DATA420-25S2 (C)

Assignment 1

GHCN Data Analysis using Spark

Due on Friday, September 12, 2025 by 5:00 PM.

If you want to discuss the assignment material you can use the Discord server where the discussion will benefit all. If you have a question that requires an official answer you can use the forum on LEARN. If you have a more personal question you can email me or contact the class rep as needed.

A reminder that the Discord server is for discussion of concepts only, not for sharing code or answers to assignment questions.

Links

Report upload (pdf)
Supplementary material upload (zip, limited to 10 MB)

Discord server
Help forum for Assignment 1

Instructions

- Your report should be submitted as a single pdf file on LEARN. Any additional code, images, and supplementary material should be submitted separately as a single zip file on LEARN. You should **not** submit any outputs as part of your supplementary material, leave these in cloud storage.

- The body of your report should be between 3,000 and 5,000 words long, excluding your cover page, table of contents, references, appendices, and supplementary material. You need to be accurate and concise and you need to demonstrate depth of understanding.

- You should make sensible choices concerning margins, font size, spacing, and formatting. For example, margins between 0.5" and 1", a sans-serif font e.g. Arial with font size 11 or 12, line spacing 1 or 1.15, and sensible use of monospaced code blocks, tables, and images.

- You should reference any external resources using a citation format such as APA or MLA, including any online resources which you used to obtain snippets of code or examples. You must reference any use of Grammerly, ChatGPT, or any other generative AI tools to **improve** the quality of your own original work.

- You **must not** use any content generated by AI directly in your report or your supplementary material. You are encouraged to use AI to help you solve problems and develop code, but you should take time to understand any content that you use so that you develop accurate depth of understanding.

GRADING

The assignment is graded across a number of categories which are summarised in the table below.

Answers	25
Reasoning	25
Tables	7
Visualizations	18
Writing	13
Coding	12

The marks for the answers, reasoning, tables, and visualizations categories are distributed across the sections in the assignment, the marks for the writing section are based on your report as a whole, and the marks for the coding section are based on your code as a whole.

Writing

- Was the report well structured overall?
- Did the report go into a suitable amount of detail and demonstrate depth of understanding?
- Was writing concise and was the report easy to understand?
- Was writing natural and professional?
- Was any code included in the report or was it only in the supplementary material?
- Were any external resources appropriately cited and referenced?
- Was any use of AI appropriately acknowledged?

Coding

- Were notebooks well structured and easy to navigate and understand?
- Were any empty cells, exceptions, or other anomalies left in the notebook?
- Was code style consistent and readable overall?
- Was code commented appropriately?
- Was supplementary material provided and was it well structured?

You should structure your assignment report based on the high level comments below, and then check that you have satisfied the grading criteria the processing, analysis, and visualization sections across each of the answers, reasoning, tables, and visualizations categories.

Structure

Your report should have the following sections within which you can also use question numbers as subheadings to group paragraphs, tables, and figures that you use to answer the questions that have been asked. You should keep your writing concise and easy to understand, and you should provide enough detail to demonstrate depth of understanding.

Background

- You should give a brief overview of the purpose of the assignment and what you have achieved or understood with your analysis and visualizations.
- You should provide a high level summary of the Global Historical Climatology Network (GHCN), similar to the first part of the data section in the assignment brief, which will provide context for the description of the structure and content of the data in the processing section below.

Processing

- You should describe the structure and content of the datasets which you can refer back to in the sections below. You should describe the steps that you took to load, join, and check the data, answer the questions that have been asked, and discuss anything else that you discovered along the way.
- You should **not** include outputs other than answers to the questions that you have been asked.

Analysis

- You should answer the questions that have been asked, give a high level summary of what you have done, and discuss any insights that you had. You should talk about any tasks that you were unable to complete and explain why.
- You should **not** just list answers that you have not explained.

Visualizations

- You should answer the questions that have been asked, present visualizations, give a high level summary of what you have done, and discuss any insights that you had.
- You should talk about any visualizations that you were not able to generate and why.

Conclusions

- You should give an overview of what you have achieved and what you have learned.

References

- You should list all references that you have used or referenced.

Processing

This section is about developing your understanding of the **structure** of the data and setting up code that you will need to load and join the datasets so that you can use them to answer questions in the analysis and visualization sections.

You should explain your methodology, your results, and any conclusions that you made as you worked through the tasks step by step. You do **not** need to include any screenshots of your output at each step but you should provide an overview of the structure of the final enriched `stations` table that you have derived. You could list each column in your table with a description of what it contains.

Answers

- Number of years in daily.
- The size and the number of rows in each dataset with appropriate units.
- How many stations collected all core elements and how many collected precipitation only.
- How many stations are in daily but not in stations vs in stations but not in daily.

Reasoning

- How the data is structured including whether files are compressed or uncompressed.
- How the size of the data changes over time.
- An estimate of how much larger daily could become when it is uncompressed.
- A comment on what data types were used.
- How expensive it would be to join all of daily and stations and if it can be done efficiently.

Tables

- A table containing the names and sizes of the datasets, including an estimate of how large `daily` could be when it is uncompressed and a comment on how the sizes of the other datasets compare to `daily`.
- A table describing the structure of the final enriched `stations` table that you have derived.

Visualizations

- A visualization of the directory tree and a brief written summary.
- A visualization of the size of each year of `daily` as compressed and a description of how the size changes over time.

Analysis

This section is about developing your understanding of the data and demonstrating that you can answer specific questions.

You should explain your methodology, your results, and any conclusions that you made as you worked through the tasks step by step. You will need to make decisions along the way such as how to compute geographical distance or how to count how many observations of TMAX do not have a corresponding observation of TMIN. You should justify these decisions carefully.

Answers

- How many stations there are in total.
- How many stations were active so far in 2025.
- How many stations there are in each of the networks referenced in `daily`.
- How many stations there are in the Southern Hemisphere.
- How many stations there are in territories of the United States around the world.
- Which stations are closest together in New Zealand.
- How many rows there are in `daily`.
- How many observations there are for each of the five core elements.
- How many observations of TMAX do not have a corresponding observation of TMIN.

Reasoning

- How to calculate spherical distance around the world **to an acceptable accuracy**. You should explain how the function takes into account that the earth is spherical and you should clearly cite any sources that you have used.
- A comment on which element has the most observations.
- How to count how many observations of TMAX do not have a corresponding observation of TMIN **in an efficient way**.

Tables

- A table containing the counts of stations above with a description of what each count represents.
- A table containing the counts of observations for each five core elements.

Visualizations

- A map showing where the stations are in New Zealand.

Visualizations

This section is about using visualizations that present information in a way that is easy to understand and visually compelling.

You should explain how you have you prepared your data for visualization and any additional processing or decisions that you have made such as how you have handled gaps in the data where no observations were reported. You will need to make decisions about how to best style and label your visualizations.

Answers

- How many observations and years of data there are for New Zealand.
- Descriptive statistics for the average rainfall in each year for each country.
- Which country has the highest average rainfall in a single year.

Reasoning

- How to smooth time series and what is an appropriate level of detail.
- How to handle gaps in data where no observations were reported.
- A clear explanation of how you have calculated average daily rainfall in each year for each country.
- A comment on the highest average rainfall in a single year and on outliers in general.
    - How many stations or observations contributed to the average.
    - How to identify and remove outliers.
- A comment on which map projection is the most suitable.
- You should demonstrate your understanding of any anomalies in your choropleth as visualized.
    - How the countries in this dataset match with countries in the open source library.
    - How rainfall is distributed around the world.
    - Any countries that stand out for having visibly high or low average rainfall.

Visualizations

- A single visualization containing subplots of TMIN and TMAX for each station in New Zealand.
- A separate visualization of the average TMIN and TMAX for the entire country.
- A choropleth of average daily rainfall in each country for 2024.