DAI-101 Assignment-1

Siddh Agrawal (23116090) , ECE

This report describes the process of data import, cleaning, exploratory data analysis (both univariate and multivariate), and visualization performed on a dataset related to stroke risk factors.

Data Import and Preparation

* Datasetimported directly from Google Drive using the Python package gdown, which downloads the CSV file into a local file named data.csv . After loading with Pandas, the first few rows are reviewed using the head() method.
* This dataset contains 5110 rows and 12 columns and checks data types of each column using df.dtypes. The columns consist of both numerical and categorical features.  
  Additionally, the code standardizes the column names by converting them to title case and later renames specific columns (for example, changing “Id” to “ID” and “Bmi” to “BMI”) for consistency.
* The script distinguishes numerical from categorical data by using Pandas’ select\_dtypes function.

Handling Missing Values and Duplicates

* **Missing Data Treatment:**  
  Missing values in numerical columns are replaced by the median of each respective column, ensuring that any skewed distribution does not overly influence the imputation. For categorical columns, the most frequently occurring (mode) value is used to fill in missing entries. These steps ensure that the dataset remains complete for subsequent analysis.
* **Duplicate Records:**  
  The code checks for and removes any duplicate rows using the duplicated() function. In this dataset, no duplicates remain after this step, ensuring uniqueness of observation.

Outlier Treatment

* **Detection Method:**  
  For each numerical column, the first (Q1) and third quartiles (Q3) are computed, and the interquartile range (IQR) is determined. Lower and upper bounds are defined using the formula:

Lower Bound=Q1−1.5×IQR and Upper Bound=Q3+1.5×IQR

These calculations help identify outliers in each variable

* **Removal Strategy:**  
  The code removes data points that fall outside the calculated lower and upper bounds from the DataFrame. As a result, after outlier elimination, the dataset’s size decreases from 5110 rows to 3788 rows. Boxplots are generated for each numerical column after treatment to visually confirm the absence of extreme values.

Text Standardization of Categorical Variables

* **Cleaning Process:**  
  The categorical columns are standardized by converting all text to lowercase and then removing any extra whitespace. Specific corrections are applied to ensure consistency in labels (for instance, converting "male" to "Male," "female" to "Female," "urban" to "Urban," and "rural" to "Rural").
* **Unique Value Inspection:**  
  For each categorical variable, the unique values are printed.

Exploratory Data Analysis (Univariate Analysis)

* **Data Summary:**  
  The notebook uses df.info() and df.describe() to provide an overall summary of the dataset’s structure, data types, and key statistical measures (mean, standard deviation, quartiles, etc.) for numerical fields.
* **Visualizations – Histograms and Countplots:**  
  Histograms for all numerical columns are plotted with 20 bins each, allowing the viewer to assess data distributions across variables such as Age, BMI, and Average Glucose Level. Similarly, count plots are generated for categorical features to visualize the frequency distribution of each category.
* **Stroke Frequency by Age:**  
  The code creates a pair of histograms for the “Age” variable, comparing individuals with (stroke = 1) and without (stroke = 0) strokes. These plots, complete with kernel density estimates (KDE), help reveal differences in age distribution between the two groups.
* **Specific Analysis on BMI:**  
  The skewness value of the BMI variable is computed (approximately 0.336), indicating a slight right-skew. A boxplot of BMI is also generated to visually inspect the central tendency and spread after outlier treatment.

Multivariate Analysis

* **Violin Plot Analysis:**  
  A violin plot is used to compare the distribution of BMI across different categories in the “Gender” column, highlighting distribution differences across male, female, and other categories.
* **Crosstab Heatmap:**  
  A heatmap based on a cross-tabulation between “Residence\_Type” and “Smoking\_Status” is plotted. This visualization provides insight into the distribution of smoking behavior across urban and rural residents.
* **Pair Plot and Correlation Heatmap:**  
  A pairplot is generated for all numerical variables, displaying their pairwise relationships, which can help in detecting trends and potential correlations. In addition, a detailed correlation heatmap (using only non-constant columns) is created with annotations, offering a clearer view of how numerical variables interrelate (for example, relationships among age, hypertension, heart disease, average glucose levels, and BMI).
* **Grouped Means Analysis:**  
  In last, the code iterates through every combination of a categorical and numerical variable to calculate and display the mean values. Barplots for each grouping are used to illustrate how numerical attributes vary when segmented by categories like Gender, Work\_Type, and Smoking\_Status. This grouped analysis can reveal important differences in average health metrics across different segments of the dataset.