1. **Case Study :** 
   1. Train ML Models to Predict Heart Attack Risk
   2. Conduct data analysis to identify key risk factors.

My case study: training ML model to predict heart attack risk and conduct data analysis to identify risk factors.

From my data set, it's clear that it is based on the health sector, and one of the main reason of me taking this case study outline I from the data set is because heart disease remains one of the leading causes of death worldwide. So it will be great that, heart attack risk detection will help to save lives , so using this case study will help predict attack risk and uncover the most significant contributing risks.

**Importance of the study**

1. Public health impact : it will help hospital and other health sector to detect attack and prevent it early.
2. Personal health awareness : it will help an individual to monitor and reduce their own risk.

**How data collection helps to answer key questions is that**

collecting raw data like blood pressure, chest pain cholesterol will help get raw fact and study the pattern of heart attack and analyse it and develop predictive models and generate insight that can inform health care decisions.

1. **Identifying the Source of the Data.**
2. My data is a structured that data because it is organized in tabular format which is rows and columes , each rows represent patient record and each column represent specific vairbale and features such as age, sex etc. And data values are numeric or categorical which is readable by machines and suitable for staistical analyisis and machine learning.
3. My data is primarily quantitatuve but contains some categorical qualitatuve elemets. My quantitative data nuemericaliy contains measurable numbers for example age conatins patients age , chol serum cholesterol. And the qualitatve which is catergorical for example cp which is chest pain type 0-3 : different categories , sex- 0= female and 1= male . so those are qualitative because the describe types or categories.

**Data set soruce**

****Kaggle Dataset Title:****  
Heart Attack Risk Prediction (Cleaned Dataset).

Three source for potential data sources for heart attack risk analysis and prediction

1. **Public datasets:**

soruce kaggle,   
URL:<https://www.kaggle.com/datasets/alikalwar/heart-attack-risk-prediction-cleaned-dataset>

**Why its relevant** : offers structured, clean and labeled health data and also ideal for building predictive machine leaning models and freely accessible and commonly used in health analytic education.

1. **UCI machine learning repository**   
   uci ml repository cleveland heart disease data set  
   **URL:** [https://archive.ics.uci.edu/ml/datasets/heart+Disease](https://archive.ics.uci.edu/ml/datasets/heart+Disease" \t "_new)  
   **why its relevant** : include ral patient records collected under controlled conditions
2. **Web scraping**   
   web scraping from medcal websits and health article for example mayo clini, healthline   
   tools : python   
   **why its relevant** : helps enrich the dataset with textual information for natural lanugage processing.
3. **Data collection stratagy**   
   **Method of collection**   
   public repository because it was gotten from kaggle an it was uploeaded there for educational and rearch purposes

**Frequency** : one time retrieval   
because the data set is static not updated in real time and it was downloaded manually and no continuous data feed automatic update is applied.   
 Tools : my data set was download using a broswer : opera browser. But there are other tools that could be of use eg. Python  
  
**Data sources and process steps**

Ali kalwar collected published the data set on kaggle.  
The data set was publicly accesed through kaggle.

And the data set was from UCI machine learning repository and was cleaned , relabeled and simplified for easier use and uploaded in .CSV format for public use.

**Storage format CSV.**

1. **Data collecting**   
   the data was collected and downloaded from kaggle and the strategy was public dataset download
2. accessed the kaggle dataset
3. Downloaded the CSV file manually
4. Verified that the dataset was structured with labeled features and was download as CSV file
5. **Data cleaning**
6. The dataset did not undergo outlier removal because Removing them could cause bias or data loss. The distribution remained consistent before and after filtering, with negligible impact on mean and variance.
7. All values in the dataset were formatted correctly at the time of download. Column types were consistent and did not require reformatting