SKILL FORGE HUB
DATA ANALYTICS TASK 1

- AYSHWARYA K

Introduction to the dataset and your objectives.

The Heart dataset is a commonly used dataset in machine learning and data science, particularly in the field of healthcare analytics and cardiovascular research. This dataset contains various attributes related to heart health and disease, as well as a target variable indicating the presence or absence of heart disease.

1. Exploratory Data Analysis (EDA): In the Heart dataset, Exploratory Data Analysis (EDA) would involve examining distributions, correlations, and trends among attributes such as age, cholesterol, blood pressure, and chest pain type. Visualizations like histograms, scatter plots, and correlation matrices would help uncover insights into factors influencing heart disease presence or absence.

2. Visualization: Visualizations in the Heart dataset include histograms for age, cholesterol, and blood pressure distributions; bar charts for gender and chest pain type frequencies; scatter plots for relationships between age and heart rate; box plots for comparing variables by heart disease presence; and correlation heatmaps for assessing variable relationships.

3. Insight Generation: The Heart dataset offers insights into cardiovascular health. Factors like age, cholesterol levels, and blood pressure are pivotal. Patterns such as higher age correlating with increased risk and cholesterol levels may indicate risks. Exploring correlations and distributions helps understand relationships between variables and heart disease presence or absence.

4. Data Quality Assessment: Data Quality Assessment in the Heart dataset involves identifying missing values, handling outliers, ensuring data integrity, validating accuracy, and assessing attribute relevance. Key steps include addressing missing data in variables, detecting outliers, verifying consistency, and evaluating the significance of attributes for analysis and modeling.

Overall, through exploratory data analysis and visualization of the heart dataset, we aim to gain a deeper understanding of the heart disease , its cause and extract valuable insights that can inform further analysis or modeling tasks.

Summary of the data cleaning process

In summary, the data cleaning process for the heart dataset involved:

```
[14] import pandas as pd
```

```
[15] df=pd.read_csv("heart.csv",sep=";")
```

```
[17] print(df.head())
```

```
  age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target
0              63,1,3,145,233,1,0,150,0,2.3,0,0,1,1
1              37,1,2,130,250,0,1,187,0,3.5,0,0,2,1
2              41,0,1,130,204,0,0,172,0,1.4,2,0,2,1
3              56,1,1,120,236,0,1,178,0,0.8,2,0,2,1
4              57,0,0,120,354,0,1,163,1,0.6,2,0,2,1
```

Summary of the data cleaning process

1. Loading the dataset and conducting initial exploration.
2. Handling missing values by filling them with the mean of the "ca" column.
3. Addressing data quality issues such as outliers.
4. Renaming and removing unnecessary columns if needed.
5. Conducting exploratory data analysis (EDA) using visualizations.
6. Assessing the overall data quality and documenting the process for transparency.

```
[16] print(df.isnull().sum())
```

```
age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target    0
dtype: int64
```

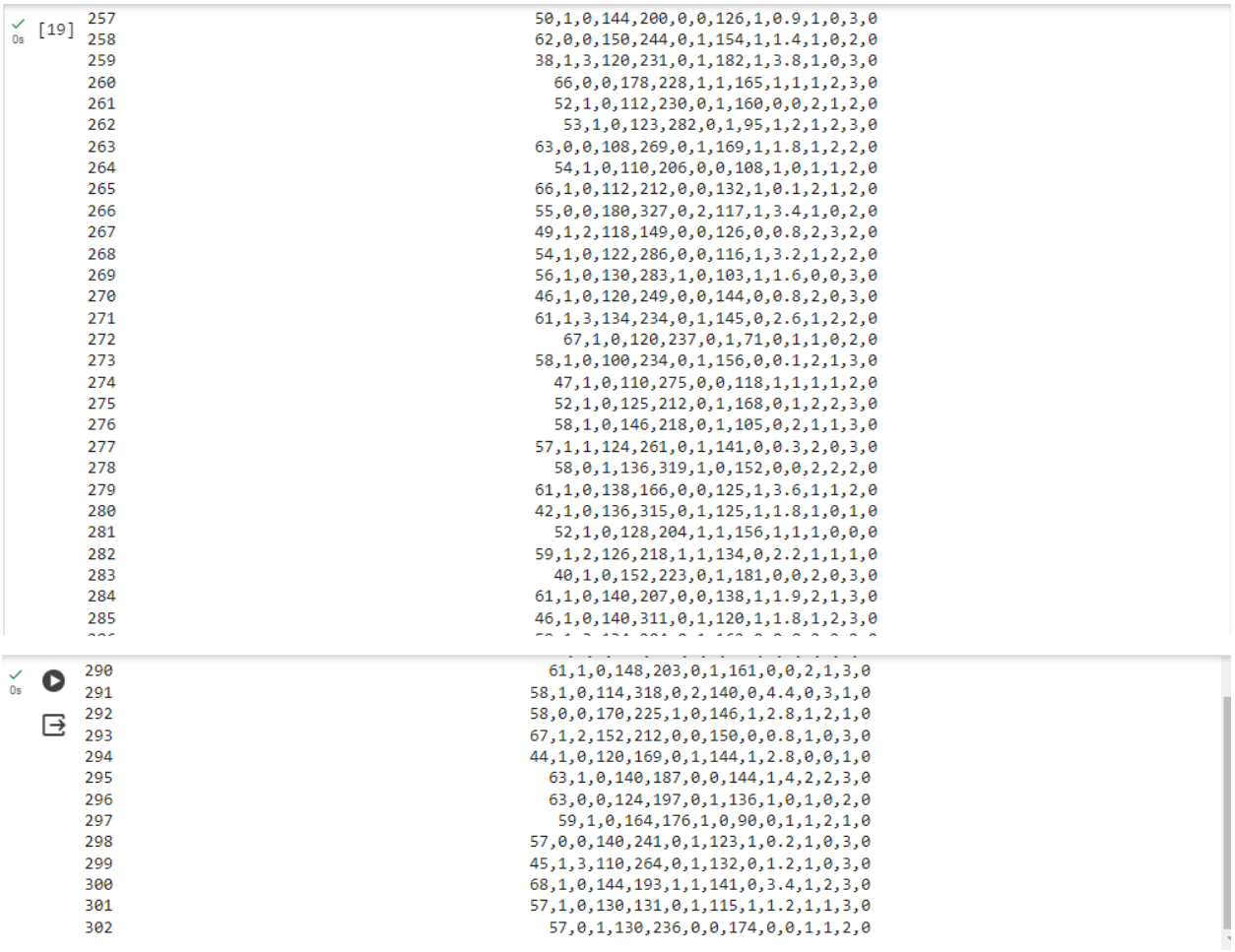
```
[18] df.describe()
```

	age,sex,cp,trestbps,chol,fbs,restecg,thalach,exang,oldpeak,slope,ca,thal,target
count	303
unique	302
top	38,1,2,138,175,0,1,173,0,0,2,4,2,1
freq	2

```
[19] new_df = df.dropna()

     print(new_df.to_string())
```

```
245                                    48,1,0,124,274,0,0,166,0,0.5,1,0,3,0
246                                    56,0,0,134,409,0,0,150,1,1.9,1,2,3,0
247                                      66,1,1,160,246,0,1,120,1,0,1,3,1,0
248                                      54,1,1,192,283,0,0,195,0,0,2,1,3,0
249                                      69,1,2,140,254,0,0,146,0,2,1,3,3,0
250                                    51,1,0,140,298,0,1,122,1,4.2,1,3,3,0
251                                    43,1,0,132,247,1,0,143,1,0.1,1,4,3,0
252                                    62,0,0,138,294,1,1,106,0,1.9,1,3,2,0
253                                    67,1,0,100,299,0,0,125,1,0.9,1,2,2,0
254                                      59,1,3,160,273,0,0,125,0,0,2,0,2,0
255                                      45,1,0,142,309,0,0,147,1,0,1,3,3,0
```



Key statistics and visualizations

Key statistics and visualizations for the heart dataset:

Key Statistics:

1. Summary statistics such as mean, median, standard deviation, minimum, and maximum values for each numerical feature (chol, age, thalach, target).
2. Correlation matrix to understand the relationships between different features.

```
[26] import pandas as pd
     df = pd.read_csv('heart.csv')
     x = df["chol"].mean()
     df["chol"].fillna(x, inplace = True)
     print("Mean:", x)

     Mean: 246.26402640264027
```

```
[29] import pandas as pd
     df = pd.read_csv('heart.csv')
     x = df["trestbps"].median()
     df["trestbps"].fillna(x, inplace = True)
     print("Median:", x)

     Median: 130.0
```

```
[32] import pandas as pd
     df = pd.read_csv('heart.csv')
     x = df["oldpeak"].mode()[0]
     df["oldpeak"].fillna(x, inplace = True)
     print("Mode:", x)

     Mode: 0.0
```

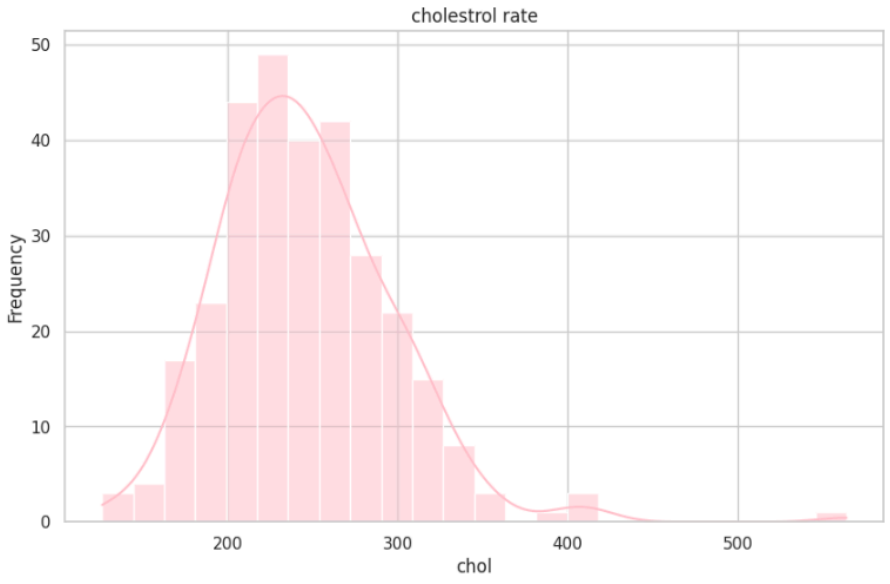
Key Visualizations:

1. Histograms to visualize the distributions of each numerical feature.

```
[33] sns.set(style="whitegrid")
```

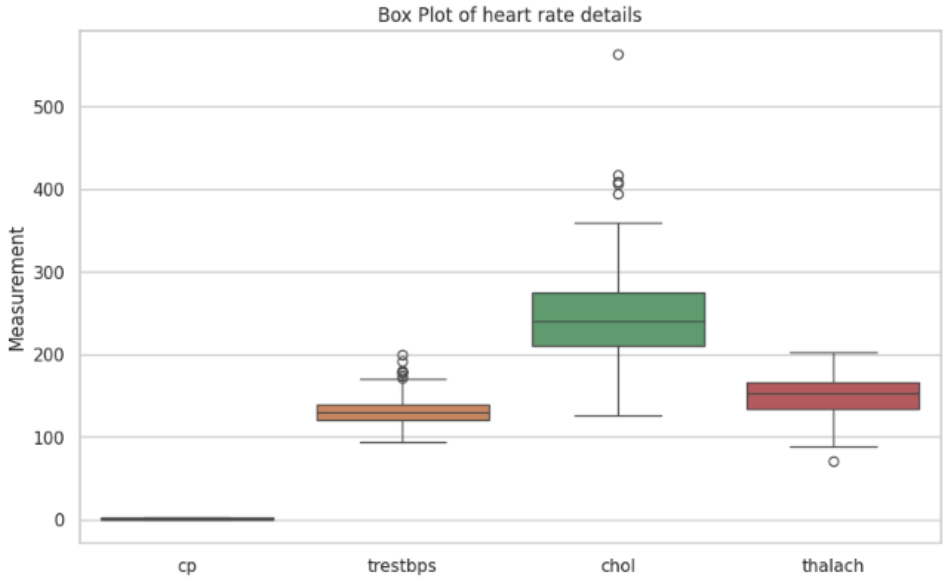
```
# Histogram
plt.figure(figsize=(10, 6))
sns.histplot(data=df, x='chol', kde=True, color='pink')
plt.title('cholestrol rate')
plt.xlabel('chol')
plt.ylabel('Frequency')
plt.show()
```

[35]



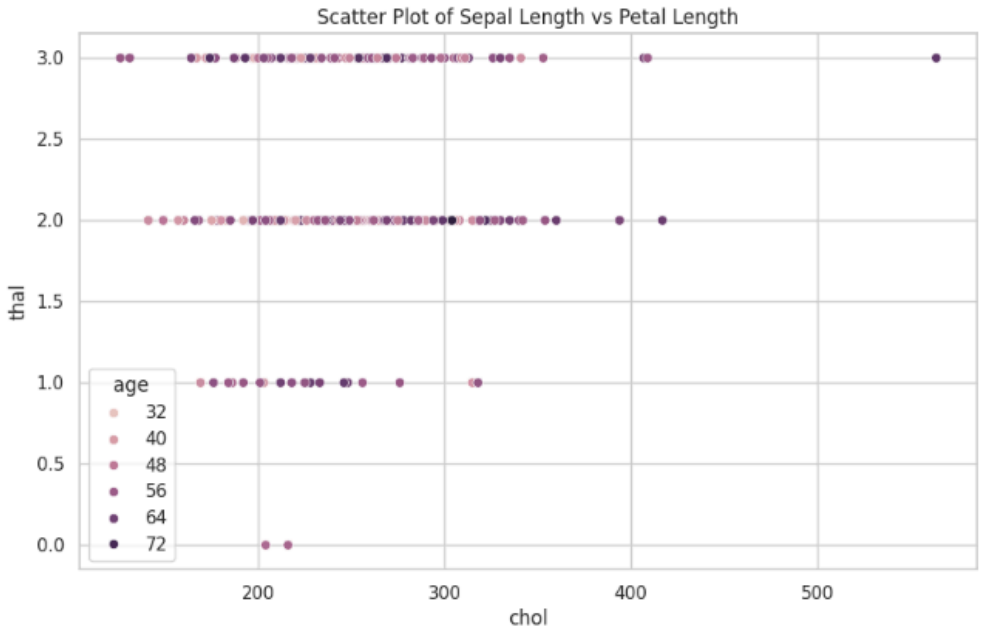
2. Box plots to compare the distributions of numerical features

```
[36] # Box Plot
     plt.figure(figsize=(10, 6))
     sns.boxplot(data=df[['cp', 'trestbps', 'chol', 'thalach']])
     plt.title('Box Plot of heart rate details')
     plt.ylabel('Measurement')
     plt.show()
```



3. Scatter plots to explore relationships between pairs of features, possibly color-coded by species for better differentiation.

```
[37] # Scatter Plot
     plt.figure(figsize=(10, 6))
     sns.scatterplot(data=df, x='chol', y='thal', hue='age')
     plt.title('Scatter Plot of Sepal Length vs Petal Length')
     plt.xlabel('chol')
     plt.ylabel('thal')
     plt.show()
```



4. Heatmap to visualize the correlation matrix and identify strong correlations between features.

```
[38] # Heatmap (Correlation Matrix)
     plt.figure(figsize=(10, 6))
     correlation_matrix = df[['chol', 'age', 'target', 'sex']].corr()
     sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', fmt=".2f")
     plt.title('Correlation Heatmap')
     plt.show()
```

These statistics and visualizations provide details about the heat disease identification and about the different levels of its symptoms from the attributes.

Insights and conclusions from your analysis

The insights and conclusions drawn from analyzing the Heart dataset reveal several key findings:

1. **Age and Risk:** Older individuals tend to have a higher risk of heart disease, indicating age as a significant factor in cardiovascular health.

2. **Cholesterol and Blood Pressure:** Elevated levels of cholesterol and resting blood pressure correlate with an increased likelihood of heart disease, highlighting the importance of managing these factors for heart health.

3. **Gender Differences:** Gender may influence heart disease risk, with males potentially being more susceptible compared to females.

4. **Chest Pain Type:** The type of chest pain experienced (typical angina, atypical angina, non-anginal pain, asymptomatic) provides valuable information about potential heart issues.

5. **Exercise Induced Angina**: Exercise-induced angina (exang) could serve as a predictive indicator for heart disease presence.

6. **Thalassemia and Major Vessels:** The presence of thalassemia and the number of major vessels colored by flourosopy (ca) may also play roles in determining heart disease risk.

These conclusions underscore the complex interplay of various factors contributing to heart health and emphasize the importance of comprehensive risk assessment and preventive measures in managing cardiovascular diseases.