Part A – Data Summarization, Data Pre-processing, and Feature selection

I have included scatterplots and boxplots of the data in my GitHub repository.
The scatterplots show how the number of covid cases, deaths, fully vaccinated and partially vaccinated people as well as the temperature vary during the 26 months, ranging from December 2020 to January 2023.

Missing values were already handled during the data staging part of the project.

The fact.csv file contains data from the fact table. However, this dataset was cumbersome to work with. Hence, I created a grouped.csv file which is a summarized version of the fact.csv file. It contains 1685 rows. It contains the weatherid for 65 distinct locations (5 from each province/territory), the date (26 months), the temperature, the proportion of fully vaccinated people and the rate of change of covid cases.

I converted the date attribute to an ordinal attribute (since there is an order to the 26 distinct months).
I also one-hot encoded the weatherid attribute. Since the weatherid column contains 65 distinct values, this resulted in 65 new columns. One-hot encoding was the best option for the weatherid attribute since it is a nominal attribute.

I decided not to normalize the numeric attributes (i.e., the temperature, rate of change of covid cases and proportion of fully vaccinated people) because the decision tree algorithms used in the subsequent part are not affected by monotonic transformation.

In addition, the grouped.csv file does not contain redundant attributes. All potential redundant attributes were already removed in the data staging process.

Github repository link for data-preprocessing: https://github.com/bsewp045/CSI4142-Phase-4