HO CHI MINH UNIVERSITY OF SCIENCE

MULTIVARIATE STATISTICS ANALYSIS - LAB02

March 18, 2024

Author

Nguyen Viet Kim - 21127333

Contents

1 Introduction **3**

2 Algorithm Description **3**

2.1 Imported Libraries 3

2.2 Total Covid-19 Cases and Deaths Per Month 3

2.3 Top 10 countries with most cases 3

2.4 Total Cases Per Month in Vietnam 4

2.5 Average Temperature vs Average Rainfall by Month 4

2.6 Average Temperature vs Average Rainfall by Month in 2009 4

3 Result **4**

3.1 Covid-19-cases.csv and Covid-19-deaths.csv 4

3.2 weatherHistory.csv 7

1 Introduction

The program use Python and Matplotlib to visualize data from 3 csv files: Covid-19cases.csv, Covid19-deaths.csv and weatherHistory.csv.

2 Algorithm Description

2.1 Imported Libraries

The required libraries for this programs are *Numpy*, *Matplotlib*, *Pandas* and *datetime* which are used in the most cases to get data from the csv files, analyze data and plotting data on charts. *sklearn.linear_model* library is also included in this program as to analyze data in one of the chart

2.2 Total Covid-19 Cases and Deaths Per Month

We first can compare the total cases per month to total death per month from the Covid-19-cases.csv and Covid-19-deaths.csv. We first store the read from the files into 2 variables: *total_cases_per_month* and *tota_deaths_per_month* by selecting from column 5 of the . Since the time is stored in a row, the program use the traversed matrix of the data for analysis.

```
# Get the data by date
cases_by_day = cases.iloc[:, 5:]
deaths_by_day = deaths.iloc[:, 5:]

# Transpose the data so that dates become columns
cases_by_day = cases_by_day.T
deaths_by_day = deaths_by_day.T
```

The program group the data by month and year, then get sum of the cases as well as the deaths.

```
total_cases_per_month = cases_by_day.groupby([cases_by_day.index.year, cases_by_day.index.month]).sum().reset_index()
total_deaths_per_month = deaths_by_day.groupby([deaths_by_day.index.year, deaths_by_day.index.month]).sum().reset_index()

# Set the index to the year and month of the data
total_cases_per_month.index = pd.to_datetime({
    'year': total_cases_per_month['level_0'],
    'month': total_cases_per_month['level_1'],
    'day': 1
})

total_deaths_per_month.index = pd.to_datetime({
    'year': total_deaths_per_month['level_0'],
    'month': total_deaths_per_month['level_1'],
    'day': 1
})

# Calculate the total cases for each row
total_cases_per_month['Total Cases'] = total_cases_per_month.iloc[:, 3:].sum(axis=1)
total_deaths_per_month['Total Cases'] = total_deaths_per_month.iloc[:, 3:].sum(axis=1)
```

The result is visualized using a bar chart with blue for total cases and red for total of deaths by each month. The program also visualizes the result by using line chart for better understanding of the trend of each data sample and the comparison between them by each month.

2.3 Top 10 countries with most cases

The program use the same method for calculating the total of cases, but instead of group by month and year, in this case we group the data by country/region and get the top 10 most cases.

```
# Calculate total cases for each country
total_cases_by_country = cases.groupby('Country/Region').sum().iloc[:, 5:].sum(axis=1)

# Find the top 10 countries with the most cases
top_10_countries = total_cases_by_country.nlargest(10)
```

For other countries, we calculate the total of cases where the country is not in the given top 10 countries.

```
# Calculate total cases for other countries
other_countries_total_cases = total_cases_by_country[~total_cases_by_country.index.isin(top_10_countries.index)].sum()

# Create a new DataFrame with top 10 countries and 'Other' category
top_countries_df = pd.concat([top_10_countries, pd.Series({'Other': other_countries_total_cases})])
```

The result is visualized by a pie chart which show the proportion of each top 10 countries comparing to others and the overall picture of total cases in the world during the period.

2.4 Total Cases Per Month in Vietnam

Using the same methods as before, the line chart of Vietnam' total cases is calculated by sorting the country/region by name, in this case is 'Vietnam', then get sum of cases grouped by month and visualized using a line chart.

2.5 Average Temperature vs Average Rainfall by Month

In this part, the data file used for visualization is the data frame weather analysis[**b1**] from Kaggle. There are various climate factors but in this program, we focus on the correlation between average tempurature and the rainfall in month.

```
# Group by month and calculate average temperature and rainfall for month
average_by_month = weather.groupby(['Month', 'Year'])[['Average temperature (°F)', 'Rainfall for month (in)']].mean().reset_index()
```

The code gets the data from the table and get the mean of the sample by each month. The program then use the scatter plot to visualize the distribution of the data by rainfall and temperature.

2.6 Average Temperature vs Average Rainfall by Month in 2009

In this section, the program get the data that is in the year of 2009 and use a weather chart to demonstrate the average temperature and the rainfall in each month.

3 Result

3.1 Covid-19-cases.csv and Covid-19-deaths.csv

In this program, we can have a look at the total Covid cases from the January 2020 to July 2021 comparing to the total number of deaths that caused by covid in the data frame.
The steps of visualizing the total number of cases and deaths by month are:

- Get the data from the columns that contains the required information in the covid-19-cases.csv and the covid-19-deaths.csv, which are columns with date and number of cases.
- Since the matplotlib would count the sum by row, we need too transpose the columns and convert the first column to datetime format. The format is chosen by looking at the format in the csv file.

- Calculate the sum of cases by column by each month and the corresponding year. But in this step, it only get the sum cases by each country, so we need to get the sum of the data along the row to get the final total number of cases.
- Visualize the cases and the deaths by two columns in the table with blue for cases and red for deaths.

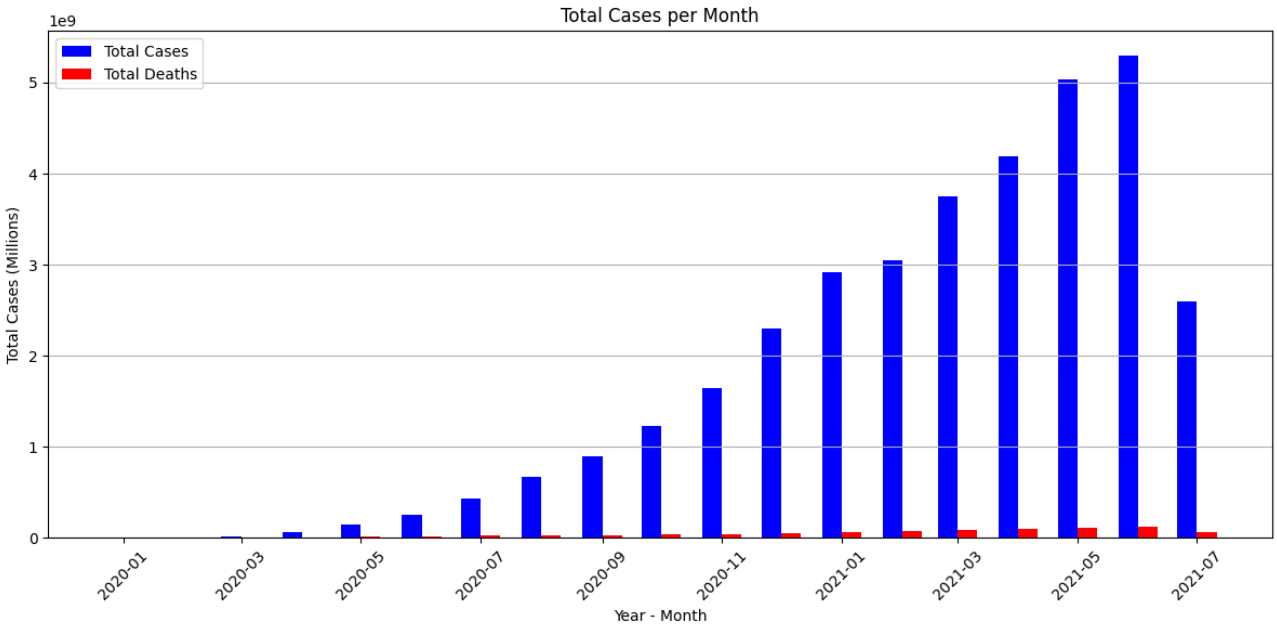

Figure 1: Bar chart of Total cases and deaths by month

From the given diagram, it is visible that the number of total cases rocketed from April 2020 to June 2021 before dramatically decreases in July 2021. Despite such huge number of infected, the total number of deaths seems small comparing to the number of cases. But they both have the same trend in this period.
For better visualization, a line diagram is created with the following code, which is better at showing trends of both cases and deaths:

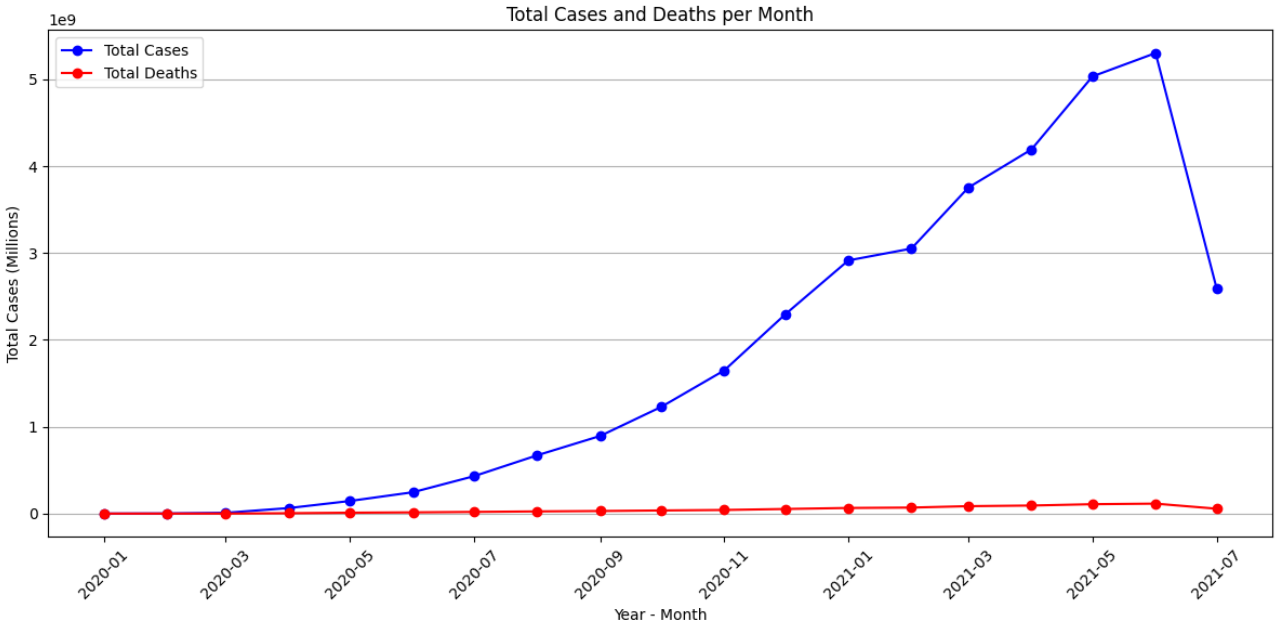

Figure 2: Line chart of Total cases and deaths by month

By calculating the total cases by country and sort in order, we can retrieve the top 10 countries that has

the highest number of total cases during the period. The code get the top 10 countries by the number of cases, and group the others in 'Other' group for plotting on the chart. In this section, a pie chart is used in order to present the portion of each countries in the total number of the world.

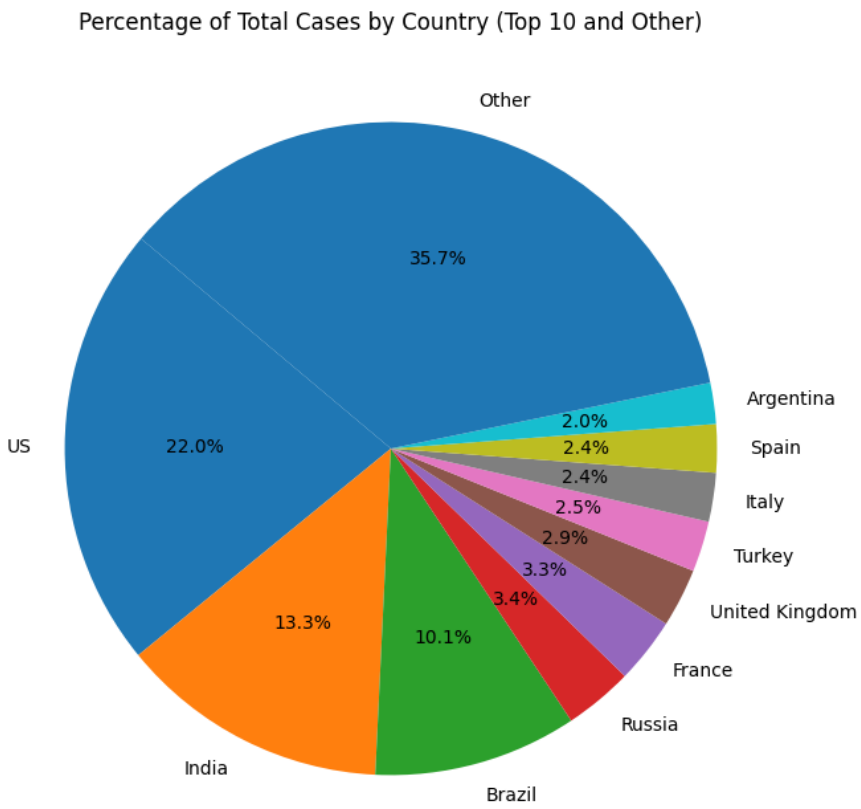

Figure 3: Pie chart of Top 10 countries with most cases

Using the same method, we can get the total cases of Vietnam during the period and use a line chart to visualize the trend:

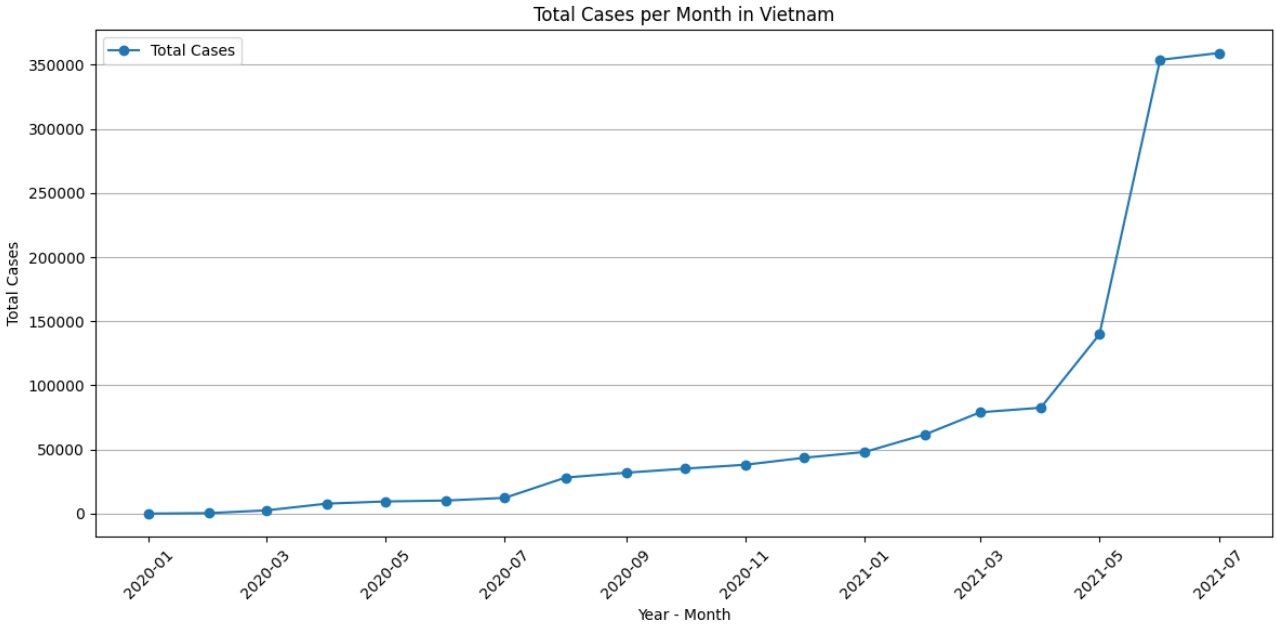

Figure 4: Line chart of Vietnam's total cases by month

3.2 weatherHistory.csv

The data frame is a csv downloaded from Kaggle website that shows information of the climate (average temperature, rainfall in month, rainfall in year, etc.) from 2009 to 2020 of Estes Park, Colarado, USA.
In this data frame, we focus on the correlation between average temperature (°F) and the rainfall in month (in). by calculating the average of the temperature and the rainfall in month, group by month, we can show the scatter plot of the trends using the following chart:

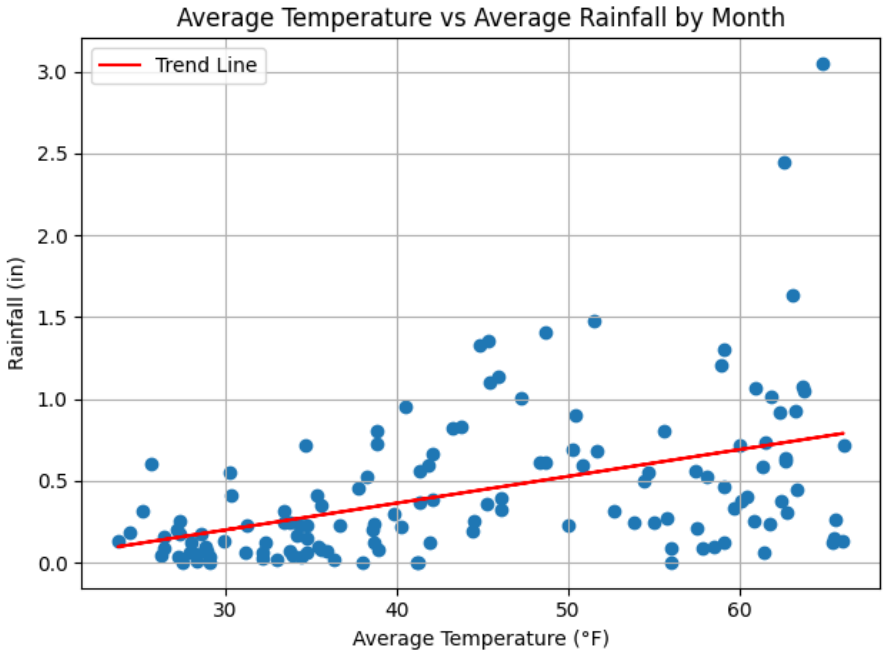

Figure 5: Scatter plot of Average temperature and rainfall by month

The trend line if drawn on the chart that show the relation between the attributes. It can be seen that the during the hot weather, the rainfall seems to be more than in cooler climate.
We can take a look at a specific year of 2009 via this rainfall chart:

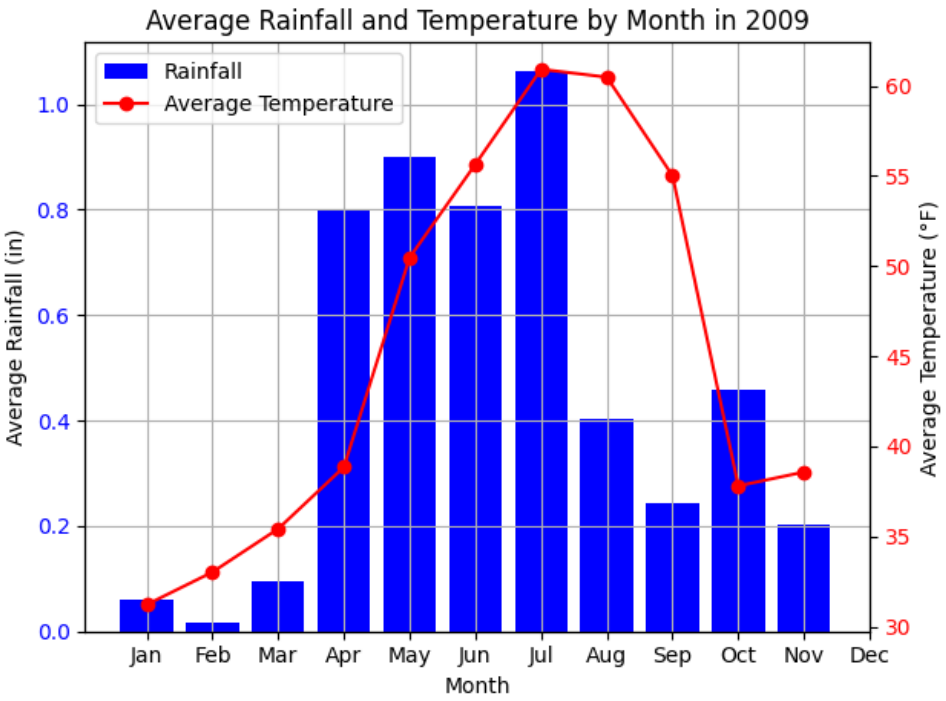

Figure 6: Rainfall chart

References

[1] Mustafa Fatakdawala. weather analysis. 2021 Kaggle. cite : https://www.kaggle.com/datasets/mastmustu/weather-analysis