Purpose

The idea of this project is to perform comprehensive data analysis on the Framingham Heart Study Dataset where we verify the previously made inferences while also making new inferences from this data by using different Statistical Methods including Logistic Regression.

Dataset

The dataset - "The Framingham Heart Study" dataset is a long term prospective study of the etiology of cardiovascular disease among a population of free living subjects in Framingham, Massachusetts.The dataset is collected from laboratory, clinic, questionnaire, and adjudicated event data with a sample size of 4434 participants,   with three examination periods, approximately 6 years apart, from roughly 1956 to 1968. This dataset contains 5209 subjects and 22 variables. Some of the useful variables that can be used in our project are described in the following table.

Variable	**Description**
SEX	Participant sex
PERIOD	Examination Cycle
TIME	Number of days since baseline exam
AGE	Age at exam (years)
SYSBP	Systolic Blood Pressure (mean of last two of three measurements) (mmHg)
DIABP	Diastolic Blood Pressure (mean of last two of three measurements) (mmHg)
BPMEDS	Use of Antihypertensive medication at exam
CURSMOKE	Current cigarette smoking at exam
CIGPDAY	Number of cigarettes smoked each day
TOTCHOL	Serum Total Cholesterol (mg/dL)
HDLC	High Density Lipoprotein Cholesterol (mg/dL)
LDLC	Low Density Lipoprotein Cholesterol (mg/dL)
BMI	Body Mass Index, weight in kilograms/height meters squared
GLUCOSE	Casual serum glucose (mg/dL)
DIABETES	Diabetic according to criteria of first exam treated or first exam with casual glucose of 200 mg/dL or more

HEARTRTE	Heart rate (Ventricular rate) in beats/min
PREVAP	Prevalent Angina Pectoris at exam
PREVCHD	Prevalent Coronary Heart Disease defined as pre-existing Angina Pectoris, Myocardial Infarction (hospitalized, silent or unrecognized), or Coronary Insufficiency (unstable angina)
PREVMI	Prevalent Myocardial Infarction
PREVSTRK	Prevalent Stroke
PREVHYP	P Prevalent Hypertensive. Subject was defined as hypertensive if treated or if second exam at which mean systolic was >=140 mmHg or mean Diastolic >=90 mmHg

In this project, we aim to verify the main findings of the framingham heart study by performing exploratory data analysis and data visualization. Some of the main findings from this study that we aim verify are:-

1. Cigarette smoking increases risk of heart disease. Increased cholesterol and elevated blood pressure increase risk of heart disease. Obesity increases it.
2. There is an inverse relationship between HDL cholesterol and coronary heart disease and vice versa is true for LDL cholesterol.
3. Elevated blood pressure increases the risk of a stroke

In addition to validating prior findings, We also expect to:

1) Find out overall, given the predictors(variables) of cardiovascular disease (CAD), is one gender more susceptible to CAD than the other and are there better predictors for different genders: difference in cardiovascular risk between men and women.

Methods:
We will use linear modelling, logistic regression and one unsupervised method to validate our results and see how different features can be used as predictors for stroke and cardiovascular disease.