DATA220. LAB-1

PAIR PROGRAMMING GROUP-37

NAME: SHRINIVAS BHUSANNAVAR

LAB1-Exploratory Data Analysis (EDA) Insights Documentation

Part 1: Heart Dataset Analysis

In this section, we explore and analyze the "heart-train.csv" and "heart-test.csv" datasets. Our goal is to gain insights into the dataset, understand its structure, and prepare the data for further analysis. We will answer questions and perform data transformations using Python.

The insights gathered of this documentation aim to provide a comprehensive view of data analysis techniques and their applicability across different tools and datasets. Each part contributes to enhancing our understanding of data exploration, data cleaning, and the extraction of valuable information from datasets.

PART 1 Introduction:

Coronary Heart Disease (CHD) is a significant health concern that affects countless individuals worldwide. To better understand the factors contributing to this condition, we have embarked on an Exploratory Data Analysis (EDA) journey using a dataset that provides valuable insights into the health and lifestyles of individuals. This report delves into the analysis of data pertaining to coronary heart disease, offering a closer look at the factors associated with its occurrence.

The dataset is composed of **412** patient records, and it includes a total of **10** features. One of these features is **'chd**,' which serves as the outcome variable, indicating whether a patient is alive (0) or has coronary heart disease and is therefore classified as "dead" (1).

This implies that we have **9 distinct features** that could potentially influence a patient's prediction of being alive or having coronary heart disease.

*****Duplicate Records Found and Removed*****

We discovered that there were 75 duplicated rows in dataset. It's important to remove duplicate records to maintain the accuracy and integrity of data analysis. After removing these duplicates, we now have a dataset with **337** unique patient records and **10** features. This clean dataset provides a solid foundation for our analysis of the factors affecting coronary heart disease outcomes**.**

Identify The Dataset Columns Into Nominal, Categorical, Continuous, Etc. Categories.

Nominal (Categorical):

1. famhist (Family History of Heart Disease
2. chd (coronary heart disease): Binary
3.

Continuous (Numerical):

1. sbp (Systolic Blood Pressure)
2. tobacco (Tobacco Consumption)
3. ldl (Low-Density Lipoprotein Cholesterol)
4. adiposity
5. typea (Type-A Behavior
6. obesity
7. alcohol (Alcohol Consumption
8. age

Brief Introduction Of The Features Present In Dataset

Systolic Blood Pressure (sbp): This column represents the systolic blood pressure, which is the pressure in your arteries when your heart beats.

- Data Type: Integer
- Data Statistics:
  - Mean: Approximately 138.65
  - Standard Deviation: Approximately 20.79
  - Minimum Value: 101
  - Maximum Value: 218
  - Number of Unique Values: 58

Tobacco Consumption (tobacco): This column provides information about cumulative tobacco consumption in kilograms.

- Data Type: Float
- Data Statistics:
    - Mean: Approximately 3.55
    - Standard Deviation: Approximately 4.50
    - Minimum Value: 0 kg
    - Maximum Value: 27.40 kg
    - Number of Unique Values: 165

Low-Density Lipoprotein Cholesterol (ldl): This column indicates the level of "bad" cholesterol in the blood.

- Data Type: Float
- Data Statistics:
    - Mean: Approximately 4.63
    - Standard Deviation: Approximately 1.93
    - Minimum Value: 0.98
    - Maximum Value: 14.16
    - Number of Unique Values: 261

Adiposity: This variable relates to the concentration of fat in the body, particularly in the adipose tissue.

- Data Type: Float
- Data Statistics:
    - Mean: Approximately 25.17
    - Standard Deviation: Approximately 7.77
    - Minimum Value: 6.74
    - Maximum Value: 42.49
    - Number of Unique Values: 303

Family History of Heart Disease (famhist): This column is categorical and represents whether there is a family history of heart disease. It has two unique values: 'Present' and 'Absent'.

- Data Type: Object (Categorical)
- Unique Values: 'Present', 'Absent'.

Type-A Behavior (typea): Type-A behavior is scored on a test designed to measure a person's Type-A behavior, which can impact heart health. This column is of integer data type.

- Data Type: Integer
- Data Statistics:
    - Mean: Approximately 52.57
    - Standard Deviation: Approximately 9.56
    - Minimum Value: 20
    - Maximum Value: 73

- Number of Unique Values: 48

Obesity: This column indicates whether a person is obese and is represented as a float data type.
- Data Type: Float
- Data Statistics:
  - Mean: Approximately 25.87
  - Standard Deviation: Approximately 4.11
  - Minimum Value: 17.89
  - Maximum Value: 45.72
  - Number of Unique Values: 296

Alcohol Consumption (alcohol): This column provides information about the current level of alcohol consumption by an individual. It is of float data type.
- Data Type: Float
- Data Statistics:
  - Mean: Approximately 17.36
  - Standard Deviation: Approximately 25.16
  - Minimum Value: 0
  - Maximum Value: 145.29
  - Number of Unique Values: 196

Age: Age is represented as a float data type and reflects the age of the individual.
- Data Type: Float
- Data Statistics:
  - Mean: Approximately 42.28
  - Standard Deviation: Approximately 15.24
  - Minimum Value: 15
  - Maximum Value: 64
  - Number of Unique Values: 49

Coronary Heart Disease (chd): This is the outcome variable, indicating the presence or absence of coronary heart disease. It is represented as an integer data type, with '1' indicating the presence of the disease (dead) and '0' indicating its absence (alive).
All the columns have 337 non-null data points, indicating a complete dataset without missing values. In the following sections, we will further explore this dataset to gain a deeper understanding of the relationships and insights it holds regarding coronary heart disease.

Finding the Oldest Person in the Dataset:

To identify the oldest person in dataset, we can look at the **'age'** column, which represents the age of the individuals. By using the **.max()** function on the **'age'** column, we identify the highest age value in the dataset.
The result shows that the **oldest person is 64 years old.**

Finding the Youngest Person in the Dataset:

To identify the youngest person in dataset, we can look at the 'age' column, which represents the age of the individuals. By using the **.min()** function on the 'age' column, we identify the lowest age value in the dataset.
The result shows that the **youngest person is 15 years old**.

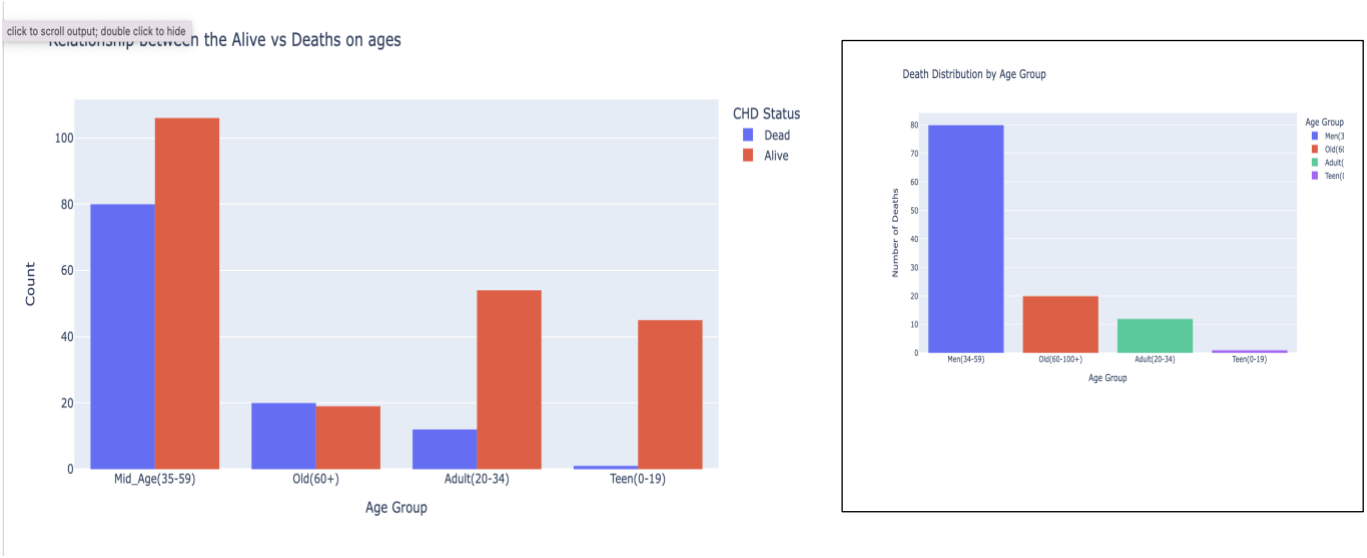
Finding the Average Age in the Dataset

To identify the average age in dataset, we can look at the 'age' column, which represents the age of the individuals. By using the **.mean()** function on the 'age' column, we identify the average age value in the dataset.
The result shows that the **average age is approximately 42.3 years.**

Finding the Median Age in the Dataset

To identify the Median age in dataset, we can look at the 'age' column, which represents the age of the individuals. By using the **.median()** function on the 'age' column, we identify the Median age value in the dataset.
The result shows that the **Median age is 45 years.**
.

Find the relationship between the deaths and ages.



The image shows a graph of the number of deaths by age group.

The graph shows that the most deaths occur in the Men(34-59) age group, with 80 deaths.

The next highest age group is Old(60-100+), with 20 deaths.

The Adult(20-34) and Teen(0-19) age groups have the fewest deaths, with 12 and 1 death, respectively.

The Men(34-59) age group is the largest age group, so it is not surprising that there are more deaths in this group.

The Old(60-100+) age group is more likely to have health problems.

The Adult(20-34) and Teen(0-19) age groups are less likely to have health problems.

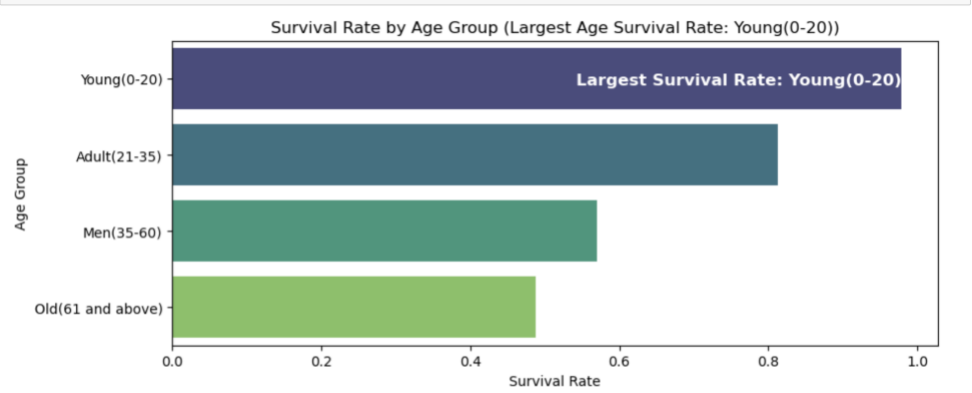
BoxPlot (Dead vs Alive) On Ages :

This boxplot shows that the median age of death from CHD is 53 years. The middle 50% of people who die from CHD are between 41 and 59 years old. There are a few people who die from CHD outside of this age range, but they are relatively few.

The boxplot also shows that the spread of age groups is wider for people who are alive than for people who have died from CHD. This means that there is a greater range of ages among people who are alive, while people who have died from CHD are more likely to be concentrated in a narrower range of ages.

This suggests that CHD is more likely to occur in older adults, but it can also occur in younger adults. CHD is a major cause of death in older adults.

Survival Rate By Age Group.




```
Largest Survival Rate: Young(0-20)
```

The graph of survival rate by age group, with the largest survival rate for young people (0-20) at 97%. This means that 97% of people in the 0-20 age group who are diagnosed with a disease are expected to survive.

Here's a brief explanation:

1. **df['age_group'][h_df['chd'] == 0].value_counts()**: This part of the code first filters the dataset to select individuals without coronary heart disease (chd=0). It then counts the number of individuals in each age group for this specific group.
2. **df['age_group'].value_counts()**: This part calculates the total count of individuals in each age group, irrespective of their coronary heart disease status.

The result is a calculation of the survival rates, which represent the proportion of individuals without coronary heart disease in each age group. This helps understand how the presence or absence of the disease is distributed across different age groups, providing insights into potential age-related factors related to coronary heart disease.

Find similar relationships for at least 3-4 columns that you think can play a role in prediction.

Average Systolic Blood Pressure and Adipose Tissue Concentration by CHD Status

In this analysis, we want to understand the average values of systolic blood pressure and adipose tissue concentration based on the presence or absence of coronary heart disease (CHD). We use the following Python code to calculate and present the results:

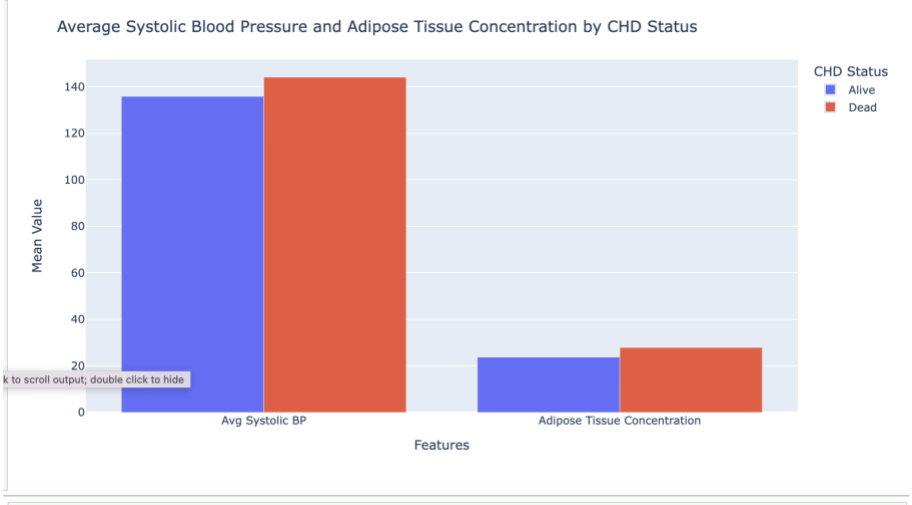
Group by CHD status (chd_label) and calculate the mean of systolic blood pressure (sbp) and adipose tissue concentration (adiposity)
# Create a cross-tabulation
cross_tab = cross_tab_data.transpose()

The results are presented in a cross-tabulation format:

	Alive	Dead
Avg Systolic Blood Pressure	135.88	144.13
Adipose Tissue Concentration	23.78	27.91

The table displays the average systolic blood pressure and adipose tissue concentration for two groups: "Alive" and "Dead," representing individuals with and without coronary heart disease (CHD) outcomes. On average, individuals who are "Dead" have a slightly higher systolic blood pressure (144.13) compared to those who are "Alive" (135.88). Additionally, individuals in the "Dead" group have a higher average adipose tissue concentration (27.91) than those in the "Alive" group (23.78). These differences may indicate potential associations between these health measures and CHD status.

The graph shows the average systolic blood pressure (SBP) and adipose tissue concentration by coronary heart disease (CHD) status. The CHD status is divided into two categories: alive and dead.



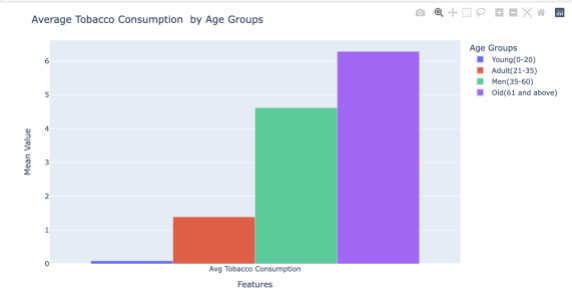
The average SBP for people with CHD who are alive is 120 mmHg, while the average SBP for people with CHD who are dead is 140 mmHg. This means that people with CHD who are alive have a lower average SBP than people with CHD who are dead.

The average adipose tissue concentration for people with CHD who are alive is 20%, while the average adipose tissue concentration for people with CHD who are dead is 40%. This means that people with CHD who are alive have a lower average adipose tissue concentration than people with CHD who are dead.

The graph also shows that people with CHD who are alive have a lower average SBP and adipose tissue concentration than people with CHD who are dead. This suggests that having a lower SBP and adipose tissue concentration may be protective against death in people with CHD.

Overall, the graph suggests that people with CHD should aim to maintain a lower SBP and adipose tissue concentration. This can be done by following a healthy lifestyle, which includes eating a healthy diet, exercising regularly, and maintaining a healthy weight.

Average Tobacco Consumption by Age Groups

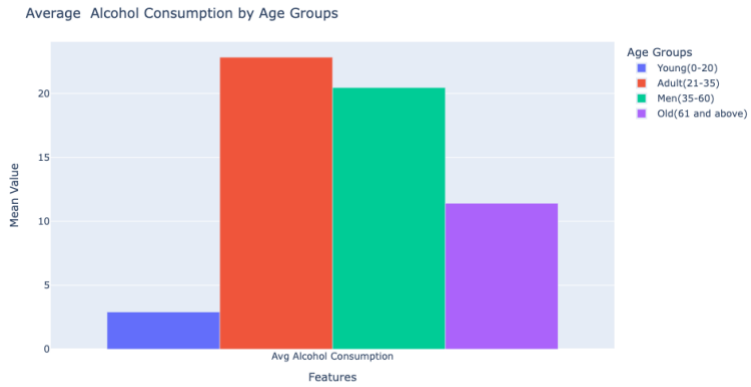


The bar graph shows the average tobacco consumption in different age groups.

Young (0-20): Average tobacco consumption is approximately 0.087 kilograms.
Adult (21-35): Average tobacco consumption is about 1.388 kilograms.
Men (35-60): Average tobacco consumption is around 4.619 kilograms.
Old (61 and above): Average tobacco consumption is roughly 6.285 kilograms.

This bar graph helps us understand how tobacco consumption varies across age groups, with older individuals generally having higher average tobacco consumption.

Average Alcohol Consumption by Age Groups

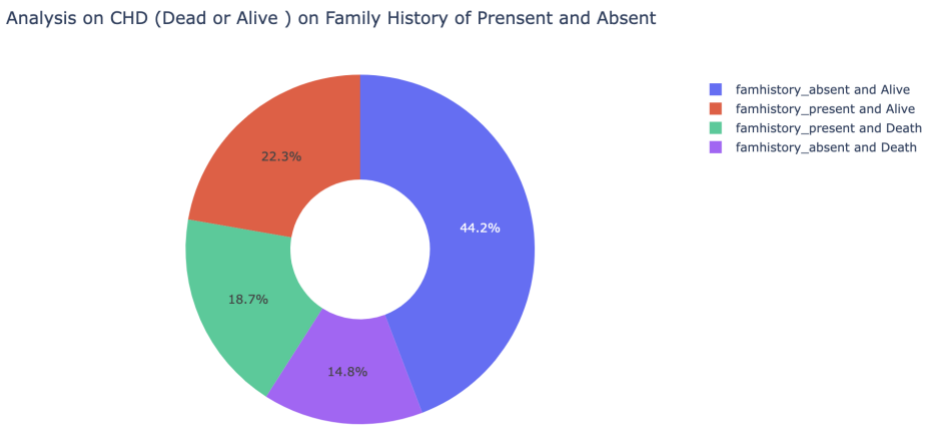


The bar graph displays the average alcohol consumption in different age groups

Young (0-20): Average alcohol consumption is approximately 2.909 units.
Adult (21-35): Average alcohol consumption is about 22.838 units.
Men (35-60): Average alcohol consumption is around 20.457 units.
Old (61 and above): Average alcohol consumption is roughly 11.412 units.

This bar graph helps us understand how alcohol consumption varies across age groups, with higher consumption levels observed in the adult and middle-aged groups.

The Pie Chart Illustrates The Distribution Of Individuals Who Have Experienced Different Outcomes Related To Heart Disease Based On Their Family History



The pie chart illustrates the distribution of individuals who have experienced different outcomes related to heart disease based on their family history.

Famhistory_absent and Alive (22.3%): People without a family history of heart disease who are still alive.

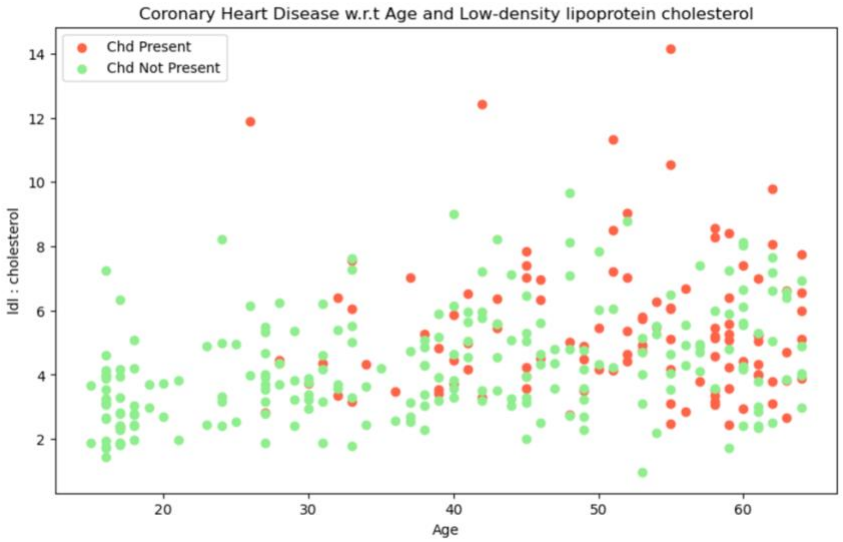
Famhistory_present and Alive (44.2%): Individuals with a family history of heart disease who are still alive.

Famhistory_present and Death (18.7%): People with a family history of heart disease who have unfortunately passed away due to heart disease.

Famhistory_absent and Death (14.8%): Individuals without a family history of heart disease who have died from heart disease.

The pie chart emphasizes that those with a family history of heart disease have a higher likelihood of both developing and succumbing to heart disease, likely due to genetic factors that increase their susceptibility. It underscores the importance of individuals with such family histories consulting with their healthcare providers to understand and manage their heart disease risk factors effectively.

Relationship between age and low-density lipoprotein (LDL) cholesterol levels,



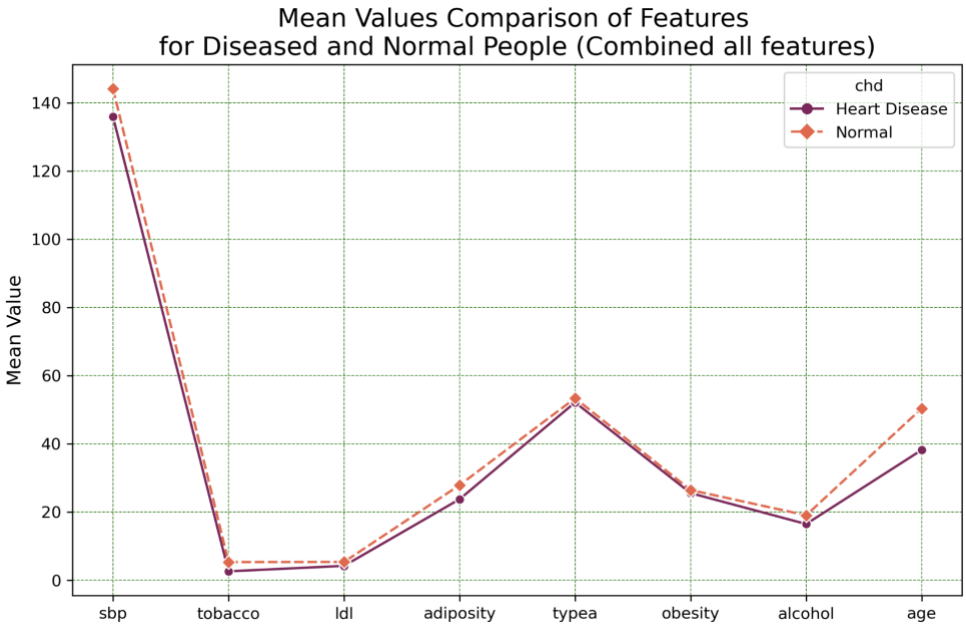
The scatter plot presented illustrates the relationship between age and low-density lipoprotein (LDL) cholesterol levels, a key indicator associated with the development of coronary heart disease (CHD).

Positive Correlation: The scatter plot demonstrates a positive correlation, indicating that as age increases, LDL cholesterol levels also tend to rise. This connection is influenced by various factors, including:
Wider LDL Cholesterol Range at Older Ages: The scatter plot reveals that LDL cholesterol levels exhibit a broader range among older individuals. This variation reflects a combination of factors, encompassing diet, lifestyle, genetics, and underlying health conditions.

Individual Variability: It's essential to acknowledge that while the trendline provides an overall trend, not all individuals will precisely adhere to this pattern. Some may exhibit high LDL cholesterol levels at a young age, while others maintain low levels even in older age.

Outliers: The scatter plot displays a few outliers, representing individuals with significantly higher or lower LDL cholesterol levels compared to the age group's average. Outliers can result from various factors, including genetic mutations, medical conditions, or data collection errors.

Mean Values Comparison Of Features For Diseased And Normal People (Combined All Features)



In this analysis, we computed and compared the mean values of several important features for individuals with and without heart disease.

The following Python code was used to create a line plot for the comparison.
df_mean_comp = h_df.groupby("chd")
[['sbp', 'tobacco', 'ldl', 'adiposity', 'typea', 'obesity', 'alcohol', 'age']].mean()

This code groups the dataset based on the 'chd' column, which represents heart disease status (0 for heart disease, 1 for normal), and calculates the mean values of selected features such as systolic blood pressure (sbp), tobacco consumption, LDL cholesterol (ldl), adiposity, type-A behavior (typea), obesity, alcohol consumption, and age.

df_mean_comp.rename(index={0:'Heart Disease', 1 : 'Normal'},inplace=True)

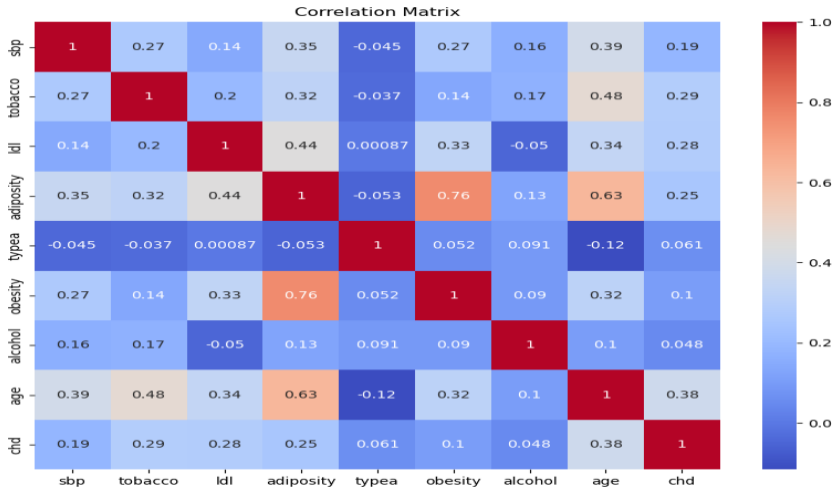
It renames the index values from 0 to 'heart disease' and 1 to 'Normal' to make the output more descriptive. t_df_mean = df_mean_comp.T: This code transposes the data to make it more suitable for plotting.

The resulting line plot visually illustrates the differences in mean values for the selected features between individuals with heart disease and those without. It provides insights into how these features may vary and contribute to the presence or absence of heart disease.

Get More Visuals On Data Distributions

Use Plotcorrelationmatrix

Code:
correlation_matrix = h_df.corr()
The heatmap visually represents the correlation between various features in the dataset, with values ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation).



Darker colors (closer to 1) indicate stronger positive correlations between variables, while lighter colors (closer to -1) suggest stronger negative correlations.
Values close to 0 indicate weak or no correlation between variables.

Top Positive Correlations

1. **Age and Tobacco Consumption (0.476452):** The highest positive correlation in the dataset is between age and tobacco consumption. As age increases, tobacco consumption tends to increase. This correlation is moderately strong, indicating that older individuals are more likely to have a history of tobacco consumption.

2. **Adiposity and Obesity (0.757865):** Adiposity (concentration of fat in the body) and obesity have a strong positive correlation. This suggests that individuals with higher adiposity levels are more likely to be obese, which is expected, as both factors relate to body fat.

3. **Age and Systolic Blood Pressure (0.392206):** Age and systolic blood pressure (sbp) have a moderate positive correlation. As individuals age, their systolic blood pressure tends to increase, indicating a connection between age and higher blood pressure.

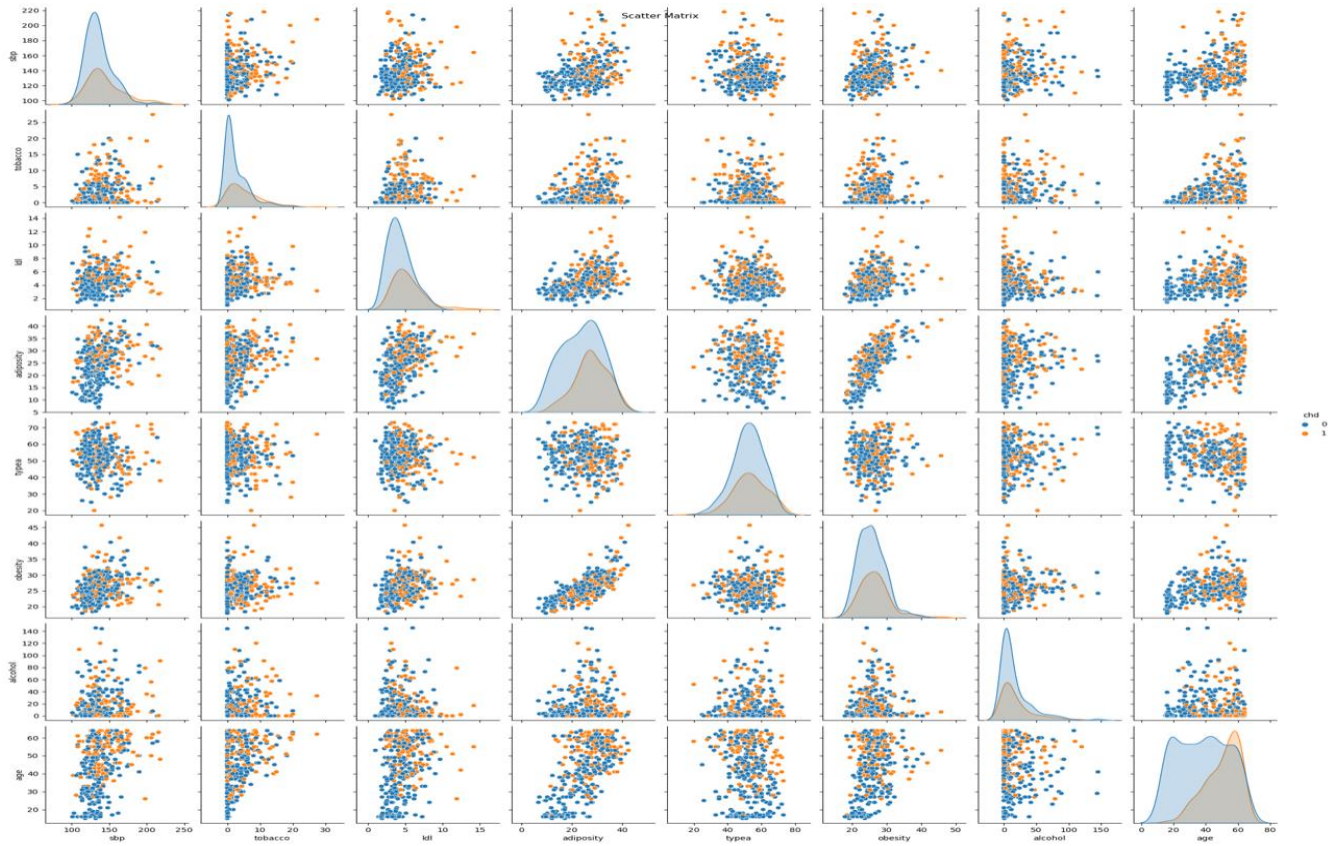
Top Negative Correlations:

1. **Type-A Behavior and Age (-0.116411):** The most negative correlation is between type-A behavior and age. As age increases, type-A behavior tends to decrease slightly. This suggests that older individuals may exhibit less type-A behavior.
2. **Adiposity and Type-A Behavior (-0.052536):** Adiposity and type-A behavior have a weak negative correlation. Higher adiposity is associated with slightly lower levels of type-A behavior.
3. **LDL Cholesterol and Alcohol Consumption (-0.050228):** LDL cholesterol and alcohol consumption also have a weak negative correlation. This implies that higher alcohol consumption is associated with slightly lower levels of LDL cholesterol.

Plotscattermatrix

The scatter plot pair plot with 'chd' as the hue variable provides a visual representation of the relationships between pairs of variables, with a focus on differentiating individuals with and without coronary heart disease (CHD). The correlation coefficients calculated from this plot help us understand the strength of these relationships.

Here's a summary of the factors with strong and moderate correlations:



Factors with Strong Correlation:

1. **Adiposity and Obesity:** These two variables exhibit a strong positive correlation, indicating that individuals with higher adiposity tend to have higher obesity levels. This correlation is significant and suggests a potential relationship between these factors

Factors with Moderate Correlation:

1. **Adiposity and Systolic Blood Pressure (SBP):** The correlation between adiposity and SBP is moderate. It implies that as adiposity increases, there is a moderate increase in systolic blood pressure.

2. **Adiposity and Low-Density Lipoprotein (LDL) Cholesterol:** Adiposity and LDL cholesterol also show a moderate correlation, indicating that as adiposity increases, there is a moderate increase in LDL cholesterol levels.

3. **Obesity and LDL:** There is a moderate correlation between obesity and LDL cholesterol. This suggests that individuals with higher obesity levels tend to have moderately higher LDL cholesterol levels.

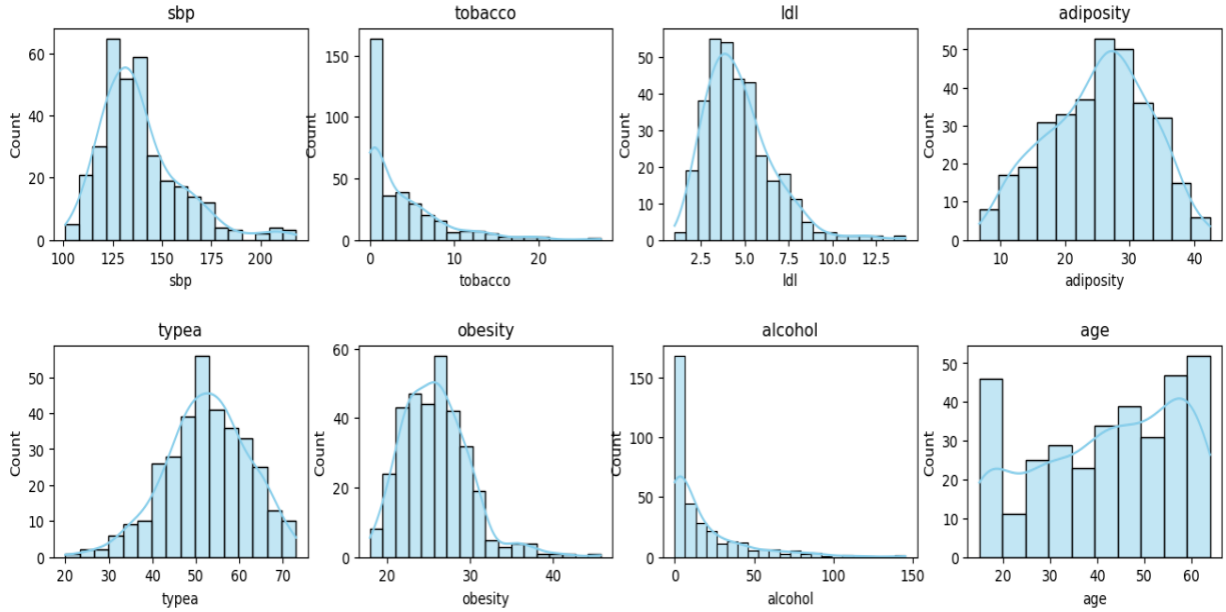
4. **Age and SBP:** Age and systolic blood pressure exhibit a moderate positive correlation, indicating that as individuals age, their systolic blood pressure tends to increase moderately.

5. **Age and Tobacco Use:** Age and tobacco use have a moderate positive correlation, implying that, on average, older individuals tend to have a slightly higher history of tobacco consumption.

6. **Age and LDL:** Age and LDL cholesterol also show a moderate correlation. As individuals age, there is a moderate tendency for LDL cholesterol levels to increase slightly.

Plot_Per_Column_Distribution:

The 8 plots in the image show the distribution for each columns The type of distribution of the data in each plot is as follows:



1. **sbp (Systolic Blood Pressure)**:
    - The distribution of 'sbp' may resemble a roughly symmetric distribution, possibly close to a normal distribution, with a mean, median, and mode that are relatively close. It might exhibit slight positive skewness if the mode is slightly higher than the median.
2. **tobacco (Tobacco Usage)**:
    - The distribution of 'tobacco' appears to be positively skewed, as the mode (0.00) is substantially lower than both the mean and median. This suggests a right-skewed distribution, which is typical for variables with many zero values and positive values.
3. **ldl (LDL Cholesterol)**:
    - The 'ldl' distribution may resemble a positively skewed distribution. The mode (3.57) is lower than the mean and median, indicating right skewness.
4. **adiposity (Adiposity Index)**:
    - The 'adiposity' distribution may exhibit a more symmetric shape, as the mean, median, and mode are relatively close. It might be approximately normally distributed.
5. **typea (Type A Behavior)**:
    - The distribution of 'typea' could be unimodal and relatively symmetric since the mode is close to both the mean and median. It may resemble a normal distribution.
6. **obesity (Obesity Index)**:
    - The 'obesity' distribution might exhibit slight positive skewness since the mode (26.09) is slightly higher than the median, which, in turn, is close to the mean.

7. **alcohol (Alcohol Consumption)**:
    - The 'alcohol' distribution seems to be right-skewed because the mode is 0.00 (indicating a significant number of non-drinkers), while the mean and median are higher, suggesting a positively skewed distribution.
8. **age (Age)**:
    - The distribution of 'age' appears to be right-skewed, with a mode (16.00) substantially lower than both the mean and median. This indicates a long tail on the right side of the distribution.

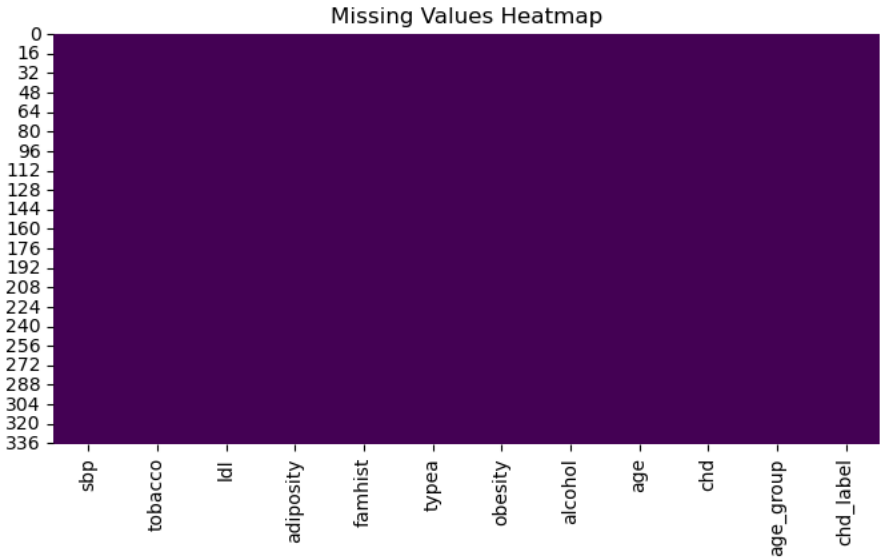
Check for missing values and What additional techniques to handle null values, excluding the drop na feature?

Code: missing_values = h_df.isnull()

1. **missing_values** is a boolean DataFrame that flags missing values in the original dataset. For each cell in the dataset, it contains **True** if the value is missing and **False** if it's not.
2. **missing_count** is created by summing the **True** values in the **missing_values** DataFrame. This operation counts how many missing values exist in each column.

In this dataset, it appears that there are no missing values (count is 0) in any of the columns, as indicated by the '0' values in the Series.

The heatmap for missing values



The heatmap for missing values is a visual representation of the dataset, where each cell corresponds to a variable. If a variable has missing values, the cell would be shaded, indicating the extent of missing data. In this specific heatmap, there are no shaded cells, which means there are no missing values in the dataset.
This is represented by the uniformity of the color, showing a consistent absence of missing data across all variables. It's a reassuring sign that the dataset is complete, and there is no need for imputation or further data cleaning.

Check for missing values and What additional techniques to handle null values, excluding the drop na feature?

Besides removing empty rows or columns, various techniques can help handle these gaps.
Some methods include:

1. **Imputation**: Mean, Median, or Mode ,Replace missing values with the mean, median, or mode of the respective column. This is a simple method and can be effective for numerical features.
2. **Prediction**: Get a little help from models.
    - Use regression to predict missing values based on related info.
    - The "K-Nearest Neighbors" approach fills gaps with nearby neighbors' values.
3. **Data-Specific Fixes**: Sometimes, there's a unique way to handle gaps.
    - Geographic data, for example, can use special methods for spatial filling.
4. **Categorical Data**: For non-numeric categories.
    - Create a new category for missing values.
    - Use the most common category for filling.
5. **Multiple Imputation**: This is like having multiple guesses.
    - Make different guesses for missing values, then combine results for a more accurate picture.
6. **Time Series Data**: When your data has a time sequence.
    - Copy the previous or next value to fill in missing points.
7. **Domain Knowledge**: Use what you know.
    - Your expertise can guide filling methods for specific data relationships.
8. **New Features**: Create hints for the gaps.
    - Add a new "missing" indicator column to show if a value was missing.
9. **Removing Rows**: If missing values are rare, drop the problematic rows.
10. **Statistical Tests**: Use data relationships to decide on filling methods.
11. **Machine Learning**: Build models to guess the missing values based on other data.
12. **Separate Datasets**: Create separate datasets for complete and incomplete data to analyze them differently and combine results.

Applying the regression models:

Process of predicting coronary heart disease (CHD) using a logistic regression model.

Model Building:

- We used a Logistic Regression model to predict CHD. This classification model is suitable for predicting binary outcomes (CHD=1 for death, CHD=0 for alive).
- We performed hyperparameter tuning using GridSearchCV to find the best combination of hyperparameters for the model. The hyperparameters included penalty, C, and max_iter.

Model Evaluation: We printed the best score achieved by the model during the hyperparameter tuning process, which represents the accuracy of the model on the training data

(best_score_: 0.7003) (70% accuracy)

Prediction: We used the trained classifier to make predictions on the test data. The predictions were assigned to the variable y_pred.

The output predictions (y_pred) contain binary values, where 1 indicates death (CHD) and 0 indicates being alive. These predictions can be compared to the actual outcomes in the test dataset to evaluate the model's performance.