This box is for the examiner only:

Question:	1	2	3	4	5	6	7	Total
Points:	10	20	16	32	30	8	37	153
Score:								

Contents

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

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

   Solution:

   Here is an incomplete list of possible answers:

   - **Scalability Issues:** Excel struggles with handling large datasets, leading to performance problems and potential data loss.
   - **Limited Advanced Capabilities:** Lacks advanced statistical, machine learning, and data manipulation features available in R, Python, and specialized data science tools.
   - **Reproducibility and Automation:** Manual operations in Excel are prone to human error, whereas R and Python enable reproducible, automated workflows.
   - **Visualization Limitations:** Excel's basic charts do not match the advanced, customizable, and interactive visualizations created with tools like Matplotlib, Seaborn, or D3.js.
   - **Integration Challenges:** Excel is not designed for seamless integration with databases, APIs, and complex data pipelines, unlike programming languages that support robust data integration.
   - **Version Control Challenges:** Excel lacks robust version control mechanisms, making it difficult to track changes and collaborate effectively on data projects compared to version control systems used in programming environments.
   - **Complexity Handling:** Managing complex data transformations and workflows can be cumbersome in Excel, whereas no-code platforms provide intuitive visual interfaces for complex data operations.
   - **Customization and Extensibility:** No-code platforms often offer more customization options and integration capabilities with other tools and services, allowing for more tailored and extensible solutions compared to Excel's limitations in customization.

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.

Solution:

1. **Artificial intelligence and machine learning technologies** enable businesses to analyze vast amounts of data, identify patterns, and make informed decisions with minimal human intervention. Impact on Data Science in Business:

- Enhanced Predictive Analytics: AI and ML algorithms can predict future trends and outcomes by analyzing historical data. This is particularly beneficial in areas such as customer behavior forecasting, supply chain optimization, and financial risk assessment.
- Automated Processes: These technologies automate routine tasks, such as data cleaning, data integration, and anomaly detection, freeing up data scientists to focus on more strategic activities.
- Personalization: AI-driven data science allows for highly personalized customer experiences through recommendation systems and targeted marketing, increasing customer satisfaction and loyalty.

2. **Big data and real-time analytics** are transforming how businesses collect, store, and analyze data. With the advent of the Internet of Things (IoT), social media, and other data-generating technologies, the volume of data available to businesses has increased exponentially. Impact on Data Science in Business:

- Improved Decision-Making: Real-time analytics enable businesses to make timely decisions based on the most current data available. This is crucial in industries like finance, healthcare, and retail, where conditions change rapidly.
- Scalable Data Management: The ability to handle Big Data allows companies to draw insights from larger datasets, offering a more comprehensive understanding of market trends, consumer behavior, and operational efficiency.
- Enhanced Operational Efficiency: Real-time data analytics can optimize business operations by providing immediate insights into system performance, customer interactions, and market conditions, allowing for swift adjustments and improvements.

3. The **evolution of computers** from mainframes to personal computers, and now to powerful cloud-based computing, has revolutionized data processing capabilities. Impact on Data Science in Business:

- Increased Processing Power: The advent of advanced computing technologies allows businesses to process and analyze large datasets much more quickly and efficiently.
- Enhanced Data Storage: Innovations in data storage, including cloud solutions, enable companies to store vast amounts of data securely and access it from anywhere.
- Complex Analytics: Improved computing power facilitates the use of complex algorithms and models, resulting in more accurate predictions and insights.

4. **Business Intelligence (BI) tools and technologies** help organizations make data-driven decisions by providing insights through data visualization and reporting. Impact on Data Science in Business:

- Improved Decision-Making: BI tools enable quick access to actionable insights and real-time data, leading to better strategic and operational decisions.
- Data Democratization: BI platforms often provide user-friendly interfaces, making it easier for non-technical stakeholders to understand data trends and insights.
- Enhanced Efficiency: Automated reporting and dashboards reduce the time spent on data analysis, allowing businesses to act on insights faster.

5. **Big Data** refers to the massive volume of structured and unstructured data generated by businesses, social media, IoT devices, and more. Impact on Data Science in Business:

- Richer Insights: Analyzing Big Data helps businesses uncover hidden patterns, correlations, and market trends that inform strategic decisions.
- Customer Understanding: Big Data analytics allow companies to gain a deeper understanding of customer behaviors, preferences, and needs, leading to better-targeted products and services.
- Operational Efficiency: Through Big Data analysis, businesses can optimize various aspects of their operations, from supply chain management to fraud detection.

6. **Internet of Things (IoT), Cloud Computing, and Blockchain** are reshaping how data is collected, stored, and secured. Impact on Data Science in Business:

- IoT: IoT devices generate real-time data from various sources like industrial equipment, consumer products, and healthcare devices, providing valuable insights for predictive maintenance, customer behavior, and health monitoring.
- Cloud Computing: Cloud platforms offer scalable and cost-effective options for data storage and processing, enabling businesses to handle large datasets and complex computations without massive upfront investments.
- Blockchain: Blockchain technology provides a decentralized and secure way to record transactions and manage data, enhancing the trustworthiness and integrity of data across various business processes.

7. **Industry 4.0** represents the fourth industrial revolution, characterized by the integration of digital transformations, automation, and smart technologies into manufacturing and other industries. By leveraging these technologies, businesses can achieve significant operational improvements, cost savings, and increased competitiveness in the market. Impact on Data Science in Business:

- Predictive Maintenance: Using data analytics, smart factories can predict equipment failures before they occur, reducing downtime and maintenance costs.
- Quality Control: Advanced sensors and analytics ensure high-quality production by constantly monitoring and adjusting the manufacturing process to meet standards.
- Efficient Production: Automation and real-time data analytics streamline production processes, reduce waste, and optimize resource use, thereby enhancing overall efficiency and productivity.

8. The rise of **remote working** has drastically changed the business environment. Enabled by advancements in software applications and increased internet speeds, remote working has become a viable and often preferred option for many organizations. Impact on Data Science in Business:

- Real-Time Collaboration: Data science tools enable remote teams to share data and insights in real-time, fostering better collaboration and quick decision-making.
- Improved Productivity: With access to collaborative platforms and analytical tools, remote teams can maintain or even improve productivity by working seamlessly regardless of their physical location.
- Data Sharing: Secure cloud-based solutions and data visualization tools allow for efficient and secure data sharing, ensuring that all team members have access to the latest information.

These advancements make it possible for businesses to remain agile and responsive, even when their teams are dispersed across various locations, thus enhancing overall productivity and innovation.

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.

Solution:

Types of Data Science Roles

1. Data Engineer

Key Responsibilities:

- Design, construct, and maintain large-scale data processing systems and infrastructures.
- Develop and optimize algorithms for data collection, storage, and retrieval.
- Ensure data pipeline integrity and manage ETL (Extract, Transform, Load) processes.

Key Skills:

- Proficiency in programming languages like Python, R, Java, and SQL.
- Experience with big data tools and frameworks such as Hadoop and Spark.
- Strong understanding of database management systems and data warehousing solutions.
- Knowledge in data architecture and data modeling.

2. Data Analyst

Key Responsibilities:

- Analyze datasets to extract actionable insights and support decision-making.
- Create visualizations, reports, and dashboards to communicate findings.
- Conduct statistical analyses to identify trends, patterns, and correlations.

Key Skills:

- Proficiency in analytics tools and software like Excel, SQL, and Tableau/Power BI.
- Strong knowledge of statistical methods and data analysis techniques.
- Ability to translate complex data sets into understandable and actionable insights.
- Good communication skills to present findings to stakeholders.

3. Machine Learning Engineer

Key Responsibilities:

- Design, implement, and deploy machine learning models and algorithms.
- Conduct experiments to test hypotheses and evaluate model effectiveness.
- Optimize machine learning models for performance and scalability.

Key Skills:

- Proficiency in machine learning frameworks and libraries like TensorFlow and PyTorch.
- Strong programming skills in languages like Python or R.
- In-depth understanding of algorithms, data structures, and statistical methods.
- Experience with model deployment and production environments, utilizing tools like Docker and Kubernetes.

4. Business Intelligence Analyst

Key Responsibilities:

- Collect, analyze, and interpret business data to provide insights for strategic and operational decisions.
- Develop and maintain BI systems and dashboards.
- Collaborate with business stakeholders to identify data needs and requirements.

Key Skills:

- Advanced proficiency in BI tools like Tableau, Power BI, or Looker.
- Strong analytical and critical thinking skills.
- Good understanding of business processes and key performance indicators (KPIs).
- Ability to communicate complex concepts to non-technical audiences.

5. Database Administrator

Key Responsibilities:

- Manage databases to ensure their optimal performance, security, and reliability.
- Regularly backup and recover data to prevent data loss.

- Monitor database performance and troubleshoot issues as they arise.

Key Skills:

- Proficiency in database management systems like Oracle and MySQL.
- Strong understanding of database architecture, indexing, and query optimization.
- Knowledge of database security measures and compliance standards.
- Experience with performance tuning and debugging.

6. Data Product Manager

Key Responsibilities:

- Oversee the development and implementation of data-driven products and solutions.
- Liaise between data science teams and business stakeholders to align product goals with business needs.
- Define product roadmaps, set priorities, and manage project timelines.

Key Skills:

- Strong project management skills and experience with Agile methodologies.
- Excellent communication and interpersonal skills to effectively work with cross-functional teams.
- Understanding of data science principles and the ability to translate business needs into technical requirements.
- Analytical mindset to evaluate product performance and make data-driven decisions.

4. **Data Literacy and Pitfalls**

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

(a) (8 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) (8 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) (8 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) (8 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.

Solution:

Here are some sketches of possible answers:

a) Reading Data Pitfall: One pitfall related to the verb "read" in data literacy is the misinterpretation of data due to insufficient context. Often, analysts may read data without understanding the full background or context in which it was collected. This can lead to incorrect conclusions or decisions based on incomplete information. For instance, assuming a sudden increase in website traffic without considering a simultaneous marketing campaign might lead to incorrect attributions of causality.

b) Understanding Data Pitfall: Understanding data involves interpreting its meaning accurately. A common pitfall here is the misuse of statistical techniques or misapplication of models. For example, applying a linear regression model to data that doesn't meet the assumptions of linearity or independence can lead to misleading interpretations. This pitfall highlights the importance of selecting appropriate analytical methods and understanding their limitations to avoid erroneous conclusions.

c) Creating Data Pitfall: Creating data refers to generating new datasets or modifying existing ones. A significant pitfall in this area is data fabrication or manipulation. This can occur unintentionally due to errors in data entry or processing, or deliberately to support a specific hypothesis or agenda. Such practices undermine the integrity and reliability of data-driven insights and can lead to invalid conclusions and decisions.

d) Communicating Data Pitfall: Communicating data effectively involves presenting information in a clear, understandable manner. A pitfall related to communication is the use of misleading visualizations or graphs. For instance, scaling a graph improperly can exaggerate differences or trends, leading to misinterpretations by stakeholders. Clear and honest communication is crucial to ensure that data is accurately represented and understood by all parties involved.

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?

Solution:

By focusing on a limited number of cherry-picked examples, researchers fail to capture the full spectrum of data patterns and do not allow drawing causal conclusions. For example, if a study selectively considers only successful case studies that highlight the factors contributing to their success, it presents a skewed view that does not reflect the entire reality. This approach often leads to a selection bias where unsuccessful cases are omitted from the analysis, distorting the interpretation of causal effects.

Therefore, attributing success solely to specific management practices observed in the selected sample may be invalid, as these practices could also be present in unsuccessful companies that are not included in the study. Biases in case selection, such as confirmation bias or survivorship bias, can skew interpretations and undermine the reliability of findings. For example, if a study focuses only on companies that survived economic downturns without considering those that failed, it may overlook important risk factors or management strategies that contribute to a comprehensive understanding of resilience in business.

(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?

Solution:

Extrapolation can lead to misleading assumptions about causality or trends, especially when the sample size is small or unrepresentative. For example, assuming a correlation between two variables based on a few observed instances without accounting for confounding factors can result in erroneous conclusions that do not hold in broader contexts.

(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?

Solution:

To mitigate these pitfalls, researchers should ensure representative sampling, consider alternative explanations for findings, and conduct sensitivity analyses to test the robustness of conclusions. Transparency in reporting methods, discussing limitations, and providing code for reproducibility can significantly enhance the reliability and validity of research outcomes, as emphasized in Jones (2020) and discussed in our course on sound data analysis principles.

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

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

[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.

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

Solution:

```
result <- 3+sqrt(9)-2^{2}*4/3
```

(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.

Solution:

```
getwd()
```

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

Solution:

```
setwd("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.

Solution:

```
?lm()
```

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

Solution:

```
install.packages("knitr")
```

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

Solution:

The assignment operator <- is primarily used in programming languages like R for assigning values to variables. Its usage involves placing the variable name on the left and the value or expression on the right. This operator helps in creating and updating variables efficiently. The main advantages include readability and clarity, as it explicitly shows the direction of data flow. Additionally, it supports both assignment and re-assignment of values, enhancing code flexibility and maintainability. The operator also aligns with the functional programming style, which emphasizes clear and concise code structure.

(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.

Solution:

```
summary(df_deaths)
```

```
##     state              deaths
##  Length:3           Min.   :12264
##  Class :character   1st Qu.:15414
##  Mode  :character   Median :18565
##                     Mean   :16773
##                     3rd Qu.:19027
##                     Max.   :19489
```

(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`.

Solution:

```
library("tidyverse")
state <- c("NRW", "BY", "BW")
cases <- c(628, 511, 374)
df_covid <- tibble(state, cases)
```

(h) (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`.

Solution:

```
df_filtered <- df_covid |>
     filter(cases > 500)
```

(i) (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`.

Solution:

```
df_sorted <- df_covid |>
  arrange(desc(cases))
```

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

Solution:

```
total_cases <- sum(df_covid$cases)

# Alternative:
df_covid <- df_covid |>
  mutate(total_cases = sum(cases))
```

(k) (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.

Solution:

```
df_covid <- df_covid |>
```

```
  mutate(cases_percentage = (cases / total_cases) * 100)
```

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

Solution:

```
library(tidyverse)
df_covid |>
  summarise(mean_cases = mean(cases),
            sd_cases = sd(cases))
```

```
## # A tibble: 1 x 2
##   mean_cases sd_cases
##        <dbl>    <dbl>
## 1       504.     127.
```

(m) (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.