**Project Report**

|  |  |
| --- | --- |
| **Course Name (WSQ)** | **Professional Diploma in Data Science** |
| Product Name (Marketing &  Sales) | **Professional Diploma in Data Science** |
| **Module Name (WSQ)** | **WSQ- Python for Data Science (SF)** |

|  |  |  |  |
| --- | --- | --- | --- |
| **Student name** |  |  | **Assessor name** |
| Lee Choon Hoe |  |  | **Dr. Rajendra** |
| **Date issued** | **Completion date** | | **Submitted on** |
| 21/10/2024 | 5/12/2024 | | 2/12/2024 |
|  |  |  | |
| **Project title** | **Worldwide COVID-19 Data Analysis** | | |

|  |
| --- |
| **Learner declaration** |
|  |
| I certify that the work submitted for this assignment is my own and research sources are fully acknowledged.        Student signature: Date: |

**Index**

1. **Project Overview:** Describe the Project along with Project Outcomes (Explain the Project in your own words in 15 – 20 lines)

**Project Overview: Worldwide COVID-19 Data Analysis**

The global COVID-19 pandemic highlighted the importance of data-driven decisions in understanding and combating public health crises. This project focuses on the comprehensive analysis of worldwide COVID-19 data sourced from the "Our World in Data" repository. The dataset includes key metrics such as confirmed cases, deaths, vaccination rates, testing numbers, and other indicators across various countries and regions.

The primary goal of the project is to leverage Python programming skills to process, analyze, and visualize the dataset effectively. By using Python libraries like Pandas, Matplotlib, and Seaborn, the project aims to extract meaningful insights, identify trends, and provide visual representations of critical data points. These insights can support decision-making for health organizations, policymakers, and researchers.

The project involves several stages, including data loading and exploration, cleaning and preprocessing, analysis, and visualization. Each step ensures the data is prepared for reliable and accurate results. Handling missing values, normalizing data formats, and converting columns (e.g., dates) to appropriate types are key preprocessing tasks.

**Project Outcomes**

1. Gain hands-on experience in working with real-world datasets.
2. Master data exploration and preprocessing techniques using Python.
3. Develop analytical skills to identify trends and patterns in COVID-19 data.
4. Create impactful data visualizations to communicate findings effectively.
5. Enhance understanding of global COVID-19 trends, aiding in better decision-making.

1. **Project Technical Environment:**  You can explain various libraries you have used.

**1. Pandas**

* **Purpose**: Pandas is the core library used for data manipulation and analysis. It provides powerful tools for handling structured data, such as CSV files.
* **Features Used**:
  + Loading datasets into DataFrames.
  + Handling missing data through imputation or removal.
  + Data cleaning, transformation, and aggregation.
  + Filtering and slicing data to derive insights.

**2. NumPy**

* **Purpose**: NumPy supports numerical computations, providing efficient array structures for handling large datasets.
* **Features Used**:
  + Performing mathematical operations on data columns.
  + Supporting Pandas DataFrame computations.

**3. Matplotlib**

* **Purpose**: Matplotlib is used for creating static, animated, and interactive visualizations.
* **Features Used**:
  + Plotting line graphs, bar charts, histograms, and scatter plots.
  + Customizing plots with titles, labels, legends, and annotations.

**4. Seaborn**

* **Purpose**: Seaborn is a Python visualization library built on top of Matplotlib. It simplifies the process of creating visually appealing and informative statistical graphics.
* **Features Used**:
  + Creating heatmaps, boxplots, and pair plots for deeper insights.
  + Enhancing aesthetics with color palettes and themes.

**5. Datetime**

* **Purpose**: The datetime module helps manage and manipulate date and time data.
* **Features Used**:
  + Converting string-based date columns to proper datetime format.
  + Extracting components like year, month, or day for trend analysis.

**6. OS Module**

* **Purpose**: The OS module is used for file path management and ensuring compatibility across different operating systems.
* **Features Used**:
  + Dynamically accessing file paths to load the dataset.

**7. Jupyter Notebook / Google Colab**

* **Purpose**: These environments provide interactive platforms for developing and executing Python code.
* **Features Used**:
  + Writing, testing, and documenting Python scripts.
  + Visualizing outputs in real-time alongside the code.

1. **Activity 1:** List the problem and screenshot for the same.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Activity 2:** List the problem and screenshot for the same.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Activity 3:** List the problem and screenshot for the same.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

A blue rectangle with white text

Description automatically generated

A blue rectangle with white text

Description automatically generated

A graph of a covid-19 virus

Description automatically generated

A red and blue squares with numbers

Description automatically generated

A graph of a number of people with blue lines

Description automatically generated with medium confidence

1. **Activity 4:** List the problem and screenshot for the same.

A screenshot of a computer

Description automatically generated

A graph of a number of people

Description automatically generated

A chart of a number of boxes

Description automatically generated

A graph of a number of red bars

Description automatically generated

A graph of blue bars

Description automatically generated

1. **Activity 5:** List the problem and screenshot for the same.

A screenshot of a computer

Description automatically generated

A graph of a graph showing the number of cases

Description automatically generated with medium confidence

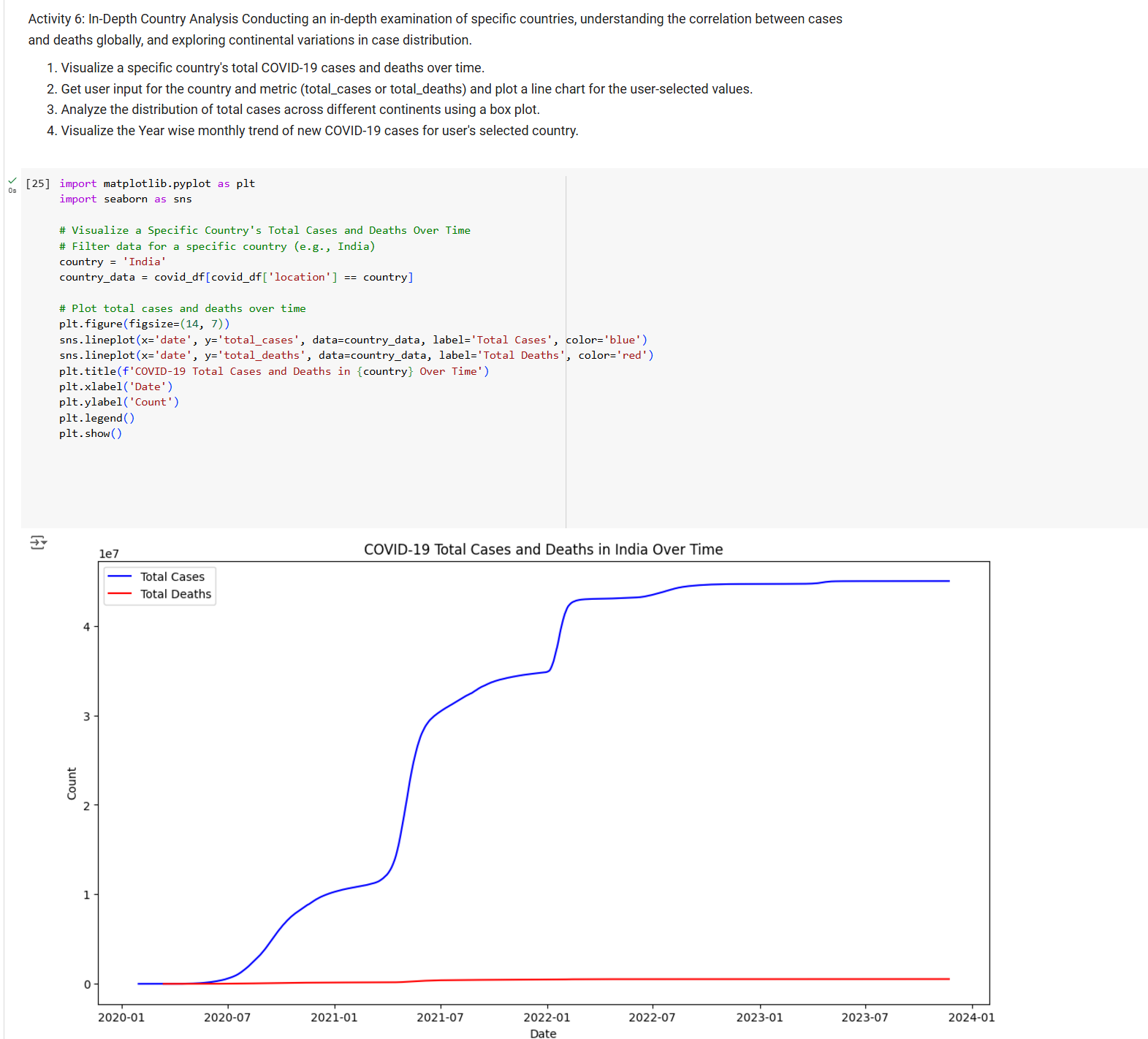
A blue lines drawn on a white background

Description automatically generated

A graph showing a number of blue and orange lines

Description automatically generated

1. **Activity 6:** List the problem and screenshot for the same.



A screenshot of a computer

Description automatically generated

A screenshot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

1. **Activity 7:** List the problem and screenshot for the same.

A screenshot of a computer

Description automatically generated

A screen shot of a graph

Description automatically generated

A graph of a graph showing the global daily news

Description automatically generated with medium confidence

A screen shot of a graph

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph with purple dots

Description automatically generated

A screenshot of a computer

Description automatically generated

A graph of a patient's death rate

Description automatically generated

1. **Conclusion and Future Improvements**

**Conclusion and Future Improvements**

Conclusion

This project has provided a comprehensive analysis of the global COVID-19 pandemic using the "Our World in Data" dataset. By leveraging data visualization and statistical techniques, we have gained valuable insights into the spread of the virus, its impact on different countries, and the effectiveness of various public health measures.

Key findings from the analysis include:

* Global Trends: The pandemic initially spread rapidly across the globe, with significant variations in case and death rates across different countries.
* Impact of Vaccination: Vaccination campaigns have played a crucial role in reducing the severity of the pandemic and mitigating its impact on public health.
* Regional Disparities: Different regions have experienced the pandemic differently, influenced by factors such as population density, healthcare infrastructure, and government policies.
* Correlation between Variables: There are strong correlations between various factors, such as testing rates, vaccination rates, and case/death rates.

Future Improvements:

1. Incorporate Real-Time Data: Continuously update the analysis with the latest data to provide timely insights.
2. Advanced Statistical Modeling: Employ more sophisticated statistical models to predict future trends and assess the impact of interventions.
3. Machine Learning: Utilize machine learning techniques to identify patterns and make data-driven predictions.
4. Geographical Analysis: Visualize the spread of the virus on interactive maps to identify hotspots and trends.
5. Social and Economic Impact: Analyze the social and economic consequences of the pandemic, including job losses, poverty, and mental health issues.
6. Ethical Considerations: Ensure that the analysis is conducted ethically and responsibly, respecting privacy and data security.

By addressing these areas, we can further deepen our understanding of the COVID-19 pandemic and inform public health policies and decision-making.

1. Annexure

(Add the Python Scripts here)

