Final Project: Data Analysis with PostgreSQL, psycopg2, and JupyterLab

Project Overview

For the final project, you will demonstrate your ability to analyze data using PostgreSQL, psycopg2, and JupyterLab. This project will involve using an existing dataset, performing various operations to analyze the data, and integrating these operations with Python. You will focus on uncovering insights and providing recommendations or highlighting interesting findings based on the data.

Project Requirements

1. **Data Analysis with SQL**:
    - Write SQL queries to perform the following tasks:
        - Retrieve specific data using SELECT statements with various conditions.
        - Perform JOIN operations to combine data from multiple tables.
        - Use GROUP BY and aggregation functions (SUM, AVG, COUNT) to analyze data.
        - Implement subqueries and nested queries.
2. **Python Integration with psycopg2**:
    - Connect to the PostgreSQL database using psycopg2 in a Jupyter notebook.
    - Retrieve data using SQL queries within Python.
    - Implement error handling for database operations.
3. **Data Visualization**:

- Use Python libraries (e.g., Matplotlib, Seaborn) to create visualizations of the data retrieved from the database.
- Include at least three different types of visualizations (e.g., bar chart, line graph, pie chart).

4. **Reporting**:
    - Document your project in a Jupyter notebook.
    - Provide a clear explanation of your SQL queries and Python integration.
    - Include your visualizations and interpret the results.
    - Discuss how the findings can be used to provide actionable recommendations or highlight interesting patterns in the data.

Online Dataset Resources

Choose one dataset from the following online resources for your project:

1. **Kaggle**:
    - Website: Kaggle Datasets
    - Description: A wide variety of datasets in different domains such as finance, healthcare, sports, and more. Requires creating a free account.
2. **UCI Machine Learning Repository**:
    - Website: UCI ML Repository
    - Description: A collection of databases, domain theories, and datasets widely used in the machine learning community.
3. **Data.gov**:
    - Website: Data.gov
    - Description: The home of the U.S. Government's open data. Find data, tools, and resources to conduct research, develop web and mobile applications, and design data visualizations.
4. **Google Dataset Search**:
    - Website: Google Dataset Search

- Description: Enables the discovery of datasets stored across the web. This includes data published on the web, but also from repositories and individual researchers.

5. **Awesome Public Datasets**:
    - GitHub: Awesome Public Datasets
    - Description: A list of high-quality open datasets in a variety of domains shared on GitHub.

Specific Goals

1. **Providing Actionable Recommendations**:
    - Identify trends or patterns that can lead to strategic decisions.
    - Analyze demographic data to tailor services or products to specific groups.
    - Evaluate performance metrics and provide suggestions for improvement.
    - Determine factors that significantly affect outcomes (e.g., patient recovery rates, game performance).
2. **Highlighting Interesting Findings**:
    - Discover correlations between different variables.
    - Identify outliers and investigate their causes.
    - Detect seasonal trends or patterns.
    - Find geographic patterns and their impacts on the studied phenomena.

Suggested Steps

1. **Select a Dataset**:
    - Browse the provided online dataset resources.
    - Choose a dataset that interests you and download it.
2. **Prepare the Database**:
    - Use pgAdmin or SQL scripts to create tables and import your selected dataset into PostgreSQL.

3. **Write SQL Queries**:
    - Retrieve specific data from your dataset using SELECT statements with various conditions.
    - Perform JOIN operations to combine data from multiple tables if applicable.
    - Use GROUP BY and aggregation functions to analyze data.
    - Implement subqueries and nested queries for more complex analysis.
4. **Python Integration**:
    - Write a Python script to connect to the PostgreSQL database using psycopg2.
    - Execute the SQL queries from the Python script and fetch the results.
    - Handle errors during database operations.
5. **Data Visualization**:
    - Create visualizations to present the insights gained from your data analysis.
    - Use Python libraries like Matplotlib and Seaborn to create at least three different types of visualizations.
6. **Documentation using Jupyter Lab**:
    - Explain your approach to data analysis.
    - Document the SQL queries and their results.
    - Include Python code snippets and visualizations.
    - Interpret the visualizations and provide insights on actionable recommendations or interesting patterns.

Submission Guidelines

1. **Jupyter Notebook**: Submit a Jupyter notebook (.ipynb file) containing:
    - SQL queries for data analysis.
    - Python scripts for database operations using psycopg2.

- Data visualizations.
- Explanations and interpretations of your results.

2. **Presentation**: Prepare a presentation (10 - 15 minutes) summarizing your project, highlighting the key aspects of your data analysis, SQL queries, and insights gained from the data.

Evaluation Criteria

- **SQL Queries (30%)**: Complexity and correctness of SQL queries.
- **Python Integration (30%)**: Correct use of psycopg2 for database operations.
- **Documentation and Data Visualization (30%)**: Quality and variety of visualizations, insights drawn.
- **Presentation (10%)**: Clarity, completeness, and professionalism.

This final project will give you hands-on experience with analyzing data in a database, integrating with Python, and visualizing the results to gain insights that can lead to actionable recommendations or highlight interesting patterns. Good luck, and enjoy the process!