ASSIGNMENT

QUESTION 1:

Choose a dataset: Browse Kaggle, World Bank, or other open-source repositories to find a dataset that interests you. Consider topics like economics, healthcare, weather, or social media.

Get your data: Download the dataset and store it in a format compatible with Python, such as CSV or Excel.

Write Python code: Create Python code to analyze the data and create visualizations. Use libraries like pandas, matplotlib, and seaborn for data manipulation, analysis, and visualization.

Include statistics: Implement descriptive statistics like describe and correlation analysis using corr from pandas to understand the data.

Create visualizations: Generate at least three different plots based on your analysis. Examples include:

Histogram/bar chart/pie chart for frequency distributions or categorical data.

Line/scatter graph for trends and relationships between variables.

Confusion matrix/heatmap/corner/box/violin plot for exploring relationships between multiple variables or model performance.

Write the report: Create a two-page PDF report summarizing your findings. Include:

Introduction: Briefly describe the dataset, research question, and methodology.

Data Exploration: Highlight key findings from descriptive statistics and correlations.

Data Visualization: Discuss each plot, explaining its purpose and key insights.

Conclusion: Summarize your findings and potential implications.

Create a GitHub repository: Host your Python code and report document in a public GitHub repository.

Link your repository: Include a clickable link to your GitHub repository in the report.

Remember:

Make your code well-documented with comments and explanations.

Ensure visualizations are clear and readable with informative labels and legends.

Proofread your report for clarity, grammar, and conciseness.

By following these steps, you can create a well-written report exploring and explaining relations within your chosen dataset. Good luck!

SOLUTION:

After browsing Kaggle, I found a suitable dataset on COVID-19 vaccination progress. The dataset provides information about vaccination progress in different countries, including the number of people vaccinated, the percentage of the population vaccinated, and various other metrics.

Dataset: COVID-19 World Vaccination Progress

Source: Kaggle

Format: CSV

This assignment will be answered in the following steps below:

- Get the Data: Download the dataset and store it in a format compatible with Python.
- Write Python Code: Create Python code to analyze the data and create visualizations using libraries like pandas, matplotlib, and seaborn.
- Include Statistics: Implement descriptive statistics like `describe` and correlation analysis using `corr` from pandas to understand the data.
- Create Visualizations: Generate at least three different plots based on the analysis.
- Write the Report: Summarize findings in a two-page PDF report, including introduction, data exploration, data visualization, and conclusion.
- Create a GitHub Repository: Host Python code and report document in a public GitHub repository.

I proceeded with downloading the dataset and writing Python code to analyze it. Then, I created the GitHub repository accordingly. Let's start with getting the data and writing the Python code.

The GitHub repository contains the Python code and report document for the analysis of the COVID-19 vaccination progress dataset. You can find it here.

[https://github.com/Teegee2000/data-set-analysis-assignmeeeent.git]

In the repository, you'll find:

- covid_vaccination_analysis.ipynb: Jupyter Notebook containing the Python code for data analysis and visualization.
- covid_vaccination_analysis.pdf: PDF report summarizing the findings of the analysis.
- data/: Directory containing the dataset file country_vaccinations.csv.
- README.md: Description of the project and instructions for running the code.

Report: Exploring COVID-19 Vaccination Progress

Introduction

The COVID-19 pandemic has prompted a global effort to vaccinate populations against the virus. In this report, we analyze a dataset on COVID-19 vaccination progress worldwide. The dataset provides information on the number of vaccinations administered, vaccination rates, and other relevant metrics across different countries.

Data Exploration

I began by loading the dataset and examining its structure. The dataset contains various columns, including country names, dates, total vaccinations, daily vaccinations, and vaccination rates. I utilized descriptive statistics to gain insights into the dataset's distribution and summary metrics.

The descriptive statistics revealed that the total number of vaccinations administered varies widely across countries, the daily vaccination rate shows fluctuations over time, as depicted in the line graph of daily vaccinations.

Data Visualization

I created three visualizations to further explore the dataset:

Histogram of Total Vaccinations: The histogram illustrates the distribution of total vaccinations administered across different countries. It provides an overview of the frequency distribution of vaccination efforts worldwide.

Line Graph of Daily Vaccinations Over Time: The line graph depicts the trend of daily vaccinations over time. It highlights the fluctuations in daily vaccination rates and allows us to identify periods of increased or decreased vaccination efforts.

Heatmap of Correlation Matrix: The heatmap visualizes the correlation matrix between variables in the dataset. It helps identify potential relationships between different metrics, such as the correlation between total vaccinations and daily vaccinations.

Conclusion

In conclusion, the analysis provides valuable insights into the COVID-19 vaccination progress worldwide. The dataset reveals significant variations in vaccination efforts across countries and highlights the dynamic nature of vaccination campaigns over time. Further analysis and exploration of the dataset can offer valuable insights for policymakers and public health officials in managing and optimizing vaccination strategies.

For the full code implementation and additional analysis, please refer to the GitHub repository linked below. https://github.com/Teegee2000/data-set-analysis-assignmeeeent.git