Stroke Data Analysis Project

Introduction

The aim of the Stroke Data Analysis Project was to analyze healthcare data for stroke cases and find the trends and relationship that exist among the key factors, including hypertension, heart disease, smoking status, and glucose levels. The dataset, healthcare-dataset-stroke-data.csv, has complete information about patient demographics, health conditions, and stroke events. The project used this dataset to derive actionable insights that can help healthcare professionals understand and reduce the risks of stroke.

Data Preprocessing

To ensure data quality, several preprocessing steps were undertaken. Missing values in the bmi column were replaced with the column's mean value, preserving the dataset's integrity while addressing inconsistencies. A thorough check for null values was conducted, and all missing values were resolved, resulting in a clean and comprehensive dataset ready for analysis.

Data Analysis

The data analysis phase brought out some critical patterns. A bar chart for stroke distribution indicated that a majority of the cases were non-stroke incidents. Further analysis indicated that hypertension, heart disease, and smoking status were major contributors to stroke cases. Surprisingly, patients who had never smoked or formerly smoked had higher stroke counts compared to active smokers. Moreover, the proportion of stroke cases was significantly higher in hypertensive patients, 13%, compared to non-hypertensive patients, 4%, indicating the grave contribution of hypertension to stroke occurrences.

Data Storage

Clean data was then saved locally and uploaded to the PostgreSQL database on AWS RDS to make it efficient for storing and retrieving data. SQL queries were used to derive insights from this data and also verify that it was intact within the cloud with security and integrity.

Visualizations

Trends and patterns were also best communicated with data visualizations. The bar chart showing stroke proportion comparing hypertensive patients to non-hypertensive patients indeed made stark differences. A bar chart then showed smoking status among stroke patients, which indicates the most affected being from the "never smoked" or "formerly smoked" categories. The bar chart showed that the average glucose level of stroke patients is significantly higher than that of non-stroke patients. Stroke cases by gender showed a little higher frequency in females compared to males.

Summary of Results

The analysis provided the important insights that hypertension was a very critical factor for stroke cases, smoking history was a crucial determinant, and high glucose level was highly found in the stroke patients. These point to the need to address the issue of hypertension, manage the habit of smoking, and control glucose levels as vital ways of preventing stroke.

Conclusion

In summary, this project illustrated how data analysis techniques are applied to gain valuable healthcare insights. Integration with Python for data manipulation, PostgreSQL for database management, and effective visualizations all provide a comprehensive, data-driven approach toward the analysis and prevention of stroke risks. These results give useful recommendations to healthcare providers in the identification of priority areas for effective interventions to achieve improved patient outcomes.