Examination Statistics
Prof. Dr. Falkenberg
Course of Study: Computer Sciences
22.6.2021
Part: Descriptive Statistics
Editing Time: 30 Minutes

Problems	1	2	3	4	5	6	7	Sum
Max. scores	2	12	4	7	7	9	9	50

Further instructions:

1. Submit all you want to be assessed (derivations, answers, interpretations, commands, diagrams, etc.).
2. You are allowed to submit totally ONE (1) computer file in every part of the exam. The file with the last time stamp will be corrected, other files NOT!!!
3. The computer file should be a .pdf-document.
4. Please notice, not only the solution but the derivation of the solution has to be given.

Good Luck!
Dr. Falkenberg

Consider the data file testing_covid.csvfrom the European Centre for Disease Prevention and Control (ECDC). It contains information about testing volume for COVID-19 by week and country and subnational region (where available).

Variable	Defintion
country	
country_code	2-letter ISO country code
year_week	
level	Whether national or subnational (regional) level data
region	2-letter ISO country code where level is national.
region_name	Country name where level is national or name of region where level is subnational
new_cases	Number of new confirmed cases
tests_done	Number of tests done
population	
testing_rate	Testing rate per 100 000 population
positivity_rate	Weekly test positivity
testing_data_source	Country API, Country GitHub, Country website, Manual webscraping, Other, Survey, TESSy:data provided directly by Member States to ECDC via TESSy

1. Import the file testing_covid.csv into a tibble called raw_data.

2. Determine the scales (nominal, ordinal, interval, ratio or absolute) of all variables

3. Determine the number of tests in Germany per week and visualize it by a lineplot.

4. Determine the sum of tests and new cases in Germany, Austria, France and Italy in december 2020. Visualize the ratio of positive tests by a pieplot.
   Note that in some countries the values are given countrywide and regionally.

5. Determine minimum, maximum and the three quartiles of the variable testing_rate for the countries Germany and France. Visualize it by a side-by-side boxplot.

6. Perform a linear regression of tests_done over the time in Germany.

   - Determine the regression coefficients tests_done = a + b $*$ time.
   - Make a prediction of tests_done in Germany in week 20 in 2021.
   - What can you say about the goodness of the linear regression?

   Hint: Sort the data tests_done in Germany by year and week and use row number as time variable. Apply row_number() to get the row number a data frame.

7. Tidy and messy data

   (a) What is a tidy data set? Is the imported dataset raw_data tidy?

   (b) Consider only the national values of tests_done per year_week, apply the spread command to generate a table containing the values of tests_done for a country and the consecutive weeks in every row.

   (c) Is the generated dataset still tidy?