REPORT OF ASSIGNMENT -1

CHANDAN KUMAR
Enrollment Number: 23116027
Branch: ECE

1. Introduction

This report provides a comprehensive data analysis of the patient dataset. The dataset contains both numerical and categorical variables, including patient demographics, health metrics, and lifestyle factors. The analysis is structured into three major parts: Data Cleaning, Exploratory Data Analysis (EDA), and Multivariate Analysis.

2. Data Cleaning

The dataset was pre-processed to handle missing values, duplicates, and outliers.

2.1 Handling Missing Values

- **Numerical Variables:** Missing values were replaced with the mean of the respective column.
- **Categorical Variables:** Missing values were filled using the mode (most frequent category).

2.2 Removing Duplicates

- Any duplicate rows in the dataset were identified and removed.

2.3 Detecting and Handling Outliers

- Interquartile Range (IQR) Method was used to detect outliers.
- Outliers were replaced with NaN, and then missing values were filled using the mean.

2.4 Standardizing Categorical Values

- String formatting issues (e.g., extra spaces, different capitalizations) were corrected.
- Typos in categorical data were fixed to ensure consistency.

3. Univariate Analysis

Univariate analysis focuses on summarizing individual variables separately.

>>> Summary Statistics

The numerical variables were analysed using descriptive statistics, including mean, median, standard deviation, and skewness. Key insights:

- Age, Height, Weight, and BMI show a normal distribution with slight variations.
- The presence of outliers in height and weight was observed.
- BMI values exhibit a wider range, indicating variability in body composition.

>>> Histograms of Numerical Variables

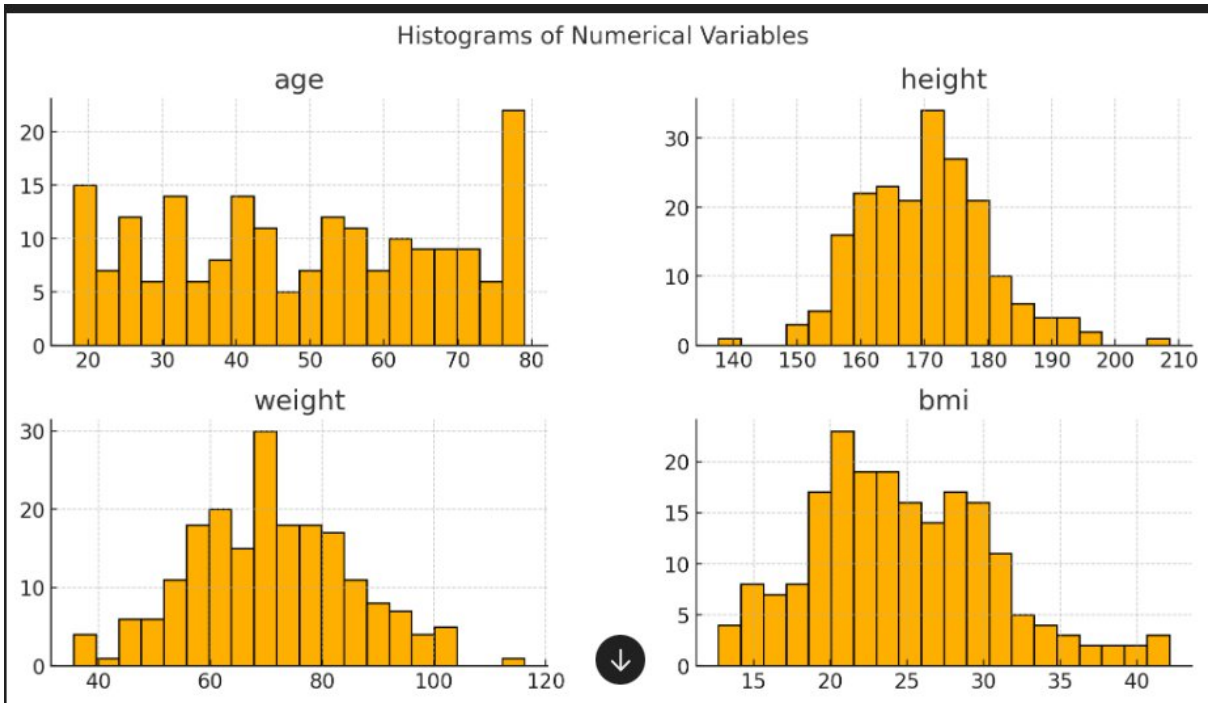
Histograms were plotted to visualize the distribution of numerical variables. Observations include:

- Age is right-skewed, indicating that there are more younger individuals in the dataset.

REPORT OF ASSIGNMENT -1

CHANDAN KUMAR
Enrollment Number: 23116027
Branch: ECE

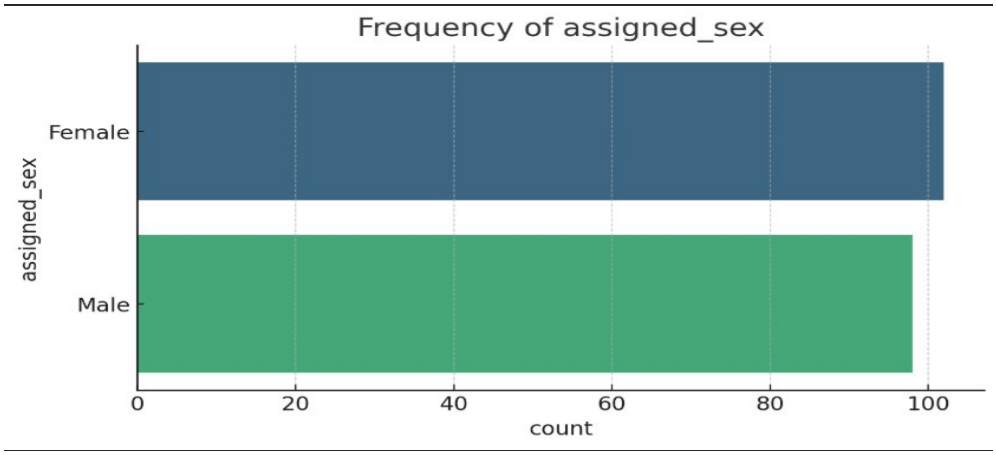
- Height and Weight follow a near-normal distribution but with some extreme values.
- BMI distribution shows a slight peak around normal weight ranges, suggesting most patients are in a healthy range.



>>> Frequency Distributions of Categorical Variables

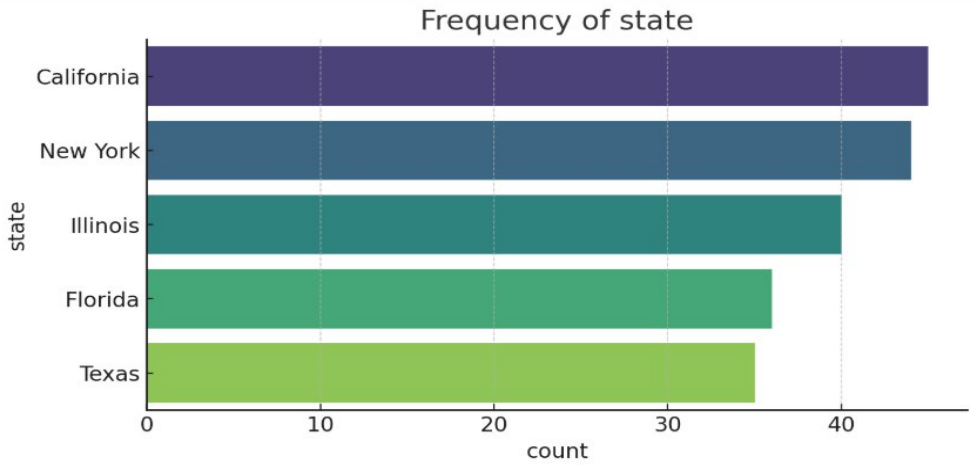
Bar charts were used to analyse categorical variables such as assigned sex, state, and country. Key observations:

- The dataset has a higher representation of one gender in the "assigned_sex" column.
- The state and country distributions show that the data is not evenly distributed across all locations, with some states appearing more frequently.



REPORT OF ASSIGNMENT -1

CHANDAN KUMAR
Enrollment Number: 23116027
Branch: ECE



4. Bivariate Analysis

Bivariate analysis examines relationships between two variables.

>>> Correlation Matrix (Numerical Variables)

A heatmap was used to visualize the correlation between numerical variables:

- Height and Weight show a strong positive correlation, meaning taller individuals tend to weigh more.

BMI and Weight have a strong correlation, as BMI is derived from weight and height.

Age has a weak correlation with BMI, suggesting that BMI does not significantly change with age.

>>> Boxplot of BMI by Assigned Sex

A **boxplot** was used to compare BMI distributions across genders. Observations:

- **Median BMI values** are slightly different for different genders.
- **Outliers are present**, indicating some individuals have extremely high or low BMI values.

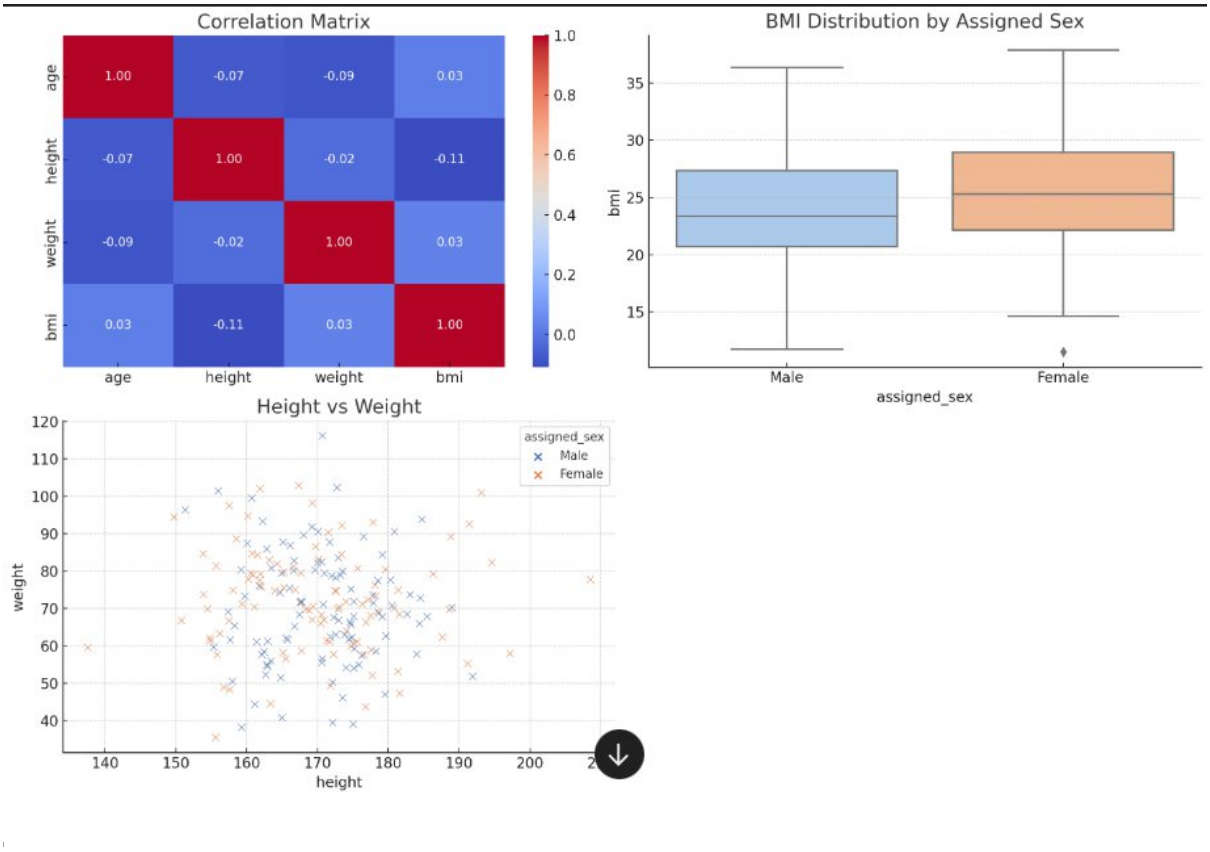
>>> Scatter Plot of Height vs. Weight

A **scatter plot** was used to analyse the relationship between height and weight:

- There is a clear **linear relationship**, confirming the correlation seen in the heatmap.
- Gender-based colour differentiation shows that height and weight distributions differ slightly between groups.

REPORT OF ASSIGNMENT -1

CHANDAN KUMAR
Enrollment Number: 23116027
Branch: ECE



5. Multivariate Analysis (Multiple Variables)

>>> Pair Plots

- Pairwise relationships between numerical variables were visualized.
- Clusters in BMI and weight were observed, suggesting distinct patient groups.

>>> Correlation Heatmap

- A heatmap visualizing correlations among multiple variables confirmed strong relationships between weight, height, and BMI.

>>> Grouped Comparisons

- Boxplots were used to analyse age, height, and weight across assigned sexes and countries.

6. Conclusion

- Data cleaning significantly improved the dataset's quality, ensuring accurate analysis.
- EDA revealed key trends in patient demographics and health indicators.
- Multivariate analysis helped identify patterns, such as the strong correlation between weight, height, and BMI.