**EDA | Logistic regression-heart disease prediction**

[**DATA-LINK**](https://www.kaggle.com/datasets/dileep070/heart-disease-prediction-using-logistic-regression)

**Dataset Description**:

The dataset contains information on various medical attributes of individuals, providing insights into cardiovascular health. It includes data from 4,238 individuals, predominantly adults, focusing on risk factors associated with coronary heart disease (CHD) over ten years. The features included are as follows:

1. male: Binary variable indicating gender (0 for female, 1 for male).

2. age: Age of the individual.

3. education: Education level of the individual.

4. currentSmoker: Binary variable indicating current smoking status (0 for non-smoker, 1 for smoker).

5. cigsPerDay: Number of cigarettes smoked per day.

6. BPMeds: Binary variable indicating whether the individual is on blood pressure medications.

7. prevalentStroke: Binary variable indicating whether the individual had a prevalent stroke.

8. prevalentHyp: Binary variable indicating prevalent hypertension.

9. diabetes: Binary variable indicating whether the individual has diabetes.

10. totChol: Total cholesterol level.

11. sysBP: Systolic blood pressure.

12. diaBP: Diastolic blood pressure.

13. BMI: Body mass index.

14. heartRate: Heart rate.

15. glucose: Glucose level.

16. TenYearCHD: Binary variable indicating the likelihood of developing CHD within ten years.

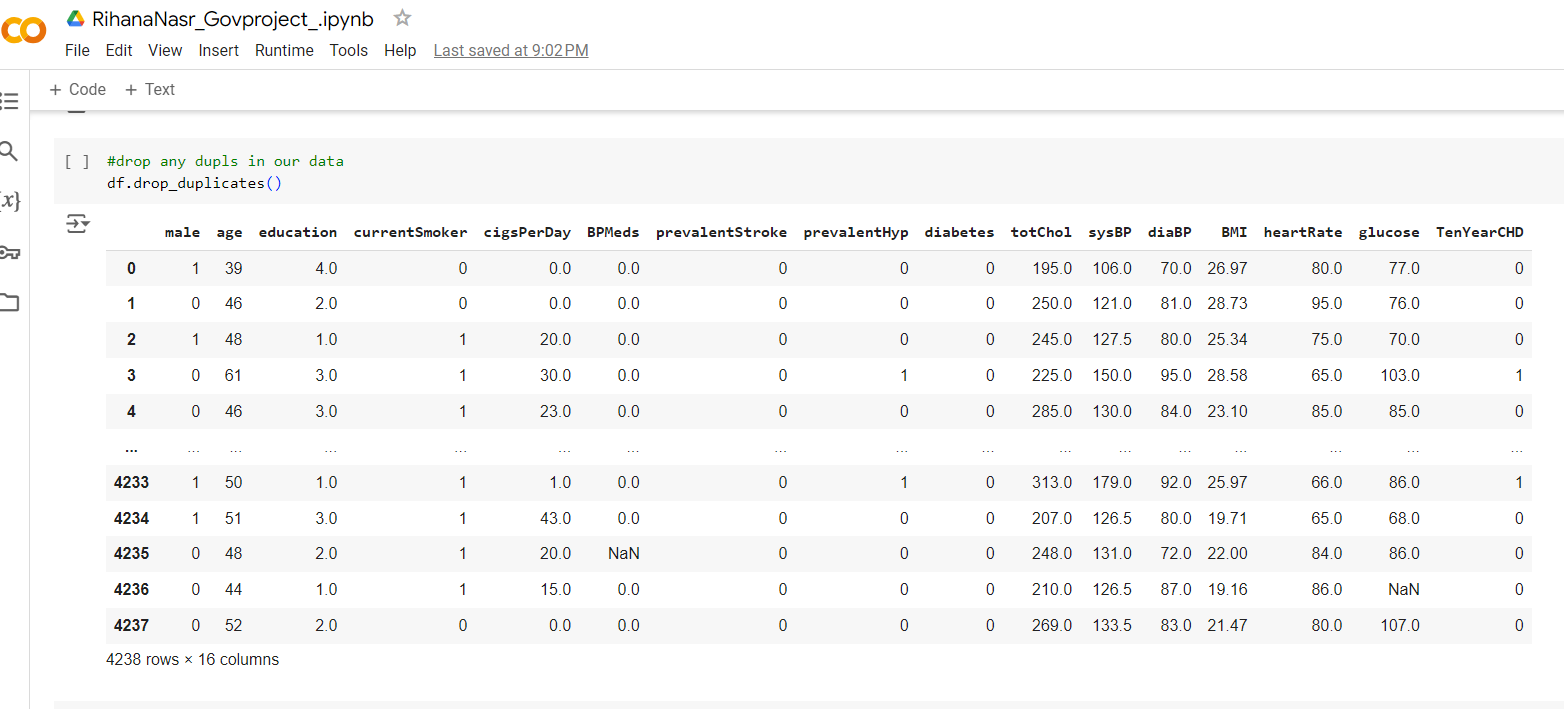
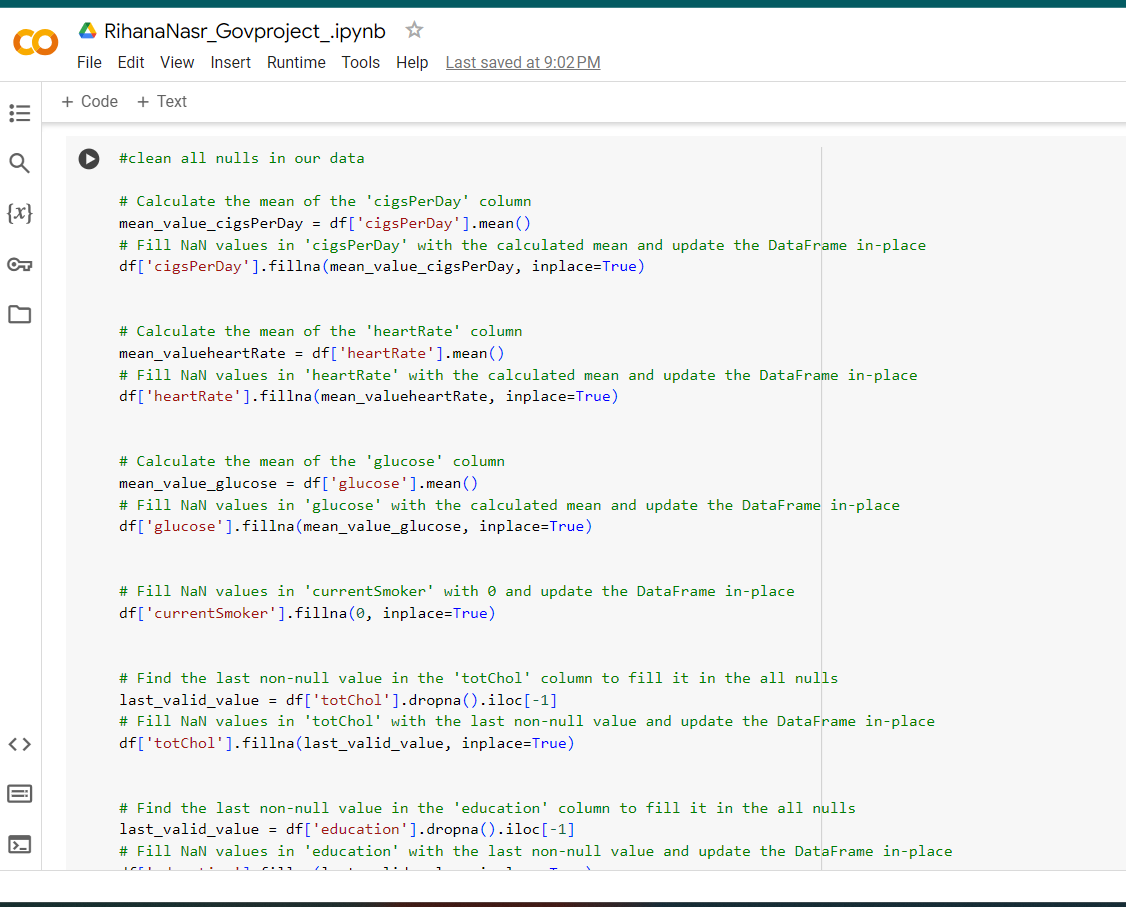
**Problem Definition**:

The dataset may contain various issues such as missing values (nullable columns), outliers, and inconsistencies. These issues can affect the accuracy and reliability of future analyses and predictions based on the dataset. For example, missing values in important features like cholesterol levels or blood pressure can skew statistical analyses and machine learning models. Similarly, outliers in numerical variables might lead to biased results.

**Methodology**:

1. Handling Missing Values: For nullable columns, we can fill missing values using appropriate methods such as mean imputation for numerical variables and mode imputation for categorical variables.

2. Data Cleaning and Validation: We will validate the data for consistency and correctness. For example, ensuring that age values are within a reasonable range and categorical variables have valid entries.



**Results**:

After applying the methodology described above, the dataset will be cleaned and prepared for analysis. Missing values will be filled, outliers will be addressed, and data integrity will be ensured. Code snippets for data preprocessing steps will be provided along with visualizations to illustrate the changes made to the dataset.

**Conclusion**:

By addressing the data issues and preparing a clean dataset, we can conduct meaningful analyses to explore the relationships between various medical attributes and the likelihood of developing coronary heart disease. This can lead to valuable insights for preventive healthcare strategies and personalized patient care.