Project Objective:

This project is an exploration of the intricate relationship between the spread of the coronavirus within a country and the corresponding levels of happiness among its residents.

Utilized Datasets:

This project relies on two crucial datasets for a comprehensive analysis:

- **COVID-19 Dataset from John Hopkins University:** This dataset captures the cumulative confirmed cases per day in various countries.
- **World Happiness Report Dataset:** An annual publication by the United Nations, this dataset encompasses scores assigned by individuals across different countries, evaluating factors like freedom, life expectancy, and social support.

Data Preparation:

- **Importing the Covid19 Dataset:** We initiate the process by creating a variable named 'corona_dataset_csv' using the pandas library's 'read_csv' method, meticulously specifying the complete file path for data loading.
- **Exploratory Data Analysis (EDA):** An initial exploration entails reviewing the structure of the dataset by examining the first 10 rows with 'head(10)' and consistently validating the dataset's shape with 'corona_dataset_csv.shape' (266 rows and 104 columns).
- **Data Cleansing:** Identification and removal of extraneous columns such as latitude and longitude are crucial for precision. The 'drop' method, with 'inplace=True,' ensures immediate changes.
- **Data Aggregation:** The data, initially organized by provinces, is aggregated by countries through the strategic use of the 'group by' method on 'country/region' and the application of the 'sum' function. This results in the 'corona_dataset_aggregated' data frame, boasting 187 rows and 100 columns.
- **Visualization:** Cumulative confirmed cases for specific countries, exemplified by China and Italy, are thoughtfully visualized, employing 'plt. legend' for optimal labelling.
- **Focus on Initial Days:** Reviewing the cumulative confirmed cases plot from January 22nd to April 30th sets the stage for subsequent detailed analysis.

Spread Measurement and Analysis:

- **Daily Changes Analysis:** Extracting and plotting data for China unveils the daily changes, a critical component for understanding the spread.

- **Maximum Infection Rate Calculation:** To quantify the spread, we identify the maximum number of new confirmed cases. Employing the 'diff' method for the first derivative and subsequently finding the maximum using 'max' unveils a single-day peak in new cases.
- **Extending Analysis:** Applying this method to different countries, such as Italy and Spain, reveals their respective maximum infection rates.
- **Loop Implementation:** A loop is adeptly employed to extend this analysis to all countries, efficiently storing the maximum infection rates in a list. This list is then seamlessly added as a new column, 'Maximum Infection Rate,' to the 'corona_dataset_aggregated' data frame.
- **Data Presentation:** The resulting data frame is a comprehensive representation, juxtaposing cumulative cases with the newly added maximum infection rate column.
- Refined Dataframe: Creating a new data frame with only pertinent columns paves the way for subsequent analysis.

Integration and Comprehensive Analysis:

- **World Happiness Report Integration:** The 'Happiness_Report_CSV' variable is instantiated to load the World Happiness Report data, with 'pd.read_csv' serving as the conduit from the 'datasets' folder. Rigorous inspection using 'head()' identifies and subsequently drops unnecessary columns related to overall rank, score, generosity, and perception of corruption.
- **Index Optimization:** To streamline the merging process, the indices of the data frame transform, utilizing 'set_index' with 'inplace=True' to set the 'Country/Region' column as the indices.
- **Datasets Merging:** Ready for integration, the two datasets ('corona_data' and 'happiness_report') are merged, resulting in the 'data' data frame. For the number of countries in 'corona_data' being more, an inner join is executed using the 'join' method, storing the result in 'data'.
- **Holistic Information:** The 'data' data frame now encapsulates crucial information, including the maximum infection rate for each country and diverse life factors scored by its residents.

Correlation Exploration:

- **Correlation Matrix:** To delve deeper into the interplay between life factors and virus spread, a correlation matrix is meticulously crafted using the 'corr' method. Higher correlation coefficients signal robust positive correlations between variables.

Visualization of Insights:

- **Insightful Discovery:** The analysis uncovers compelling correlations, highlighting positive associations between the maximum infection rate and all life factors within our dataset.
- **Visualization Importance:** Acknowledging that analysis remains incomplete without visualization, the 'data' data frame, enriched with life factors and the maximum infection rate, serves as the canvas for plotting GDP per capita against the maximum infection rate.
- **Enhanced Visualization:** Seaborn's 'scatterplot' method is thoughtfully employed, with log scaling applied to the y-axis using NumPy's 'log' method to address scale differences.
- **Regression Line Application:** Elevating the visualization, Seaborn's 'regplot' method fittingly incorporates a regression line to the scatter plot, facilitating a nuanced understanding of the positive correlation between GDP per capita and the maximum infection rate.

Conclusion:

The nuanced analysis suggests a heightened susceptibility to COVID-19 infections among individuals residing in more developed countries compared to their less developed counterparts.