Student Name: Ramneek Kaur Riar Student ID: 4075995

Data Preparation

In this section, we will be looking at how the data was prepared to provide efficient data exploration. The dataset used for this assignment was given as a CSV (comma-separated values) file containing a total of 3865 records and the following 4 attributes: Year, Month, Day, and Rainfall amount (millimetres). To begin with the data preparation, verification was performed to confirm whether the loaded data was equivalent to the data in the source CSV file. This was done by loading the data twice into separate data frames and then compared with one another to ensure the data was consistent.

Error 1: Data Types

The first error encountered were the inconsistent data types found in the different columns. Attribute 'Month' and 'Day' were found to have the data type as object instead of a numerical data type such as integer or a float. To accurately visualize and perform calculations on this data set, it was important to make this interval data be in a numerical format. The data in those columns was converted to a numerical format using the pandas to_numeric function with the errors coerced (NumFOCUS, Inc., 2024b). This meant that any errors found within the data in the column that does not align with being numerical, would be replaced with the NaN (Not a Number) value. The data types of all the columns were checked again and all the columns returned numerical data types as expected. As 'Month' and 'Day' are interval data types, the columns were then further converted to type int to make for accurate analysis and exploration.

Error 2: NaN (Not a Number) Values

Throughout cleaning the data set, it was necessary to resolve the NaN values within the data that may have existed prior and after converting the data to numerical values. After checking for the number of NaN values, it was confirmed that the 'Month' column contained 2 NaN values, the 'Day' column contained 3 NaN values, while 'Rainfall amount (millimetres) contained 5 NaN values. To ensure that the data used within our data exploration was consistent and accurate, the rows containing these values were dropped as to not skew the rest of the data. After dropping these rows, it was confirmed that no more NaN values were present in the data.

Error 3: Drop Duplicates

Another safeguarding technique used was to drop any duplicate data that may exist within the file. This means that any rows which contain identical information, such as the rainfall which occurred on the same day in the same month. Having duplicate data can skew the data when it comes to exploring, analysing, and drawing conclusions from the data so it is beneficial to remove it.

Error 4: Typos and Outliers

Upon reviewing the data manually, a typo was found in the 'Year' column where 2017 was written as 2027 in a row of data. Given the data surrounding this row was all for the year of 2017, this typo was safely replaced in place with the year 2017 to ensure accuracy of the data. Another error found through looking at the data was an outlier where the recorded rainfall in millimetres was unusually high at 100,000 mm. This was significantly different to the surrounding rainfall pattern and would skew the data. As a result, the data here was replaced with the value of 0 in place.

Error 5: Missing Data for Year 2013

The final error observed was related to the data found for the year 2013, which only had information recorded from June 2nd till the end of the year. This was unlike the rest of the years within the dataset which had rainfall information present for the entire year. This discrepancy if not addressed, would pose a significant problem since it would provide an inaccurate exploration and visualization of the data with only half of the data for 2013 being taken into consideration. According to the Australian Bureau of Meteorology, the average total rainfall in 2013 was 428 millimeters (ABC News, 2014). To allow for accurate data representation, when creating sample data for 2013 this was the figure used to be the total rainfall sum. Since the data in 2014 is the closest data available, the data from January 1st, 2014, to June 1st, 2014, was used as a guide for when rainfall could potentially occur.

First, the total rainfall that occurred in the latter half of 2013 was calculated, along with the additional rainfall that was required to reach a total sum of 428 millimeters. Then, the data from the beginning half of 2014 was retrieved. The scaling factor was calculated by dividing the additional rainfall required to reach a sum of 428 millimeters by the total rainfall that occurred in the first half of 2014. Then, the values in the first half of 2014 were multiplied by this scaling factor and thus scaled down. All these values were rounded to one decimal point to ensure that it followed the same pattern as the rest of the values under the 'Rainfall amount (millimeters)' column. Lastly, the year value was updated to be 2013 and the new, modified data was concatenated with the main data set so that a thorough and more accurate analysis of the data could follow. The clean data set was saved into the file "cleaned_version.csv" and was ready to be used for analysis.

Data Exploration

Task 2.1

For the first task as part of the data exploration, the data for the year 2014 was asked to be converted into a pandas DataFrame with the rows being the days in a month, and the columns as the months in the year. After all the data within the year 2014 was obtained, the pivot_table function was used to "pivot" the rows and columns, essentially reshaping the data (NumFOCUS, Inc., 2024b). The number of days equals the number of rows within the data frame and the months equals the number of columns. This data frame will be filled with the amount of rainfall in millimetres occurring at the given day in the given month. After this,

the values for the maximum amount of rainfall that occurred in each month of 2014 were obtained. As shown in Figure 2.1, the month of November in 2014 had the highest amount of rainfall at 36.6mm, while July had the least amount of rainfall at its maximum of 5.0mm.

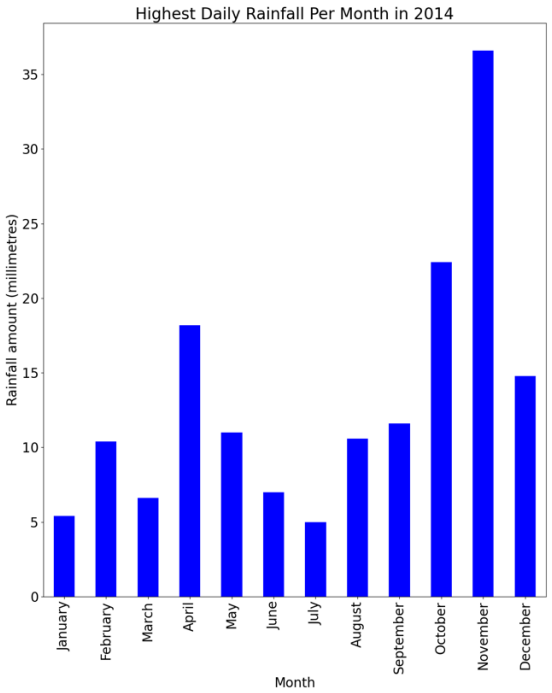

Figure 2.1: Highest Daily Rainfall Per Month in 2014

Task 2.2

In this task, the data between 2015 and 2017 was explored and analysed on a yearly and monthly basis. To do this, the data between 2015 and 2017 was retrieved and stored in a data frame called 'data_btwn_2015_to_2017'. The yearly analysis was performed by taking a sum of the total rainfall that occurred within a year, and then plotting this information within a bar graph. As shown below in Figure 2.2.1, the range of rainfall occurring every year within years 2015 to 2017 fell between 400mm to 600m. The year 2016 accumulated the most rainfall at 599.6mm, whereas the year 2015 accumulated the least amount of rainfall at 439.2mm.

After, the data between 2015 and 2017 was examined on a monthly basis. This analysis was performed by taking a sum of the total rainfall that occurred within each month throughout the years. This helps to recognize seasonal patterns of rainfall, showing a pattern of which months are more likely to accumulate rainfall versus others. This information was then plotted onto a bar graph. As shown in Figure 2.2.2, December had the greatest amount of rainfall at 214.8mm, while March had the least at 85.0mm. This information follows the seasonal structure within an average city in Australia, where winters are most likely to experience high rainfall, and summers are relatively dry.

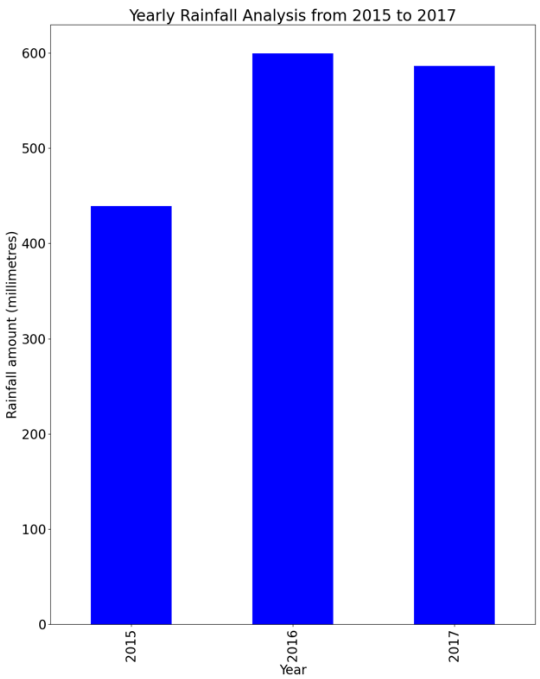

Figure 2.2.1: Yearly Rainfall Analysis from 2015 to 2017

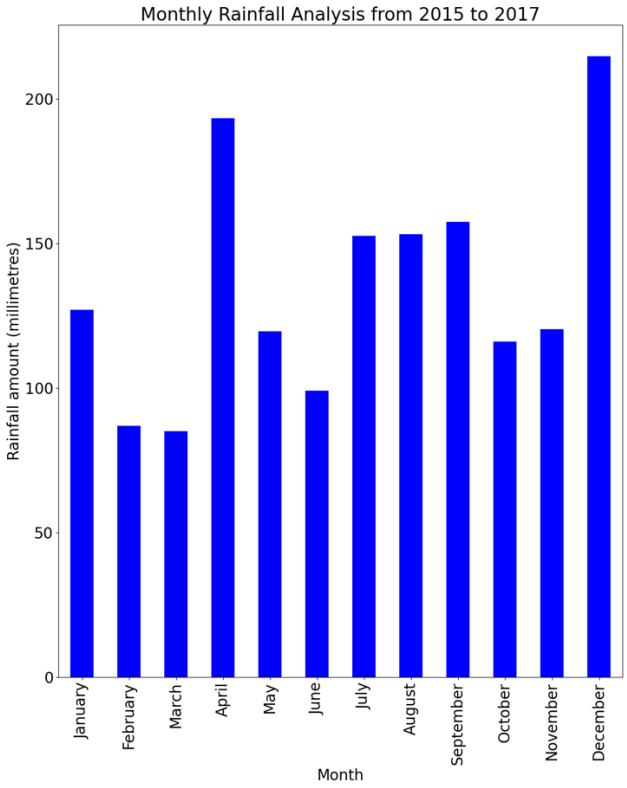

Figure 2.2.2: Monthly Rainfall Analysis from 2015 to 2017

Task 2.3

In this section, we explore and analyse the data to determine the top 3 years with the highest rainfall amount and compare that with the top 3 years with the lowest rainfall amount. The data was grouped by year and then a total sum of rainfall amount was taken for each year. This data frame called 'yearly_rainfall' was then sorted by rainfall amount in descending order. Using the head function, the top 3 years with the highest rainfall amount were retrieved, while using the tail function aided in retrieving the top 3 years with the lowest rainfall amount. From these results, it is evident that the range of rainfall sits between 432.0mm and 786.8mm, indicating that in ABC City there has always been periods of rainfall. The highest rainfall amounts were recorded in years 2020-2022, meanwhile the lowest rainfall amounts were recorded in past years such as 2013, 2014 and 2019. Figure 2.3 displays the comparison of these results, with the red bars indicating the years with the least rainfall, and the blue bars with the highest rainfall. By analysing the graph, the rainfall amount has a general upward trend, suggesting that rainfall amount has relatively increased over time considering the years 2020-2022 contain the highest amounts of rainfall.

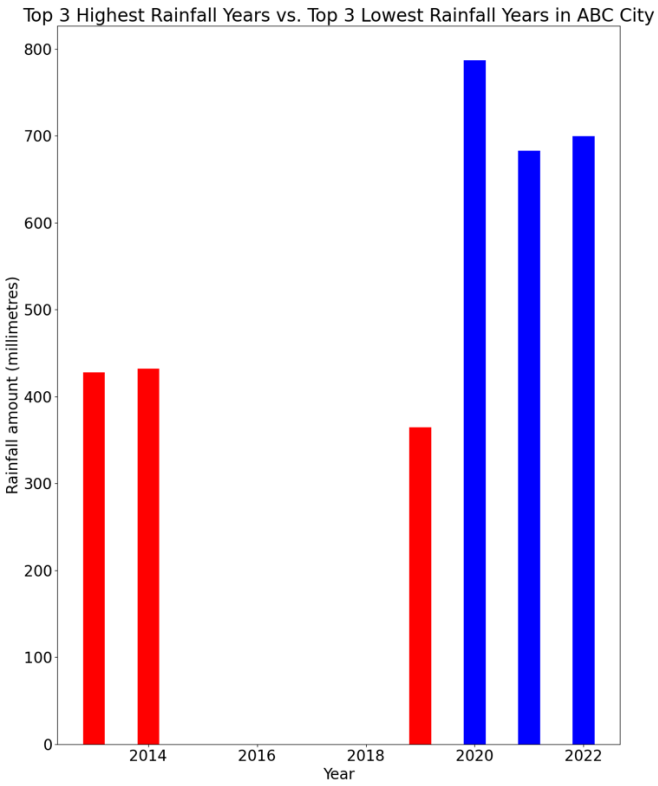

Figure 2.3: Top 3 Highest Rainfall Years vs. Top 3 Lowest Rainfall Years in ABC City

Task 2.4

In this last section, the changes of rainfall within ABC City in the last 10 years were explored and plotted onto a graph. The data frame called 'yearly_analysis' previously created was used for this analysis, sorting it by year in ascending order to accurately view the changes overtime. After gathering the data, this was plotted onto a line graph to visualize the drastic

changes and trends that may have occurred. Figure 2.4 displays a sharp incline in rainfall from 2015 to 2016, and then an overall downward trend until 2019. This changes in 2020 where a sharp incline is noticed and hits a peak again, following the same downward trend we noticed earlier. If the data were to be extrapolated beyond the past 10 years and into the future, a similar pattern is assumed where after the downward trend another sharp incline appears reaching a peak before repeating the downward pattern in rainfall amount.

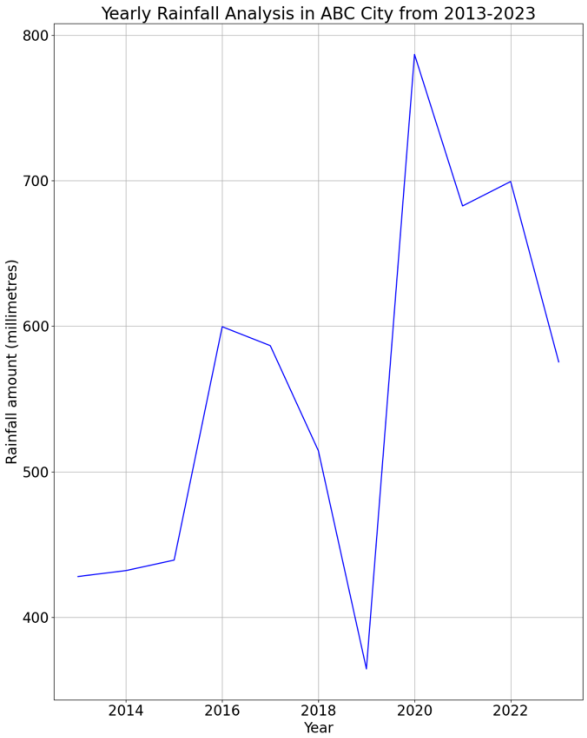

Figure 2.4: Yearly Rainfall Analysis in ABC City from 2013-2023

References

ABC News. (2014, January 3). Heat maps: Australia's Hottest Year explained. https://www.abc.net.au/news/2014-01-03/australia27s-climate-in-2013/5183378

NumFOCUS, Inc. (2024a). Pandas.pivot_table#. pandas.pivot_table - pandas 2.2.2 documentation. https://pandas.pydata.org/docs/reference/api/pandas.pivot_table.html

NumFOCUS, Inc. (2024b). *Pandas.to_numeric#*. pandas.to_numeric - pandas 2.2.2 documentation. https://pandas.pydata.org/docs/reference/api/pandas.to_numeric.html