Shaikh Mohammad Rahil

Data Cleaning and Summarising
Exploring Daily Rainfall Climate Data: Insights from ABC City

Introduction

This report thoroughly prepares and explores Daily Rainfall Climate Data from the Australian Government's Bureau of Meteorology. It meticulously inspects the data, addresses issues through data cleaning, and explores various aspects to derive insights. Data preparation involves handling entry errors, impossible values, and missing values, ensuring data formatting. Subsequent exploration includes analysing yearly and monthly rainfall patterns, comparing rainfall across years, and examining trends over the last decade. These findings offer valuable insights into ABC City's rainfall patterns, aiding in climatic data understanding and management.

Data Preparation

In this section, we meticulously prepare the Daily Rainfall Climate Data sourced from the Australian Government's Bureau of Meteorology for ABC City. This comprehensive preparation involves several critical steps aimed at ensuring the reliability and integrity of the dataset.

Error 1: Data Entry Errors -

Data entry errors are common in datasets and can significantly impact the reliability of analyses. In our dataset, we identified several data entry errors that required correction:

- Error 1.1: Incorrect Year Entry: One common error is the inclusion of future year values, such as '2027' instead of '2017'. To rectify this, we utilize the replace() function in pandas, correcting the entries accordingly.
- Error 1.2: Incorrect Month Entry: Non-numeric entries like 'Jan' or 'April' in the month column are converted to their respective numerical representations using the replace() function.
- Error 1.3: Incorrect Day Entry: Invalid day values, such as '48', 'nine', or '200', are identified and replaced with appropriate alternatives ('48' with '10', '200' with '20', 'nine' with '9').

Error 2: Impossible Values -

Negative values: We identified negative values in the 'Rainfall_mm' column, which are physically impossible for rainfall measurements. To rectify this, negative values were replaced with the mean rainfall amount using the mean() function, ensuring data integrity.
Excessive Rainfall Values: Instances of rainfall values exceeding 100000 mm were found. A threshold of 100 mm was established based on the second largest value. Values exceeding this threshold were replaced with the mean rainfall value to mitigate outliers.

Error 3: Missing Values -

In our dataset, missing values in the 'month', 'day', and 'Rainfall_mm' columns pose potential challenges, risking biases and inaccuracies in analysis. Addressing these gaps required a systematic approach:

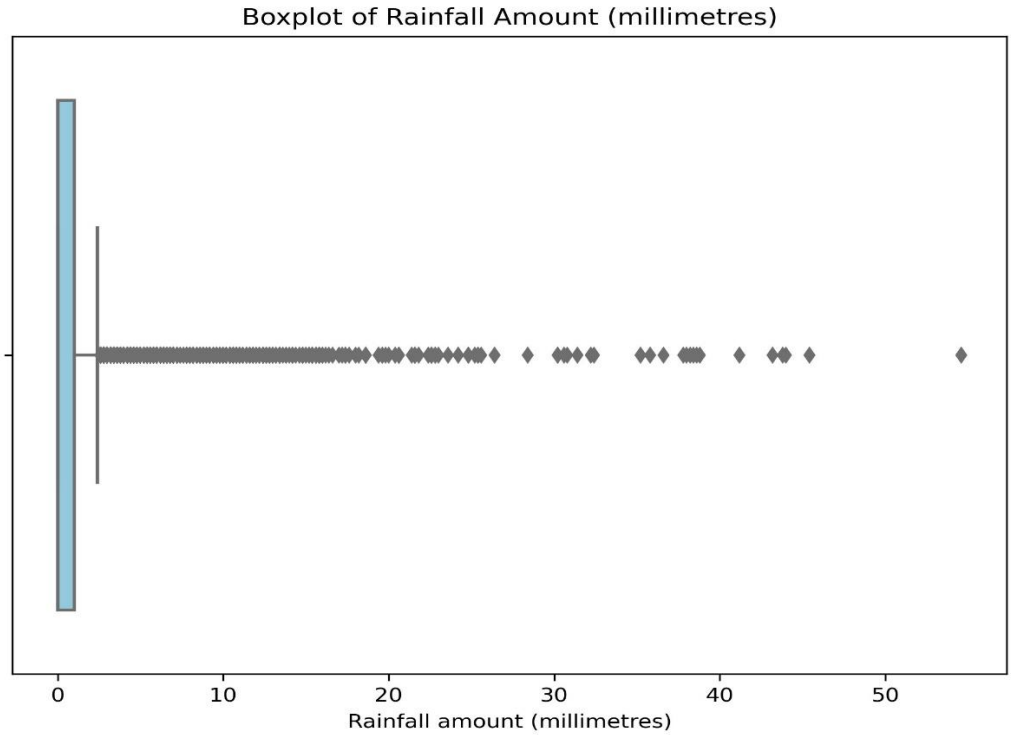
- 'month' Column: Upon scrutiny, we detected a missing value at index 2067. Given the sequential pattern of surrounding records, we reasonably imputed '1' (January).
- 'day' Column: Similarly, at index 1736, a missing value was evident. Inferred from adjacent data points, we filled this gap with '4'.
- 'Rainfall_mm' Column: Before handling missing values, we assessed the distribution of rainfall amounts. Considering the skewness, we opted to impute missing values with the median to maintain distribution integrity and counter outlier effects.

Error 4: Data Formatting -

Data formatting ensures that data types are appropriate for each column, enhancing data consistency and analysis efficiency. In our dataset, we converted the 'month' and 'day' columns to integer format to ensure uniformity and compatibility with other numerical operations.

Error 5: Handling Outliers -

Outliers in datasets, indicating potential errors or extreme values, require attention for accurate analysis. In our focus on the 'Rainfall_mm' column, we employ boxplots to detect outliers visually.



Interpretation of the boxplot highlights numerous outlier records, primarily due to the prevalence of lower rainfall amounts compared to higher ones. Further analysis and treatment of these outliers may be necessary based on specific analysis objectives.

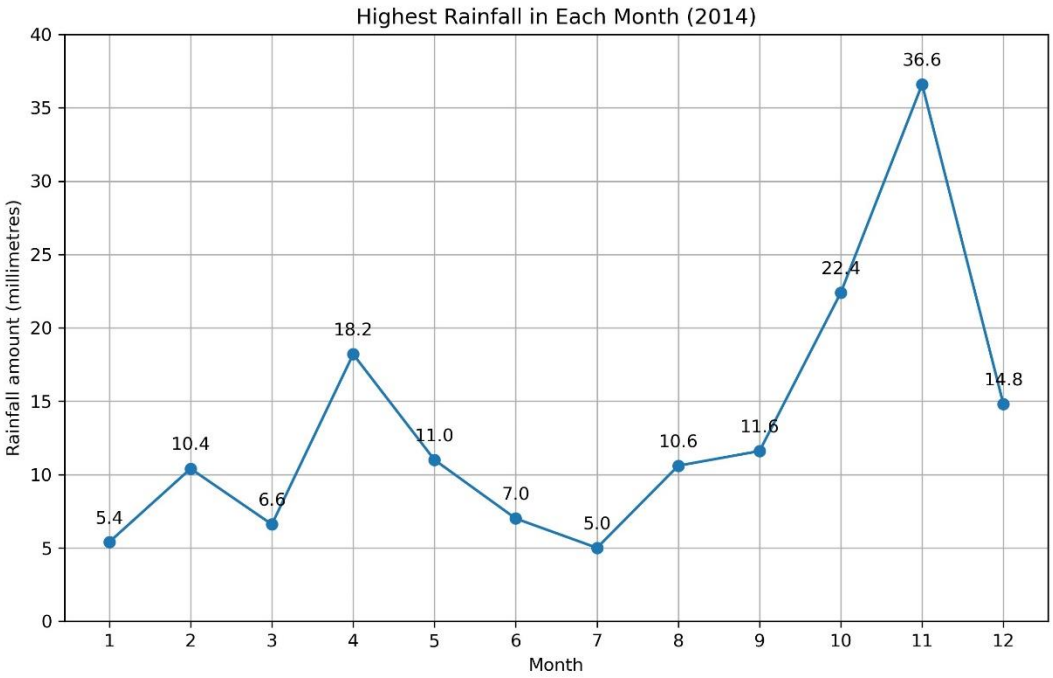
Subsequently, ensuring dataset cleanliness and consistency, we write the cleaned data into a CSV file named 'cleaned version.csv'.

Data Exploration

In this section, we explore the Daily Rainfall Climate Data to gain insights into rainfall patterns in ABC City. We analyse yearly and monthly rainfall patterns, identify trends, and compare rainfall across different years to discern any significant variations.

Task 2.1 Exploring highest daily rainfall in each month of 2014.

To achieve this, we filtered the dataset for year 2014. Create a pivot table with days as rows and months as columns and then get the max per each month. Next, we look at the descriptive statistics and plot the highest daily rainfall in each month. The plot is shown below,

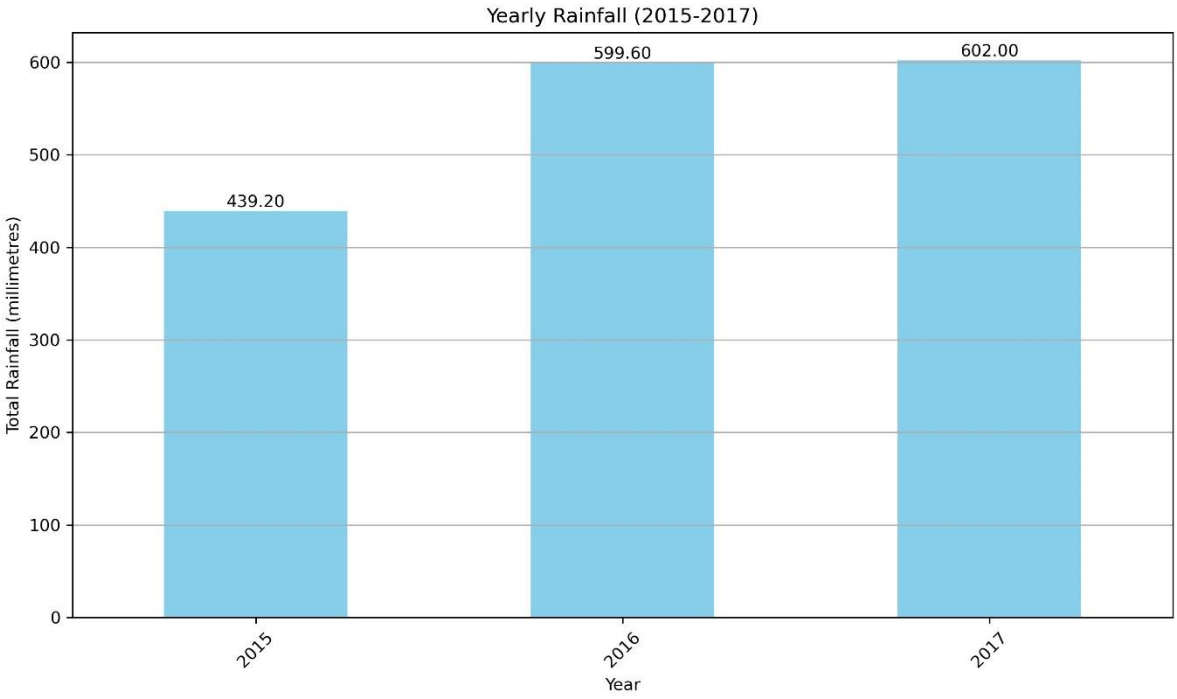


Interpretation:

The plot clearly illustrates that the peak daily rainfall per month occurred predominantly during the final three months of the year 2014. July exhibited the lowest rainfall, while November experienced the highest rainfall.

Task 2.2 Analysing Yearly/Monthly Rainfall Patterns (2015 – 2017)

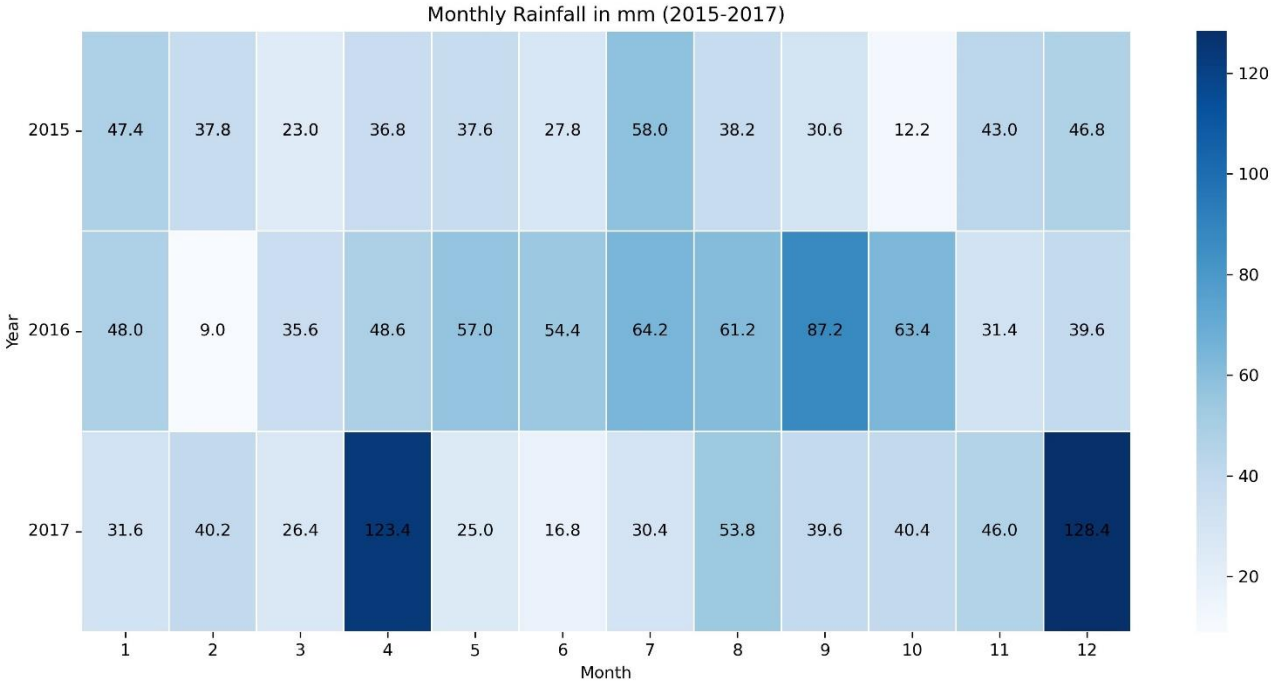
Next, we shift our focus to analysing Yearly and monthly rainfall patterns over the years 2015 to 2017, to understand seasonal variations and trends in precipitation levels. Let's look at the total yearly rainfall amounts across 2015 and 2017,

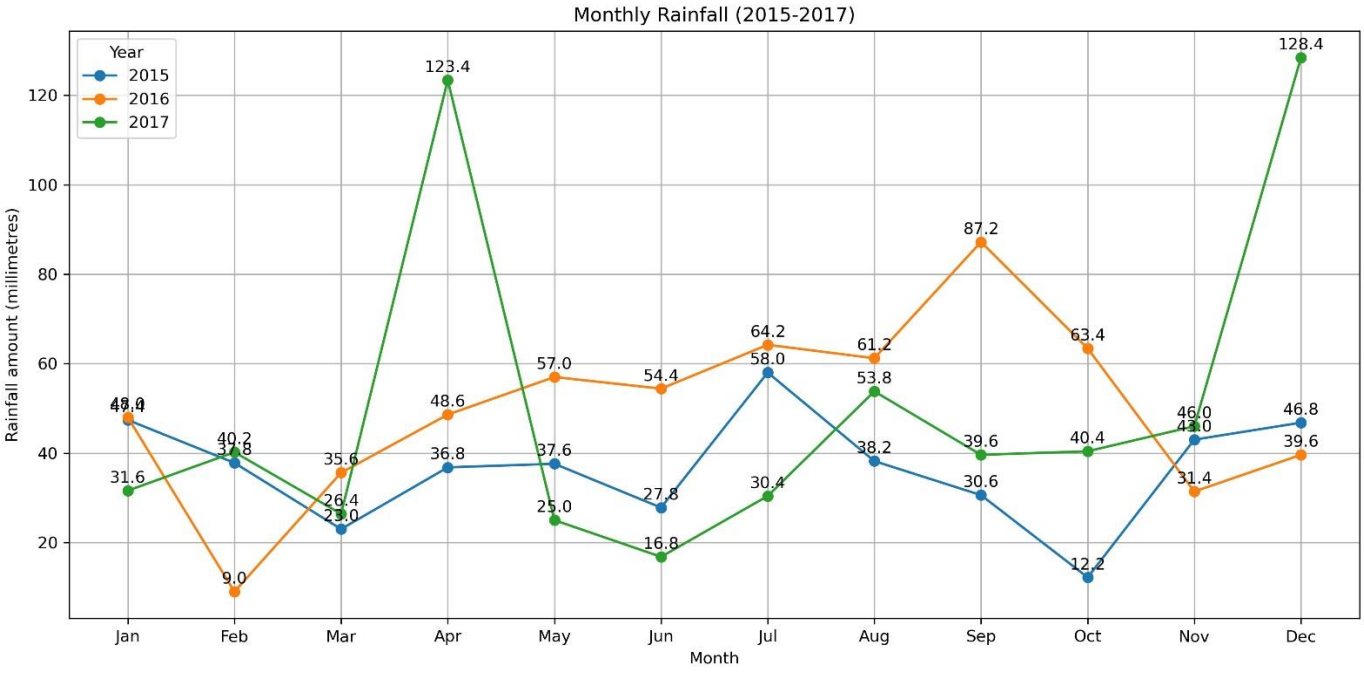


Interpretation:

Based on the bar chart and the descriptive statistics, in 2017, the total rainfall for the year was the highest at 602 mm, just slightly more than the 599 mm recorded in 2016. In contrast, 2015 had a lower rainfall total of only 439 mm. The average rainfall over the 3 years was 547 mm.

Next, let's look at the monthly rainfall amounts across 2015 and 2017. We used heatmaps and trend charts,





Interpretation:

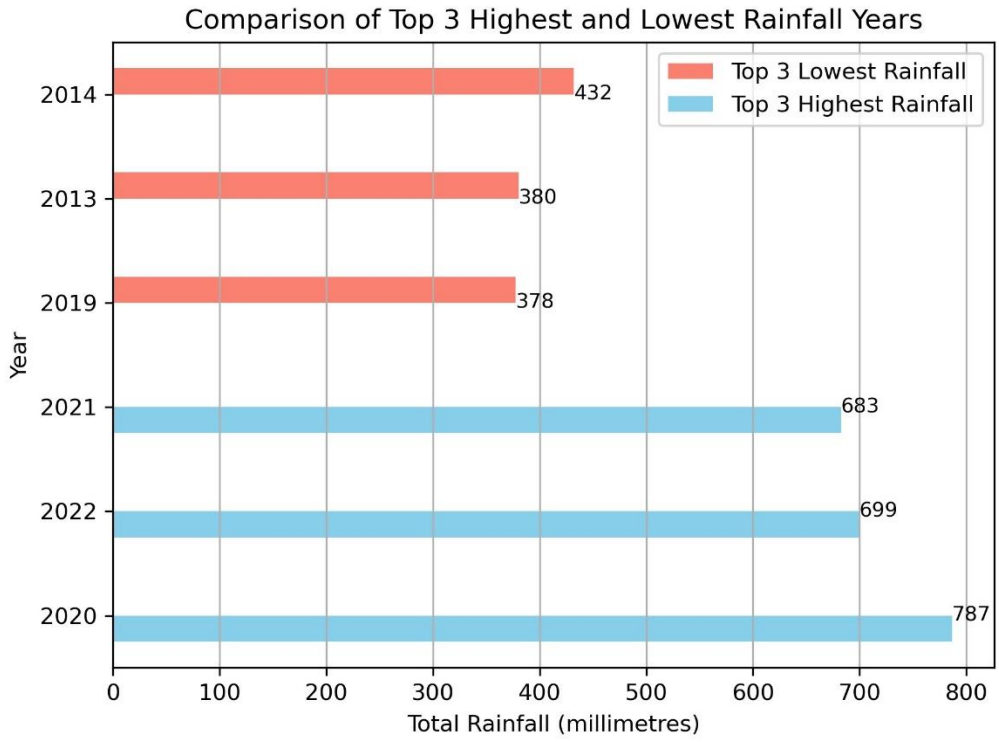
Based on the heatmap, trend charts and descriptive statistics,

- December 2017 recorded the highest monthly rainfall, reaching 128 mm, followed closely by April of the same year with 123 mm.
- Conversely, February 2016 experienced the lowest rainfall, registering only 9 mm.
- The average monthly rainfall over the 3 years was 45 mm.

Task 2.3 Top 3 highest and lowest yearly rainfall amounts

To find the top 3 years with the highest and lowest rainfall amounts and compare them, we will :

- Group the data by year and calculate the total rainfall for each year.
- Sort the years based on rainfall amounts to find the top 3 highest and lowest rainfall years.
- Perform descriptive analysis of the rainfall amounts for the top and bottom years.
- Visualize the comparison using bar plots.



Interpretation:

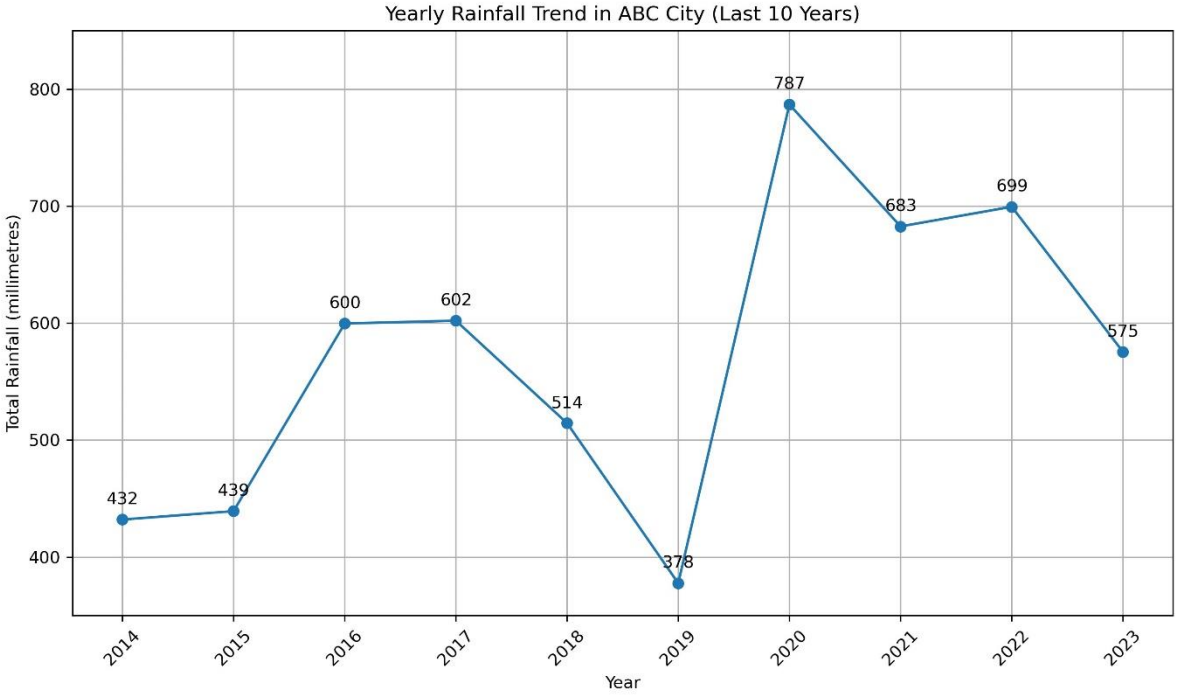
Based on the horizontal bar chart and descriptive statistics,

- The years 2019, 2013, and 2014 exhibited the least total rainfall, measuring 378 mm, 380 mm, and 432 mm, respectively.
- Conversely, the years 2020, 2022, and 2021 recorded the highest rainfall amounts, reaching 787 mm, 699 mm, and 683 mm, respectively.
- The average highest rainfall among the top three years stands at 722 mm, while the average lowest rainfall among the bottom three years stands at 396 mm. mm.

Task 2.4: Analysing Seasonal Trends over last 10 years

In this task, to explore the changes in rainfall in ABC City over the last 10 years, lets first filter the data to include only the last 10 years. Group the data by year and calculate the total rainfall for each year. Visualize the yearly rainfall

trend over the last 10 years.



Interpretation:

Based on the trend chart and descriptive statistics,

- Throughout the 10-year period, the maximum annual rainfall occurred in 2020, reaching 787 mm, while the minimum was observed in 2019, at 378 mm.
- The variation in total annual rainfall over the 10 years does not exhibit a discernible pattern.
- On average, the yearly rainfall over the decade amounted to 570 mm.
- The spread of data around the mean was measured to be 131 mm.

References

- Australian Government Bureau of Meteorology (2022). State of the Climate 2022, BoM Website, accessed 23rd Apr 2024. http://www.bom.gov.au/state-of-the-climate/