Matriculation number:						

Examination Assignment

Module: Data Analysis and Statistics

Exam part: Data Analysis and Statistics

Examiner: Prof. Dr. Schwind, Dipl.-Biol. Ralf Darius

Deadline for the submission: 31.08.2019, 11:59 pm

Study program	Begin of studies	Last name, First name
Information Engineering and Computer Science (M.Sc.)		

Assessment criteria and number of points that can be achieved:

Maximum number of points	Skills and Expertise	Systematic and scientific Quality	Quality of the results	Presentation of the results
100	45	15	30	10

Result:

Points	Mark	Skills and Expertise	Systematic and scientific Quality	Quality of the results	Presentation of the results

The assessment is only open to students who are enrolled in the study course "Information Engineering and Computer Science (M.Sc.)" and have successfully registered for the exam in Data Analysis and Statistics at the end of the summer semester 2019.

The assessment consists of the assignment that is given in this document. It involves a practical task and the subsequent preparation of a scientific report. Grading will be based on both parts. Carefully read the whole document before you start working on the assignment!

1. Introduction

In this assignment, you will examine a number of factors that relate to wages for a group of males from the Atlantic region of the United States. In particular, you will try to understand the association between an employee's age, an employee's educational attainment, as well as the calendar year and an employee's wage. (G. JAMES ET AL., 2013)

2. The Data

The data that you will analyze are income survey data for males in the central Atlantic region of the USA. The original data were manually assembled by Steve Miller, of Open BI, from the March 2011 Supplement to Current Population Survey data. Check e.g. the following websites for more information:

http://thedataweb.rm.census.gov/TheDataWeb

https://www.census.gov/programs-surveys/cps/technical-documentation/methodology.html

The data that you will work on will be provided as a sample (sample size n=3000) of the original data mentioned above. More information on how the sample is provided will be given in section 4 (Practical Tasks).

3. References

G. James et al., An Introduction to Statistical Learning: with Applications in R,Springer Texts in Statistics, DOI 10.1007/978-1-4614-7138-7 1, © Springer Science+Business Media New York 2013.

4. Practical Tasks

You will investigate how the calendar year as well as the age and educational attainment of males in the central Atlantic Region affect the wages of those males. The investigation will be done based on survey data for males in the central Atlantic region by means of R and RStudio.

A) Read carefully the material in sections 1- 3 above.

B) Read through the tasks in this section.

C) For your practical task you will have to prepare an R script. The final version of the script will have to be part of your assessment submission. Create a new R script in RStudio an save it. Use the following pattern to name the file:

Your Matr.-Number_ **DataAnalysis_Assessment_2019.R**

D) In order to get the data that you will have to work on in this assignment, go to the moodle course "*M-IE_1.02 Data Analysis / Statistics, SS 2019*" and download the R script "*Assessment Data – R Script* " from section "*Assessment*".

Open the script in RStudio. Copy the content and paste it into the script "*Your Matr.-Number*_ DataAnalysis_Assessment_2019.R".

Run the copied code. This code will provide you with your own sample of the original data. Do not share this sample or the results of the analysis of these sample data with any other participant of the exam.

Running the copied code will create the data.frame "assessment_dataframe" in your R environment. assessment_dataframe holds the data you have to analyse.

E) **The data.frame "assessment_dataframe" provides values for 12 different variables. The columns year, age, education and wage contain the records for the variables that represent calendar year, age of a male employee, educational attainment of a male employee and wage of a male employee. Extract the data of the columns year, age, education and wage and work on those only.**

F) **Describe the provided data appropriately. Visualize the data and obtain suitable statistics by means of R in order to describe and analyze it in a meaningful way.**

G) **Read through the assessment components in the next section before you carry out any tasks.**

5. Assessment

The results of the tasks given in section 4 – Practical Tasks - have to be compiled into a scientific report. The report and the R script (see section 4 C) that contains the R code of your work will be due on 31.08.2019, 11:59 pm!

Your report should comprise the following elements:

- **Cover page**: **The first page of this assignment paper has to be used!!**
- A signed statement of authorship. You can copy the following text into your report and sign it:

  This report is the result of my own work. Material from the published or unpublished work of others, which is referred to in the report, is credited to the author in the text.
- Table of contents.
- Introduction to the overall subject of the report and the particular tasks covered in it (research question/s, motivation, goals, context, approach in brief).
- A description of the approach of the work and the rationale behind it. Describe the tools and methods you use and in which way you use them in order to solve the given tasks. Your description should enable the reader not only to fully understand but also reproduce what you do.

- A detailed description of the results.
- An interpretation and discussion of your results **and the methods** used. You might consider the following questions when approaching the analysis and interpreting and discussing your results:
    - Is there a relationship between age/education/calendar year and wage? Our first goal should be to determine whether the data provide evidence of an association between age/education/calendar year and wage.
    - How strong are the relationships? Assuming that there are relationships, we would like to know the strength of those.
    - Given a certain age/education/calendar year, can we predict wage with a high level of accuracy? This would be a strong relationship. Or is a prediction of wage based on age/education/calendar year only slightly better than a random guess? This would be a weak relationship
    - Which factors contribute to wage? Do all three factors — age, education and calendar year — contribute to wage, or do just one or two of the factors contribute?
    - Is the relationship linear? If there is approximately a straight-line relationship between wage and age/education/calendar year, then linear regression is an appropriate tool. How well does the linear model fit the data? If the relationship is not linear, what could be considered?
    - Are there interaction effects?
- A list of the references used in your report.
- Fully commented RStudio code

The report and your R script (see section 4 C) at the latest have to be turned in on Aug 31, 2019, 11:59 pm. A report that will not have turned in by then will automatically be graded as failed! **The date of receipt applies!**

Your R script has to be submitted via Email to ralf.darius@hochschule-rhein-waal.de **!**

There are different options for the delivery of your report:

(1) Hand out a printed copy personally to Mr Ralf Darius (room 02 00 405).

(2) Post a printed copy to:

Ralf Darius
Hochschule Rhein-Waal
Friedrich-Heinrich-Allee 25
D-47475 Kamp-Lintfort

Alternatively you can simply drop your report in the post office box on the left side of the entrance hall of building 02 when entering through the main entrance. Use the POB labelled “DARIUS”.

(3) Send a digital copy via Email to ralf.darius@hochschule-rhein-waal.de**.**

Grading of the module Data Analysis / Statistics is done based on the report and the R script. It will be based on a point scale with a maximum of 100 points and factor in 4 different aspects of the report and R script. These aspects will be considered and graded separately. An overview of the grading scheme is given in table 1.

Maximum number of points	Skills and Expertise	Systematic and scientific Quality	Quality of the results	Presentation of the results
100	45	15	30	10