The "CapstoneAwolaja.ipynb" notebook performs an exploratory data analysis (EDA) and predictive modeling on a COVID-19 dataset. Here's a breakdown of the analysis, focusing on the visualizations, models, and key findings you've specified:

1. Data Loading and Initial Exploration

- The notebook starts by importing necessary libraries such as pandas, numpy, matplotlib and seaborn.
- The dataset is loaded from a CSV file named "*worldometer_data.csv*" into a pandas DataFrame.
- The notebook displays the first few rows of the DataFrame using ***df.head()*** to understand the data structure.
- It uses ***df.info()*** to show the data types and non-null counts for each column, revealing missing data in columns like 'NewCases', 'TotalDeaths', 'NewDeaths', and others.
- The notebook uses ***df.describe()*** to provide descriptive statistics of the dataset including the mean, standard deviation, and quartiles.
- Missing values are filled with 0 using ***df.fillna(0, inplace= True).***

2. Exploratory Data Analysis (EDA) Visualizations

- **Total Deaths and Recovered by Continent:**
  - The notebook calculates the total deaths and total recovered cases for each continent.
  - A bar chart is generated to visualize the total deaths and total recovered cases for each continent, allowing for a comparison of the pandemic's impact across different regions.
- **Top 10 Countries by Total Cases:**
  - The notebook identifies the top 10 countries with the highest number of total cases.
  - A bar chart is generated to display these top 10 countries, providing a clear visualization of which countries were most affected by the pandemic.
- **Top 5 Countries with the Least Cases:**
  - The notebook also identifies the 5 countries with the least number of total cases.
  - A bar chart is generated to visualize the countries with the least total cases, in order to provide a contrast to the most affected countries.
- **Top 10 Countries by Total Deaths:**
  - The notebook identifies the top 10 countries with the highest total deaths.
  - A bar chart is generated to display the top 10 countries with the highest death tolls.
- **Top 10 Countries by Total Recovered Cases:**
  - The notebook identifies the top 10 countries with the highest total recovered cases.

  - A bar chart is generated to visualize these top 10 countries, highlighting which countries have seen the most recoveries.

- **Correlation Heatmap:**
  - A heatmap is created to show the correlation between 'TotalCases', 'TotalDeaths', and 'TotalRecovered', helping visualize relationships between these variables.
- **Line Plot of Total Cases and Deaths Over Time:**
  - A line plot visualizes the progression of total cases and total deaths over the dataset's index, giving an overview of how these variables trended.
- **Filled Area Plot of Total Cases and Recovered:**
  - A filled area plot is created to show the relationship between total cases and total recovered over the dataset's index, showing both trends in the same plot.
- **Scatter Plots with Regression Lines:**
  - A scatter plot with a regression line is used to explore the relationship between 'Population' and 'TotalCases'.
  - Another scatter plot with a regression line is used to explore the relationship between 'Tot Cases/1M pop' and 'Deaths/1M pop'.
- **Kernel Density Estimate (KDE) Plots:**
  - KDE plots are generated to visualize the distributions of 'NewDeaths', 'NewRecovered', 'ActiveCases' and 'NewCases'.

3. Data Transformation and Feature Engineering

- The notebook calculates "Daily Growth Rate of Cases (%)" and "Daily Growth Rate of Deaths (%)" using the "NewCases", "TotalCases", "NewDeaths", and "TotalDeaths" columns.
- It calculates "Cases per 1M Population" using "TotalCases" and "Population".
- The notebook calculates "DeathRate" and "RecoveryRate" using "TotalDeaths", "TotalRecovered", and "TotalCases".
- A new feature named HighMortalityRisk is created based on whether a country's TotalDeaths exceed 480.

4. Predictive Modeling

- **Data Preparation:** The data is split into training and testing sets using train_test_split, with 50% of the data used for testing. The features used for the model are 'Population', 'TotalCases', 'TotalDeaths', and 'TotalRecovered' and the target variable is 'HighMortalityRisk'.
- **Linear Regression:**
  - A linear regression model is trained and used to predict HighMortalityRisk.

    - The notebook computes the mean squared error (MSE) and R-squared score for the linear regression model.
    - It is important to note, that this is not an appropriate model for this task.
- **Logistic Regression:**
    - A logistic regression model is trained to predict HighMortalityRisk.
    - The notebook calculates the Mean Squared Error, Mean Absolute Error and the R-squared Score.
    - A classification report is also printed for the logistic regression model.
    - The accuracy score of the model is also evaluated.
    - The notebook uses a confusion matrix and heatmap to evaluate the model's performance.
    - The notebook uses a bar plot to visualize the prediction results, showing the counts of 'Yes' and 'No' predictions.
- **Decision Tree Classifier:**
    - A decision tree classifier is trained to predict HighMortalityRisk.
    - The notebook calculates the Mean Squared Error and the accuracy of the model.
    - The prediction results are visualized using a pie chart.
- **Random Forest Classifier:**
    - A random forest classifier is trained to predict HighMortalityRisk.
    - The notebook calculates the accuracy score and classification report.
    - The prediction results are visualized using a pie chart.

5. Key Findings

- The EDA visualizations provide an overview of the pandemic's impact across different continents and countries.
- The modeling results suggest that predictive models can be used to predict countries at high risk of mortality, with the decision tree classifier and random forest classifier providing better results than logistic regression, and linear regression being inappropriate for this task.

In summary, "CapstoneAwolaja.ipynb" provides a comprehensive analysis of COVID-19 data, combining data visualization, statistical analysis, and predictive modeling to explore the patterns and relationships within the data. The notebook identifies key trends, geographical disparities, and factors related to mortality risk using a variety of methods.