Bachelor Informatik:

Mock Up Exam – Descriptive Statistics

WS 2020/21

Prüfer: Falkenberg, Schrader

Name:

Matriculation Number:

|  |  |
| --- | --- |
| I hereby confirm that I am physically fit to take the examination and that I have revised the exam alone without the help of third parties. |  |

Some hints:

* Please mention, not only the solution but the derivation of the solution has to be given.
* Insert your solutions to the individual questions at the appropriate places in the document and upload the document to Moodle at the end of the exam.
* If you want to add a hand written solution, please copy the sheet with hand written solution by your smartphone and copy the corresponding file in the document.
* To copy a by R generated diagram click on the plots window “Export” and then click “Copy to Clipboard”. Copy the diagram by pressing the “Copy Plot” button in the “Copy Plot to Clipboard” windows. With ctrl-v you can paste the diagram in the document.

# Descriptive Statistics

# Task I

Dataset from “Freie Universität Berlin” with the variables

* Identification information: ID
* Personal details: sex, year of birth, height, weight
* Information on education: school leaving certificate, highest vocational qualification
* Information on the education of the parents: school leaving certificates of the father and
* mother and highest vocational qualifications of the father and mother.
* Information on employment: type of occupation, hours worked, income
* Information on health: visits to the doctor and whether one is a smoker
* Indication of life satisfaction

The csv file Umfragedaten\_v1\_an.csv contains the dataset.

1. Import the csv file into a tibble survey. Use the R function read\_cvs() to do this.

Write your code here!

1. Specify the type and scale for all variables in the dataset.

Please write your answer here!

1. Change the values of the variables GESCHL to male and female and change the values of the variable RAUCH to yes and no and add the new variables income and work.time to the tibble survey. The variables are defined as follows:

* depending on the values of NETTO the values of income are: low <= 1200, medium (1200, 2500], high > 2500
* depending on the values of BERUFSTAETIG the values of work.time are: "NICHT ERWERBSTAETIG" -> "none", "HAUPTBERUFL.HALBTAGS" -> "half day", "HAUPTBERUFL.GANZTAGS" -> "full time", "NEBENHER BERUFSTAE." -> "part time", empty-> "NA"

Write your code here!

Display the first 10 lines of the modified dataset here!

1. Calculate the no of observations, quartiles, mean, standard deviation, min, max for the variable NETTO grouped by BERUSTAETIG.

Write your R-Code here!

Copy your R-ouput here!

1. Create a side by side boxplot of the variable height with respect to the variable gender.

Write your R-Code here!

Copy your diagram here!

1. How do the values of the variables NETTO differ with respect to GESCHL in terms of location and variability? Use the side by side boxplot to answer the question!

Write your answer here!

1. Create a histogram of the variable NETTO where the classes given by the categories of income.

Write your R-Code here!

Copy your diagram here!

1. Plot a scatterplot of all pairs (ARBEITSSTD,NETTO).

Write your R-Code here!

Copy your diagram here!

1. Which type of association could exist between the variables ARBEITSTD and NETTO?

Write your answer here!

1. Calculate the covariance the coefficient of correlation to get further information about the relationship.

Covariance: Klicken oder tippen Sie hier, um Text einzugeben.

Coefficient of Correlation: Klicken oder tippen Sie hier, um Text einzugeben.

1. Determine the parameters of the linear regression line NETTO = a + b\*ARBEITSSTD.

Write your R-Code here!

Coefficients:

1. Interpret the parameters a and b of the regression line.

Write your answer here!

1. Calculate the coefficient of determination and interpret the value of it.

Write your answer here!

1. Add the regression line NETTO = a + b\*ARBEITSSTD to the scatterplot of the pairs (ARBEITSSTD, NETTO).

Write your R-Code here!

Copy your diagram here!

# Task II

The csv file corona\_march\_2020 contains the number of new Corona infections and Corona deaths in Germany and Spain on March 2020.

1. Import the csv-file corona\_march\_2020.csv into a tibble called messy.data.

Write your R-Code here!

1. Why ist he dataset messy.data not tidy?

Write your R-Code here!

1. Transform the tibble messy.data into a tidy dataset.

Write your R-Code here!

Display the tidy dataset here!