Assignment – 1: Practical Data Science

Student Name: Athul Varghese Thampan
Student ID: s3958556

Data Preparation

To begin, I imported all the necessary libraries required such as pandas and matplotlib. Then, I loaded the CSV file into a variable called ‘daily_rainfall’ and verified the data types of each column.

Next, I examined each attribute to identify potential errors in entries. Once I detected the errors, I either manually corrected them or I removed the rows containing impossible values. By removing impossible and missing values rather than replacing them, I avoid causing any inaccurate estimations.

I also verified that only months with 31 days actually contained 31 days in the dataset and ensured that February included 29 days only in leap years. This check was conducted to prevent incorrect entries of days for each month.

Error 1: Missing values.

I checked for any null values in the dataset, and it returned true. Therefore, I calculated the total number of null values in each column. I removed the rows containing null values and then rechecked to ensure they were removed.

Error 2: Impossible values.

Upon reviewing the unique years in the dataset, I noticed one entry for the year 2027. Since the dataset should contain rainfall data from the years 2013 to 2023, I removed the row with the year 2027 as it was invalid.

I performed a similar check for the ‘days’ column and discovered values such as 48 and 200, which are impossible. Consequently, I deleted the rows with these values.

For rainfall, I examined whether any values were below 0 (negative rainfall is impossible) or above 1500 (the highest rainfall recorded in Australia was 1,147mm [Source: Australia Weather News]). I identified the entries 100000.0 and -10.0, both of which were impossible. Therefore, I removed the rows containing these invalid values.

Error 3: Mistakes during data entry.

When inspecting the unique months, I noticed two entries labelled ‘April’ and ‘Jan’, which differed from the rest of the entries that were stored as numbers. To maintain consistency, I replaced ‘April’ with ‘4’ and ‘Jan’ with ‘1’.

Similarly, upon reviewing the unique days, I found an entry labelled ‘nine’, which was inconsistent with the other entries that were numerical. I corrected this by replacing ‘nine’ with ‘9’.

Error 4: Incorrect datatypes.

While inspecting the datatypes of each attribute, I noticed that the columns 'Day' and 'Month' were stored as strings, although their values were numerical. Therefore, I converted them to integer values after correcting other errors. This was done to ensure that the days and months are displayed in numerical order.

Data Exploration

Task 2.1

I first created a data frame containing just the rainfall of the year 2014. Then I grouped it by Day and Month to ensure that the rows are days and columns are months. Then I replaced all null values with 0 to account of varying month lengths. For example, February in 2014 does not contain 29th, 30th and 31st. Similarly, April contains only 30 days and so on. Then I calculated the maximum rainfall and plotted a bar graph to visualise the maximum daily rainfall in each month of 2014.

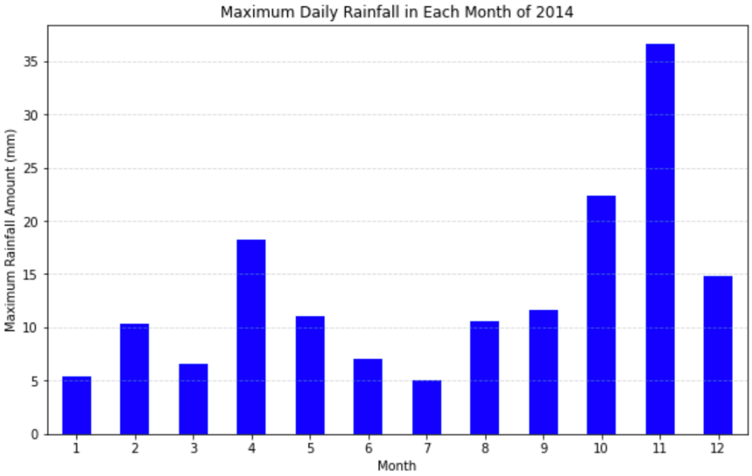

Figure 1

From the bar graph (Figure 1), we can see that November had the highest daily rainfall (~37 mm) in 2014, while July had the least maximum daily rainfall (~5 mm) in the same year.

Task 2.2

For this task, I created a data frame containing rainfall for the years 2015 to 2017. Then I grouped the data by year for the yearly analysis. I plotted the bar graph for the total annual rainfall for the years 2015 to 2017.

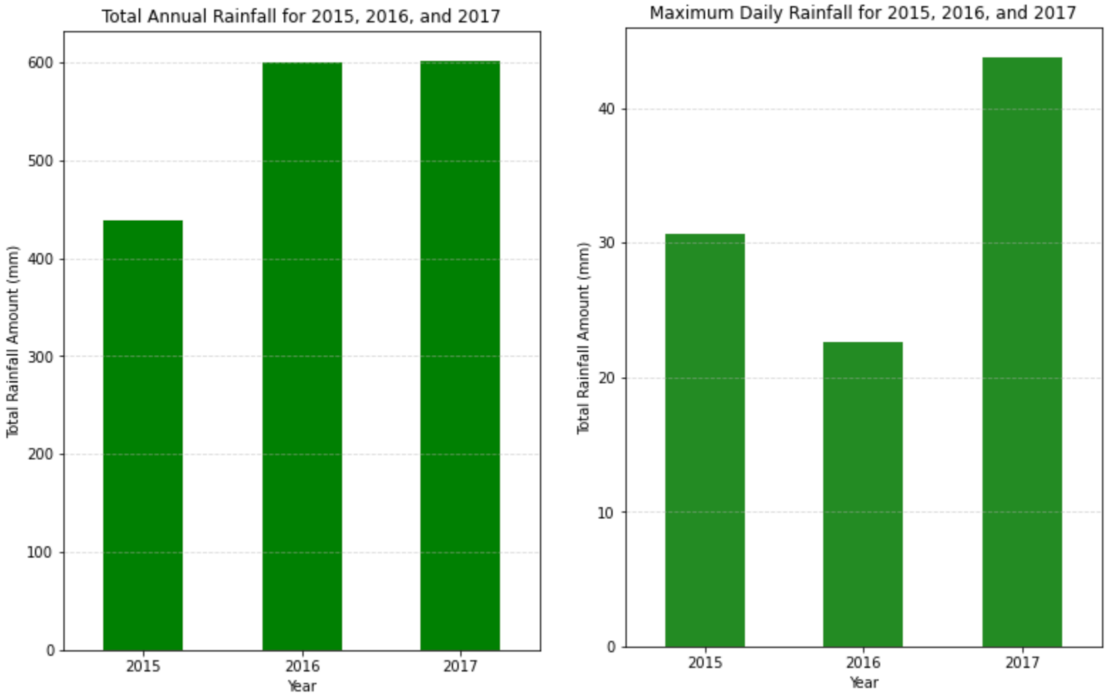

Figure 2

Figure 3

From the bar graph in Figure 2, we can see that 2015 had the lowest amount of rainfall (~440 mm), while both 2016 and 2017 had the same amount of rainfall (~600 mm).

Next, I created a data frame containing maximum daily rainfall for the years 2015 to 2017 and plotted it. From the resulting graph (Figure 3), it is apparent that although 2016 and 2017 had the highest total rainfall, 2016 experienced the lowest maximum daily rainfall (~22 mm), with 2017 having the highest (~45 mm).

Then, I created a data frame containing the daily rainfall of the years 2015 to 2017 but omitted days with 0.0 mm of rain (i.e., days with no rain) so that I can find the mean rainfall during these years.

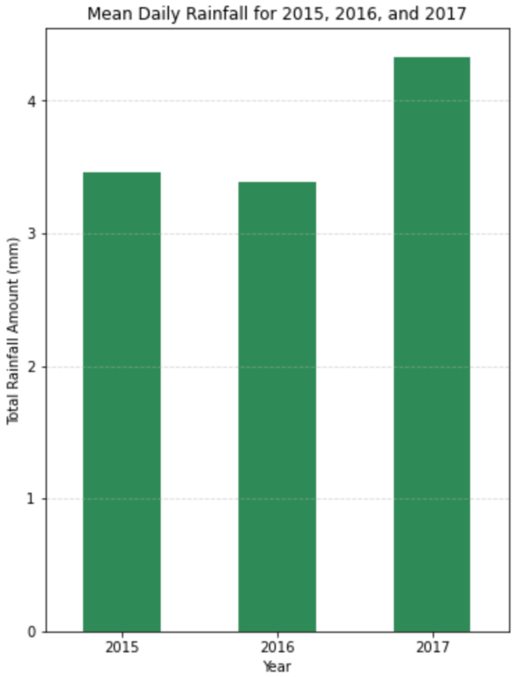

Figure 4

The plotted graph (Figure 4) indicates that 2017 had the highest mean daily rainfall, while 2016 had the lowest mean daily rainfall.

Analysing these three graphs collectively suggests that 2017 was the rainiest year among the selected years.

Following this, I plotted the graphs for monthly analysis. First, I plotted the graph for total monthly rainfall for the years from 2015 to 2017.

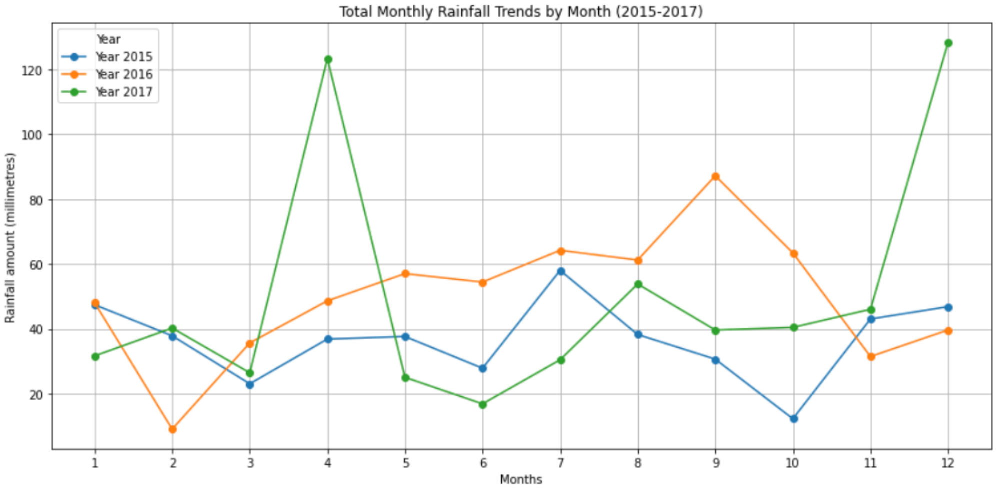

Figure 5

The line graph (Figure 5) highlights April and December of 2017 as the months with the highest rainfall, while July was the wettest month in 2015 and September in 2016. Conversely, February, October, and June experienced the least rainfall in 2016, 2015, and 2017, respectively.
Next, I plotted a graph showcasing maximum daily rainfall by month for the years 2015 to 2017.

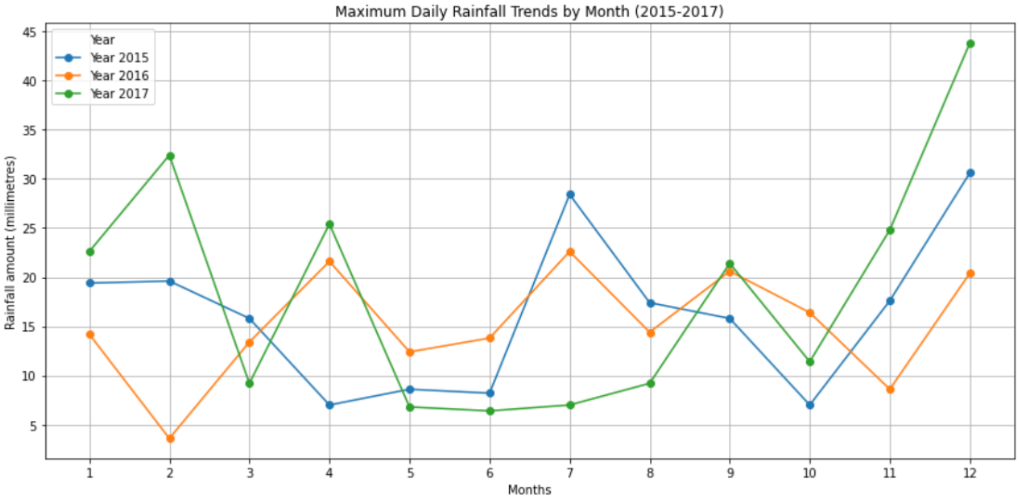

Figure 6

From Figure 6, we can see that the highest maximum daily rainfall for the years 2015 and 2017 was in December at ~31 mm and ~44 mm respectively. For 2016, the highest daily rainfall was in July at ~23 mm. The driest month was February of 2016 with ~3 mm of rainfall.

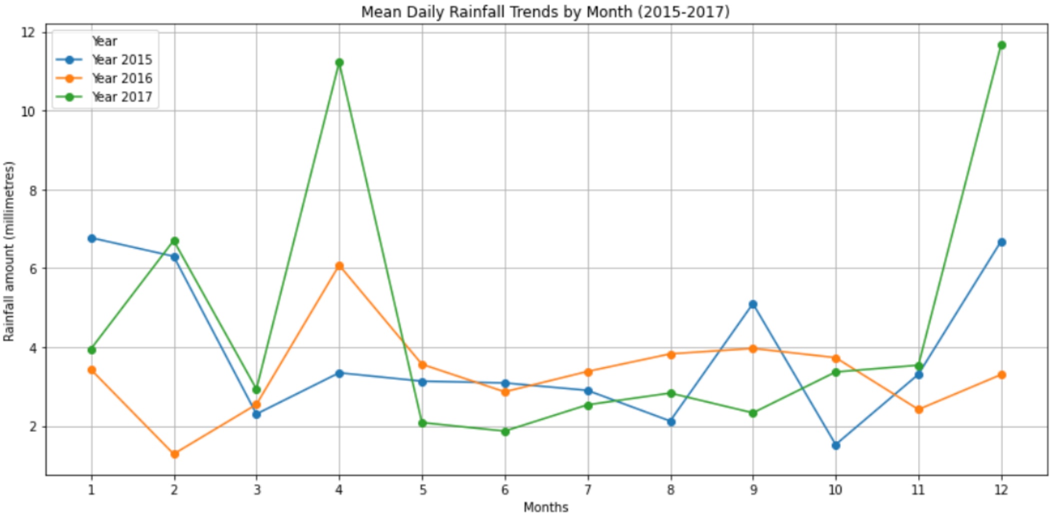

Figure 7

Similarly, I created a graph (Figure 7) for the mean daily rainfall and was able to find that 2017 had the highest mean daily rainfall and 2016 had the least mean daily rainfall. I avoided plotting a graph for minimum daily rainfall as the minimal amounts would be negligible and would not provide any significant insights.

Task 2.3

For this task, I created a data frame grouped by the years. Then I created one with top three years with most rainfall and another with bottom three years with least rainfall. I concatenated these into a single data frame, and I plotted the results as follows.

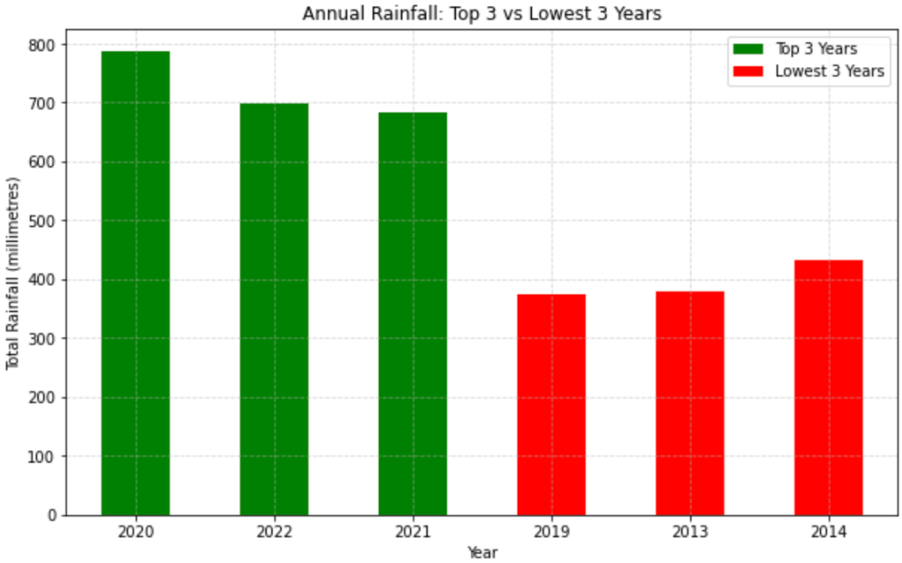

Figure 8

The graph (Figure 8) illustrates that 2021, 2022 and 2020 had the highest rainfall in that order. On the other hand, 2014, 2013 and 2019 had the least rainfall in that order. The year with the most rainfall was 2020 with ~790 mm, while 2019 had the least rainfall which was ~380 mm.

Task 2.4

For this task, I grouped the data by year and found the total annual rainfall which I have plotted as shown below.

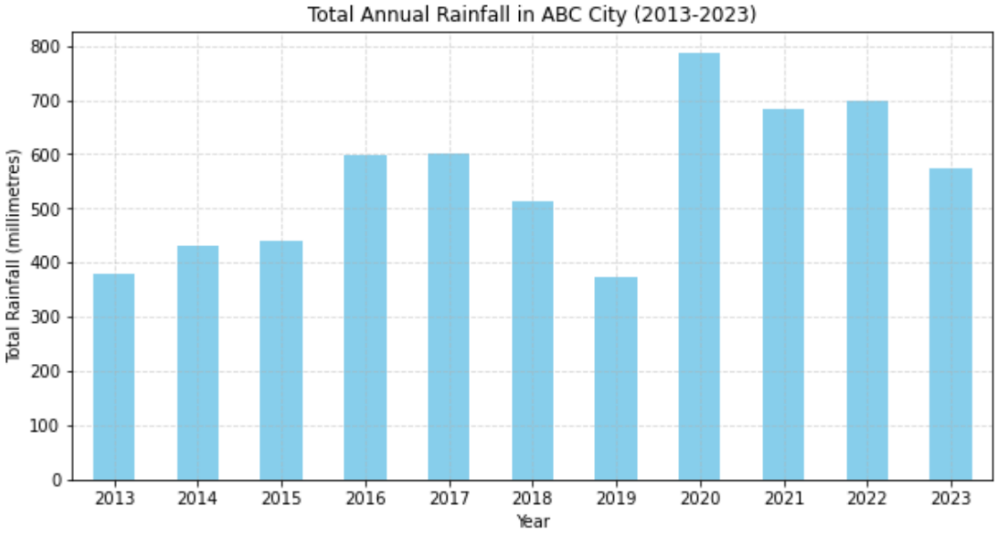

Figure 9

From Figure 9, we can see that 2020 had the highest annual rainfall with ~790 mm of rainfall, while 2019 had the least annual rainfall with ~380 mm. We can also see from the graph that the amount of rainfall drastically increased from 2019 to 2020.

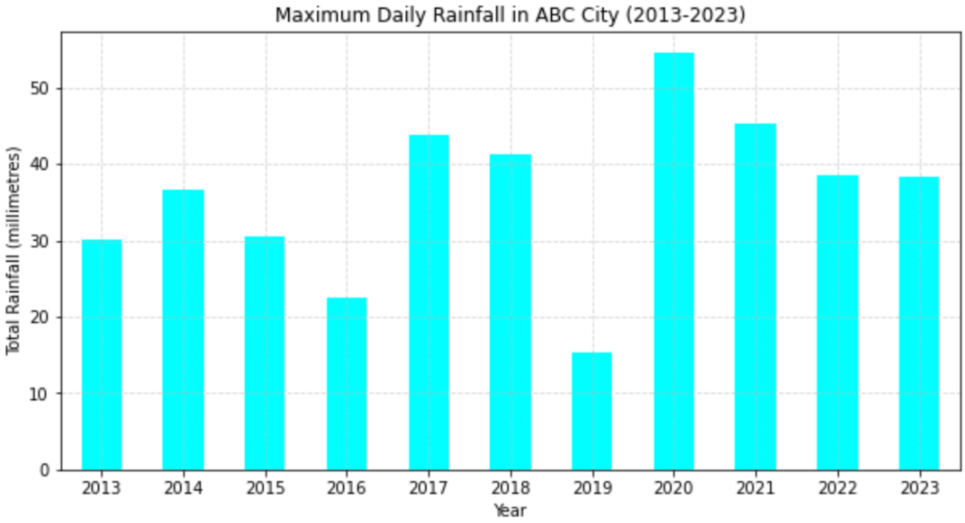

Figure 10

From Figure 10, we can see that the maximum daily rainfall was the highest in 2020 with ~55 mm of rain. The least maximum daily rainfall was in 2019 with ~15 mm.

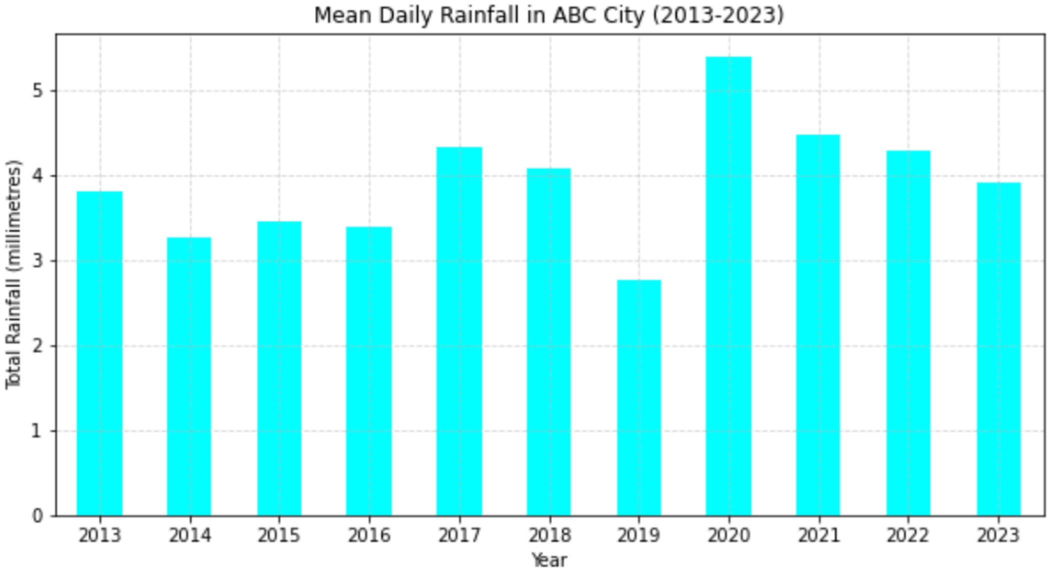

Figure 11

From Figure 11, we can see that the highest mean daily rainfall was in the year 2020 with ~5.5 mm of rain. The least was in 2019 with ~2.7 mm of rain. I did not calculate the minimum daily rainfall since it would be a negligible amount.

Overall, the analysis reveals that 2019 stood out as the driest year, whereas 2020 emerged as the wettest, despite being consecutive years.

Bibliography

1. Australia Weather News (no date) City's Wettest Days. Available at: https://www.willyweather.com.au/news/5909/city's+wettest+days.html#:~:text=The%20state%20record%20rainfall%2024,on%204th%20of%20January%201979 (Accessed: 17 April 2024)