# Exploratory Data Analysis (EDA) for Patient Diagnostic

*Sarthak Khandelwal*

## 1. Introduction

The Jupyter Notebook titled **eda.ipynb** conducts an **Exploratory Data Analysis (EDA)** on a patient diagnostic dataset. The analysis includes loading the data, cleaning and exploring demographic and diagnostic data, identifying a prospective study cohort, and investigating the relationship between hypertension and diabetes.

## 2. Loading and Cleaning

The script begins by loading three datasets: **demo.csv**, **dx.csv**, and **dx_map.csv**. The data is loaded using the Pandas library, and initial information is displayed using the info() method. It is noted that **birth_date** and **death_date** columns need to be converted to datetime format, which is promptly addressed.

## 3. Descriptive Statistics

Descriptive statistics are provided for both demographic and diagnostic data using the describe() method. The 5 number summary statistics help understand the central tendency and distribution of the data.

## 4. Data Merging

The three datasets are merged into a single dataframe (df_merged) using patient IDs and diagnostic codes. Duplicate columns are removed, and appropriate column renaming is performed.

## 5. Prospective Study Cohort

A prospective study cohort is identified, consisting of living patients aged 18 or older. The script calculates age for each patient and filters the dataset accordingly. A box plot is used to visualize the age distribution, revealing an outlier aged 129. This outlier is removed based on the assumption of erroneous data.

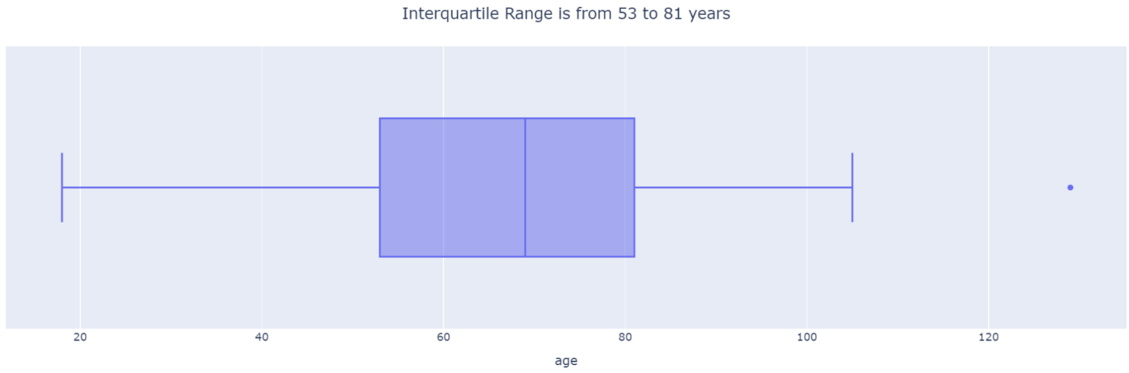

Fig.1: Age Distribution of the cohort of living patients

## 6. Data Exploration

The script explores missing data, removes unnecessary columns, checks for duplicates, and corrects case sensitivity in the gender column. Visualization includes a pie chart of gender distribution, bar charts for site and dx_method distribution, and a box plot for age distribution.

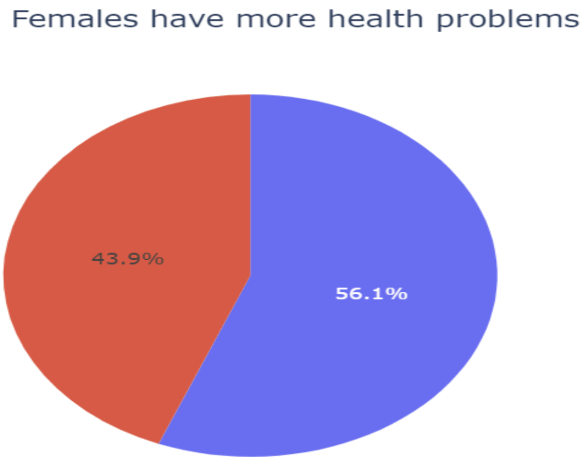

Fig.2: Gender Distribution of the Cohort

## 7. Relationship Between Hypertension and Diabetes

The script identifies patients with hypertension, diabetes, or both by filtering the dataset. A bar chart visually represents the overlap between patients with hypertension and diabetes. Additionally, a count of patients with both diagnoses is calculated, resulting in 158 individuals.

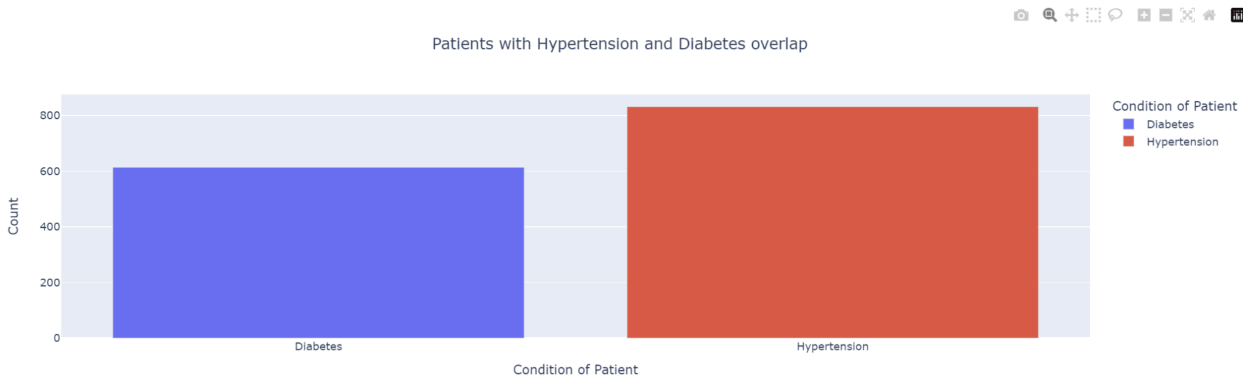

Fig.3: Distribution of Patients with Diabetes and Hypertension

# 8. Age Distribution for Diabetes and Hypertension

Two separate dataframes (Diabetes_df and Hypertension_df) are created for patients with diabetes and hypertension, respectively. Histograms are plotted to visualize the age distribution for both conditions.

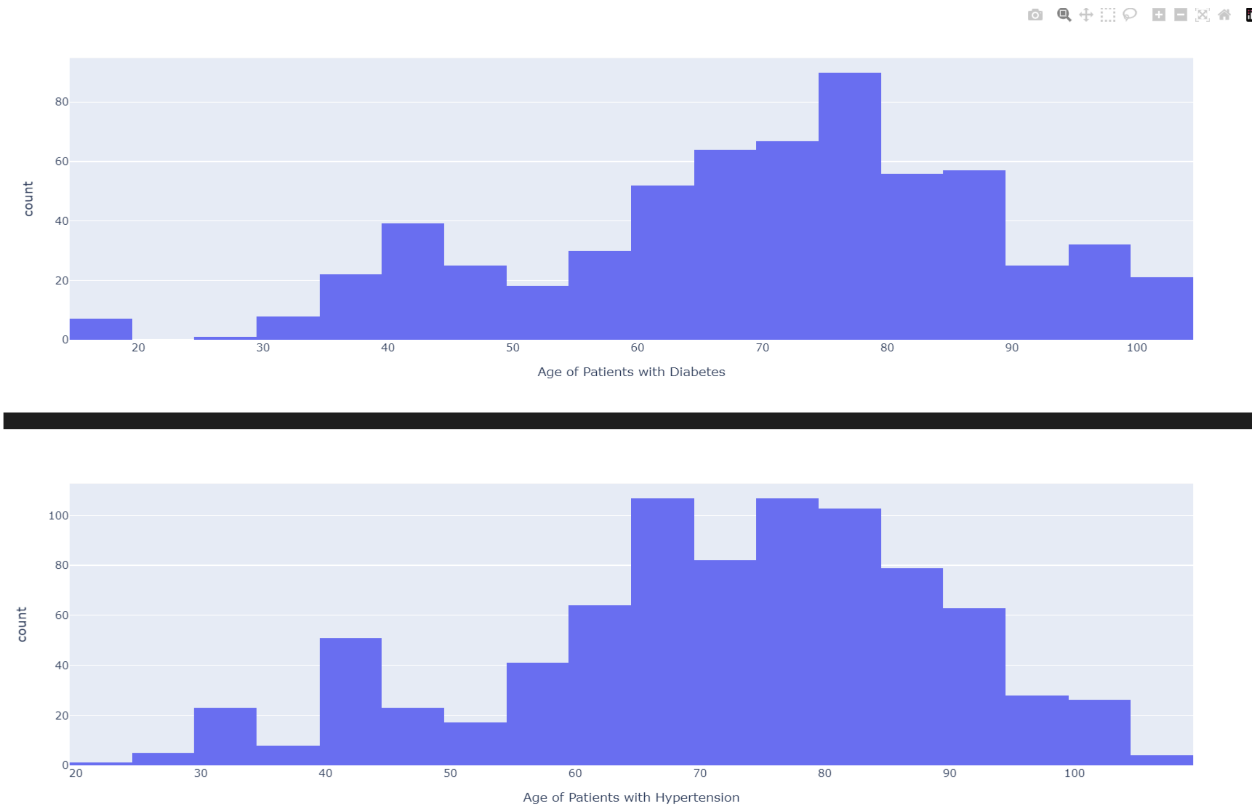

Fig.4: Distribution of Age of Patients with Diabetes and Hypertension

## 9. Conclusion

The EDA script successfully explores the patient diagnostic dataset, identifies a prospective study cohort, and investigates the relationship between hypertension and diabetes. Visualizations and statistical summaries enhance the understanding of the data, providing valuable insights for further analysis and decision-making.

## 10. Future Work

Future work includes tasks such as examining the number of people with hypertension and diabetes compared to those with other diseases. This could lead to a more comprehensive understanding of the dataset and potential correlations between various health conditions.