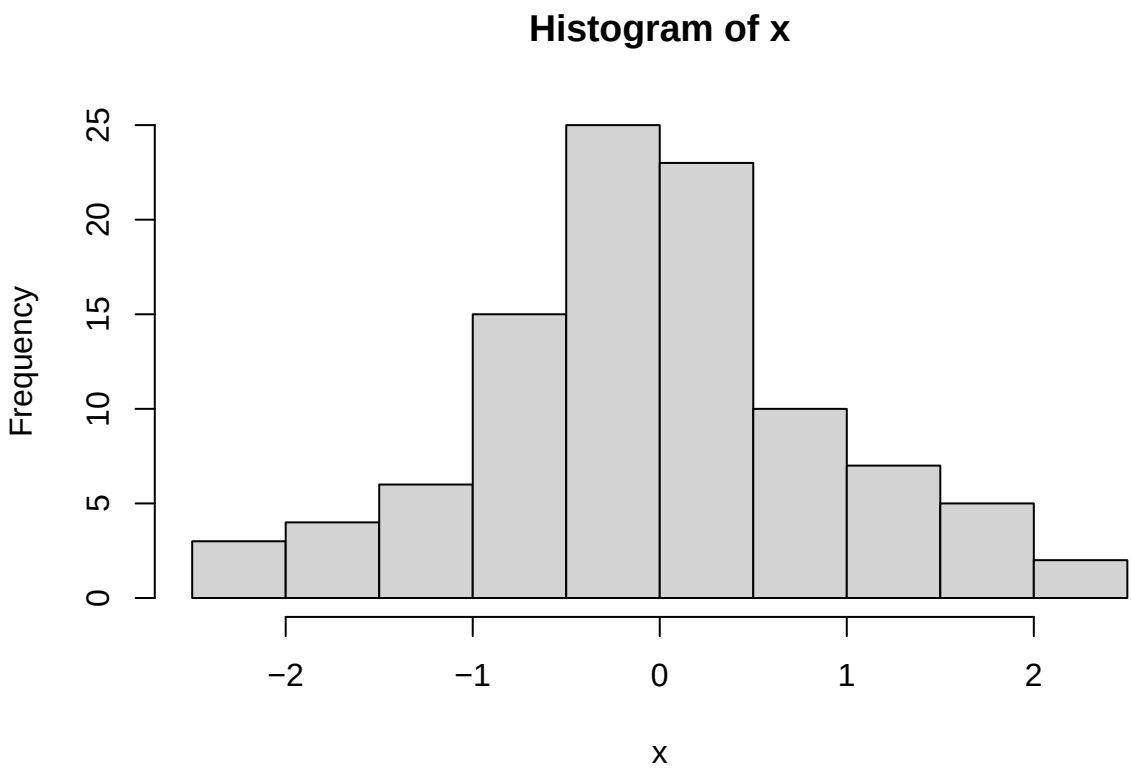
Solution for question 0.1:

```
x=rnorm(100,0,1)
print(x)
```

```
##   [1] -0.0004800434  2.1032443629 -0.9715771329  0.0688182704  0.0713783569
##   [6] -0.9449268278  0.3898447252 -0.2939145050 -1.1230372986 -0.9942495039
##  [11]  0.0698194274 -0.0881932432  0.8248811793 -0.2747765569  1.9303069973
##  [16] -1.6444944997  0.1289055031  0.2326756145 -1.4268388850 -0.3356922928
##  [21] -0.2698366645  1.3554671214 -0.4530292273 -0.3749730730  0.3762447534
##  [26]  0.3336555200 -0.5136537909 -0.0871932801  0.9952475290 -0.1266698267
##  [31]  0.2652165884 -0.4026453771 -1.0585623094 -1.0639087709  1.5390743509
##  [36] -0.6571804657  0.2403832330 -0.3878291102 -0.2359783834  0.9455464540
##  [41]  1.4459499373 -0.6942828133  0.8972222983  0.7694925275  0.4304087872
##  [46] -2.3844073853  0.2447215229  1.4030788748  0.6802353363 -0.7968145796
##  [51] -0.0531475315  1.5037955800  0.9798449814 -0.4528498136 -0.5228384641
##  [56]  0.1104354964 -0.1910919776  2.0306631425 -0.4497572941 -0.6636192093
##  [61] -0.7320420631  0.0766524904 -0.7408163023 -0.2037513583 -1.9988402309
##  [66]  0.3517501560  0.0947020832  1.2837792244 -0.7044620148  0.2570959587
##  [71]  0.2083396680  0.5543814846  0.9466559181  0.3850104412  0.1861474604
##  [76] -1.2397799291 -2.2283697702  0.4069448391 -0.0267264872 -0.0524905128
##  [81]  0.5382596114  1.8529178941 -0.3322765777 -1.6064164714 -1.1420341629
##  [86] -0.1474263289  0.1005690730  1.0919408865 -0.4341931949 -2.1192032576
##  [91]  0.3075768394  1.2321595018 -0.5570607780  1.5453885855 -0.8729729585
##  [96]  1.2080944114 -0.3398421468 -0.3401486009 -1.5986273618 -0.6859550373
```

Solution for question 0.2:

```
hist(x)
```



You **do not** need to explain your R code. For example, you do not need to write: "the function hist() was used to produce the histogram." Your answers to the questions should be the R code that you used to produce the output.

What do you need to submit as a solution for the exam?

You need to submit the following materials:

1. R markdown program that can be used to conduct the analysis.
2. PDF file version of the solution (produced using the R markdown program).
3. HTML file version of the solution (produced using the R markdown program).
4. In Question 4, you are asked to produce a presentation using R markdown. For this question you need to submit:

- R markdown program that was used to produce the presentation.
- PDF file of the presentation.

5. Other files (xls, txt etc.) if you are asked to produce these files in a specific question.

What you do not need to write?

You **do not** need to interpret the results!!! For example, if the question is to fit a One-Way ANOVA model, you do not need to formulate the model and to interpret the results. This means, for example, that you do not need to write "the p-value is 0.007 indicating on a significant effect of the factor."

When do you need to submit the solution?

- Date: 10/01/2025.
- Time: 17:00.

How to submit the solution?

You should upload the solutions in BB. You will receive information about the submission **by email**.

Part 2 of the exam

The second part of the exam (part 2) will be available online in BB on **13/01/2025** at **08:00-11:30**.

Oral exam

The oral exam will take place on 13/01/25, 14/01/25, and 15/01/25. You will receive information about your exam date and time by email. The schedule is available online in BB.

Part 1 : the Hitters data

For the analysis of this part we use the data Hitters which is a part of the R package ISLR. This is the major league baseball data from the 1986 and 1987 seasons. More information can be found in https://rdrr.io/cran/ISLR/man/Hitters.html. The code below can be used to access the data

```
library(ISLR)
data(Hitters)
names(Hitters)
```

```
##  [1] "AtBat"     "Hits"      "HmRun"     "Runs"      "RBI"       "Walks"
##  [7] "Years"     "CAtBat"    "CHits"     "CHmRun"    "CRuns"     "CRBI"
## [13] "CWalks"    "League"    "Division"  "PutOuts"   "Assists"   "Errors"
## [19] "Salary"    "NewLeague"
```

Question 1

In this question we focus on the player's division at the end of 1986 (the variable Division) and the number of runs in 1986 (the variable Runs).

1. How many observations are there in each category of the variable Division?
2. Produce the table below.

```
## # A tibble: 2 x 2
##   Division median_runs
##   <fct>          <int>
## 1 E                 49
## 2 W                 46
```

3. Conduct a Wilcoxon test for two independent samples to test if the number of runs (the variable Runs) is equal across the divisions.
4. Produce Figure 1.1. Note that the points in red are the sample means.

Solution Q1.1

Solution Q1.2

Solution Q1.3

Solution Q1.4

Question 2

In this question we focus on the variable number of walks in 1986 (the variable Walks) in addition to the variables from Q1.

1. Produce Figure 1.2, 1.3, and 1.4.

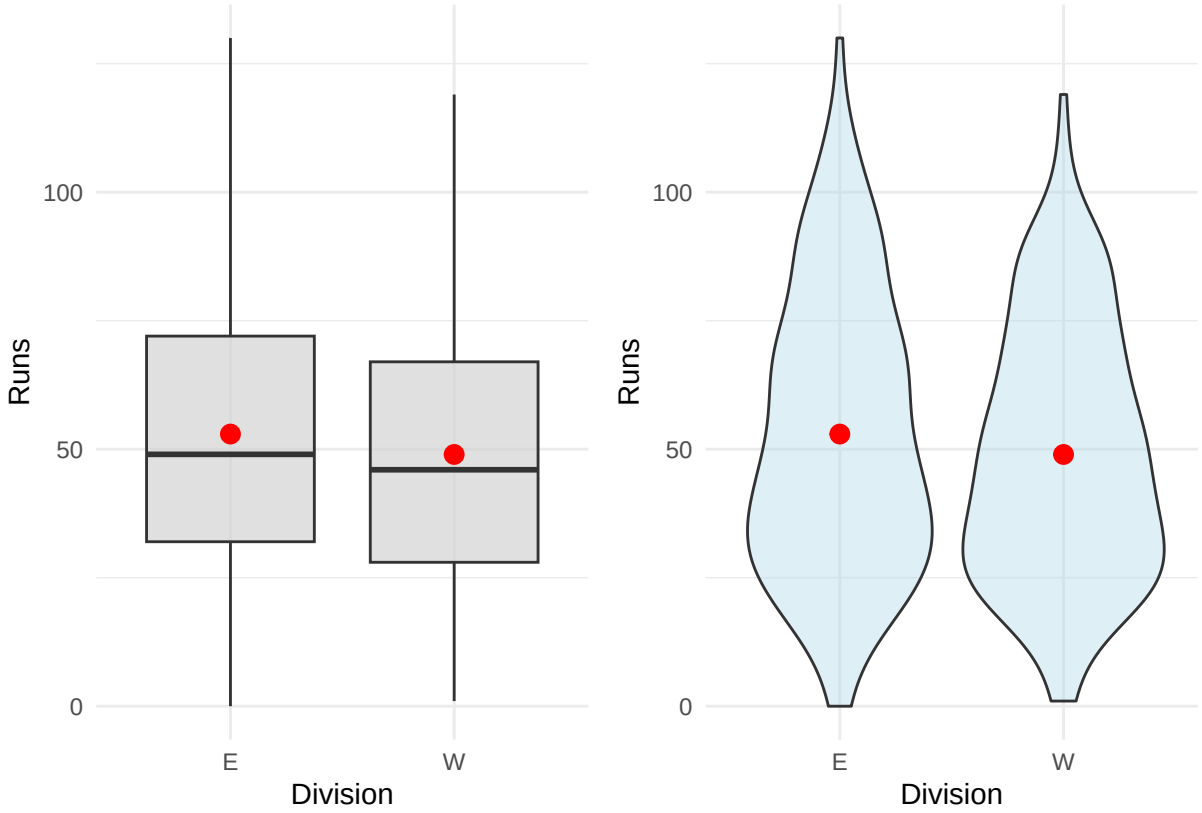


Figure 1.1

2. Produce the table below (correlation between the variables Walks and Runs by division group).

```
## # A tibble: 2 x 2
##   Division Correlation
##   <fct>          <dbl>
## 1 E              0.745
## 2 W              0.718
```

3. Fit the following linear regression model:

$$\text{Runs}_i = \beta_0 + \beta_1 \text{Walks}_i + \beta_2 \text{Division}_i + \varepsilon_i$$

.

4. Define a R object, fit.coef, in which you store the parameter estimates of the coefficients and print the object.

5. Let e_i the residual obtained for the regression model in Q2.3. Let es_i the standardized residual given by:

$$es_i = \frac{e_i}{MSE}$$

Check if the standardized residuals follow a standard normal distribution using a qq normal probability plot shown in Figure 1.5.

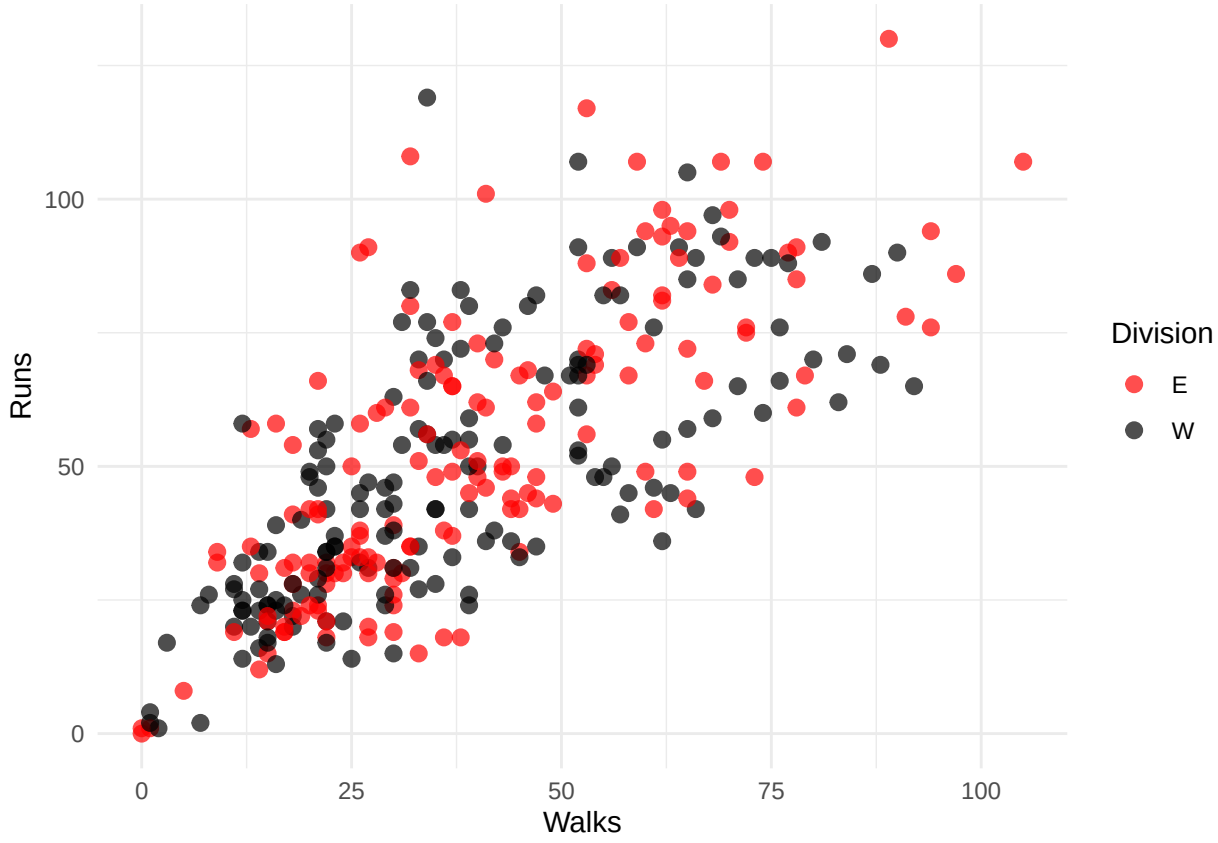


Figure 1.2

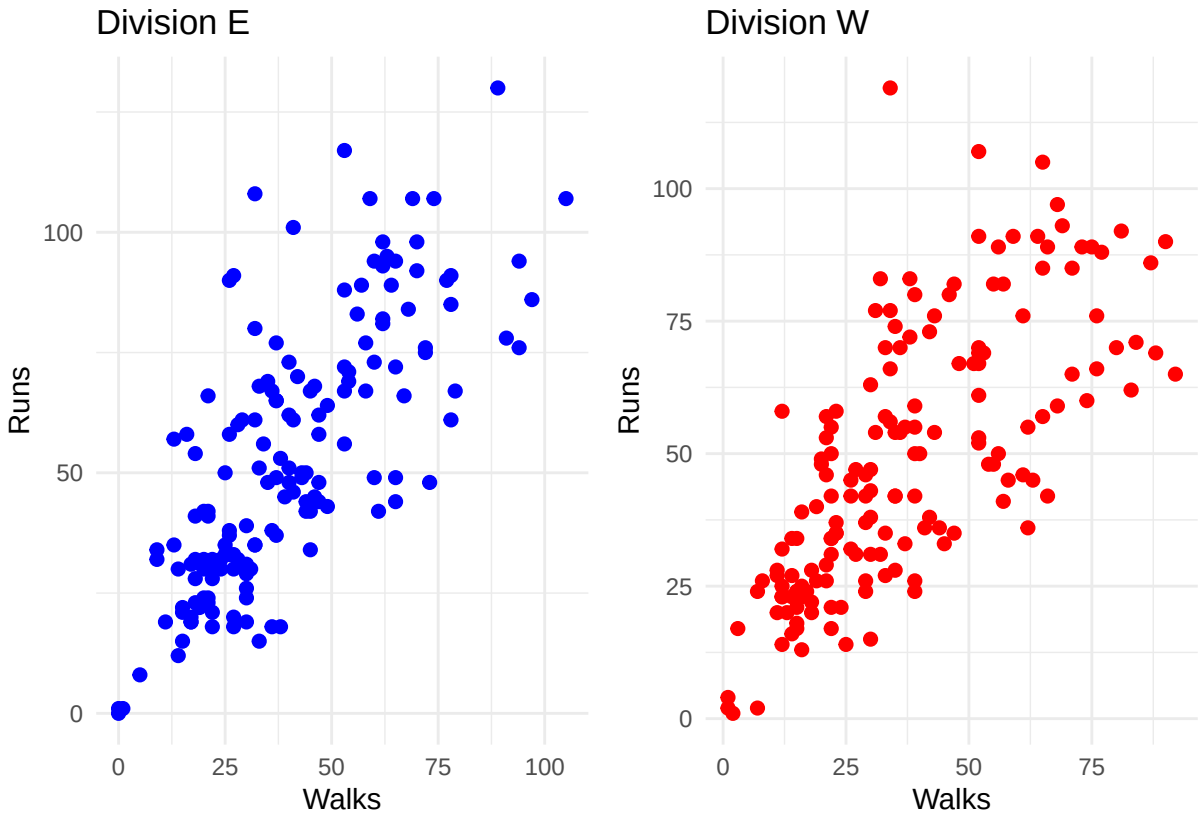


Figure 1.3

Solution Q2.1

Solution Q2.2

Solution Q2.3

Solution Q2.4

Solution Q2.5

Question 3

Create a new dataset in which only observations with number of runs in 1986 greater than 30 are included. The following variables should be included in the dataset: Hits, HmRun, Runs, Walks and Division.

1. How many observations are included in the new dataset?
2. Sort the new data according to the variable the number of hits in 1986 (the variable Hits). Print the top 5 observations in each division.
3. Export the new dataset that was created in Q3.1 as an excel file (and include the data in out output that you submit as a solution for the exam).

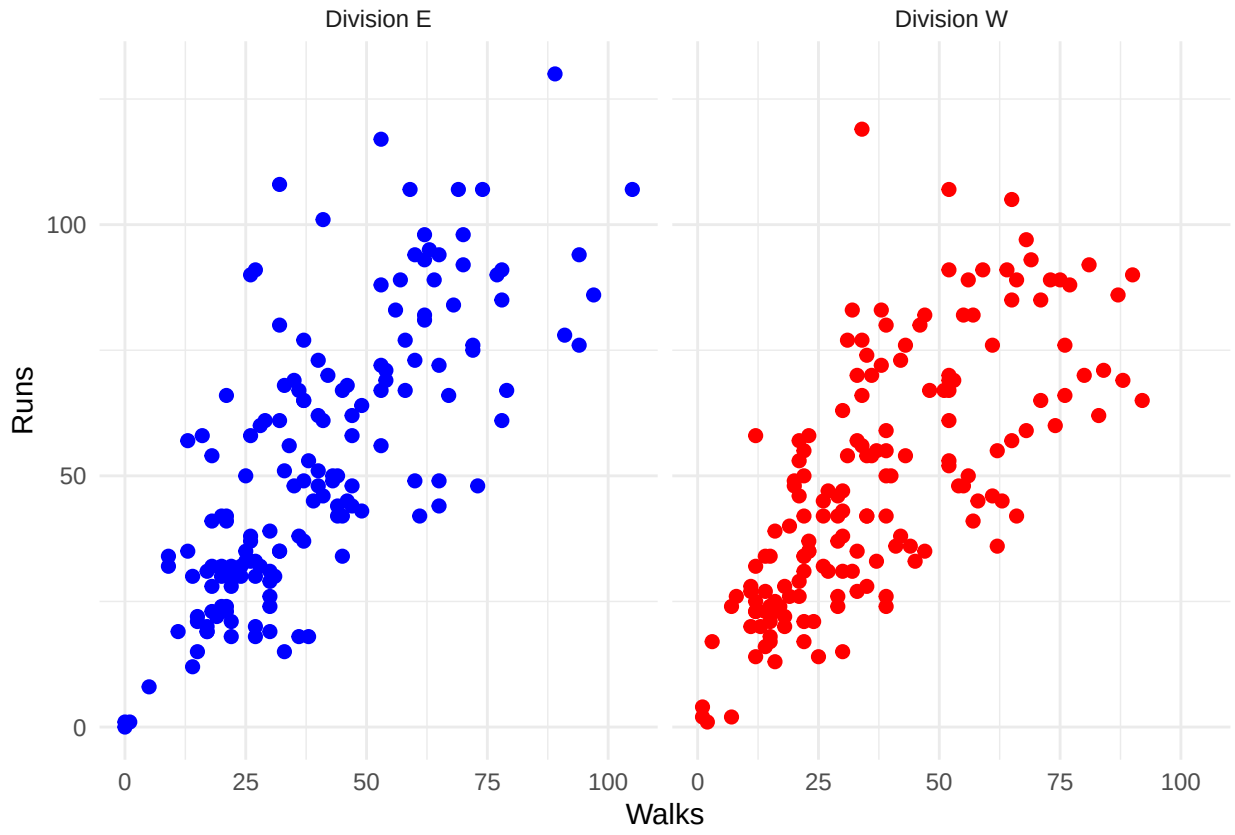


Figure 1.4

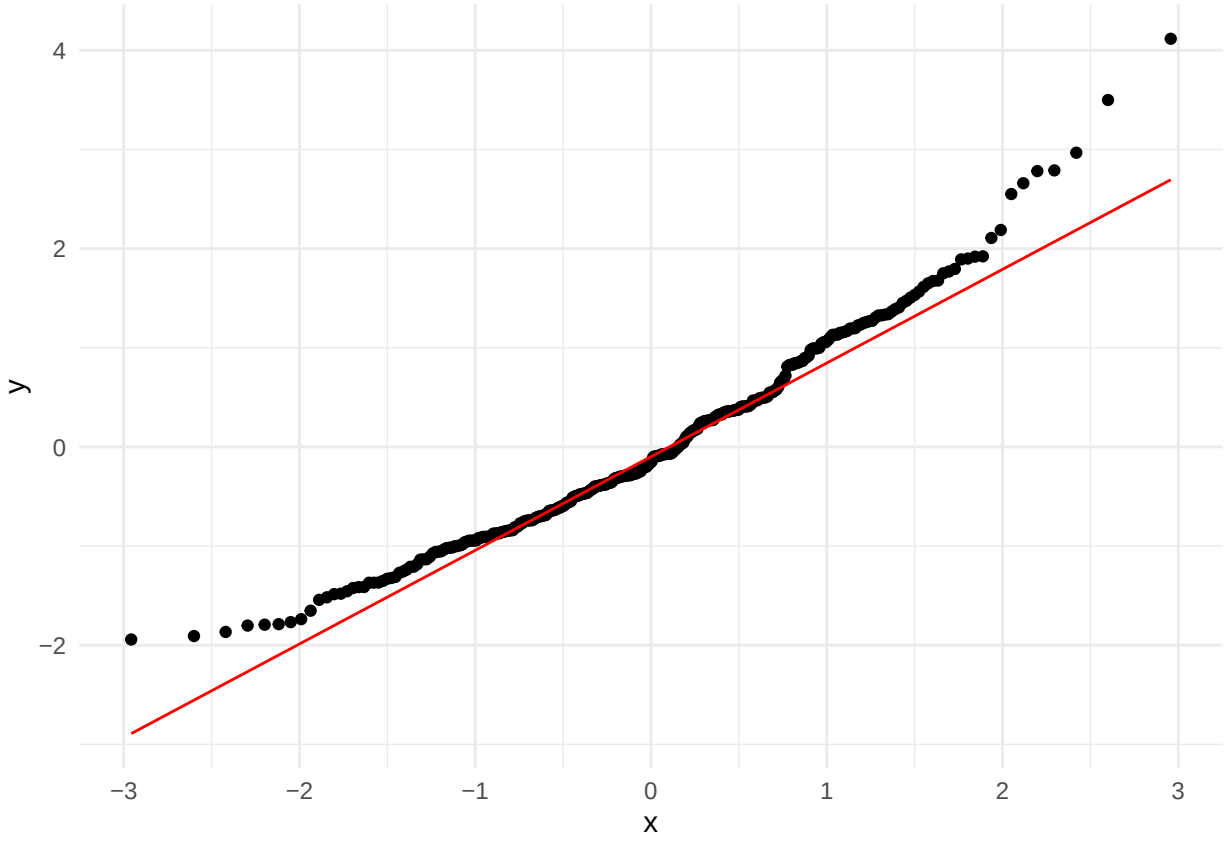


Figure 1.5

Solution Q3.1

Solution Q3.2

Solution Q3.3

Part 2 : the Hitters data

Question 4

In this part you need to prepare a presentation of 5 slides **using R markdown** about the analysis that you conducted in part 1.

1. Your presentation should include a title page, at least one slide with text, at least one slide with a Graphical display and at least one slide in which you print an output from a statistical model or a statistical test applied to the data.
2. Create from the presentation a PDF file and add this file with the Rmd file that was used to produce the presentation to your solution's output.

Part 3: the Boston data

In this part of the exam, we focus on the Boston dataset which is a part of the MASS R package. To access the data you need to install the package. More information can be found in https://www.statology.org/boston-dataset-r/. Use the code below to access the data.

```
library(MASS)
data(Boston)
names(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "black"   "lstat"   "medv"
```

Question 5

1. How many observations and variables are included in the dataset? How many missing values, per each variable, are there in the dataset?
2. Calculate the minimum and maximum for the variables crim, zn and indus across the levels of the variable chas. Produce the panel below.

```
## # A tibble: 2 x 7
##    chas crim_min crim_max zn_min zn_max indus_min indus_max
##   <int>    <dbl>    <dbl>  <dbl>  <dbl>     <dbl>     <dbl>
## 1     0  0.00632    89.0       0    100      0.46      27.7
## 2     1  0.0150      8.98      0     90      1.21      19.6
```

3. Count the number of homes that are near the Charles River (i.e., observations with chas equal to 1) vs. those that are not located near to the Charles river (observations with chas equal to zero).
4. For each level of the variable chas, calculate the average number of rooms per dwelling (the variable rm). Sort the data according to the average number of rooms per dwelling. Print the average number of rooms per dwelling.

Solution Q5.1

Solution Q5.2

Solution Q5.3

Solution Q5.4

Question 6

1. Create a new data frame, Boston2, for which the crime rate (the variable crim) is lower than 5 and the proportion of lower-status population (the variable lstat) is lower than 10. How many observations are included in this data frame?
2. What are the average median home value (the variable medv) and the average number of rooms per dwelling (the variable rm) for the dataset created in Q6.1.
3. Visualize the relationship between the variables medv and rm for the dataset created in Q6.1 as shown in Figure 6.1.

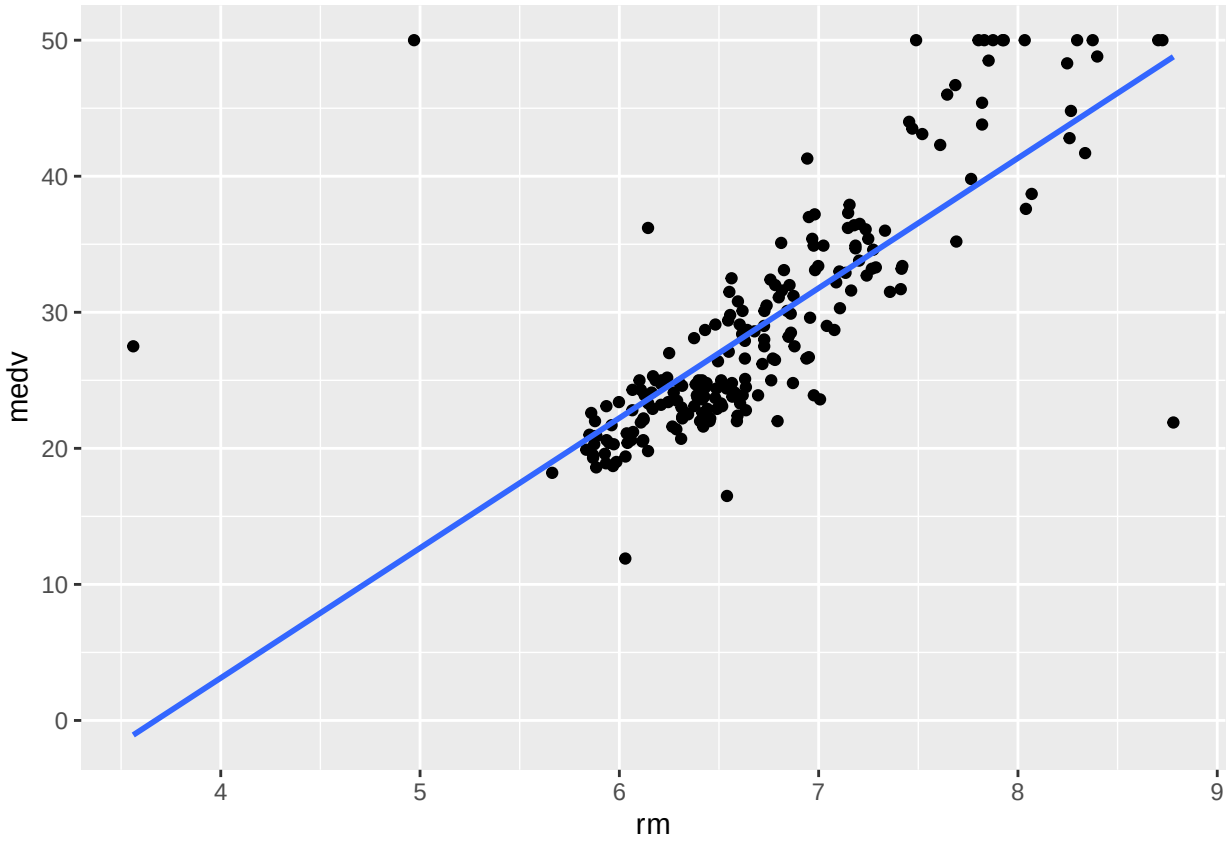


Figure 6.1

4. Identify the outlying observations for which the average number of rooms per dwelling (the variable rm) is smaller than 5 or higher than 8.75 in the dataset created in Q6.1. Add the value of the average number of rooms per dwelling to the figure (inside the frame) as shown in Figure 6.2.

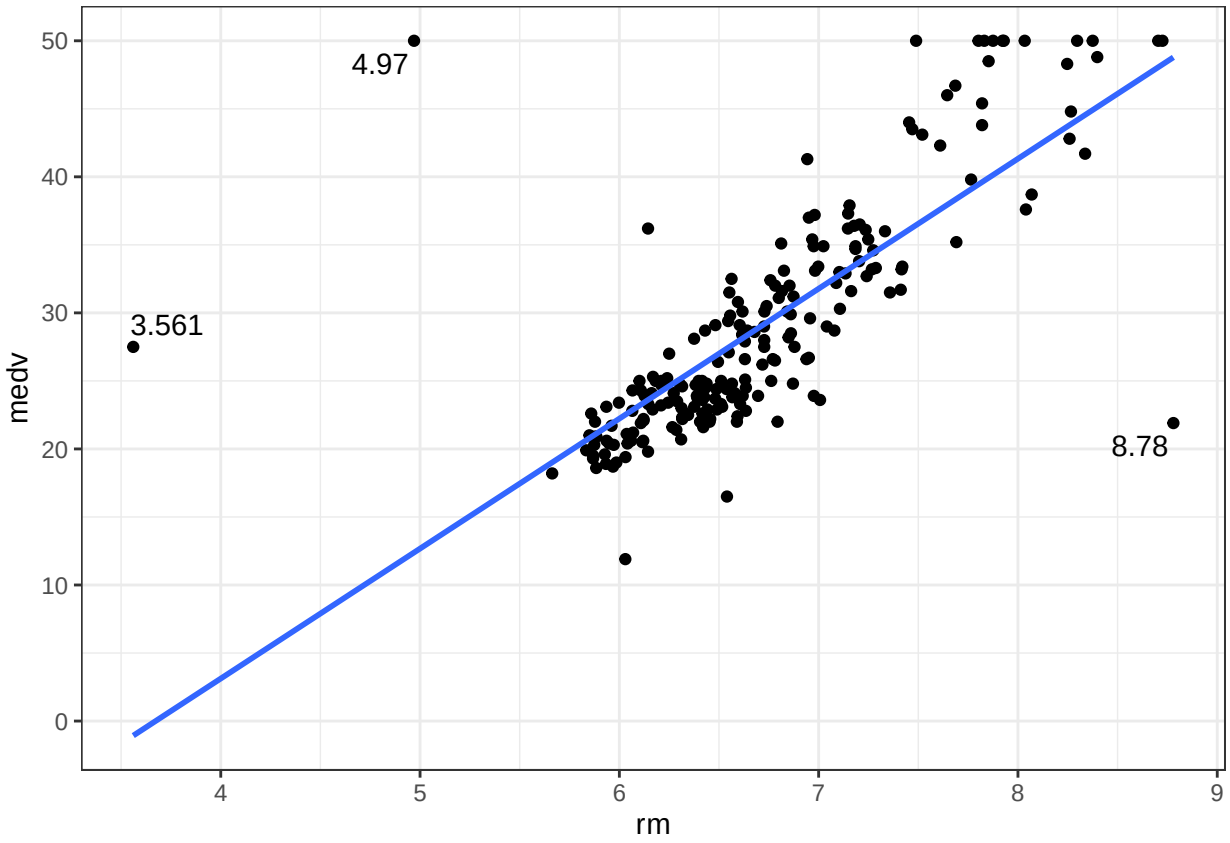


Figure 6.2

5. Exclude the three outliers that was marked in Figure 6.2 from the data and produce Figure 6.3, label points with extreme values of medv i.e, observations for which the value of the variable medv is above the 90th percentile.

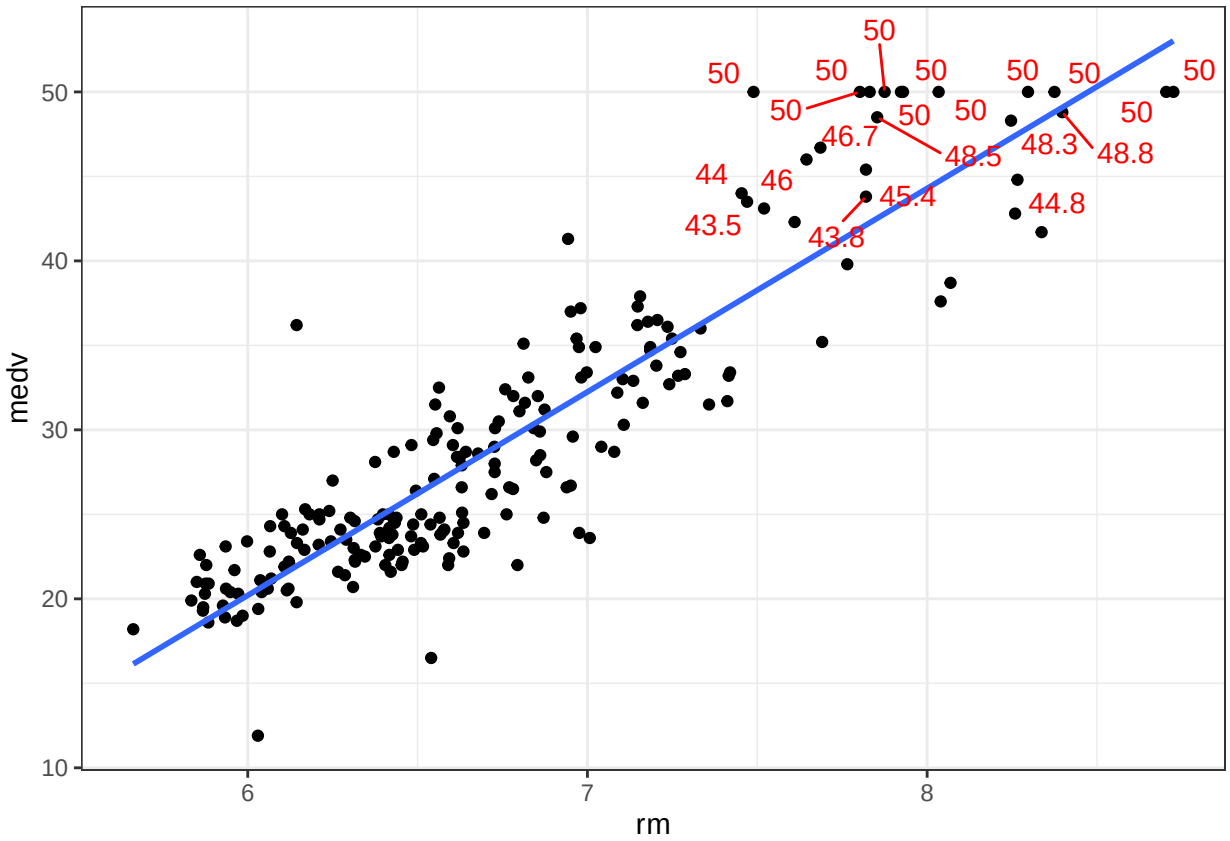


Figure 6.3

Solution Q6.1

Solution Q6.2

Solution Q6.3

Solution Q6.4

Solution Q6.5

Question 7

1. Define a new categorical variable crim_cat in the following way: Re-code the variable crim into three categories:

   crim <5: Low.
   crim 5-15: Medium.
   crim >15: High.

Count how many observations are included in each category.

2. Produce the pie plot and the barplot in a figure with one row of two columns, as presented in Figure 7.1.

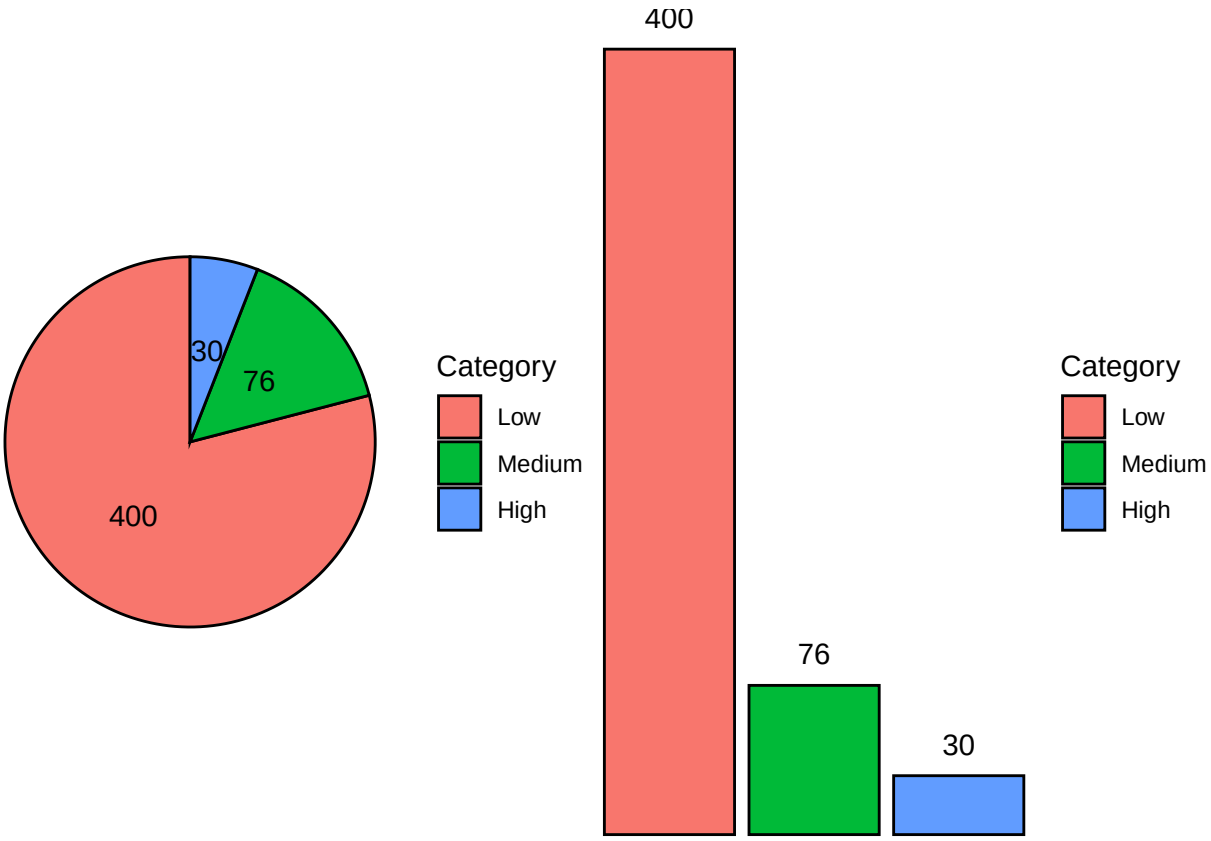


Figure 7.1

3. Produce the frequency table below and Figure 7.2 (the figure shows the proportion of each category of the variable crim_cat across the levels of the variable chas).

```
##
##     Low Medium High
##   0 370     71   30
##   1  30      5    0
```

4. Use a chi-square test to test the hypothesis whether or not a home is near the Charles River (the variable chas) and the categorical crime variable crim_cat are independent.


Proportion of crim_cat across levels of chas
Proportion
0.75
0.50
0.25
0.00
Low
Medium
High
Crime Category (crim_cat)
chas
0
1

Figure 7.2

Solution Q7.1

Solution Q7.2

Solution Q7.3

Solution Q7.4

Question 8

1. Calculate the correlation between the proportion of Black residents by town (the variable black) and the status of the population (the variable lstat) using the R function cor.test().
2. Use the R package corrplot to produce the heatmap of correlations between variables shown in Figure 8.1. Note that the categorical variables are excluded from the data in this figure.

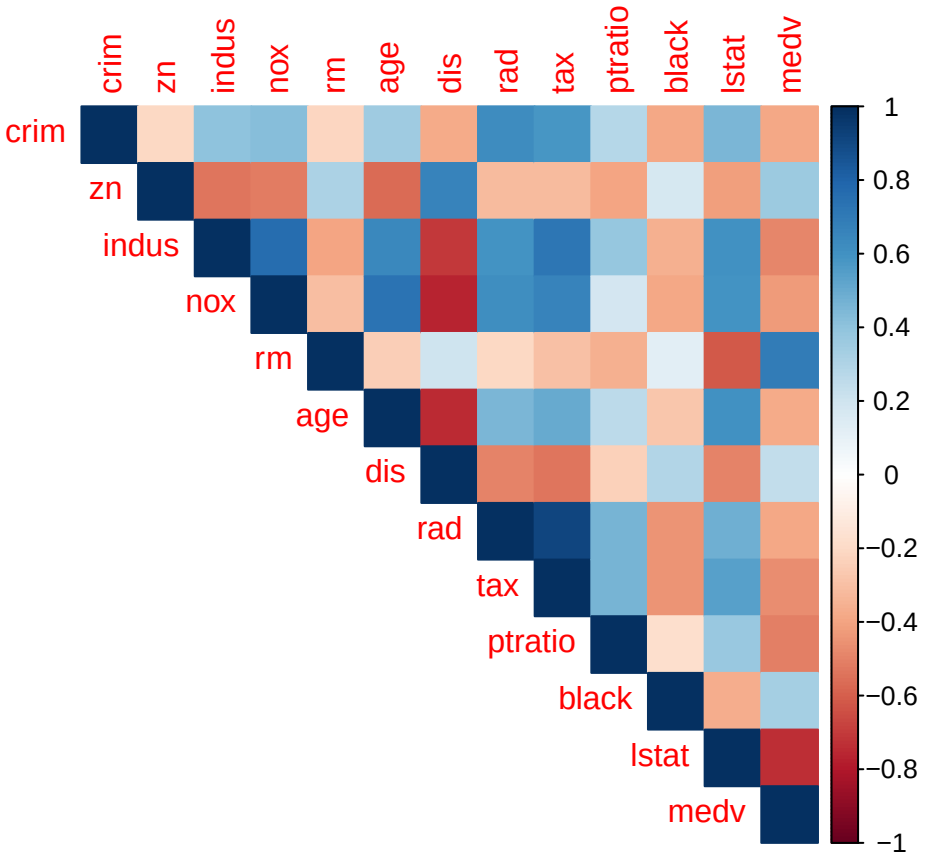


Figure 8.1

Solution Q8.1

Solution Q8.2

Question 9

1. Define a new categorical variable, medv_cat, in the following way: Re-code the variable medv into three categories:

medv $<$ 15: Low.
medv 15-25: Medium.
medv $>$ 25: High.

Include the new variable in the Boston dataset and produce the the table below.

```
## # A tibble: 3 x 4
##   medv_cat  mean    SD     N
##   <chr>    <dbl> <dbl> <int>
## 1 High      35.3  7.88   124
## 2 Low       11.6  2.64    94
## 3 Medium    20.6  2.69   288
```

2. Produce Figure 9.1.

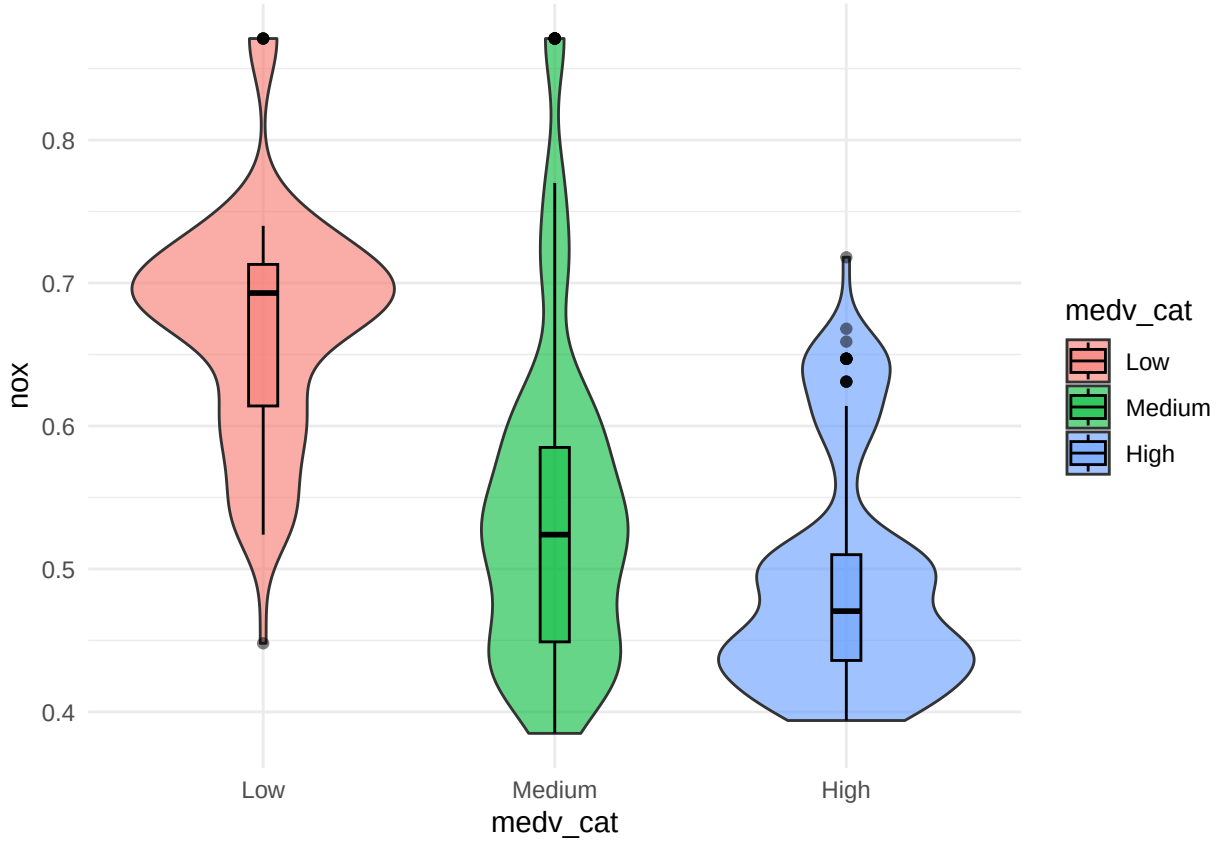


Figure 9.1

3. Test if the means of the variable nox across the three groups of the variable medv_cat are equal using the Kruskal-Wallis test. Print the output from the Kruskal-Wallis test.

4. Produce Figure 9.2. To make the plot, you can use the function ggline() of the package ggpubr or any other R package/function that you wish. Do not forget to add the error bars around the sample means to the figure.

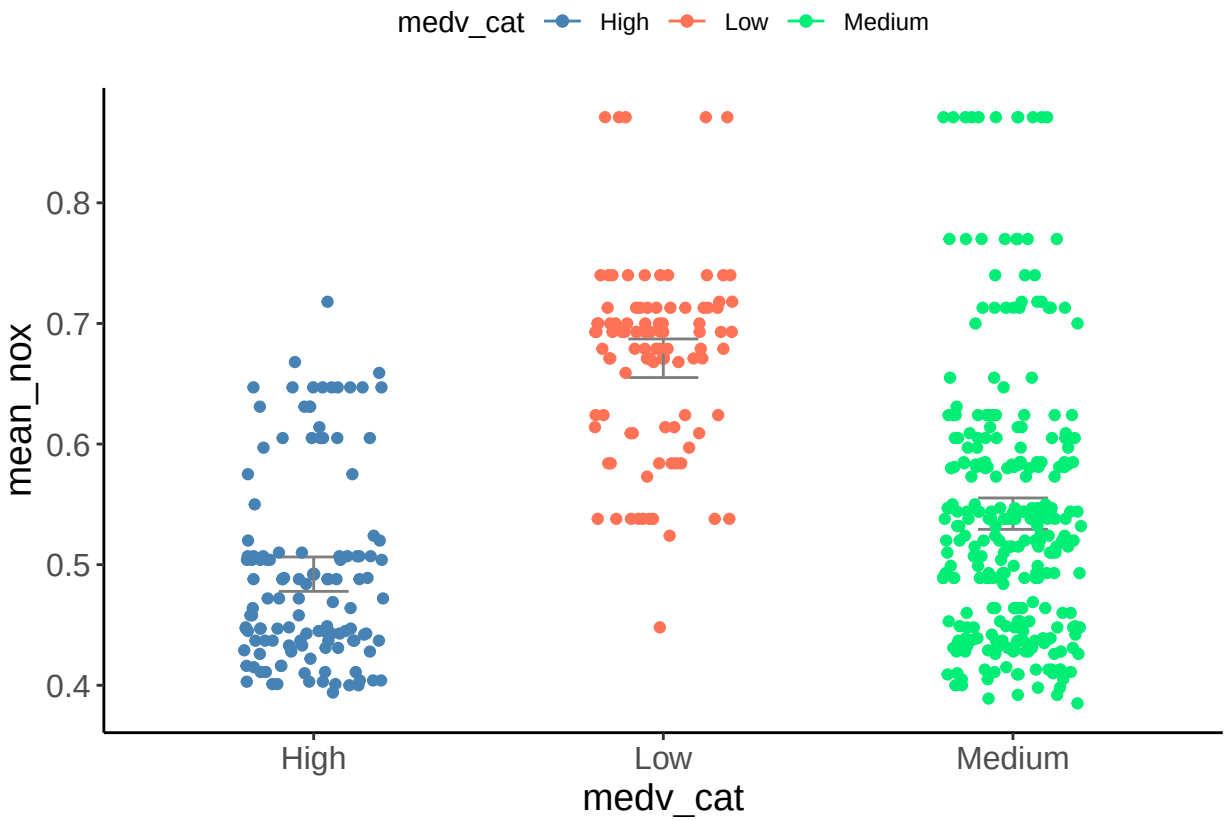

medv_cat
High
Low
Medium
mean_nox
0.8
0.7
0.6
0.5
0.4
High
Low
Medium
medv_cat

Figure 9.2

Solution Q9.1

Solution Q9.2

Solution Q9.3

Solution Q9.4

Question 10

For this question, use the version of the Boston dataset produced in Q9.1.

1. Categorize the crim and nox variables into three levels ("Low," "Medium," and "High") and two levels ("Low","High"), respectively, based on their quantiles as below:

crim_level: Represents the crime rate divided into three groups:

```
"Low": Bottom 33% of crime rate values.
"Medium": Middle 33% of crime rate values.
"High": Top 33% of crime rate values.
```

nox_level: Represents nitric oxide concentration divided into two groups:

```
"Low": Bottom 50% of nitric oxide concentration values.
"High": Top 50% of nitric oxide concentration values.
```

2. Produce the plot in Figure 10.1 which shows separate density plots of medv for each category of nox_level (Low,High), with each density plot color-coded by crim_level.

3. For subjects with Low level of crim, conduct a t-test to test the hypothesis that the mean medv is equal between the low and high nitrogen oxides concentration groups.

4. Create a R object that contain the 95% confidence interval for the mean difference. **DO NOT** use object=c(3.932519,8.891779). Print the object.

Solution Q10.1

Solution Q10.2

Solution Q10.3

Solution Q10.4

Part 4: the PimaIndiansDiabetes2 data

In this part of the exam, the questions are focused on the PimaIndiansDiabetes2 dataset which is a part of the mlbench R package. To access the data you need to install the package. More information about the dataset and variables can be found in https://search.r-project.org/CRAN/refmans/mlbench/html/PimaIndiansDiabetes.html. Use the code below to access the data.

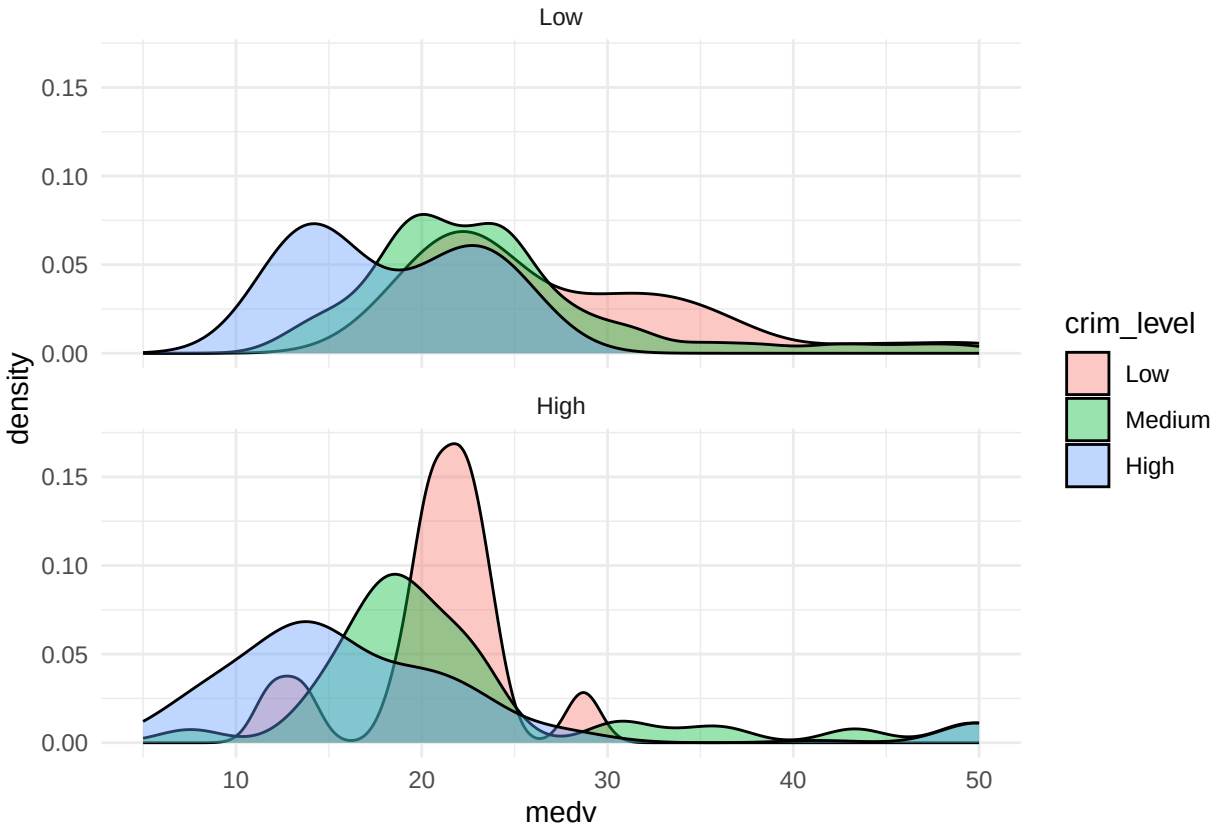

Low
High
density
0.15
0.10
0.05
0.00
10
20
30
40
50
medv
crim_level
Low
Medium
High

Figure 10.1

```
library(mlbench)
data(PimaIndiansDiabetes2)
names(PimaIndiansDiabetes2)
```

```
## [1] "pregnant" "glucose"  "pressure" "triceps"  "insulin"  "mass"     "pedigree"
## [8] "age"      "diabetes"
```

Question 11.

1. Filter out observations with missing values and define a new dataset: new_PimaIndiansDiabetes2. How many observations are included in the new dataset?
2. For each level of diabetes status (the variable diabetes), identify the top 5 patients with highest glucose and lowest mass values and produce the object bellow.

```
## # A tibble: 10 x 9
## # Groups:   diabetes [2]
##    pregnant glucose pressure triceps insulin  mass pedigree   age diabetes
##       <dbl>   <dbl>    <dbl>   <dbl>   <dbl> <dbl>    <dbl> <dbl> <fct>
##  1        4     197       70      39     744  36.7    2.33     31 neg
##  2        1     193       50      16     375  25.9    0.655    24 neg
##  3        3     191       68      15     130  30.9    0.299    34 neg
##  4        3     180       64      25      70  34      0.271    26 neg
##  5        0     173       78      32     265  46.5    1.16     58 neg
##  6        0     198       66      32     274  41.3    0.502    28 pos
##  7        2     197       70      45     543  30.5    0.158    53 pos
##  8        1     196       76      36     249  36.5    0.875    29 pos
##  9        8     196       76      29     280  37.5    0.605    57 pos
## 10        7     195       70      33     145  25.1    0.163    55 pos
```

Solution Q11.1

Solution Q11.2

Question 12.

In this question we use the PimaIndiansDiabetes2 dataset.

1. Remove missing values from the variables glucose and mass and create a new dataset new_PimaIndiansDiabetes3. How many observationas are included in the new dataset?
2. Create a new variable glucose_level which categorizes the variable glucose as “Low”, “Normal”, or “High” based on quantiles: bottom 25% as Low, 25%-75% as Normal, and top 25% as High.
3. Define a new R objectes, the mean (mean_mass) and standard deviation (sd_mass) of the variable mass within each glucose_level category and produce the folowing table:

```
## # A tibble: 3 x 3
##   glucose_level mean_mass sd_mass
##   <chr>             <dbl>   <dbl>
## 1 High               35       6.8
## 2 Low                30.4     6.6
## 3 Normal             32.2     6.8
```

Solution Q12.1

Solution Q12.2

Solution Q12.3

Question 13.

In this question we use the dataset new_PimaIndiansDiabetes3 that was created in Q12.1.

1. Figure 13.1 shows multi boxplot displaying the distribution of the variable mass by age groups (age_group), separated by diabetes status. Note that the data points in Figure 13.1 are colored according to the glucose_level. Produce Figure 13.1.

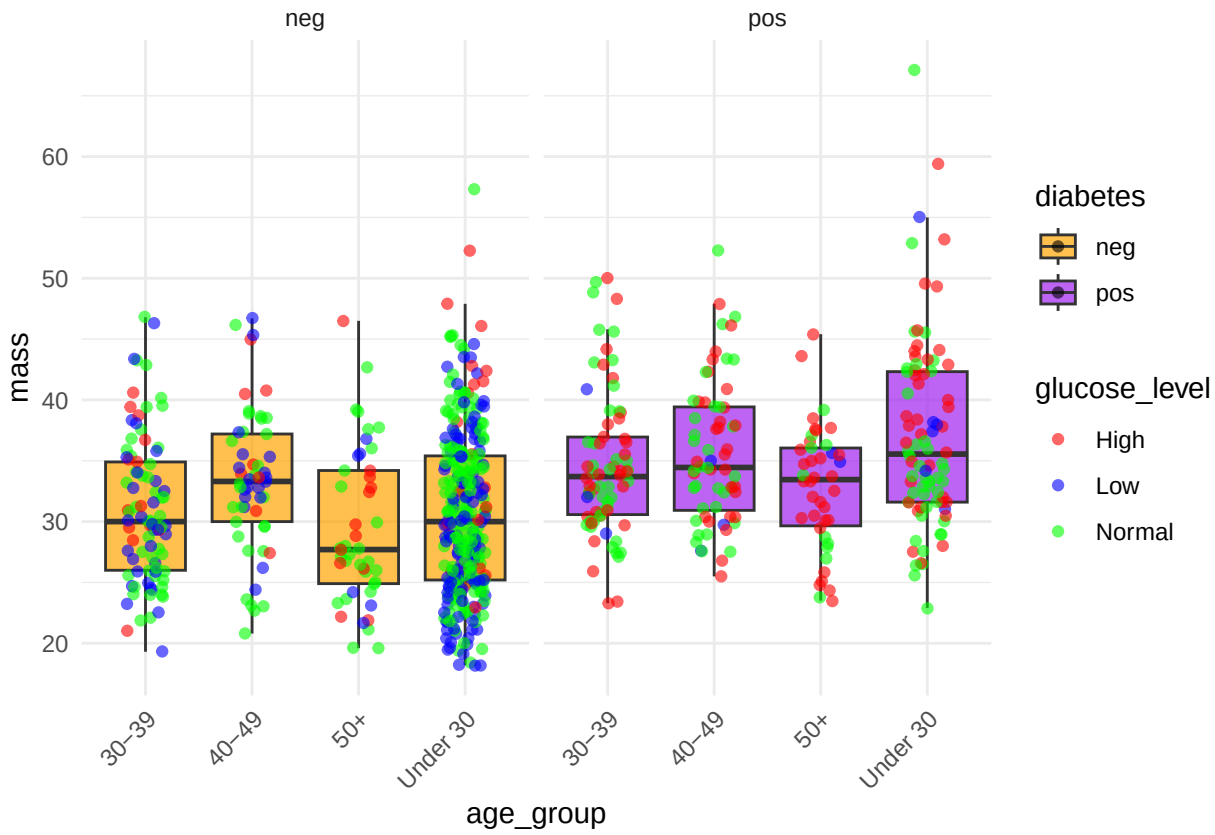


Figure 13.1

Solution Q13.1

Question 14.

In this question we use dataset new_PimaIndiansDiabetes3 created in Question Q12.1.

1. Create a new variable, age_adjusted_risk, based the equation below.

$$\text{age\_adjusted\_risk} = \sqrt{\frac{0.5 \times \text{glucose} + 0.3 \times \text{mass} + 0.2 \times \text{pressure}}{\text{age}}}.$$

2. Produce the Figure 14.1. Note that the information that is provided in the title is an output from a two sample t-test of the adjusted risk (defined in Q14.1) across the diabetes groups (the variable diabetes).

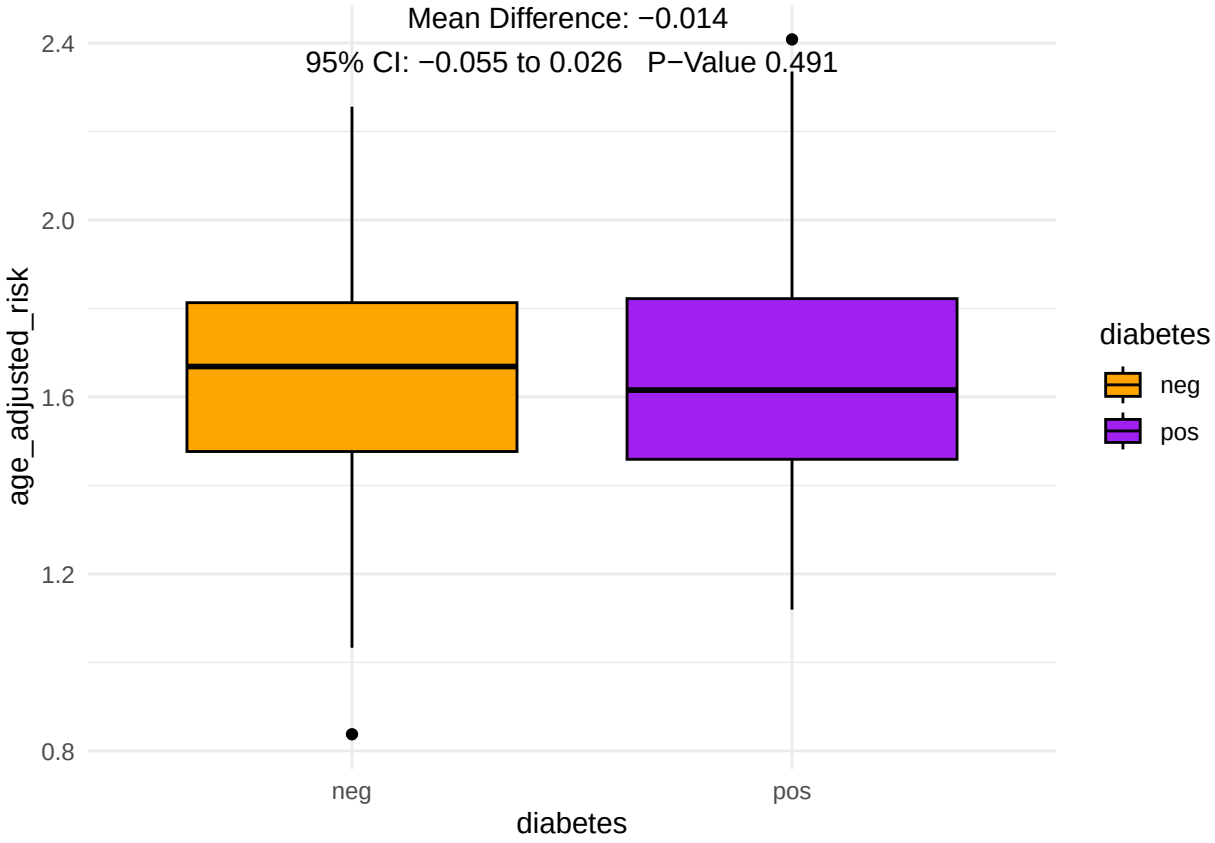


Figure 14.1

Solution Q14.1

Solution Q14.2

Question 15.

1. Create a function that receives as an input: (1) a data frame (data), (2) a names of a categorical variable (group_col) and (3) all numeric variables (numeric_cols). The function output should be a table with the mean, median, standard deviation, and IQR for all the numeric variables in the data frame across the level of the categorical variable.

2. Apply the function that you wrote in Q15.1 to the data frame (PimaIndiansDiabetes2), use the variable (diabetes) as the categorical variable and all the numerical variables in the data set. Print the mean of two numeric variables only (glucose, insulin).

Solution Q15.1

Solution Q15.2

Question 16.

1. Create a new dataset, PimaIndiansDiabetes5, from the original dataset PimaIndiansDiabetes2 and remove missing values.
2. Using the new dataset created in Question 16.1, create another dataset , median_insulin_data, by including only patients over 40 years old (the variable age) with blood pressure above 70 (the variable pressure), and select only patients whose insulin levels (the variable insulin) is above the median insulin level. How many observations are included in the new dataset?
3. Using the dataset created in Question Q16.2, create a new dataset selected_median_insulin which includes the variables age, pressure, insulin, pedigree, pregnant, diabetes variables. Display the last six observations of the dataset.

Solution Q16.1

Solution Q16.2

Solution Q16.3

Question 17.

In this question we use the dataset created in Q16.1 (PimaIndiansDiabetes5).

1. Create a new dataset filtered_primadia5 by including patients with the variable mass above the median and the variable pedigree above the mean. How many observations are included in the new dataset?
2. Using the new dataset, produce an animated dynamic plot, shown in Figure 17.1 (Note the plot is colored by the variable age). You need to look at this plot in the HTML file of the exam. Produce an identical plot. Note that it should be produced in the HTML version of your solution, on the PDF file of the solution it will be a static file.

Figure 17.1

3. Export Figure 17.1 that was produced in Q17.2 as an external file. Name the file 3D_Bubble_Plot.html and make sure to add this exported file to the solutions folder. Note that in the PDF solution of the exam, Figure 17.1 will not be dynamic.

Solution Q17.1

Solution Q17.2

Solution Q17.3

Question 18.

Create a new dataset, prima_data, without missing data data using PimaIndiansDiabetes2.

1. How many observations are included in the new dataset?

2. Recode the variable pregnant in the following way: 0 = "0", 1 = "1", 2 = "2" and >= 3 = "3+". Name the new variable pregnant_grouped. Add the new variable to the dataset prima_data. Print the first 5 observations for whom the number of pregnancies is equal to 2.

3. Use the prima_data created in Q18.1, create a new dataset mean_mass_data containing the mean of the variable (mass) for each combination of grouped pregnancy levels (pregnancy_grouped) and diabetes status (the variable diabetes). What is the dimension of the new dataset? Print the new dataset.

4. Produce Figure 18.1 (blood pressure VS. pregnancies).

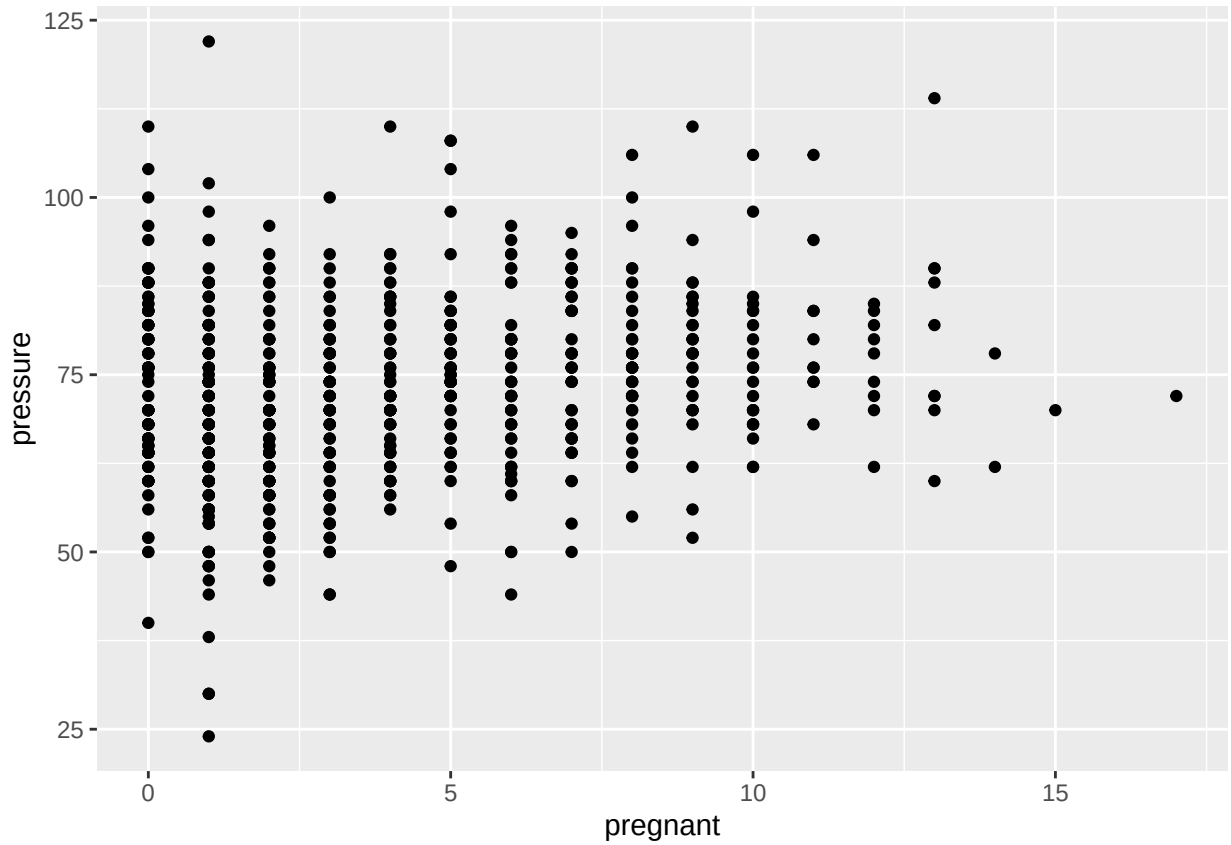


Figure 18.1

5. Produce the table below that shows the mean and SD of age by number of pregnancies.

```
## # A tibble: 17 x 3
##    pregnant mean_age sd_age
##       <dbl>    <dbl>  <dbl>
##  1        0     27.6   9.69
##  2        1     27.4   8.11
##  3        2     27.2   9.55
##  4        3     29.0   8.10
##  5        4     32.8  11.0
##  6        5     39.0  12.5
```

```
##  7        6     39.3  12.0
##  8        7     41.1   7.93
##  9        8     45.4  10.7
## 10        9     44.2  10.4
## 11       10     42.7   9.37
## 12       11     44.5   6.19
## 13       12     47.4   7.76
## 14       13     44.5   5.84
## 15       14     42     5.66
## 16       15     43    NA
## 17       17     47    NA
```

6. Produce Figure 18.2. Note that the black dots are the mean age at each pregnancy group.

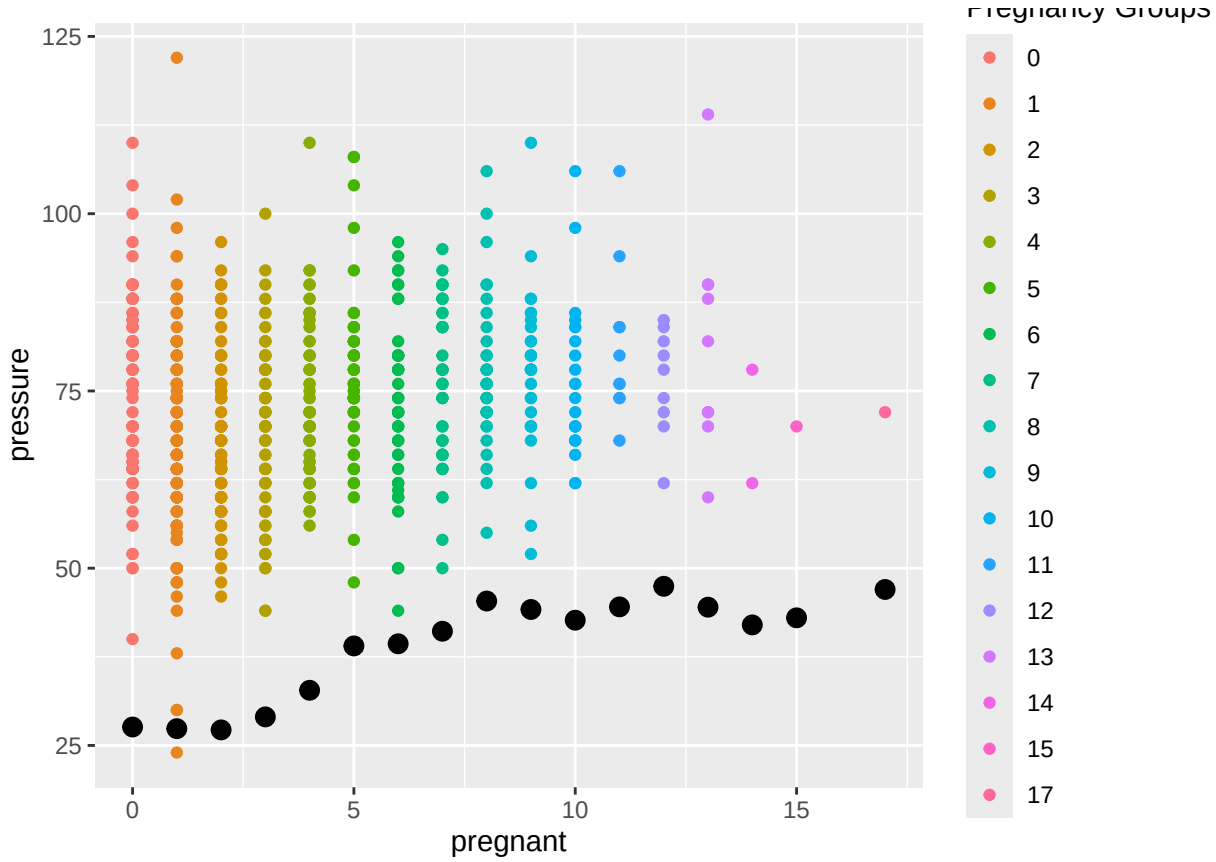


Figure 18.2

7. Produce the interactive plot which is shown in Figure 18.3. You need to look at the HTML version of the exam to see Figure 18.3 in an interactive form.

Figure 18.3

Solution Q18.1

Solution Q18.2

Solution Q18.3

Solution Q18.4

Solution Q18.5

Solution Q18.6

Solution Q18.7

Question 19.

In this question we use the dataset PimaIndiansDiabetes5 that was created in Q16.1.

1. Use a **for loop** in which you calculate, at each step of the loop, the correlation between glucose level (the variable glucose) and blood pressure (the variable pressure) for each level of number of pregnancies (the variable pregnant). This mean that in step 1 you calculate the correlation between glucose and pressure for observation with pregnant=0 etc. Produce the data frame bellow.

```
##    pregnancies correlation
## 1            0  0.04483389
## 2            1  0.18459352
## 3            2  0.21800751
## 4            3  0.21480046
## 5            4 -0.03648928
## 6            5  0.19994282
## 7            6  0.44004193
## 8            7 -0.19804490
## 9            8  0.25516453
## 10           9  0.70865117
## 11          10 -0.74739750
## 12          11 -0.02292352
## 13          12  0.17512432
## 14          13  0.93676591
## 15          14          NA
## 16          15          NA
## 17          17          NA
```

2. In the above panel, correlations between the variables for observations with number of pregnancies higher than 13 is NA. Change the for loop on Q19.1 so the NA values will not be included in the panel.

Solution Q19.1

Solution Q19.2

Question 20.

In this question we use the dataset prima_data created in Q18.1.

1. Produce the interactive plot which is shown Figure 20.1. You need to look at the HTML version of the exam to see the figure in an interactive format.

Figure 20.1

Solution Q20.1