Data analysis on Dengue and Climate in Singapore

By Li Yongjie (2342377)

DAAA/FT/1B/01

Datasets

Rainfall - Monthly Total

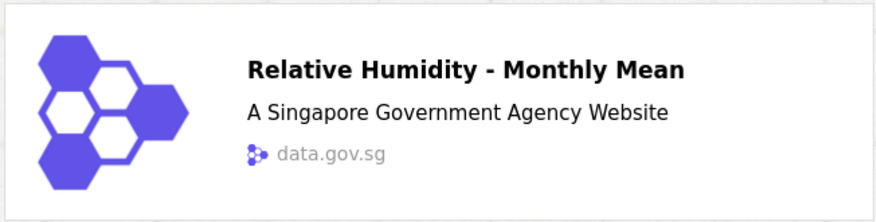
https://beta.data.gov.sg/collections/1399/view

Weekly number of Dengue and Dengue Haemorrhagic Fever Cases

https://beta.data.gov.sg/collections/509/view

Relative Humidity - Monthly Mean

https://beta.data.gov.sg/collections/1404/view



Surface Air Temperature - Monthly Mean

https://beta.data.gov.sg/collections/1419/view

Information on Dataset

Rainfall - Monthly Total

510 Rows, 2 Columns

Columns:

- month (object), no NA values, 501 unique values
- total_rainfall (float64), no NA values, 467 unique values

Relative Humidity - Monthly Mean

510 Rows, 2 Columns

Columns:

- month (object), no NA values, 501 unique values
- mean_rh(float64), no NA values, 142 unique values

Weekly number of Dengue and Dengue Haemorrhagic Fever Cases

530 Rows, 4 Columns

Columns:

- year (int64), no NA values, 5 unique values
- eweek (int64), no NA values, 53 unique values
- type_dengue (object), no NA values, 2 unique values
- number (float64), 8 NA values, 169 unique values

Surface Air Temperature - Monthly Mean

510 Rows, 2 Columns

Columns:

- month (object), no NA values, 501 unique values
- mean_temp (float64), no NA values, 42 unique values

Initialising data set and libraries

```python
# Importing Pandas, Matplot and Numpy
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import calendar

# Import calendar to get name of months
import calendar

# Loading the Data
humid_df = pd.read_csv('./Healthcare-Dataset/RelativeHumidityMonthlyMean.csv')
rain_df = pd.read_csv('./Healthcare-Dataset/RainfallMonthlyTotal.csv')
temp_df = pd.read_csv('./Healthcare-Dataset/SurfaceAirTemperatureMonthlyMean.csv')
dengue_df = pd.read_csv('./Healthcare-Dataset/WeeklyNumberofDengueandDengueHaemorrhagicFeverCases.csv')
```

✓ 0.0s Python

```python
# See available styles for
print(plt.style.available)
```

✓ 0.0s Python

```
['Solarize_Light2', '_classic_test_patch', '_mpl-gallery', '_mpl-gallery-nogrid', 'bmh', 'classic', 'dark_background', 'fast', 'fivethirtyeight', 'ggplo
```

```python
# Set the style for plots
plt.style.use('seaborn-v0_8-darkgrid')
```

✓ 0.0s Python

Data Manipulation (Climate)

Based on our EDA the 3 climate-related datasets all have the same number of rows and share the same column 'month' we can easily merge them.

As seen from the range of years, the dengue dataset only has the years containing 2014-2018, hence we will have to clean our other datasets as well to contain only the range of 2014-2018

```
# Merging the datasets
humid_temp_df = pd.merge(humid_df, temp_df, on = 'month')
rain_humid_temp_df = pd.merge(humid_temp_df, rain_df, on='month')

# Subsetting the range to 2014-2018
index_of_2014_01 = (rain_humid_temp_df.loc[rain_humid_temp_df['month'] == '2014-01'].index[0])
index_of_2018_12 = (rain_humid_temp_df.loc[rain_humid_temp_df['month'] == '2018-12'].index[0])

# Create new data frame called climate_df
climate_df = rain_humid_temp_df.loc[index_of_2014_01:index_of_2018_12].reset_index(drop=True)
```

✓ 0.0s

```
# First and last month of climate_df
analysis_df.iloc[[0,-1],:]
```

✓ 0.0s

	month	Mean RH	Mean Temp	Total Rainfall	Dengue Cases	year
0	2014-01	78.5	26.2	75.4	1886.0	2014
59	2018-12	81.5	27.6	172.6	554.0	2018

Data Manpulation (Dengue)

There are some underlying issues with our dataset as seen from our EDA, there are 8 missing values for the count column and the dataset is based on weeks rather than months. Hence, we will have to manipulate it such that we can concat our datasets together for easier readability and plotting

Grouping weeks into months, I divided the index by the number of epi weeks by the total number of months so that each month would have an equal number of weeks. However, this could lead to several problems such as an inaccurate representation of the true number of dengue cases which is needed to be taken into account.

```python
# Removing DHF from dengue type, using foward fill to replace missing values and reseting the index
filtered_dengue = dengue_df[dengue_df['type_dengue'] == 'Dengue'].drop(dengue_df[['type_dengue']], axis = 1)
                                                                 .ffill()
                                                                 .reset_index(drop=True)

# Grouping the weeks together such that it is split by it's index equally (This assumes all months have equal days)
grouped_dengue = filtered_dengue.groupby(filtered_dengue.index//(265/60))[['number']].sum()

# Creating a new dataframe for analysis by concatting our Climate Dataframe along with our Dengue Dataframe
analysis_df = pd.concat([climate_df, grouped_dengue], axis=1).reset_index(drop=True)
analysis_df = analysis_df.rename(columns={'number': 'Dengue Cases'})
analysis_df['year'] = pd.to_datetime(analysis_df['month'], format='%Y-%m').dt.year
analysis_df = analysis_df.rename(columns={"mean_rh":"Mean RH",
                        "mean_temp": "Mean Temp",
                        "total_rainfall": "Total Rainfall",
                        "dengue_cases": "Dengue Cases"})

number_of_year = len(analysis_df['year'].unique())
```

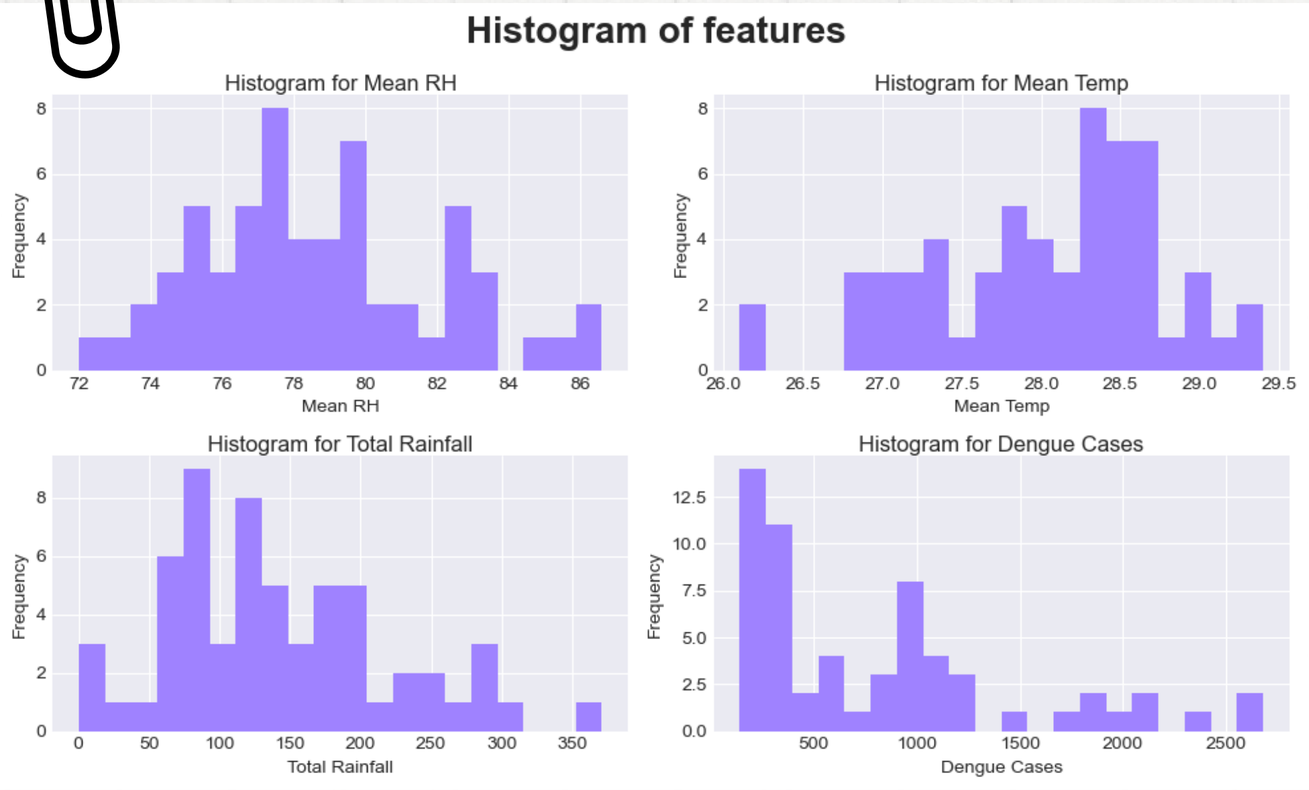
✓ 0.0s

```python
analysis_df.iloc[list(range(5)) + list(range(-5, 0))]
```

✓ 0.0s

	month	Mean RH	Mean Temp	Total Rainfall	Dengue Cases	year
0	2014-01	78.5	26.2	75.4	1886.0	2014
1	2014-02	74.5	27.2	0.2	1021.0	2014
2	2014-03	76.0	27.9	66.0	1108.0	2014
3	2014-04	80.0	28.2	110.0	1012.0	2014
4	2014-05	80.2	28.6	125.8	1893.0	2014
55	2018-08	77.0	28.5	121.6	317.0	2018
56	2018-09	77.1	28.0	144.4	203.0	2018
57	2018-10	79.7	27.9	234.4	334.0	2018
58	2018-11	83.2	27.3	169.6	397.0	2018
59	2018-12	81.5	27.6	172.6	554.0	2018

Graph 1: More information of our the features



- It seems that the ditribution for relative humidity is almost symmetrical with the highest frequency of 8 for 77

- Looking at rainfall and dengue cases, the graphs are positively-skewed. For rainfall, months usually recieve around 85mm of rainfall and certain months that are outliers recieve higher rainfalls which could be due to moonsoon seasons. Meanwhile, dengue cases frequent at about 0-375 cases, where some outliers cases have cases of over 1500 - 2600, which could be due to dengue outbreaks.

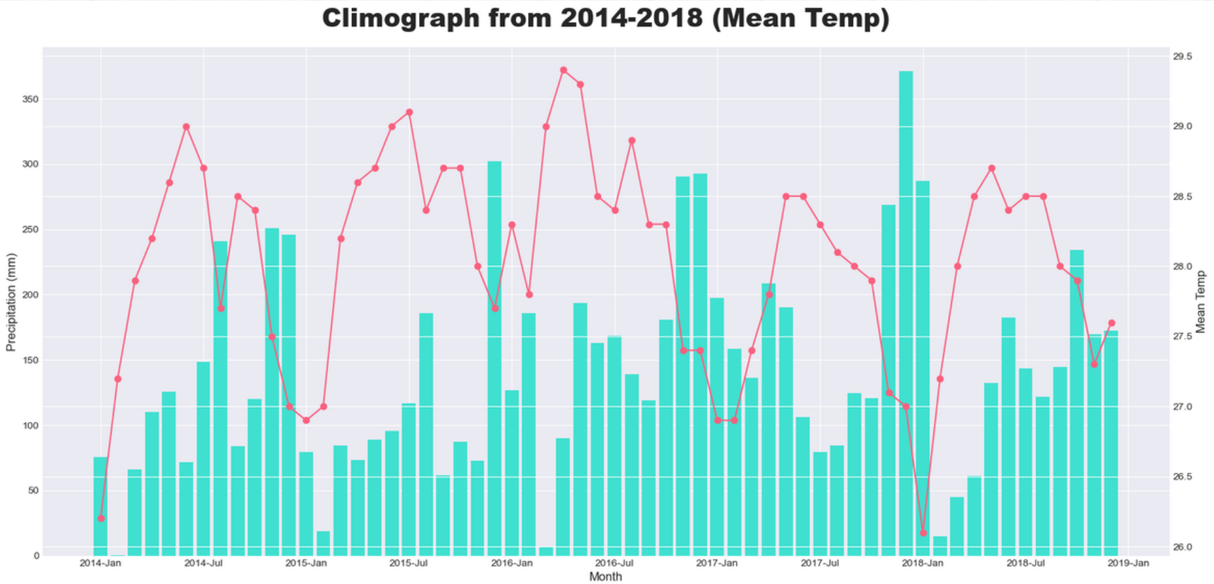
- While for the temperature, it is negatively-skewed and most months have a temperatur of about 28.5 degree celsius but there are a few months with lower degree of about 26 degree celsius

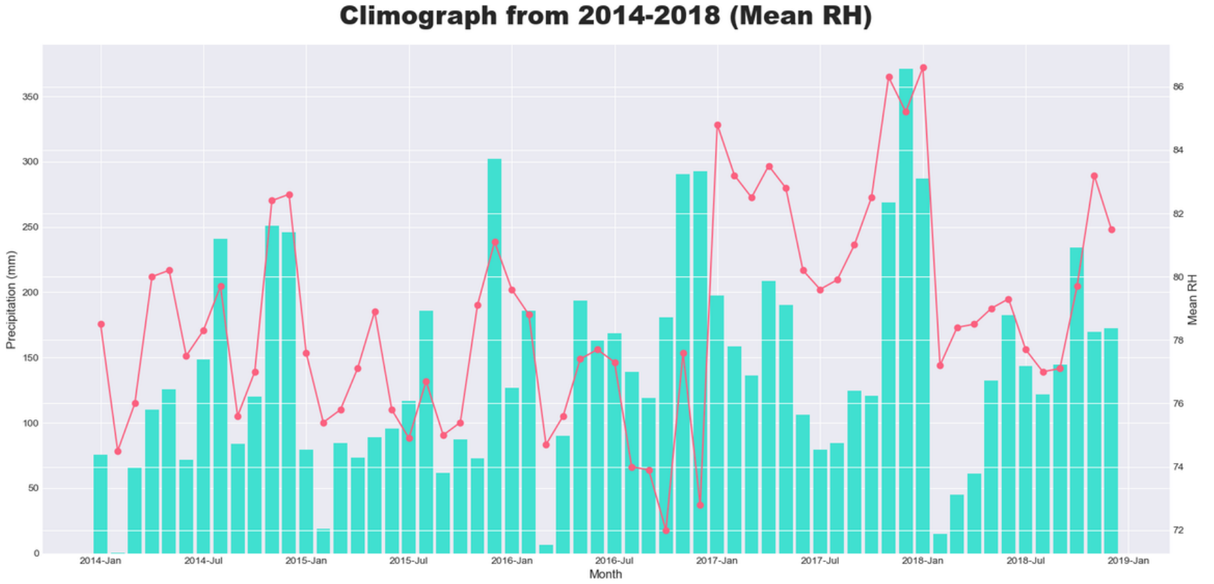
Understanding climate data

it is known that temperature, rainfall, and relative humidity are related to one another, therefore, by plotting a climograph we can observe the trend for these variables over 2014-2018.

```
fig, ax = plt.subplots(2,1,figsize=(20,20))

col_names=['Mean Temp', 'Mean RH']
for i in range(2):
    ax[i].bar(analysis_df['month'].values, analysis_df['Total Rainfall'].values, label ='Precipitation', color='#40e0d0')
    ax[i].set_xlabel('Month')
    ax[i].set_ylabel('Precipitation (mm)')
    ax[i].tick_params('y')

    ax2 = ax[i].twinx()
    ax2.plot(analysis_df['month'].values, analysis_df[col_names[i]].values, label='col_names', c='#fb607f', marker='o')
    ax2.set_ylabel(col_names[i])

    custom_xtick = np.linspace(0,60,(2019-2014)*2+1)
    custom_xticklabels = []
    for year in range(2014, 2019):
        custom_xticklabels.extend([f'{year}-Jan', f'{year}-Jul'])
    custom_xticklabels.append('2019-Jan')

    ax[i].set_title(f"Climograph from 2014-2018 ({col_names[i]})", fontsize =25, fontweight=b)
    ax[i].set_xticks(custom_xtick)
    ax[i].set_xticklabels(custom_xticklabels)

plt.show()
```

Graph 2: Understanding climate data





A general trend observed is that as total rainfall increases, the mean temperature of the month decreases while for relative humidity, it increases. This suggest a linear relationship between relative humidity and rainfall and an inverse relationship for temperature and rainfall.

The lowest total rainfall recieved was in Feburary in 2014 with amost 0mm and the corresponding temperature and relative humidty were 27.3 degree celsisus and abt 74.5 respectively. While the highest rainfall recieved was in December of 2017 with 375mm of precipitation and in the following month in January 2018 the lowest temperature of 26.1 and highest relative humidty of 86.5 was recorded.

We can also observe that usually, rainfall and relative humidity is highest during the end of the year around December/ January while temparture is usually low during this period. However in 2014, rainfall is highest in the middle of the year in July.

Over the years, there is a general trend where temperature has been increasing, however, there was a drop observed from 2016 to 2017

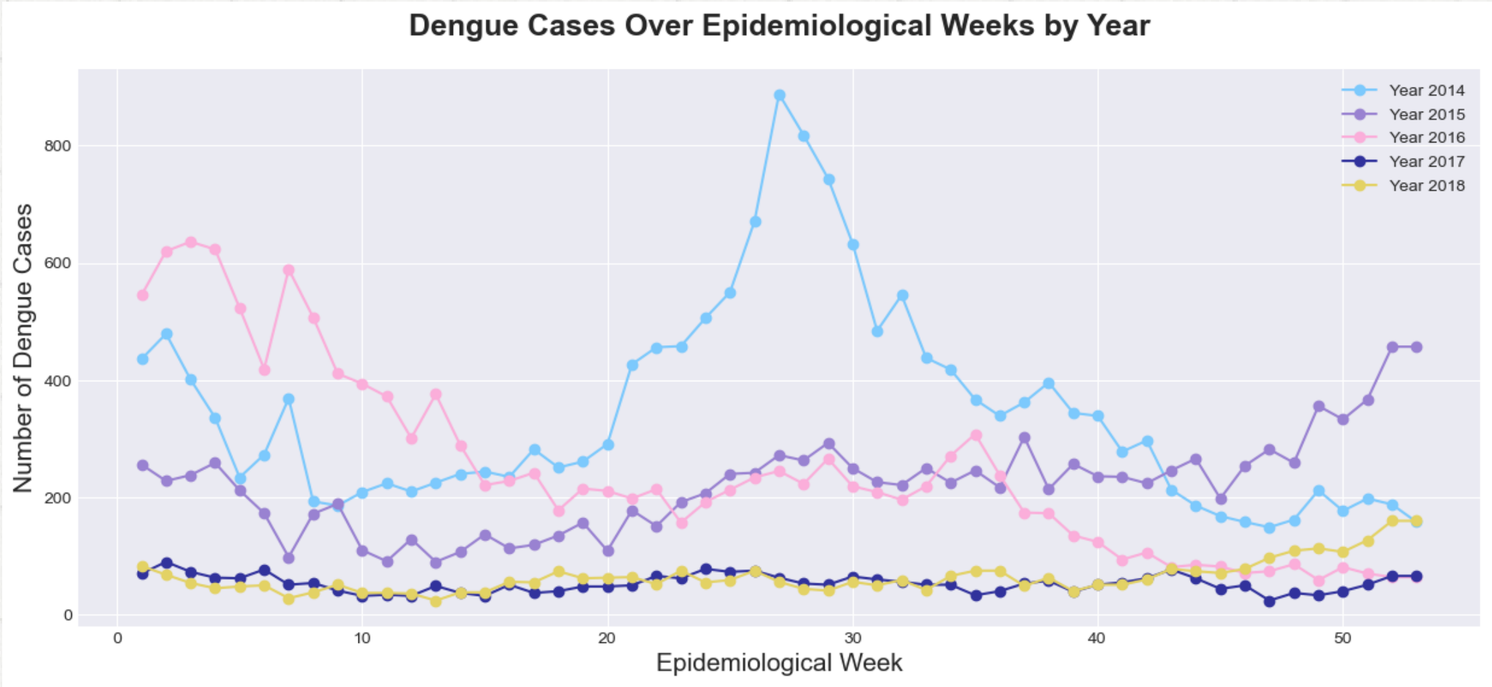
Graph 3: Understanding dengue data

A general trend observed is that over the years, the number of dengue cases has been reducing which could be due to the efforts of government and its citizen in realsing the threat of dengue mosqitoes and doing wipeouts.

We can see that dengue cases are usually high towards the start and end of the year which observed from our climograph is when rainfall is the highest. There is an anomaly in 2014 where there is a significantly high number of cases in the middle of the year around June/ July which was also where we observed an anomaly in the climograph where the middle of the year also had the highest rainfall.

This could suggest a relationship between rainfall and dengue cases.

```
color = ['#7cc9fd', '#9982d1', '#fbaeda', '#31329c', '#e3d264']

plt.figure(figsize=(15,6))
for i in np.linspace(0, len(filtered_dengue)-53, number_of_year, dtype=int):
    x = filtered_dengue[i:i+53]['eweek']
    y = filtered_dengue[i : i+53]['number']
    plt.plot(x, y, label=f'Year {int(filtered_dengue.iloc[i]["year"])}', marker='o', color= color[int(i/53)])

plt.xlabel('Epidemiological Week', fontsize = '15')
plt.ylabel('Number of Dengue Cases', fontsize = '15')
plt.title('Dengue Cases Over Epidemiological Weeks by Year', fontsize = '18', fontweight ='bold')
plt.legend()
plt.show()
```



Looking for correlation between varables

```
fig, ax = plt.subplots(3, 1, figsize=(10,14), layout ='constrained')

col_names = analysis_df.iloc[:, 1:-2].columns
colors = analysis_df['year'].apply(lambda x:'#7cc9fd' if x==2014 else
                                            '#9982d1' if x ==2015 else
                                            '#fbaeda' if x==2016 else
                                            '#31329c' if x== 2017 else
                                            '#e3d264')

for i, (col, ax) in enumerate(zip(col_names, ax)):
    custom_legend_entries = []

    for year in analysis_df['year'].unique():
        mask = (analysis_df['year'] == year)
        x = analysis_df.loc[mask, col]
        y = analysis_df.loc[mask, 'Dengue Cases']
        ax.scatter(x, y, label=f'{year}', s=200, alpha=0.8, c=colors[mask])

    ax.legend(handles=custom_legend_entries, title='Year', loc='upper right')
    ax.set_xlabel(f'{col}')
    ax.set_ylabel('Dengue Cases')
    ax.set_title(f'Correlation between Dengue Cases and {col}', fontweight = 'bold', fontsize ='13', pad = '5')

    x = analysis_df[col]
    y = analysis_df['Dengue Cases']

    m,b = np.polyfit(x, y, deg=1)
    ax.plot(x, m*x + b,'#f55141')
    corr_coefficient = np.corrcoef(x, y)[0, 1]
    text_x = x.iloc[-1]
    text_y = y.iloc[-1]
    ax.text(text_x+1, 1500,
            f'Correlation Coefficient: {corr_coefficient:.2f}',
            color='black',
            bbox={'facecolor': '#ffffe3','pad':10}
            )
    ax.legend(title='Year')

plt.show()
```

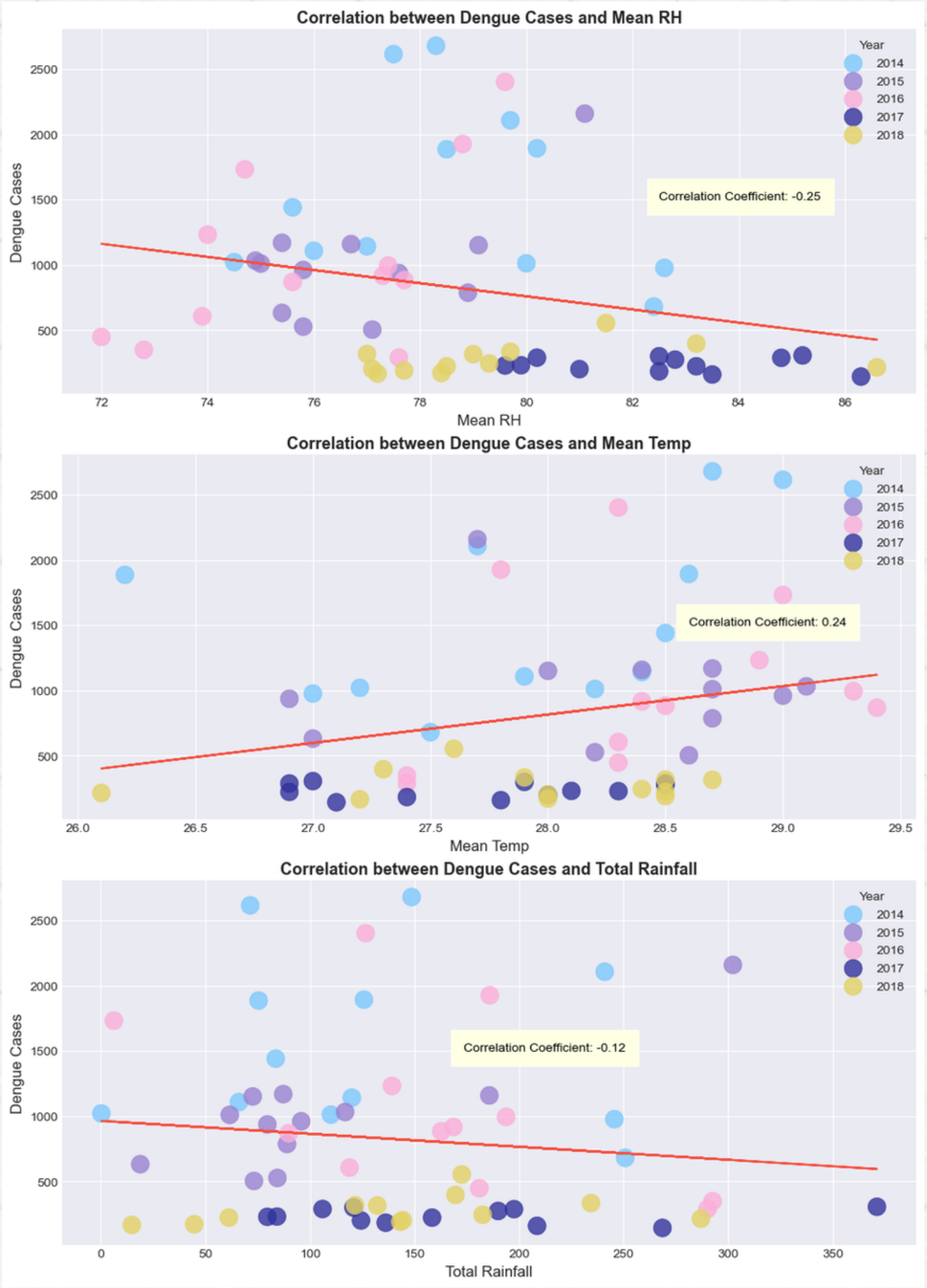
Graph 4: Looking for correlation between varables

On the contrary to what we observed in our graphs earlier, it seems that there is a weak correlation between our climate variables and the number of dengue cases.

For rainfall and relative humidity, there is a negative but weak correlation suggesting that as rainfall/relative humidity decreases, dengue cases increase. While the opposite is true for temperature, where although there is still a weak correlation, it is positive, suggesting that as temperature increases, the number of cases increases as well.

There are some reasons why this may be unreliable, such as:

- There are other factors affecting dengue cases in a given month
- Some outliers cause the correlation to drop
- Our data has mutated due to the change from weeks to months
- As we are plotting for multiple years, certain years that have fewer dengue cases may not give a true representation



Looking for correlation between variables (by year)

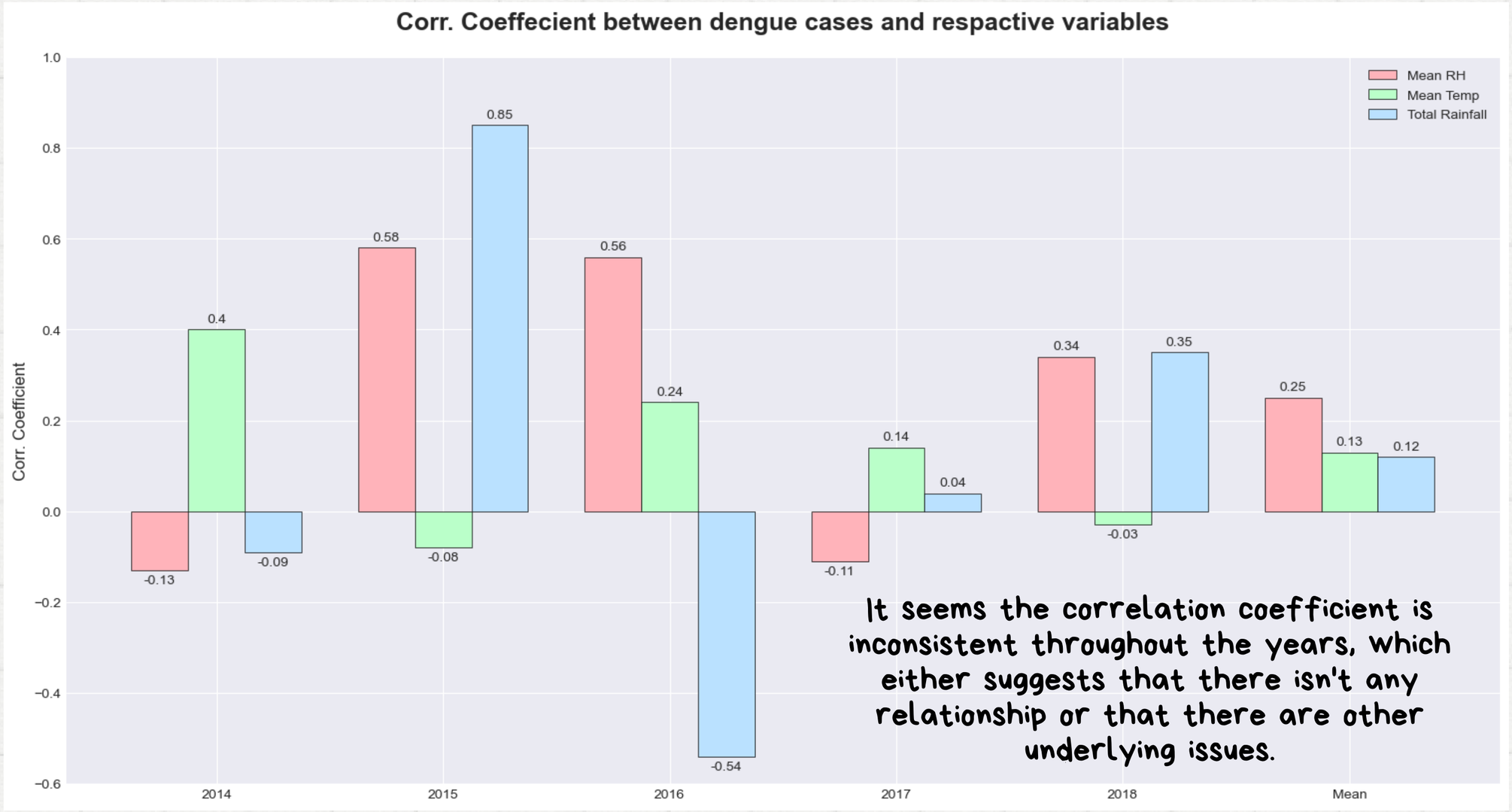
```
results = []
for i in np.linspace(0, len(analysis_df)-12, number_of_year, dtype=int):
    temp_df = analysis_df[i:i+12]

    for i, col in enumerate(col_names):
        x = temp_df[col]
        y = temp_df['Dengue Cases']
        corr_coefficient = round(np.corrcoef(x, y)[0, 1],2)
        results.append(corr_coefficient)

results_df = pd.DataFrame(np.array(results).reshape(5,3), columns=col_names, index=analysis_df['year'].unique())
results_df = pd.concat([results_df, pd.DataFrame([round(results_df.mean(),2)], index=['Mean'])])
```

✓ 0.0s

```
x = np.arange(len(results_df))  # the label locations
width = 0.25  # the width of the bars
multiplier = 0
colors = ['#ffb3ba', '#baffc9', '#bae1ff']

fig, ax = plt.subplots(figsize = (15,8), layout='constrained')

for i, col in enumerate(results_df):
    offset = width * multiplier
    rects = ax.bar(x + offset, list(results_df[col]), width, label=col, edgecolor = 'black', color = colors[i])
    ax.bar_label(rects, padding=3)
    multiplier += 1

ax.set_ylabel('Corr. Coefficient')
ax.set_title('Corr. Coeffecient between dengue cases and respactive variables', fontsize='18', fontweight='bold')
ax.set_xticks(x + width, results_df.index)
ax.legend()
ax.set_ylim(-0.6, 1)

plt.show()
```

✓ 0.5s

Graph 5: Looking for correlation between variables (by year)



Looking at our dataset by Year

By grouping the dataset into months we can look at the data by year and we might be able to see a clearer trend if our dataset has mutated

```
# Grouping the dataset by months
grouped_months = analysis_df.copy()
grouped_months['month'] = pd.to_datetime(analysis_df['month'], format='%Y-%m').dt.month
grouped_months = grouped_months.sort_values(by = 'month')
```

✓ 0.0s

```
# Pivoting the dataset for easier time plotting
rh_months = grouped_months.pivot(index='year', columns='month', values='Mean RH')
temp_months = grouped_months.pivot(index='year', columns='month', values='Mean Temp')
rf_months = grouped_months.pivot(index='year', columns='month', values='Total Rainfall')
dengue_months = grouped_months.pivot(index='year', columns='month', values='Dengue Cases')
```

✓ 0.0s

Before pivot:

```
grouped_months.head()
```

✓ 0.0s

	month	Mean RH	Mean Temp	Total Rainfall	Dengue Cases	year
0	1	78.5	26.2	75.4	1886.0	2014
24	1	79.6	28.3	126.6	2402.0	2016
36	1	84.8	26.9	197.6	288.0	2017
12	1	77.6	26.9	79.6	937.0	2015
48	1	86.6	26.1	287.0	215.0	2018

After pivot:

```
dengue_months
```

✓ 0.0s

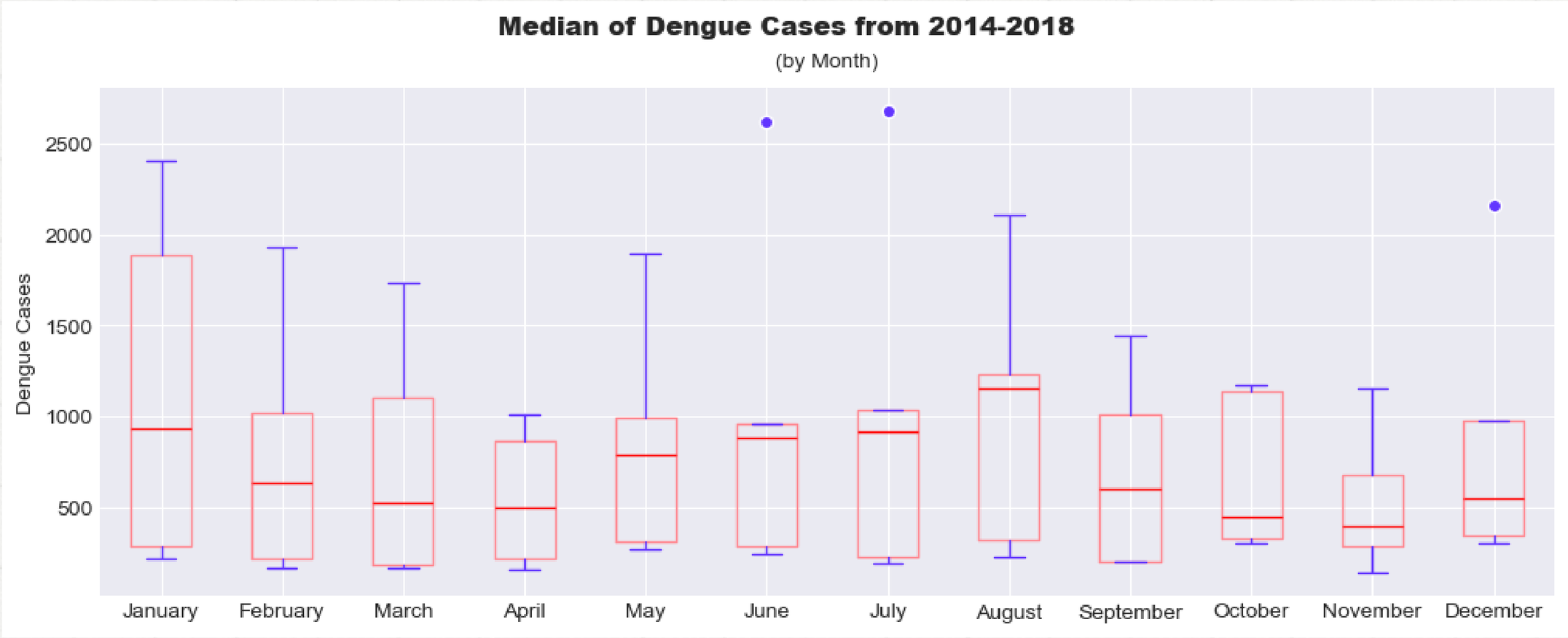
month	1	2	3	4	5	6	7	8	9	10	11	12
year												
2014	1886.0	1021.0	1108.0	1012.0	1893.0	2615.0	2679.0	2107.0	1441.0	1142.0	681.0	977.0
2015	937.0	632.0	527.0	504.0	787.0	961.0	1032.0	1159.0	1010.0	1169.0	1151.0	2159.0
2016	2402.0	1926.0	1732.0	869.0	995.0	884.0	917.0	1232.0	606.0	448.0	290.0	349.0
2017	288.0	223.0	184.0	160.0	274.0	289.0	229.0	231.0	201.0	300.0	144.0	306.0
2018	215.0	167.0	172.0	223.0	316.0	245.0	191.0	317.0	203.0	334.0	397.0	554.0

Graph 6: Boxplot (Dengue)- By Year

We can observe that out of all the months, August and January the has the highest median number of cases of 1250, and 900 respectively, while the lowest was recorded in November and October of about 400 each.

```
month_name = calendar.month_name[1:]

fig, ax = plt.subplots(figsize = (10,4), layout='constrained')
ax.boxplot(dengue_months,
           boxprops= dict(color = '#ff8b94'),
           flierprops=dict(marker='o', markerfacecolor='#6235ff',
           medianprops=dict(color='red'),
           capprops=dict(color='#6235ff'),
           whiskerprops=dict(color='#6235ff'))

plt.suptitle("Median of Dengue Cases from 2014-2018", fontweight=b
ax.set_title("(by Month)", fontsize=10, pad=10, fontweight='light'
ax.set_ylabel("Dengue Cases", fontsize=10)
ax.set_xticklabels(month_name)
plt.show()
```

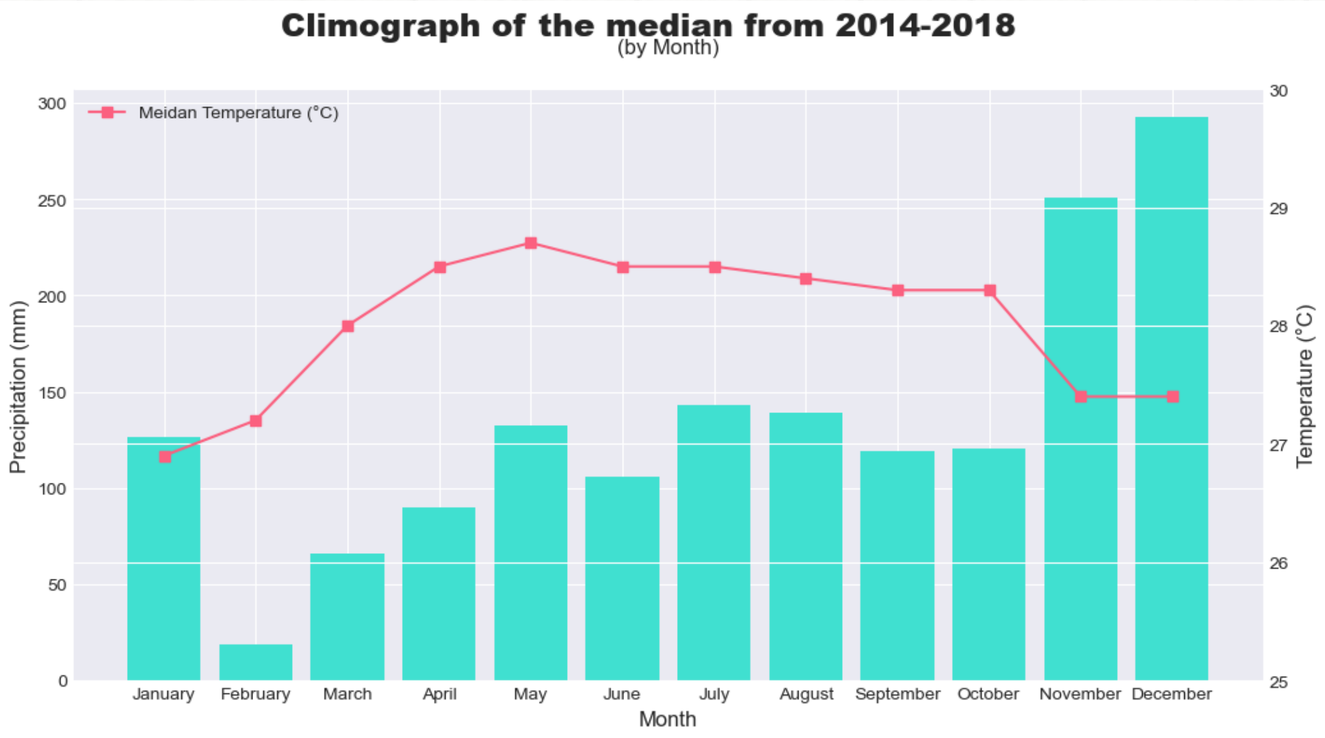
✓ 0.3s



Graph 7: Climograph - By Year

The highest ranfall experiecned was in December with 300mm while lowest experienced was in Feburary with about 20mm while lowest temperatrure was in Janurary with about 27 degree and highest in May with about 28.5 degrees.

The fact the one of the highest number of cases was in January and the highest amount of rainfall collected was in the month before in December, suggesting that our theory at the start that there is a relationship between rainfall and dengue cases is true.




```
fig, ax1 = plt.subplots(figsize=(12,6))

ax1.bar(month_name, list(rf_months.median()),
        label ='Precipitation', color='#40e0d0')
ax1.set_xlabel('Month')
ax1.set_ylabel('Precipitation (mm)')
ax1.tick_params('y')

ax2 = ax1.twinx()
ax2.plot(month_name, list(temp_months.median()),
         label= 'Meidan Temperature (°C)',
          c = '#fb607f', marker='s')
ax2.set_ylabel("Temperature (°C)")
ax2.set_ylim(25,30)

plt.suptitle("Climograph of the median from 2014-2018",
             fontsize =18, fontweight=b)
ax1.set_title("(by Month)", fontsize=12)

plt.legend(loc='upper left')
plt.show()
```

✓ 0.3s

Results

Overall, from our graphs, we can see that there is a high chance that there is an increase in dengue cases as there is more rainfall. Although we couldn't find much correlation between temperature and relative humidity, since these variables are closely related to rainfall, we can assume that as temperature decrease dengue cases increase while as relative humidity increase, dengue increases.

This shows that during periods with high rainfall which happen typically towards the start/end of the year, Singapore experiences the monsoon season, we should take more precautions towards dengue to try and minimize them from breeding and causing more cases.

We can also see that as years go by, the number of dengue cases has decreased despite a general increase in both rainfall and temperature, this is especially true after 2014 which is when a dengue outbreak occurred causing Singapore to take more precautions the following years which has proven to be making an impact.

Issues

There are several problems when dealing with the dataset, the biggest problem was when I was trying to convert epi weeks into months, and even after finding a possible solution, it may have impacted the actual representation of our data and the correlations.

Takeaways

This assessment has taught me the importance of finding good datasets that are relatable and how to manipulate them.

It also gave me a deeper understanding of how to use Matplotlib as there were many things that I found out when researching how to plot graphs which I would not have found out if I only looked at the syllabus.

Overall, it was really fun trying to explore and having to find ways to deal with the dataset

Refferences

Dengue outbreak:
Auto, H. (2021) Dengue cases approaching record high in Singapore as infections continue to soar, The Straits Times. Available at: https://www.straitstimes.com/singapore/health/dengue-cases-likely-to-hit-a-record-high-this-week-as-infections-continue-to-soar (Accessed: 14 December 2023).

Weather:
Climate of Singapore (no date) Climate of Singapore |. Available at: https://www.weather.gov.sg/climate-climate-of-singapore/#:~:text=Seasons,Monsoon%20from%20June%20to%20September. (Accessed: 14 December 2023).