Mock Midterm Exam
Introduction to Applied Data Science

Name: _____ Student No: _____

2026-05-17

Instructions

This mock midterm exam contains **20 multiple choice questions** and **5 open questions**, and has the same format and duration as the actual midterm exam. For the MC questions, encircle the letter of your chosen answer. All MC questions carry equal weight (1 pt each). The open questions are worth 3 pts each. The **total is 35 points**. The duration of this exam is **2 hours**. You may only ask clarifying questions during the first 30 minutes of the exam. No computers, phones, or internet access are permitted. Don't forget to fill in your name and student number on this page. Good luck!

Part A: Multiple Choice Questions (20 points)

1. In empirical research, what is the primary focus of the **prediction** objective?

A) Estimating the causal effect of one variable on another, holding all other factors constant
B) Minimizing the gap between the predicted values ($\hat{y}$) and the actual observed values (y)
C) Accurately measuring an economic quantity, such as a country's GDP
D) Building formal structural models of economic behavior

---

2. Which of the following violates the **tidy data** principle?

A) Each row is one student; each column records one variable (name, age, grade)
B) A single column called `scores` contains entries like "Math: 8, English: 7" for each student
C) Each row is one country-year observation, with separate columns for GDP and population
D) A dataset has one column for city names and a separate column for country names

---

3. You want to briefly inspect a `data.frame` called `gdp_data` to see the data types of all its columns and a preview of the first few values. Which function is most useful for this?

A) `head()`
B) `dim()`
C) `class()`
D) `str()`

---

4. What is the fundamental difference between `x <- c(1, "hello", TRUE)` and `y <- list(1, "hello", TRUE)` in R?

A) `x` will cause an error; `y` will work correctly
B) `x` will coerce all elements to the character type; `y` preserves each element's original type
C) Both `x` and `y` store all elements with their original types
D) `x` creates a data frame; `y` creates a vector

---

5. What is the result of `length(list(a = 1, b = c(10, 20, 30), c = "hello"))` in R?

A) 5
B) 6
C) 3
D) 1

---

6. Which R function returns the path of your current working directory?

A) `setwd("C:/Users/Me/Documents")`
B) `list.files()`
C) `getwd()`
D) `dir.exists()`

Have a look at the following `data.frame`:

```
         name gdp_pc eu_member
1 Netherlands  52341      TRUE
2     Germany  48203      TRUE
3      France  43187      TRUE
4       Italy  35982      TRUE
```

7. Which expression returns the `name` column as a **data frame** (not a plain vector)?

A) `countries$name`
B) `countries[["name"]]`
C) `countries[, "name"]`
D) `countries["name"]`

---

8. Which expression correctly filters the `countries` data frame to keep only rows where `gdp_pc` is greater than 45000?

A) `countries[countries$gdp_pc > 45000]`
B) `countries[, countries$gdp_pc > 45000]`
C) `countries[countries$gdp_pc > 45000, ]`
D) `countries["gdp_pc" > 45000, ]`

---

9. What is the purpose of the `na.rm = TRUE` argument commonly used in functions like `mean()`, `sum()`, or `sd()` in R?

A) It removes entire rows containing at least one NA before the calculation
B) It replaces NA values with zero before performing the calculation
C) It excludes NA values from the calculation, performing it on the remaining values
D) It renames columns that contain NA values

---

10. Your working directory is `~/project/reports/`. You want to read the file located at `~/project/data/unemployment.csv`. Which relative path is correct?

A) `read_csv("./data/unemployment.csv")`
B) `read_csv("~/data/unemployment.csv")`
C) `read_csv("../data/unemployment.csv")`
D) `read_csv("/data/unemployment.csv")`

---

11. What is the primary advantage of using `here("data", "gdp.csv")` instead of `"./data/gdp.csv"` inside a Quarto document?

A) `here()` reads files faster than relative paths

B) `here()` always resolves paths relative to the project root, regardless of where the script or `.qmd` file is located
C) `here()` automatically downloads missing data files from the internet
D) `here()` only works with `.csv` files

---

12. Which function from the `readr` package reads a comma-separated values (CSV) file into R?

A) `read.csv()`
B) `read_excel()`
C) `read_csv()`
D) `fromJSON()`

---

13. You need to read the sheet named `"Trade_2023"` from an Excel file called `eurostat.xlsx`. Which function call is correct?

A) `read_excel("eurostat.xlsx", page = "Trade_2023")`
B) `read_excel("eurostat.xlsx", tab = "Trade_2023")`
C) `read_excel("eurostat.xlsx", sheet = "Trade_2023")`
D) `read_excel("eurostat.xlsx", range = "Trade_2023")`

---

14. When you call `GET()` from the `httr` package to make an API request, what type of HTTP operation are you performing?

A) Deleting a resource from the server
B) Uploading data to the server
C) Retrieving data from the server
D) Updating an existing resource on the server

---

15. Which HTTP status code indicates that your API request was **successful** and data was returned?

A) 404
B) 401
C) 429
D) 200

---

16. You call an API and receive HTTP status code 401. What is the most likely cause?

A) The server is temporarily overloaded or unavailable
B) The requested endpoint does not exist on the server
C) Your authentication credentials (e.g., API key) are missing or invalid
D) You have exceeded the allowed number of requests in a given time period

---

17. In the URL `https://api.example.com/v2/weather?city=Utrecht&units=metric`, what does the `?` symbol indicate?

A) It marks the start of the secure HTTPS protocol
B) It separates the domain from the endpoint path
C) It marks the beginning of the query parameters
D) It indicates a wildcard for pattern matching in the URL

---

18. What is the primary function of the `rvest` package in R?

A) To perform complex statistical modeling
B) To create publication-quality data visualizations
C) To efficiently perform web scraping
D) To interface with SQL databases

---

19. You are using `rvest` to scrape a webpage. After selecting an `<a>` tag element, you want to extract the URL it links to (stored in the `href` attribute). Which function should you use?

A) `html_text()`
B) `html_name()`
C) `html_children()`
D) `html_attr("href")`

---

20. What does the CSS selector `div.news > p` match?

A) All `<p>` elements anywhere inside a `<div class="news">`, at any level of nesting
B) All `<p>` elements that are **direct children** of a `<div class="news">`
C) All `<div class="news">` elements that contain a `<p>`
D) All elements that have both the tag name “news” and tag name “p”

Part B: Open Questions (15 points)

Question 1 (3 pts)

Consider the following file structure:

```
my_project/
├── data/
│   └── gdp.csv
├── scripts/
│   └── analysis.R
└── report.qmd
```

Your working directory is currently set to `my_project/scripts/`. Write the R function call to read `gdp.csv` using a **relative path**. Then write the equivalent call using the `here()` function (assuming `my_project` is your Positron project root). What is the key difference between the two approaches?

Question 2 (3 pts)

Have a look at the following `data.frame`:

```
  student_id  name grade passed
1          1  Anna   7.5   TRUE
2          2 Boris   8.8   TRUE
3          3  Chen   6.2  FALSE
4          4 Diana   9.1   TRUE
```

What is the difference in output between `students["grade"]` and `students[["grade"]]`? What class does each expression return? Why does this distinction matter in practice?

Question 3 (3 pts)

Have a look at the following URL:

```
https://api.worldbank.org/v2/country/nl/indicator/NY.GDP.MKTP.CD?format=json&
per_page=10
```

Identify and name each component of this URL: the **protocol**, the **domain**, the **endpoint**, and all **query parameters**. Briefly explain what each query parameter likely does.

Question 4 (3 pts)

Have a look at the following HTML snippet:

```
<div class="article">
  <h2 class="headline">EU Economy Grows by 2.5%</h2>
  <p class="summary">The European economy showed strong growth...</p>
  <a href="https://news.example.com/eu-growth">Read more</a>
</div>
<div class="article">
  <h2 class="headline">Inflation Remains Stable</h2>
  <p class="summary">Consumer prices held steady in October...</p>
  <a href="https://news.example.com/inflation">Read more</a>
</div>
```

(a) Write one CSS selector to select **all elements with class** `"headline"`. (b) Write one CSS selector to select **only** `<a>` **elements** that are **inside** a `<div class="article">`. How would you use `rvest` in R to extract the `href` values from those links?

Question 5 (3 pts)

Have a look at the following `rvest` code:

```
library(rvest)

page     <-  read_html("https://en.wikipedia.org/wiki/List_of_countries_by_GDP_
(PPP)")
result <- page |>
  html_element("table.wikitable") |>
  html_table()
```

Describe step-by-step in plain English what each line (after `library(rvest)`) does. What type of R object does `result` contain? What would change if you replaced `html_element()` with `html_elements()`?

End of Exam