6.1. **Sourcing Open Data**

Daria Navrotska

1) Data Source:

The **Heart Attack Risk Prediction Dataset** was sourced from Kaggle. It is a synthetic dataset created for research and educational purposes. It likely draws inspiration from established clinical datasets such as the Cleveland Heart Disease dataset or other public health repositories.

Heart attacks, or myocardial infarctions, continue to be a significant global health issue, necessitating a deeper comprehension of their precursors and potential mitigating factors. This dataset encapsulates a diverse range of attributes including age, cholesterol levels, blood pressure, smoking habits, exercise patterns, dietary preferences, etc. These variables are commonly used in medical research and diagnostics to assess heart disease risk, making the dataset a useful proxy for educational and exploratory analysis.

This dataset provides a comprehensive array of features relevant to heart health and lifestyle choices, encompassing patient-specific details such as age, gender, cholesterol levels, blood pressure, heart rate, and indicators like diabetes, family history, smoking habits, obesity, and alcohol consumption. Additionally, lifestyle factors like exercise hours, dietary habits, stress levels, and sedentary hours are included. Medical aspects comprising previous heart problems, medication usage, and triglyceride levels are considered. Socioeconomic aspects such as income and geographical attributes like country, continent, and hemisphere are incorporated. The dataset, consisting of **8763 records** from patients around the globe.

Dataset Glossary (Column-wise):

1. **Patient ID:** Unique identifier for each patient.
2. **Age:** Age of the patient (numerical).
3. **Gender:** Sex of the patient (categorical: Male/Female).
4. **Cholesterol Level:** Cholesterol levels of the patient (numerical), **mg/dL**
5. **Blood Pressure:** Blood pressure of the patient (categorical or numerical depending on the format, likely represented as **systolic/diastolic**).
6. **Heart Rate:** Heart rate of the patient (numerical).
7. **Diabetes:** Whether the patient has diabetes (binary: Yes/No).
8. **Family Heart Problems:**
   Family history of heart-related problems (binary: 1 for Yes, 0 for No).
9. **Smoking:** Smoking status of the patient (binary: 1 for Smoker, 0 for Non-smoker).
10. **Obesity:** Obesity status of the patient (binary: 1 for Obese, 0 for Not obese).
11. **Alcohol Consumption:**
    Level of alcohol consumption by the patient (categorical: None/Light/Moderate/Heavy).
12. **Exercise Hours Per Week:** Number of exercise hours per week (numerical).
13. **Diet:** Dietary habits of the patient (categorical: Healthy/Average/Unhealthy).
14. **Previous Heart Problems:**
    Previous heart problems of the patient (binary: 1 for Yes, 0 for No).
15. **Medication Use:** Medication usage by the patient (binary: 1 for Yes, 0 for No).
16. **Stress Level:** Stress level reported by the patient (numerical, scale 1-10).
17. **Sedentary Hours Per Day:** Hours of sedentary activity per day (numerical).
18. **Income:** Income level of the patient (numerical).
19. **Body Mass Index:** BMI of the patient (numerical). The BMI is expressed in kg/m^2, resulting from patient mass in kilograms and height in metres.
20. **Triglyceride Levels:** Triglyceride levels of the patient (numerical), **mg/dL**

21. **Physical Activity Days Per Week:** Days of physical activity per week (numerical).
22. **Sleep Hours Per Day:** Hours of sleep per day (numerical).
23. **Country:** Country of the patient (categorical).
24. **Continent:** Continent where the patient resides (categorical).
25. **Hemisphere:** Hemisphere where the patient resides (categorical).
26. **Heart Attack Risk:** Target Variable.
    Presence of heart attack risk (binary: 1 for Yes, 0 for No).

Heart Attack Risk Prediction Dataset could be accessed here: https://www.kaggle.com/datasets/ghnshymsaini/heart-attack-risk-prediction-dataset?resource=download

Why Was This Dataset Chosen?

This dataset offers a unique opportunity to explore one of the most pressing public health challenges: identifying patterns of heart attack risk across a population. Rather than focusing solely on individual predictions, the dataset allows analysts to uncover broader trends such as which age groups, lifestyle profiles, or clinical markers are most associated with elevated cardiovascular risk. These insights can inform targeted awareness campaigns, preventive strategies, and resource allocation in healthcare systems.

What makes this problem especially compelling is its intersection of medicine, behavior, and data science. The dataset includes both physiological metrics (like cholesterol and blood pressure) and behavioral factors (such as smoking and exercise), enabling a holistic view of health risk. In short, this dataset transforms a personal health issue into a population-scale analytics challenge, making it ideal for data scientists who want to make a meaningful impact through their work.

2) Data Profile

Data Cleaning Summary

- **Initial Data Inspection** Loaded the dataset and previewed the first few rows using df.head() to understand its structure.
- **Dataset Dimensions** Used df.shape to check the number of rows and columns.
- **Data Types and Structure** Called df.info() to review column types and non-null counts.
- **Statistical Overview** Applied df.describe() to generate summary statistics for numerical columns.
- **Missing Values Check** Verified with df.isnull().sum() that there were no missing values in the dataset.
- **Duplicate Records Check** Used df.duplicated().sum() and confirmed there were no duplicate rows.
- **Unique Identifier Validation** Checked that the 'Patient ID' column contains unique values using df['Patient ID'].is_unique.
- **Age Range Validation** Ensured all age values fall within a realistic range using df['Age'].between(0, 120).all().
- **Column Renaming** Renamed several columns for clarity using df.rename() with a dictionary of new names.
- **Post-Cleaning Preview** Re-checked the updated dataset with df.head() to confirm changes.
  Saving Cleaned Data Defined a new file path and saved the cleaned dataset using df.to_csv() with index=False.

Data Limitations:

- *No verified source or clinical validation*: The dataset lacks documentation about its origin, making it unsuitable for real-world medical use.

- *Synthetic or anonymized data*: May not reflect the full complexity or diversity of actual patient populations.
- *Static records only*: No time-based or longitudinal data to track health changes over time.
- *Limited demographic detail*: Missing ethnicity, location, or socioeconomic indicators restrict population-level insights.
- *Binary outcome oversimplifies risk*: Heart attack risk is treated as yes/no, ignoring the spectrum of severity.
- *Unclear feature definitions*: Some variables lack units or medical context, which can affect interpretation.

Ethical Considerations

- *Privacy and Anonymity* - ensure no personal data is exposed, even if the dataset is synthetic.
- *Responsible Use of Models* - avoid using predictions for real medical decisions without proper validation.
- *Bias and Fairness* - check for unequal performance across demographic groups to prevent discrimination.
- *Transparency and Accountability* - clearly state the dataset's limitations when sharing results or models.
- *Health Equity Implications* - use insights to support inclusive care, not to stigmatize or exclude individuals.

3) List of Questions to Explore

1. **Which features are most correlated with heart attack risk?**
Use feature importance plots to identify strong predictors.

2. **Are there any outliers or anomalies in key health indicators?**
Use box plots to detect unusual values in features like blood pressure or heart rate.

3. **How does heart attack risk vary across age groups?**
Create age distribution plots segmented by risk status to uncover age-related trends.

4. **Is there a relationship between cholesterol levels and heart attack risk?**
Use scatter plots and box plots to visualize cholesterol across risk categories.

5. **Do lifestyle factors (e.g., smoking, exercise) significantly impact risk?**
Analyze categorical variables with bar charts and group comparisons.

6. **What is the distribution of heart attack risk across genders?**
Use count plots and pie charts to compare risk prevalence between male and female participants.

7. **How do multiple risk factors interact to influence outcomes?**
Use multivariate visualizations to explore combined effects.