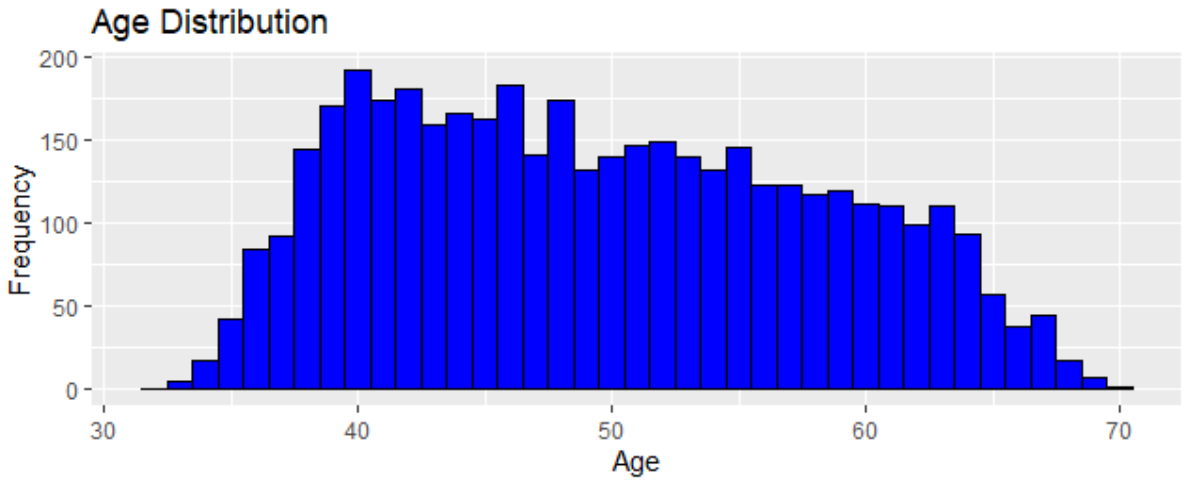
# I. Exploratory Data Analysis

## Introduction:

The Framingham Heart Study has been pivotal in elucidating the relationships between lifestyle, clinical measures, and cardiovascular health. This Exploratory Data Analysis (EDA) report delves into the wealth of data collected as part of this longitudinal study, with the primary objective of uncovering patterns, trends, and correlations within the cardiovascular health domain. The dataset encompasses a wide array of variables, from demographic information to detailed medical history and biometric data, making it a comprehensive source for understanding the risk factors associated with coronary heart disease (CHD). The EDA serves as the bedrock of this endeavor. It is the analytical process that allows us to summarize the main characteristics of the data, often using visual methods. Through EDA, we aim to discover underlying structures, detect outliers and anomalies, test assumptions, and check for preliminary relationships between variables. Such insights not only pave the way for more advanced statistical analysis and predictive modeling but also help inform clinical understanding and public health strategies. In this report, we will scrutinize the distribution of participants' ages, examine total cholesterol levels across genders, and explore the relationship between age and total cholesterol. Each of these factors—age, cholesterol, and gender—has been identified as a significant contributor to the risk of developing CHD, and by thoroughly analyzing their interplay, we aim to contribute to the collective understanding and ongoing discussion in cardiovascular health research.

## Age Distribution:

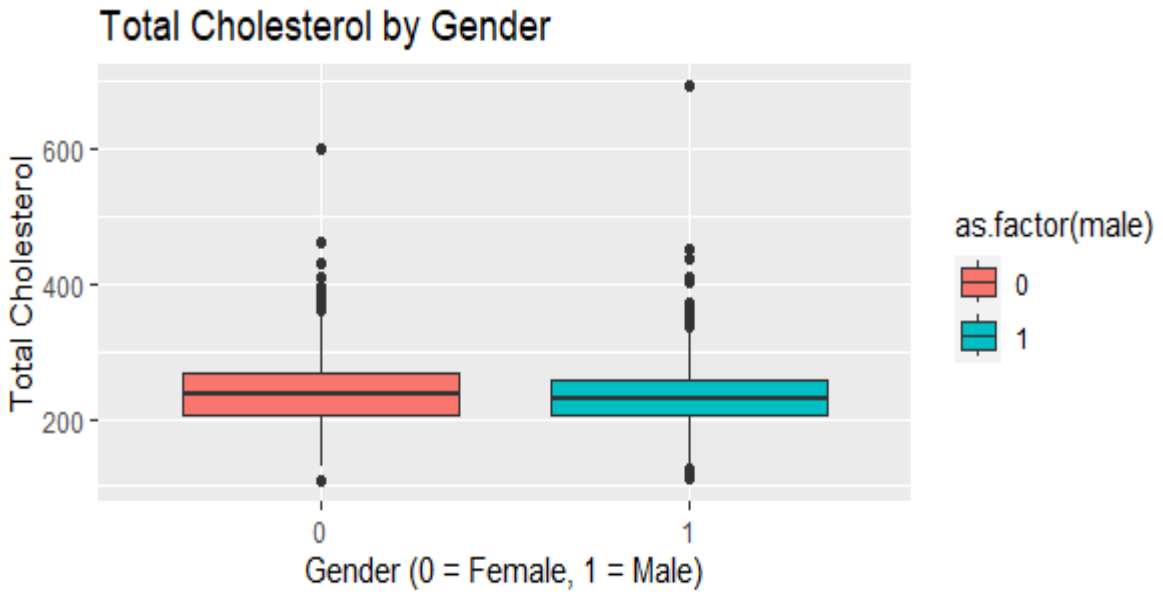


The histogram depicting the age distribution of the Framingham Heart Study dataset provides some important insights:

- **Overall Distribution**: The histogram displays a roughly bell-shaped distribution with a slight right skew. This skewness indicates that there are more older individuals in the dataset compared to younger ones.
- **Mid-Age Representation**: There's a noticeable peak in the frequency for ages in the mid-range (around 50-60 years). This suggests that middle-aged adults are well-represented in the dataset, which is a critical age range for the onset of CHD risk factors.
- **Older Age Tail**: The right skew or tail suggests a gradual decrease in frequency as age increases past the median, which is expected due to the decreasing population size in higher age brackets.
- **Implications for CHD Risk**: Since age is a significant risk factor for CHD, the representation of a higher number of older individuals might mean the dataset could have a higher prevalence of CHD or CHD risk factors. This age distribution is important for understanding the generalizability of the study's findings to the broader population. If the study aims to predict CHD risk in the general population, the model would need to account for this skew and potentially use age stratification or weighting in the analysis.
- **Considerations for Modeling**: For predictive modeling, the age distribution's shape could affect the performance of the model. Models that assume a normal distribution of continuous predictors may need to adjust for this skew. Also, given that the dataset doesn't have many younger individuals (less than 40 years old), the predictive accuracy for this age group might be lower.

The histogram serves as a crucial visual check for the data's distribution. It can inform not only the choice of statistical tests and models that assume normality but also the need for age-adjustment strategies when considering the risk of CHD across different age groups.

**Total Cholesterol by Gender :**



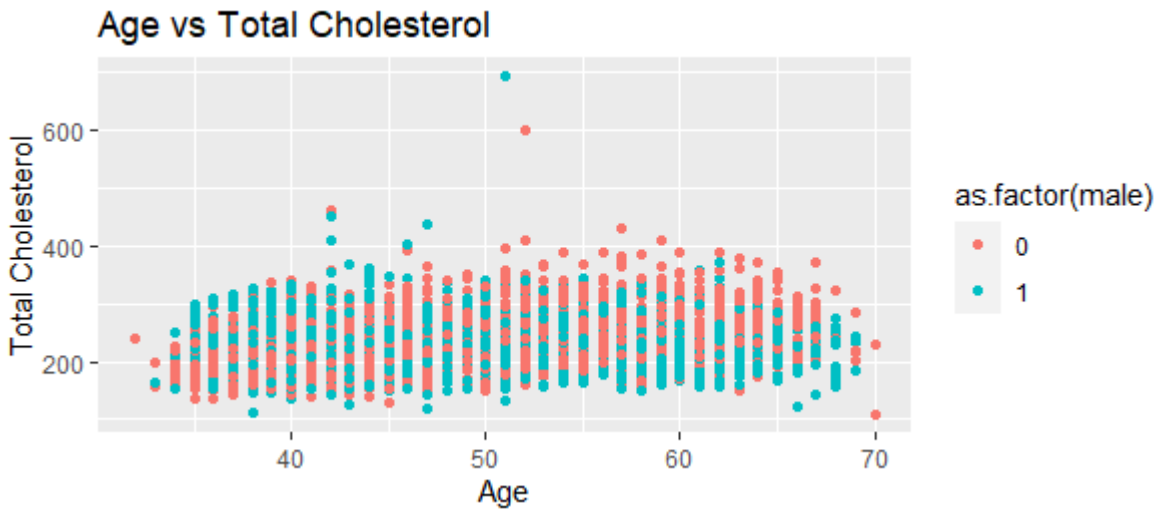

The box plot of total cholesterol by gender from the Framingham Heart Study dataset offers valuable insights into how cholesterol levels, a critical risk factor for coronary heart disease (CHD), vary between genders:

- **Gender Differences**: The median total cholesterol level for both genders appears to be around the same value, suggesting that the central tendency for cholesterol is similar for males and females in this cohort. However, the box plot can show whether there's a significant difference in the variability of cholesterol levels between genders.
- **Interquartile Range**: If one gender shows a wider interquartile range (the box's height), it indicates more variability in cholesterol levels within that group. For example, a wider box for females could suggest a broader distribution of cholesterol levels compared to males.
- **Outliers**: The presence of outliers (individual points beyond the whiskers of the box plot) indicates individuals with exceptionally high or low cholesterol levels. These cases could represent individuals with genetic predispositions, dietary habits, or other factors that lead to atypical cholesterol levels. Outliers can affect the average and need to be taken into consideration in risk prediction models.

- **Clinical Implications**: Since total cholesterol is a factor in developing CHD, understanding its distribution across genders can influence gender-specific screening strategies or interventions.
- **Predictive Modeling Considerations**: For the predictive model, differences in cholesterol distributions by gender may warrant separate models or inclusion of an interaction term between cholesterol and gender to accurately estimate CHD risk.

This box plot is critical for identifying not just the central trends, but also for assessing the spread and outliers in cholesterol levels, which may have implications for personalized medicine and public health policies targeting CHD prevention.

## Age vs. Total Cholesterol




The scatter plot illustrating the relationship between age and total cholesterol, with points colored by gender, provides several key insights:

- **Age Relationship**: The distribution of points does not indicate a strong or clear trend between age and total cholesterol levels. This lack of a visible trend suggests that, within this population, age may not be a linear predictor of cholesterol levels. However, this does not rule out age as a risk factor for CHD, as the relationship could be non-linear or influenced by other factors not displayed in the plot.

- **Gender Distribution**: The coloring distinguishes between male (blue) and female (red) participants. There doesn't appear to be a distinct pattern separating the genders, indicating that within this cohort, both genders share similar cholesterol profiles when viewed against age.
- **Variability and Outliers**: There is substantial overlap in cholesterol levels among the age groups, with a consistent spread of values from low to high. Notables are several potential outliers with very high cholesterol levels, which could be of particular interest in understanding extreme risk factors or could indicate data errors.
- **Clinical Interpretation**: Clinically, total cholesterol levels can vary due to a combination of genetic, dietary, and lifestyle factors. The absence of a strong correlation with age in this plot underscores the importance of considering multiple variables when evaluating the risk of CHD. Additionally, the overlap between genders suggests that sex-specific differences in cholesterol levels may not be pronounced in this population.
- **Implications for CHD Risk Modeling**: When developing a predictive model for CHD risk, it would be important to consider the interaction between age and cholesterol levels, and possibly other variables, to capture the complexity of CHD risk factors. Models might benefit from including non-linear terms or stratification to account for the variability in cholesterol levels across ages.

This visualization reinforces the importance of multivariate analysis in understanding CHD risk, as it suggests that the relationship between cholesterol levels and age is not straightforward and is likely influenced by a multitude of factors.

## EDA Conclusion:

Our exploratory data analysis has provided insightful glimpses into the age and cholesterol variables within the Framingham dataset, reinforcing their complex relationship with coronary heart disease risk. The distributions and patterns observed set the stage for more sophisticated modeling and emphasize the importance of considering multiple factors in CHD risk assessment. Moving forward, these findings will guide the development of predictive models aimed at improving cardiovascular health outcomes.

## II. Data preparation and cleaning

The data preparation and cleaning processes applied to the "framingham.csv" dataset were executed through a series of carefully planned steps in R. Initially, the script addressed missing values by employing median imputation for continuous variables like 'cigsPerDay', 'BPMeds', 'totChol', 'BMI', 'heartRate', and 'glucose', ensuring robustness against outliers. For the categorical variable 'education', mode imputation was utilized, replacing missing entries with the most frequent value. To mitigate skewness and normalize distributions, log transformations were applied to 'cigsPerDay' and 'glucose', with a small constant added to avoid undefined log values. Furthermore, a capping strategy was employed for variables such as 'sysBP', 'diaBP', 'BMI', 'heartRate', and 'totChol', where extreme outliers were capped at clinically or statistically significant upper limits. These steps were crucial in refining the dataset, enhancing its consistency and suitability for subsequent analysis and modeling, ultimately leading to the creation of the cleaned and prepared version named "framingham_cleaned.csv".

## III. Statistical Methods for Feature Selection and Analysis

This paragraph outlines the process and findings of applying statistical methods for feature selection and analysis on the Framingham Heart Study dataset. The objective was to identify significant features that contribute to the prediction of a 10-year risk of Coronary Heart Disease (CHD).

### Univariate Analysis

Univariate analysis was conducted to evaluate each feature's individual impact on the risk of CHD. Continuous variables were assessed using logistic regression, while categorical variables were analyzed using the Chi-square test. The following table summarizes the results of the univariate analysis:

| Feature | Statistic/Score | Significance (p-value) / Accuracy |
|---|---|---|
| male | 32.62 | 1.12e-08 |
| age | 0.85 | N/A |
| education | 31.05 | 8.29e-07 |
| currentSmoker | 1.50 | 2.21e-01 |
| cigsPerDay | 0.85 | N/A |
| BPMeds | 30.27 | 3.75e-08 |
| prevalentStroke | 14.03 | 1.80e-04 |
| prevalentHyp | 132.46 | 1.19e-30 |

| | | |
|---|---|---|
| diabetes | 38.48 | 5.53e-10 |
| totChol | 0.85 | N/A |
| sysBP | 0.85 | N/A |
| diaBP | 0.85 | N/A |
| BMI | 0.85 | N/A |
| heartRate | 0.85 | N/A |
| glucose | 0.85 | N/A |

### Multivariate Analysis

Multivariate analysis was performed using a logistic regression model, incorporating all features to understand their collective impact on the CHD risk. The model highlighted the challenge of accurately identifying CHD cases, especially in the presence of class imbalance. The classification report is as follows:

| | precision | recall | f1-score | support |
|---|---|---|---|---|
| 0 | 0.85 | 0.99 | 0.92 | 1077 |
| 1 | 0.44 | 0.04 | 0.07 | 195 |
| accuracy | | | 0.85 | 1272 |
| macro avg | 0.64 | 0.51 | 0.49 | 1272 |
| weighted avg | 0.79 | 0.85 | 0.79 | 1272 |

### Conclusion of the Statistical Methods for Feature Selection and Analysis

The statistical analysis provided valuable insights into the relationship between various features and the risk of CHD. Both univariate and multivariate analyses suggested areas for further model improvement, especially in handling imbalanced classes. These findings lay a foundation for developing a predictive model to estimate the 10-year risk of CHD.

## IV. Development and Evaluation of a Logistic Regression Model for Predicting 10-Year CHD Risk

In developing a predictive model to estimate the 10-year risk of Coronary Heart Disease (CHD), a logistic regression approach was employed on the Framingham Heart Study dataset. The dataset was first preprocessed, involving the balancing of class distribution using the Synthetic Minority Over-sampling Technique (SMOTE) implemented via the ROSE package

in R, to address the inherent class imbalance typical in medical datasets. The logistic regression model was trained using this balanced dataset, ensuring a more robust learning from both the minority and majority classes. Upon evaluating the model's performance, the confusion matrix revealed the following: 749 true negatives (correctly identified non-CHD cases), 70 false negatives (CHD cases incorrectly identified as non-CHD), 320 false positives (non-CHD cases incorrectly identified as CHD), and 133 true positives (correctly identified CHD cases). The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) score was calculated to be 0.7134, indicating a moderate ability of the model to distinguish between patients who will develop CHD in 10 years and those who will not. While the model demonstrates reasonable discriminative ability, the relatively high number of false positives suggests a potential area for improvement. Future work might involve further feature engineering, experimenting with other machine learning algorithms, or hyperparameter tuning to enhance the model's predictive accuracy and reduce the rate of false positives.