This box is for the examiner only:

Question:	1	2	3	4	5	6	7	Total
Points:	10	20	16	40	30	8	39	163
Score:								

Contents

1. (10 points) **Excel and low-code applications**

   Discuss why Microsoft Excel is not the first choice of data scientists.

2. (20 points) **Emerging Trends in a Data-Driven Business Environment**

   Identify and discuss briefly two emerging trends in a data-driven business environment. Explain how these trends have impacted the use of data science in business.

3. (16 points) **Types of Data Science Roles**

   Describe two distinct roles within data science teams. Discuss the key responsibilities and skills required for each role.

4. **Data Literacy and Pitfalls**

   Data literacy is the ability to (a) read, (b) understand, (c) create, and (d) communicate data as information.

   (a) (10 points) Explain one pitfall as described in the book of Jones (2020) that relates to the verb "read". Explain the linkage and provide an example of the pitfall.

   (b) (10 points) Explain one pitfall as described in the book of Jones (2020) that relates to the verb "understand". Explain the linkage and provide an example of the pitfall.

   (c) (10 points) Explain one pitfall as described in the book of Jones (2020) that relates to the verb "create". Explain the linkage and provide an example of the pitfall.

   (d) (10 points) Explain one pitfall as described in the book of Jones (2020) that relates to the verb "communicate". Explain the linkage and provide an example of the pitfall.

5. **Over-Interpreting Empirical Results**

   Consider a research paper that over-interprets its empirical results by cherry-picking a few cases and extrapolating them to explain the broader picture.

   (a) (10 points) Describe the potential pitfalls of over-reliance on selected cases in empirical research. How does this practice undermine the validity and generalizability of findings and how might biases in case selection influence the interpretation of results?

(b) (10 points) Discuss the implications of extrapolating[1] findings from limited empirical evidence to broader contexts. What are the risks and limitations associated with this approach?

(c) (10 points) Relate your critique to specific principles of sound data analysis discussed in our course and in Jones (2020). How can researchers mitigate these pitfalls to ensure more reliable and comprehensive conclusions?

6. **R: As a calculator**

(a) (4 points) One of the most basic ways to use R is as a calculator. Suppose you want to calculate

$$3 + \sqrt{9} - 2^2 \cdot \frac{4}{3}$$

What code would you need to type in the R-console to ***calculate and store*** the result in an object?

(b) (4 points) What code would you need to type in the R-console to **calculate and store** the result of the following calculation:

$$\left(\sqrt{8} \cdot \frac{3}{4} + 7^{-2}\right) \cdot 9$$

7. **The programming language R**

(a) (1 point) Write down the code to get the path of your current working directory in R.

(b) (1 point) Write down the code to set the path of your current working directory in R to `C:/Users/me/workwithr`

(c) (1 point) The `lm()` function is used to fit linear models in R. Write down the code to get the help documentation of R through the console.

(d) (1 point) Write down the code you would need to install the `knitr` package.

(e) (4 points) Explain briefly the usage and advantages of the assignment operator "<−".

(f) (2 points) The following table shows COVID deaths for three states in Germany as stored in the dataframe `df_deats`.

state	deaths
BY	18,565
BW	12,264
NRW	19,489

Write down the code you would need to put into the R-console to show the following descriptive statistics: Minimum, 1st Quantile, Median, Mean, 3rd Quantile, Maximum.

[1] Extrapolation refers to the process of estimating, predicting, or projecting beyond known or observed values. It involves extending existing data or trends to make predictions about future values or outcomes.

(g) (4 points) The following table shows COVID cases for three states in Germany, rounded in thousands.

state	cases
NRW	628
BY	511
BW	374

Write down the code you would need to put into the R-console

- to store the variable *state* in a vector,
- to store the variable *cases* in a vector, and
- to store both vectors in a data frame with the name `df_covid`.

(h) (2 points) Show the following descriptive statistics of the COVID dataset `df_covid` described above: Minimum, 1st Quantile, Median, Mean, 3rd Quantile, Maximum.

(i) (4 points) Using the `dplyr` package, write the code to filter the dataset to include only the rows where the number of cases is greater than 500. Save the filtered data in the dataframe `df_filtered`.

(j) (4 points) Write the code to sort the dataset in ascending order based on the number of cases. Save the sorted data in the dataframe `df_sorted`.

(k) (4 points) Calculate the total number of cases across all states.

(l) (4 points) Calculate the percentage share of each state's cases out of the total cases. Add this as a new column `cases_percentage` to the dataframe.

(m) (4 points) Make a table that shows the mean and the standard deviation of of cases. Use the `dplyr` package and the function `summarise()`.

(n) (3 points) Suppose your friend shows you the following excerpt from his R console:

```
rm(list = ls())
library(tidyverse)
library(dplyr)
data("mtcars")
```

You see that your friend works with the mtcars dataset. As your friend is interested in energy efficient cars only, he seeks to save cars that can drive more than 20 miles per gallon `mpg` in a new dataframe. Write down the code your friend needs to put into his R-console next. Use therefore the *filter()* function.

References

Jones, B. (2020). *Avoiding data pitfalls: How to steer clear of common blunders when working with data and presenting analysis and visualizations*. John Wiley & Sons.