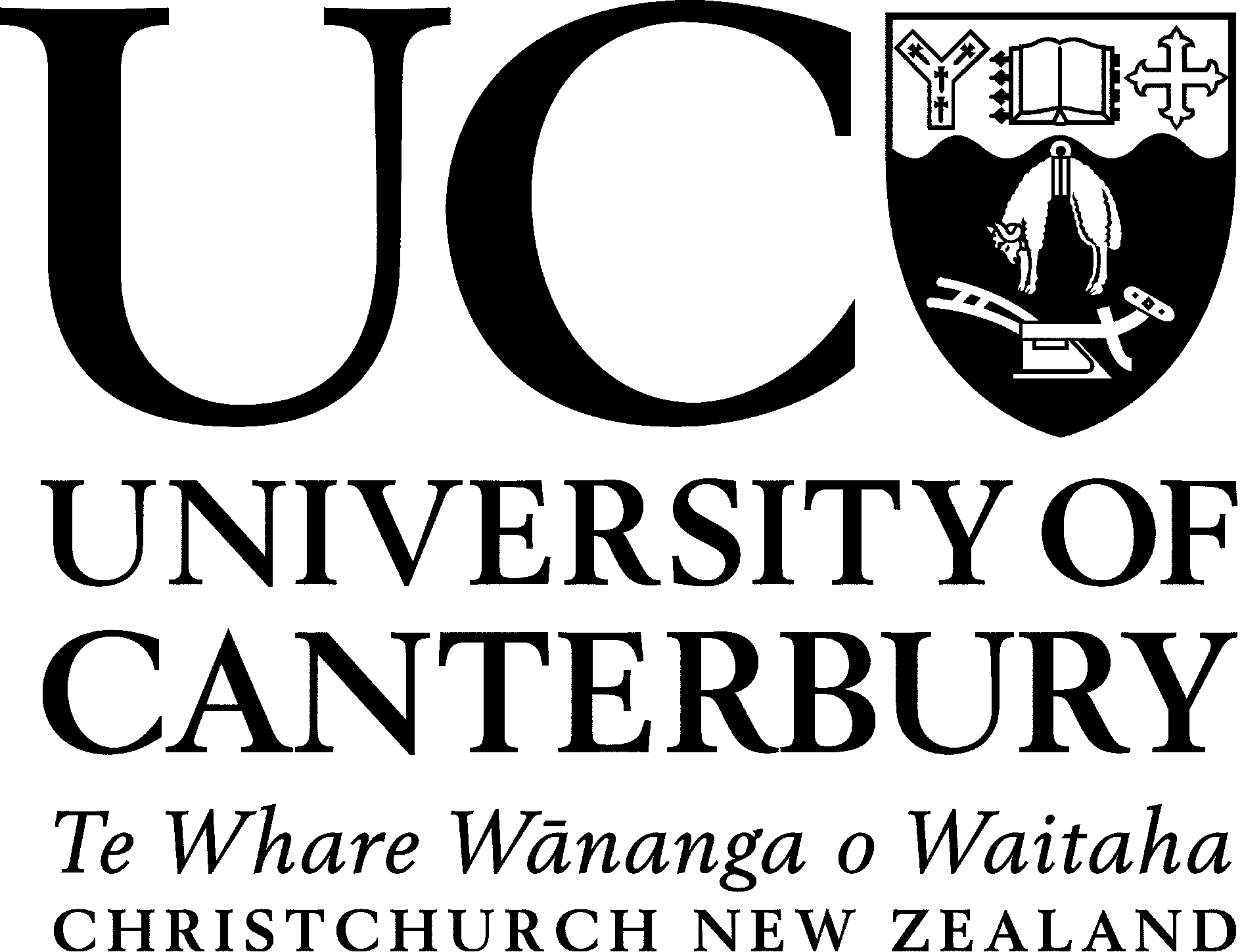
****

**DATA420**

**Scalable Data Science**

**Assignment 1**

**GHCN Data Analysis using Spark**

**Yu Xia 62380486**

**September 14, 2025**

## 1. Background

* Purpose of the assignment
* Overview of what was achieved (processing, analysis, visualization)
* High-level introduction of **GHCN Daily**: scope, scale, variables, metadata
* Context for why Spark was used (scale, distributed computing)

## 2. Processing

* **Description of datasets**: daily, stations, states, countries, inventory
  + Data formats (CSV, fixed-width, compressed)
  + Schema definition for each dataset
* **Steps performed**:
  + Loading datasets into Spark (schemas, casting, parsing fixed-width)
  + Exploration (row counts, years, compressed/uncompressed size)
  + Joins to build *enriched stations table* (country/state info + inventory)
  + Data quality checks (missing stations, missing values, negative precipitation placeholders)
* **Results**:
  + Counts of rows & sizes in each dataset
  + Structure of enriched stations table (columns + descriptions)
* **Tables & Figures**:
  + Dataset sizes table (compressed/uncompressed estimate)
  + Structure of enriched stations table
  + Directory tree figure
  + Size of daily per year (compressed) visualization

## 3. Analysis

* **Stations**:
  + Total stations; active in 2025
  + Counts per network (GSN, HCN, CRN)
  + Stations in Southern Hemisphere
  + Stations in US territories (excluding mainland US)
  + Which stations in NZ are geographically closest (explain spherical distance calculation, haversine formula, cite sources)
* **Daily observations**:
  + Total number of rows in daily
  + Counts for each of the five core elements (TMAX, TMIN, PRCP, SNOW, SNWD)
  + Which element has the most observations
  + How many TMAX records lack a TMIN partner (methodology for efficient count)
* **Tables & Figures**:
  + Counts of stations table
  + Counts of observations per element table
  + Map of NZ stations

## 4. Visualizations

* **New Zealand Temperature**:
  + Time series of TMIN and TMAX for each NZ station (subplot figure)
  + Smoothed/aggregated to reduce noise, discussion of handling gaps
  + Average TMIN and TMAX for entire NZ (single larger figure)
* **Global Precipitation**:
  + Group precipitation by year + country → average daily rainfall
  + Outlier detection (boxplots, clipped + log-scale)
  + Descriptive statistics (summary table)
  + Identify country-year with max rainfall; discuss sensibility
  + 2024 Choropleth of rainfall by country (Robinson projection, grey hatched = No data)
  + Discussion:
    - Outliers & anomalies
    - Matching GHCN FIPS codes with GeoPandas ISO3 codes
    - Suitable projection choice (Robinson)
    - Color scale choices
    - Countries with unusually high/low rainfall

## 5. Conclusions

* Summary of findings and what was learned
* Key insights from processing, analysis, visualizations
* Limitations (e.g., missing data, mapping issues, extreme outliers)
* Possible future improvements (more robust mapping, handling missing observations, larger scale Spark optimizations)

## 6. References

* GHCN-Daily README
* Natural Earth (GeoPandas world dataset)
* PySpark, GeoPandas, Matplotlib/Seaborn/Cartopy docs
* Any external references (e.g., Haversine distance formula)
* AI usage acknowledgment (if required)

## 7. Appendix (Optional)

* Supplementary tables/figures not included in main body
* Code excerpts (but bulk code should go to supplementary material zip, per grading rules )

这样一个框架，已经 **覆盖了评分细则里要求的 Answers, Reasoning, Tables, Visualizations, Writing, Coding 六类** 。

你可以直接在 Word 里用这个结构展开，逐步把我们生成的表格、图形、代码输出嵌入其中。

要不要我帮你把 **每一节需要的图表/表格清单**再单独列出来，作为“写作 checklist”，这样你在写报告时不会漏掉任何评分点？