Programming For Data Science

Subject Code: CSL 225

Project Report



Faculty name:

Ms. Akanksha Kaushik

Team:

* Purab Gupta (23csu252)
* Ishika Avasthi (23csu131)
* Raghav Bhalla (23csu255)

**Department of Computer Science and Engineering The NorthCap University, Gurugram- 122001, India**

Project Title:

COVID-19 DATA ANALYSIS



**ABSTRACT**

This project uses Python to analyze COVID-19 case data alongside socio-economic indicators from the World Happiness Report. The goal is to explore how factors like GDP per capita, social support, healthy life expectancy, and freedom of choice correlate with the spread of COVID-19 across countries.

The analysis begins with data cleaning and aggregation, followed by the calculation of daily infection rate changes. Statistical analysis and data visualization, including scatter plots and regression lines, highlight the relationships between these variables.

The findings reveal a positive correlation between wealthier nations and higher infection rates, while social support and healthcare systems also influence pandemic impact. This study offers insights for policymakers seeking to understand the broader societal effects of COVID-19.

**INDEX**

**1. Code Overview**

**2. User Options**

**3. Summary**

**4. Tools used**

**OBJECTIVES**

1. Perform data wrangling on COVID-19 and happiness datasets.

2. Aggregate and visualize maximum infection rates for each country.

3. Identify correlations between socio-economic indicators and infection rates.

4. Communicate findings through intuitive plots.

**TASKS**

**Task 1: Import and Inspect Data**

1. Imported and examined the shape and structure of **Covid19\_Confirmed\_dataset.csv** and **worldwide\_happiness\_report.csv**.
2. Dropped irrelevant columns (e.g., Latitude, Longitude).
3. Aggregated data by country.

**Task 2: Maximum Infection Rate Calculation**

1. Used differentiation to compute daily changes in infection rates.
2. Identified the maximum daily infection rate for each country.

**Task 3: Data Integration**

1. Performed an inner join on COVID-19 and Happiness datasets to merge relevant indicators.
2. Adjusted data indices and ensured consistency.

**Task 4: Statistical Analysis**

1. Computed correlation matrices to quantify relationships between indicators like GDP and infection rates.

**Task 5: Visualization**

1. Generated scatterplots and regression lines to display relationships:
   1. **GDP per capita vs. Max Infection Rate**
   2. **Social Support vs. Max Infection Rate**
   3. **Healthy Life Expectancy vs. Max Infection Rate**
   4. **Freedom to Make Life Choices vs. Max Infection Rate**
2. Log-transformed infection rates to improve interpretability of visualizations.

**SUMMARY**

The analysis revealed several key insights into the relationship between socio-economic factors and COVID-19 infection rates:

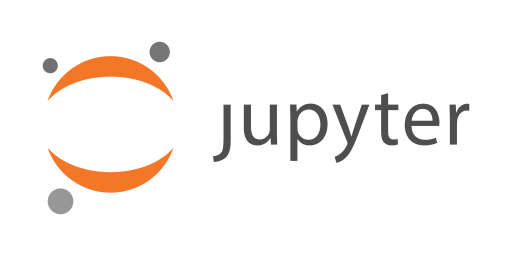
1. **GDP per Capita and Infection Rates**: A positive correlation was observed between higher GDP per capita and higher maximum infection rates. Wealthier countries tend to report more cases, which could be attributed to increased testing capabilities, better reporting infrastructure, and denser populations. However, this also suggests that economic resources might enable countries to better manage health crises, even though higher case numbers could be recorded due to more comprehensive data collection.
2. **Social Support and Infection Rates**: The analysis showed a moderate positive correlation between social support and infection rates. Countries with stronger social safety nets and healthcare systems, which are often reflected in higher social support scores, tended to fare better in controlling the virus's spread. This correlation hints at the protective effect of well-established healthcare systems and social policies in mitigating the spread of the virus.
3. **Healthy Life Expectancy and Infection Rates**: There was a moderate positive correlation between healthy life expectancy and the infection rates. Countries with higher healthy life expectancy often had more robust healthcare infrastructures, which likely contributed to lower fatality rates despite higher infection numbers. This suggests that healthcare quality plays a crucial role in managing both the spread of the virus and its consequences on public health.
4. **Freedom to Make Life Choices and Infection Rates**: The weakest correlation was found between the freedom to make life choices and maximum infection rates. This result suggests that, although freedom and personal autonomy are vital for overall well-being, they may not have a direct or strong influence on infection rates. The spread of COVID-19 may be more closely linked to other factors like healthcare accessibility and social support, rather than the degree of personal freedom.

**TOOLS USED**

THE TOOLS THAT WE HAVE USED IN THIS PROJECT ARE AS FOLLOWS:

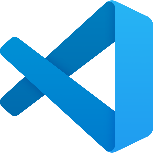
1. JUPYTER NOTEBOOK

FOR CODE COMPILATION



2. VSCODE

ADDITIONAL CODING TASKS AND DEBUGGING



3. MICROSOFT EXCEL

PREPROCESSING DATASETS



4. GITHUB

CREATING REPOSITORY AND DOCUMENTING THE PROJECT

