National University of Computer and Emerging Sciences

Heart Disease Prediction

Project Deliverable 1

Team

Muhammad Hammad | 21L-5388

Muneeb Khalid | 21L-1782

Jamshed Siddiqui | 21L-5298

Shahzaib Ahmad | 21L-5220

FAST School of Computing

National University of Computer and Emerging Sciences

Lahore, Pakistan

1. Introduction

This project aims to predict the likelihood of heart disease in individuals based on health-related data. Accurate prediction can help with early intervention, potentially saving lives. The dataset used contains self-reported information from individuals, which includes lifestyle habits, medical conditions, and general health assessments.

Data wrangling is a crucial part of any data science pipeline. Cleaning, transforming, and understanding the data ensures that the models built later are both reliable and insightful.

2. Dataset Overview

- Dataset: Heart Disease Prediction Dataset
- Rows: 352,028
- Columns: 33
- Target Variable: `HeartDisease` (Binary – Yes/No → Encoded to 1/0)

Features Snapshot:

- Demographic & Lifestyle: Gender, AgeCategory, SleepHours, PhysicalActivities
- Health Metrics: BMI, HeightInMeters, WeightInKilograms, GeneralHealth
- Medical Conditions: HadStroke, HadAngina, HadAsthma, HadDepressiveDisorder, etc.
- Vaccination & Testing: FluVaxLast12, HIVTesting, PneumoVaxEver
- Disabilities & Limitations: DifficultyWalking, DifficultyConcentrating, etc.

3. Data Loading and Exploration

- Dataset was successfully loaded into a Pandas DataFrame.
- Head and summary statistics (`.head()`, `.info()`, `.describe()`) were used to get an initial look at the structure and data types.

```
# Show the first few rows
df.head()
```

	Gender	GeneralHealth	PhysicalHealthDays	MentalHealthDays	PhysicalActivities	SleepHours	HeartDisease	HadAngina	HadStroke	HadAsthma	...	AgeCategory	HeightI
0	Female	Very good	0.0	0.0	No	8.0	No	No	No	No	...	Age 80 or older	
1	Female	Excellent	0.0	0.0	No	6.0	No	No	No	No	...	Age 80 or older	
2	Female	Very good	2.0	3.0	Yes	5.0	No	No	No	No	...	Age 55 to 59	
3	Female	Excellent	0.0	0.0	Yes	7.0	No	No	No	Yes	...	NaN	
4	Female	Fair	2.0	0.0	Yes	9.0	No	No	No	No	...	Age 40 to 44	

5 rows × 33 columns

```
[3] # Basic info about the DataFrame
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 352028 entries, 0 to 352027
Data columns (total 33 columns):
 #   Column                    Non-Null Count   Dtype
---  ------                    --------------   -----
 0   Gender                    352028 non-null  object
 1   GeneralHealth             351070 non-null  object
 2   PhysicalHealthDays        343221 non-null  float64
 3   MentalHealthDays          344757 non-null  float64
 4   PhysicalActivities        351112 non-null  object
 5   SleepHours                347627 non-null  float64
 6   HeartDisease              349512 non-null  object
```

```
[4] df.describe()
```

	PhysicalHealthDays	MentalHealthDays	SleepHours	HeightInMeters	WeightInKilograms	BMI
count	343221.000000	344757.000000	347627.000000	329268.000000	318744.000000	313341.000000
mean	4.330396	4.372219	7.020824	1.702928	83.100792	28.529419
std	8.675360	8.391346	1.513752	0.107235	21.483552	6.554388
min	0.000000	0.000000	1.000000	0.910000	22.680000	12.020000
25%	0.000000	0.000000	6.000000	1.630000	68.040000	24.130000
50%	0.000000	0.000000	7.000000	1.700000	80.740000	27.440000

---

4. Data Cleaning

- **Data Type Conversion:** The datatypes of the data (non-numeric) were listed as object and they were first converted into categorical type.
- **Handling Missing Data:** Missing fields were filled with their median in case they were numerical data as median is less prone to outliers and they were filled with mode if they were categorical data.

- **Duplicate Removal:** A total of 5679 rows were duplicates and they were dropped using `df.drop_duplicates()`
- **Outlier Detection:** Outliers were detected using the IQR method because it is a better approach as compared to the z-score method because it does not rely upon mean.
- **Outlier Removal:** The outliers were not removed but they were capped to the upper and lower limit by IQR method.

---

5. Data Transformation

- **Feature Scaling:** Applied Standardization (Z-score) on numerical features using `StandardScaler`.
- **Encoding Categorical Variables:** Each feature was checked for how many unique values it had in its column . If they were 2 or less than 2, then label encoder was used and if they were greater than 2, then one-hot encoding was used.

  The following categories were label encoded and the rest were one-hot encoded

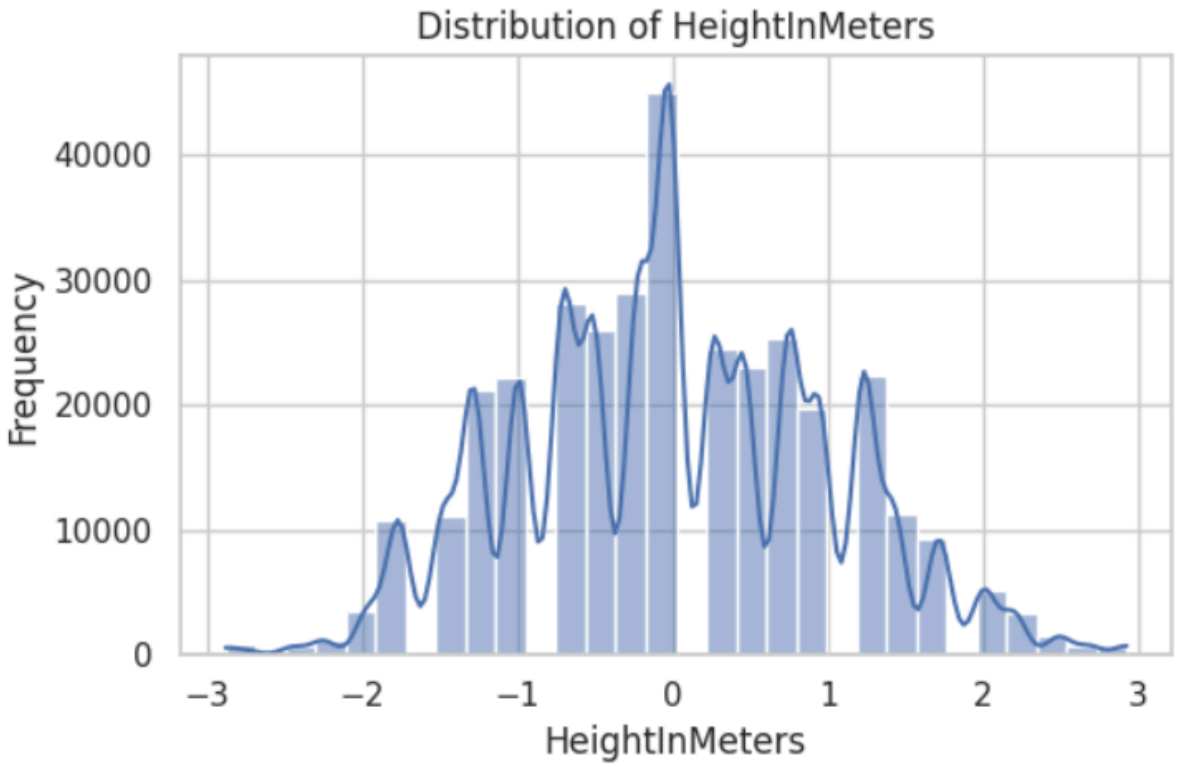
---

6. Exploratory Data Analysis (EDA)

Univariate:

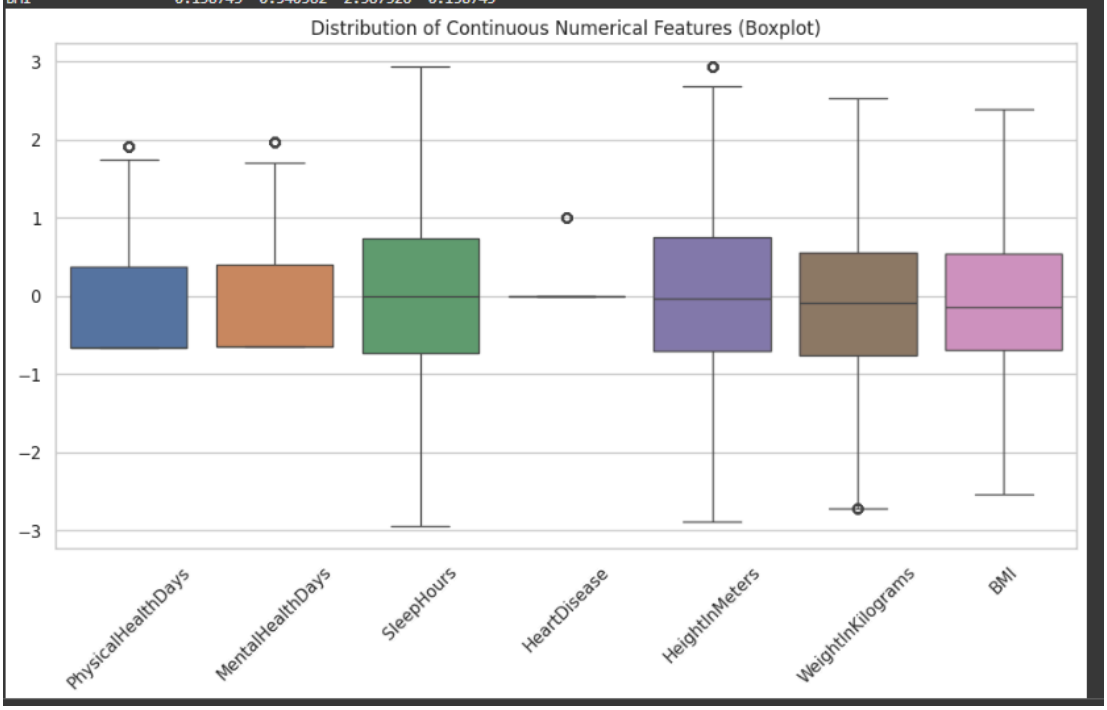
- Histograms are used to analyze the distributions of the numerical columns in univariate analysis and the columns do not indicate any significant skewness

e.g



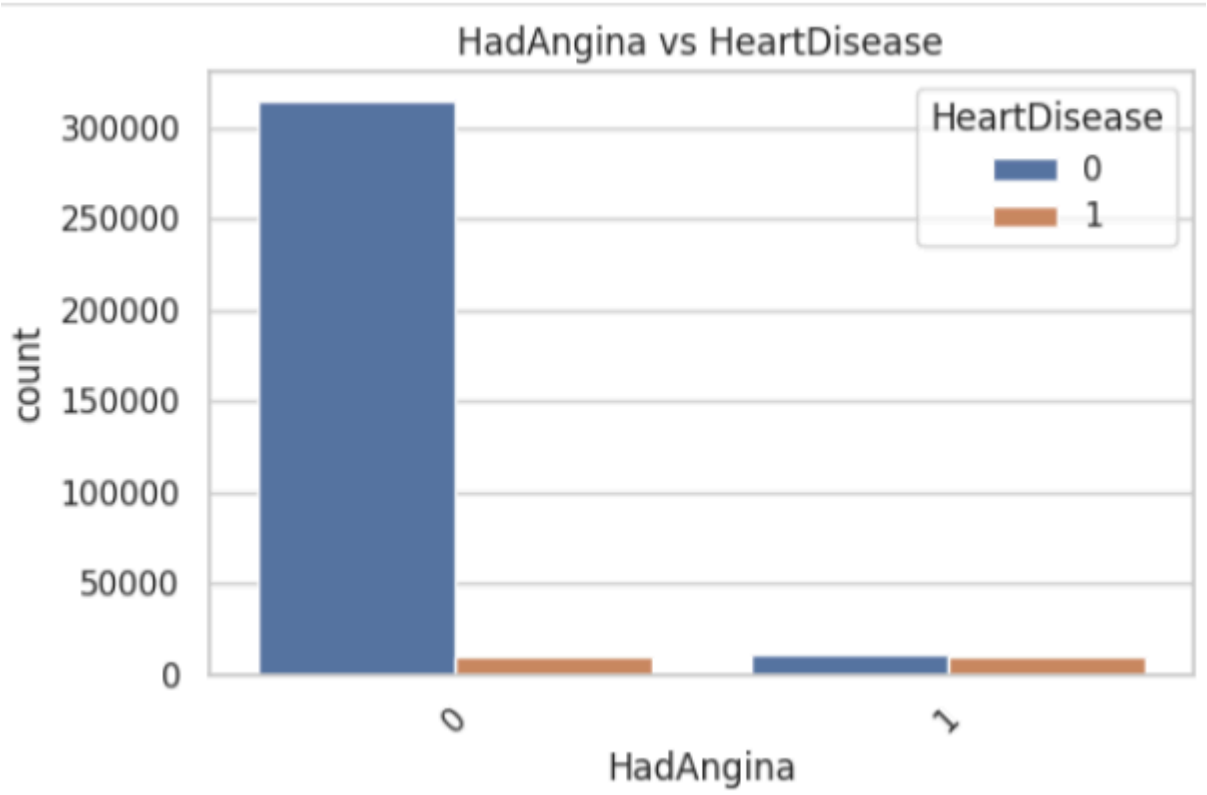
- Boxplot were used to get an idea about the numerical statistics of the columns as shown in the graph below. The values of the mean median mode etc are also printed in the image.

	count	mean	std	min	25%	50%	75%	max	mode
PhysicalHealthDays	346349.0	6.449986e-17	1.000001	-0.659603	-0.659603	-0.659603	0.367192	1.907383	-0.659603
MentalHealthDays	346349.0	4.816975e-17	1.000001	-0.649756	-0.649756	-0.649756	0.393624	1.958694	-0.649756
SleepHours	346349.0	-1.346209e-16	1.000001	-2.934741	-0.733947	-0.000349	0.733248	2.934042	-0.000349
HeartDisease	346349.0	5.812923e-02	0.233988	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000
HeightInMeters	346349.0	1.931590e-15	1.000001	-2.882239	-0.705078	-0.027740	0.746362	2.923523	-0.027740
WeightInKilograms	346349.0	1.173307e-15	1.000001	-2.722334	-0.757025	-0.090106	0.553181	2.518490	-0.090106
BMI	346349.0	-1.316257e-15	1.000001	-2.537326	-0.690581	-0.138743	0.540582	2.387326	-0.138743

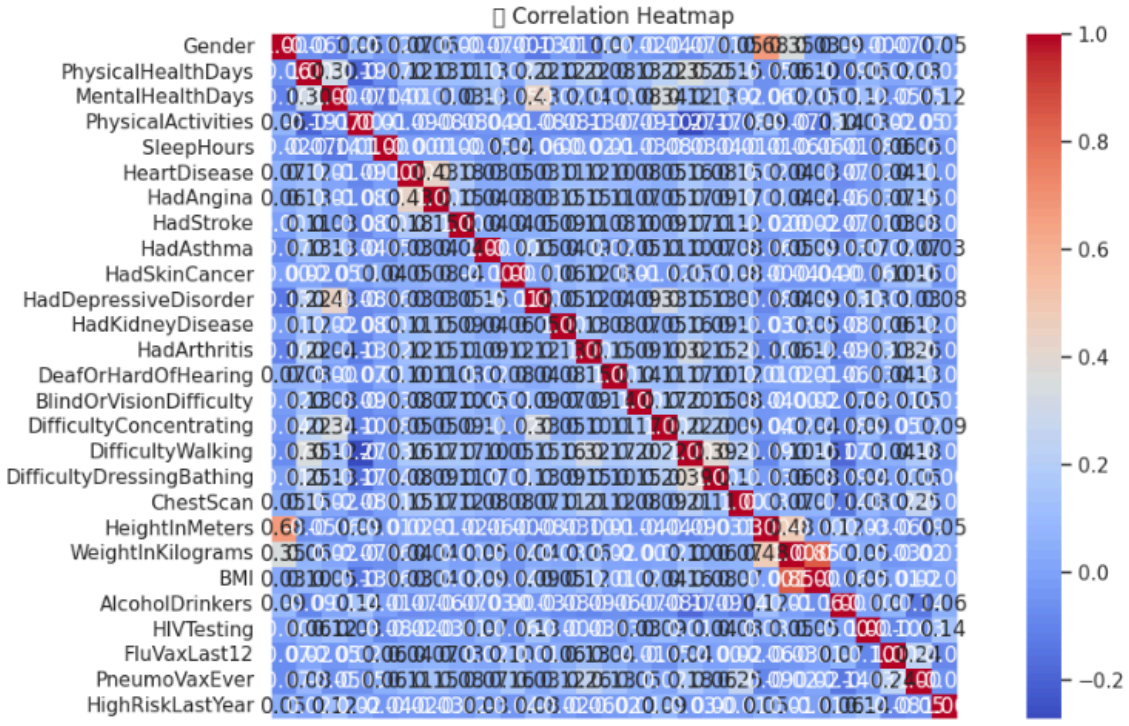


Bivariate:

- Count plots were used for comparing the relationship between all the variables to the target variable (Heart Disease) and some features such as HadAngina and HadKidenyDisease showed a strong relation between having a heart disease.

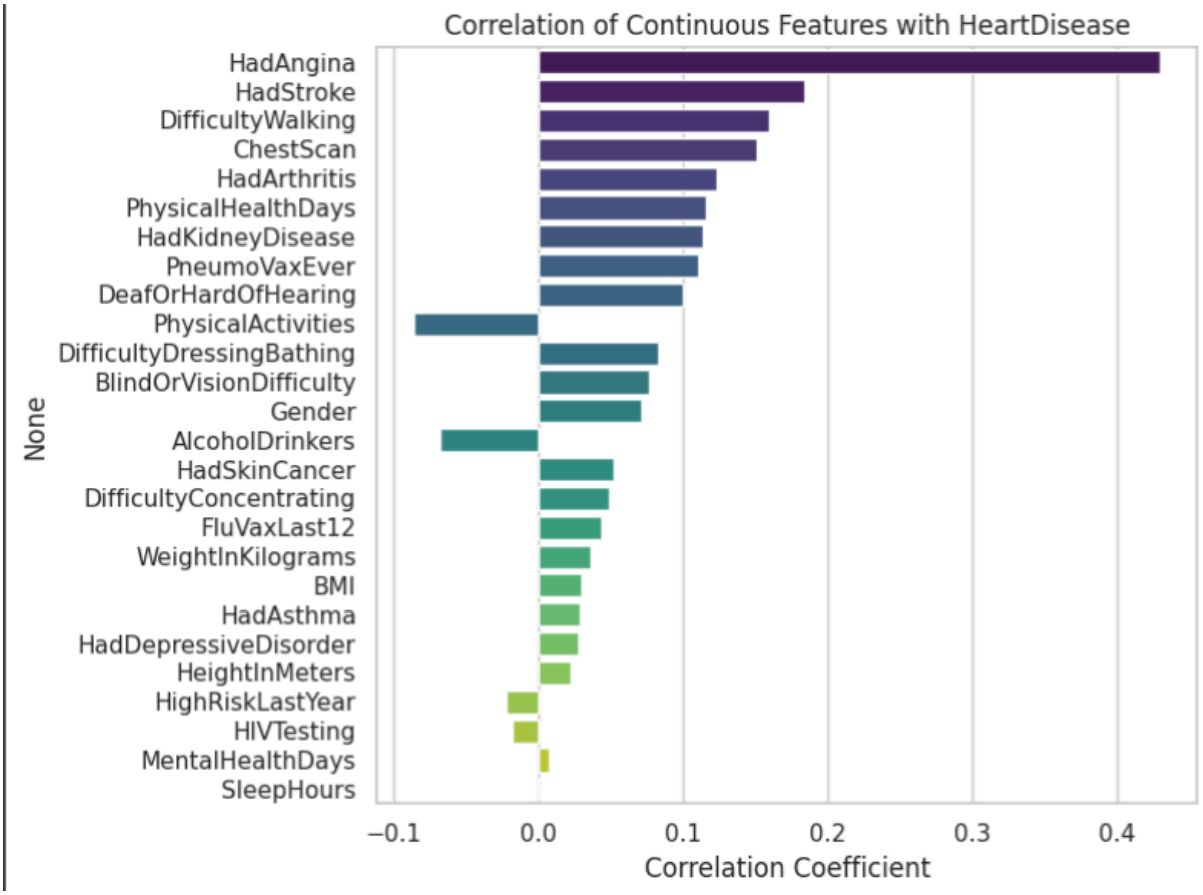


- A correlation matrix was plotted showing the relationship between various variables and here as well the angina and heart disease showed good corelation

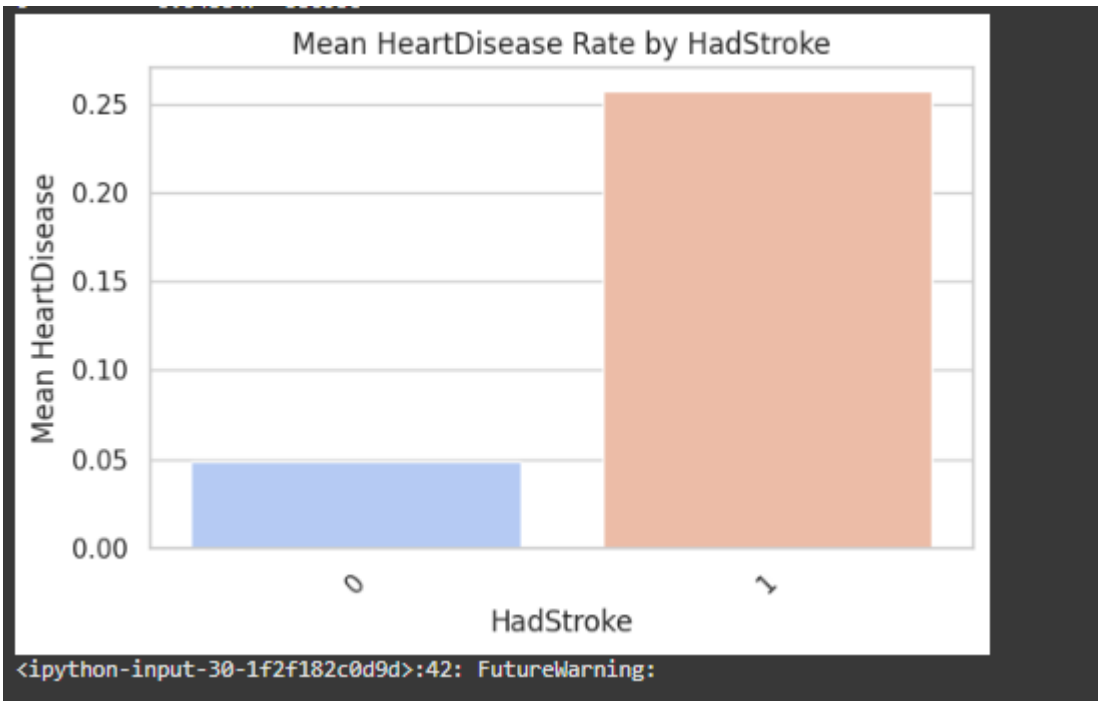


7. Feature Analysis

- A correlation graph was plotted showing the relationship between the heart disease (target) and other features. A positive corelation was found between HadAngina and HeartDisease and a medium negative correlation was found between PhysicalActivities and HeartDisease.



- To analyze the impact of categorical features on the target variable (*HeartDisease*), we used group-by summaries to calculate the mean occurrence of heart disease within each category. This approach helps identify significant trends and relationships between categorical predictors and heart disease risk. For instance, the analysis of the *HadStroke* feature revealed a substantial disparity in heart disease rates between individuals who had previously experienced a stroke and those who had not. The mean heart disease rate for individuals with no history of stroke was approximately **5%**, whereas it increased sharply to **27%** for those who had suffered a stroke. This suggests a strong positive correlation. This image can confirm it. Similar graph exists for other features as well in the notebook.


Mean HeartDisease Rate by HadStroke
Mean HeartDisease
0.25
0.20
0.15
0.10
0.05
0.00
0
1
HadStroke
<ipython-input-30-1f2f182c0d9d>:42: FutureWarning: