

# BCB744: Intro R Theory Test

A. J. Smit  
University of the Western Cape

2026-02-13

## Instructions

The Intro R Theory Test will start at **12:30** on **26 March, 2026**. You have until **15:30** to complete it.

Your answer should demonstrate a comprehensive understanding of the theoretical concepts and techniques required to read and comprehend R code.

Only answer what is explicitly stated. For example, if the question asks for only a graph as the final output, *only* the graph will be assessed, not the reasoning that brought you there. Anything extra will not amount to extra marks, so save yourself the time and produce the most concise answer possible given the content of the question. **What is required will *always* be explicitly stated.**

**This is a closed book assessment.** Below is a set of questions to answer. You must answer **all** questions in the allocated time of **3-hr**. Please write your answers neatly in the answer book provided. Structure your answers logically.

### Question 1 [10 marks]

Please translate the following code into English by providing an explanation for each line. At the end, state what kind of data structure this pipeline produces and how that output could flow into a `ggplot2` figure.

```
1 library(tidyverse)
2 monthlyData <- dailyData |>
3   mutate(t = asPOSIXct(t)) |>
4   mutate(month = floor_date(t, unit = "month")) |>
5   group_by(lon, lat, month) |>
6   summarise(temp = median(temp, na.rm = TRUE)) |>
7   mutate(year = year(month)) |>
8   group_by(lon, lat) |>
9   mutate(num = seq(1:length(temp))) |>
10  ungroup()
```

In your answer, refer to the line numbers (1-9) and provide an explanation for each line. Then add one final sentence describing the resulting dataset and a likely `ggplot2` use.

### Question 2 [5 marks]

What are the three properties of **tidy data**? Briefly explain each one, and then state why tidy data work especially well with `ggplot2`.

### Question 3 [20 marks]

Using the penguin data provided in Table 1, please produce the figure produced by the code block.

**Table 1:** Penguin Sample (n = 4 per Species x Island)

Species	Island	Bill length (mm)	Bill depth (mm)	Flipper length (mm)	Body mass (g)	Sex	Year
Adelie	Biscoe	37.8	20.0	190.0	4,250.0	male	2009
Adelie	Biscoe	37.9	18.6	193.0	2,925.0	female	2009
Adelie	Biscoe	37.9	18.6	172.0	3,150.0	female	2007
Adelie	Biscoe	40.1	18.9	188.0	4,300.0	male	2008
Adelie	Dream	41.5	18.5	201.0	4,000.0	male	2009
Adelie	Dream	41.1	19.0	182.0	3,425.0	male	2007
Adelie	Dream	36.0	17.9	190.0	3,450.0	female	2007
Adelie	Dream	38.1	18.6	190.0	3,700.0	female	2008
Adelie	Torgersen	40.2	17.0	176.0	3,450.0	female	2009
Adelie	Torgersen	42.9	17.6	196.0	4,700.0	male	2008
Adelie	Torgersen	33.5	19.0	190.0	3,600.0	female	2008
Adelie	Torgersen	39.3	20.6	190.0	3,650.0	male	2007
Chinstrap	Dream	45.2	17.8	198.0	3,950.0	female	2007
Chinstrap	Dream	49.3	19.9	203.0	4,050.0	male	2009
Chinstrap	Dream	46.5	17.9	192.0	3,500.0	female	2007
Chinstrap	Dream	45.5	17.0	196.0	3,500.0	female	2008
Gentoo	Biscoe	50.4	15.7	222.0	5,750.0	male	2009
Gentoo	Biscoe	49.5	16.1	224.0	5,650.0	male	2009
Gentoo	Biscoe	52.5	15.6	221.0	5,450.0	male	2009
Gentoo	Biscoe	49.3	15.7	217.0	5,850.0	male	2007

```
1 pen_long <- pen |>
2   pivot_longer(
3     cols = c(bill_length_mm, bill_depth_mm, flipper_length_mm),
4     names_to = "measurement_type",
5     values_to = "value_mm"
6   ) |>
7   group_by(species, island, measurement_type) |>
8   summarise(
9     mean_mm = mean(value_mm, na.rm = TRUE),
10    sd_mm = sd(value_mm, na.rm = TRUE),
11    .groups = "drop"
12  )
13
14 ggplot(pen_long, aes(x = measurement_type, y = mean_mm, fill = measurement_type)) +
```

```

15 geom_col(width = 0.7, colour = "black", linewidth = 0.2, show.legend = FALSE) +
16 geom_errorbar(
17   aes(ymin = mean_mm - sd_mm, ymax = mean_mm + sd_mm),
18   width = 0.2,
19   linewidth = 0.2
20 ) +
21 facet_grid(species ~ island) +
22 scale_x_discrete(
23   labels = c(
24     bill_length_mm = "Bill length", bill_depth_mm = "Bill depth",
25     flipper_length_mm = "Flipper length"
26   )
27 ) +
28 labs(x = "Measurement", y = "Mean length (mm) ± SD") +
29 theme_bw(base_size = 11) +
30 theme(
31   strip.text = element_text(face = "bold"),
32   axis.text.x = element_text(angle = 20, hjust = 1)
33 )

```

Marks will *only* be assigned for the figure that the code produces.

#### Question 4 [5 marks]

Why do we prefer to use R over Excel for data analysis and statistics?

#### Question 5 [10 marks]

By way of example, please explain some key aspects of R code conventions. For each line of code you write (**neatly and legibly so each intended style item is visible**), explain also in English what aspects of the code are being adhered to.

For example:

**a <- b** is not the same as **a < -b**. The former is correct because there is a space preceding and following the assignment operator (<-, a less-than sign immediately followed by a dash to form an arrow); this has a different meaning from the latter, which is incorrect because there is no space between the less-than sign, the dash and reading as “a is less than negative b”.

#### Question 6 [15 marks]

You are a research assistant who have just been given your first job. You are asked to analyse a dataset about patterns of extreme heat in the ocean and the possible role that ocean currents (specifically, eddies) might play in modulating the patterns of extreme sea surface temperature extremes in space, time.

Being naive and relatively inexperienced, and misguided by your exaggerated sense of preparedness as young people tend to do, you gladly accept the task, and start by exploring the data. You notice that the dataset is quite large and you have no idea what is happening, what you are doing, why you are doing it, or what you are looking for. Ten minutes into the job you start to question your life choices. Your feeling of bewilderment is compounded by the fact that, when you examine the data (the output of the `head()` and `tail()` commands is shown below), the entries seem confusing.

```

1  fpath <- "/Volumes/OceanData/spatial/processed/WBC/misc_results"
2  fname <- "KC-MCA-data-2013-01-01-2022-12-31-bbox-v1_ma_14day_detrended.csv"
3  data <- read.csv(file.path(fpath, fname))

```

```

1  > nrow(data)
2  [1] 53253434
3
4  > head(data)
5           t      lon  lat    ex  ke
6  1 2013-01-01 121.875 34.625 -0.7141 2e-04
7  2 2013-01-01 121.875 34.625 -0.8027 2e-04
8  3 2013-01-02 121.875 34.625 -0.8916 2e-04
9  4 2013-01-02 121.875 34.625 -0.9751 2e-04
10 5 2013-01-03 121.875 34.625 -1.0589 3e-04
11 6 2013-01-03 121.875 34.625 -1.1406 3e-04
12
13 > tail(data)
14           t      lon  lat    ex  ke
15 53253429 2022-12-29 174.375 44.875 0.4742 -0.0049
16 53253430 2022-12-29 174.375 44.875 0.4856 -0.0049
17 53253431 2022-12-30 174.375 44.875 0.4969 -0.0050
18 53253432 2022-12-30 174.375 44.875 0.5169 -0.0050
19 53253433 2022-12-31 174.375 44.875 0.5367 -0.0051
20 53253434 2022-12-31 174.375 44.875 0.5465 -0.0051

```

You resign yourself to admitting that you do not understand much, but at the risk of sounding like a fool when you go to your professor, you decide to do as much of the preparation you can do so that you at least have something to show for your time.

a. What will you take back to your professor to show that you have prepared yourself as fully as possible?

For example:

- What is in your ability to understand about the study and the nature of the data?
- What will you do for yourself to better understand the task at hand?
- What do you understand about the data?
- What will you do to aid your understanding of the data?
- What will your next steps be going forward?
- Etc. (Anything else you can think about doing to convince the professor you thought about the data?)

[/10 marks]

b. What will you need from your professor to help you understand the data and the task at hand so that you are well equipped to tackle the problem? [/5 marks]

### Question 7 [15 marks]

Name the general characteristics of ASCII-type data files, then name and explain three common variations of these tabular data files.

**Question 8 [20 marks]**

Explain each of the following in the context of their use in R. For each, provide an example of how you would construct them in R:

- a. A vector
- b. A matrix
- c. A dataframe
- d. A list

**TOTAL MARKS: 100**

**- THE END -**

**Bibliography**