Stroke Risk Analysis

Exploratory Data Analysis and Dashboard Summary

1. Dataset Overview

This project analyzes a healthcare dataset containing demographic, lifestyle, and clinical variables related to stroke occurrence. The objective is to clean the data, create interpretable analytical features, and explore patterns associated with stroke risk using visual analytics.

2. Data Cleaning and Feature Engineering

After understanding the dataset structure and variables, a data-cleaning process was applied while preserving all original categories and meanings.

- Missing values were identified using standard structural checks (`info`, `describe`, `isnull`).
- Missing values in **Body Mass Index (BMI)** were imputed using the **median**, as BMI is a continuous variable with a skewed distribution.
- The **age** variable was divided into five age groups (Child, Young Adult, Adult, Middle Age, Senior) to improve interpretability.
- A **Risk Score** was computed by aggregating key clinical risk factors (hypertension, heart disease, high glucose level, high BMI, and older age).
- Based on this score, individuals were classified into **Low**, **Medium**, and **High** risk levels.

No semantic changes were made to categorical variables (e.g., gender, smoking status), ensuring data integrity and reproducibility.

3. Exploratory Data Analysis

Exploratory analysis was conducted to understand distributions and relationships without introducing additional transformations.

- Descriptive statistics were examined for numerical variables such as age, BMI, and average glucose level.
- Frequency distributions were reviewed for categorical variables including gender, work type, smoking status, and residence type.
- Correlation analysis using the Pearson coefficient was applied to numeric variables to explore linear relationships.
- Visual analysis focused on **rates and proportions** rather than raw counts to allow fair comparison across groups.

4. Dashboard Insights

The Tableau dashboard highlights several key patterns:

- Stroke risk increases sharply with age, with individuals aged **65 years and older** showing the highest stroke rates.
- Stroke rates among **males and females are similar**, suggesting that clinical and age-related factors are stronger determinants than gender alone.
- A higher number of stroke cases is observed among individuals working in the **private sector**, largely reflecting the larger size of this group in the dataset rather than higher individual risk.
- The **Risk Score** effectively summarizes multiple health conditions and clearly differentiates low-, medium-, and high-risk populations.

5. Dashboard Title

Stroke Risk Analysis Dashboard

6. Notes on Interpretation

- All findings are observational and descriptive.
- BMI is used as a population-level indicator and does not represent a clinical diagnosis.
- Comparisons between groups are based on proportions to avoid bias from unequal group sizes.

References

- World Health Organization – Stroke Fact Sheet
  https://www.who.int/news-room/fact-sheets/detail/stroke
- World Health Organization – Body Mass Index (BMI)
  https://www.who.int/data/gho/data/themes/theme-details/GHO/body-mass-index-(bmi)
- Pandas Documentation – Handling Missing Data
  https://pandas.pydata.org/docs/user_guide/missing_data.html
- Centers for Disease Control and Prevention – Data Visualization Guidelines
  https://www.cdc.gov/dataviz/index.html