Heart Disease Prediction

Milestone: Data Exploration and Visualization

Group 3

Mugdha Sanjay Parbat - 002142372

Pranav Chandrakant Pulkundwar - 002121679

Telephone

+1 (617) 901-8417

+1 (617) 901-8418

Email ID

parbat.m@northeastern.edu

pulkundwar.p@northeastern.edu

Percentage of contribution by Student 1: 50%

Percentage of contribution by Student 2: 50%

Submission Date: 14th March 2022

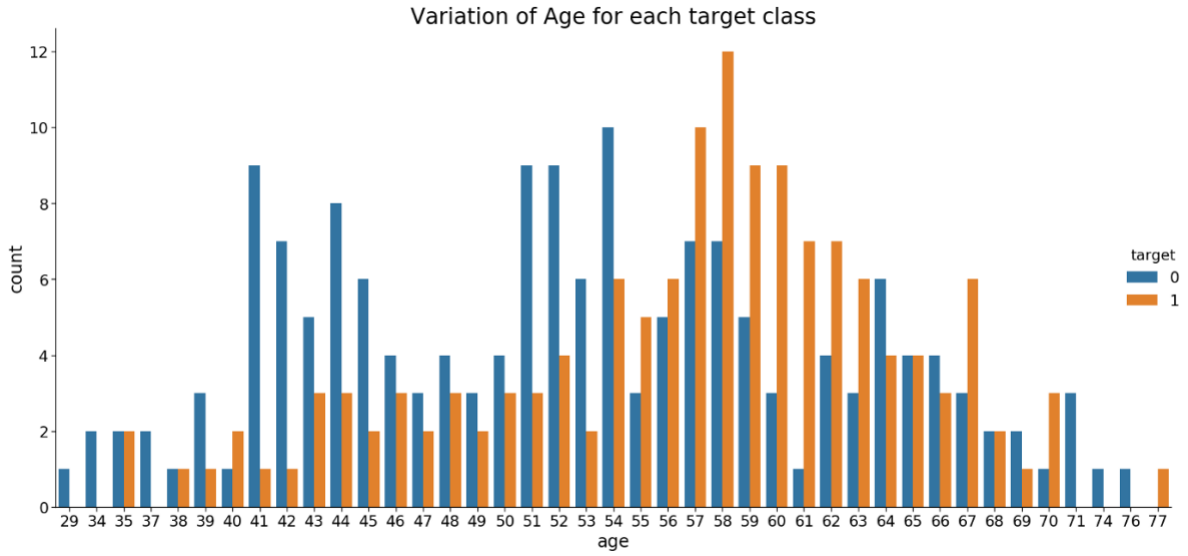
The dataset consists of 303 rows and 14 columns of data. The 14 features or columns considered in this visualization.

The code and the visualizations of the data are given below

```
> import numpy as np
> import pandas as pd
> import seaborn as sns
> import matplotlib.pyplot as plt
> %matplotlib inline
> df = pd.read_csv('heart.csv')
> df.head()
```

The plot for variation of age for each target class is given below

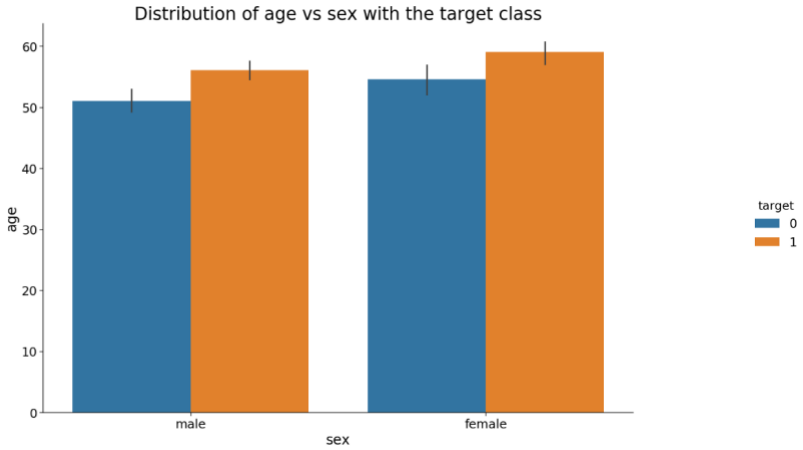
```
> sns.set_context("paper", font_scale = 2, rc = {"font.size":
  20,"axes.titlesize": 25,"axes.labelsize": 20})
> sns.catplot(kind = 'count', data = df, x = 'age', hue = 'target', order
  = df['age'].sort_values().unique())
> plt.title('Variation of Age for each target class')
> plt.show()
```



In the plot titled as 'Variation of Age for each target class', target equals 1 means that the person is suffering from heart disease and target = 0 says that the person is not suffering.
It is evident from the chart that majority of people suffering from the heart disease are 58 years old, and second highest are 57 years old. Overall, people who are 50+ years old suffer from heart disease.

Bar-plot of age vs sex with color differentiation of target

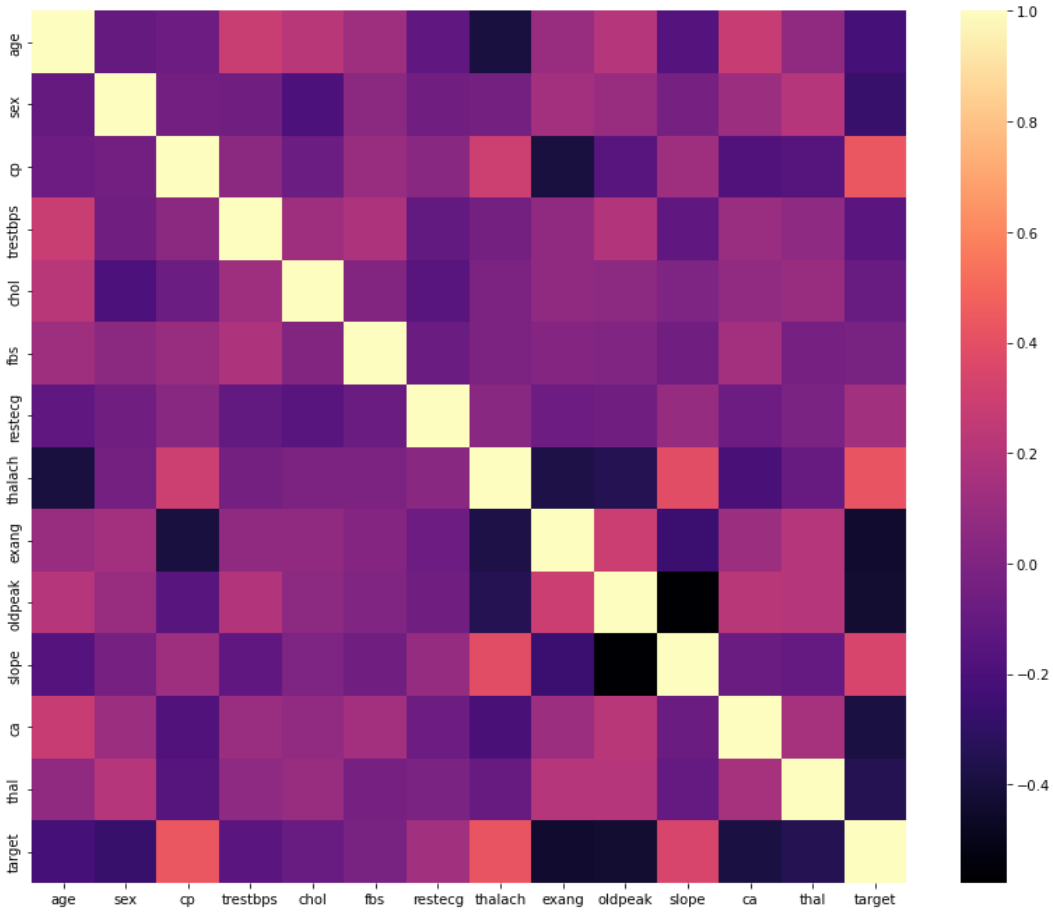
```
> sns.catplot(kind = 'bar', data = df, y = 'age', x = 'sex', hue =
  'target')
> plt.title('Distribution of age vs sex with the target class')
> plt.show()
```



Now the next plot, bar-plot of age vs sex represents distribution of age and gender for each target. Age of males who are suffering from the heart disease are younger than the females.

The heat-map of the feature of the dataset is given below.

```
> hc = df.corr()
> fig, ax = plt.subplots(figsize = [14,12])
> sns.heatmap(hc, ax = ax, cmap = 'magma')
```

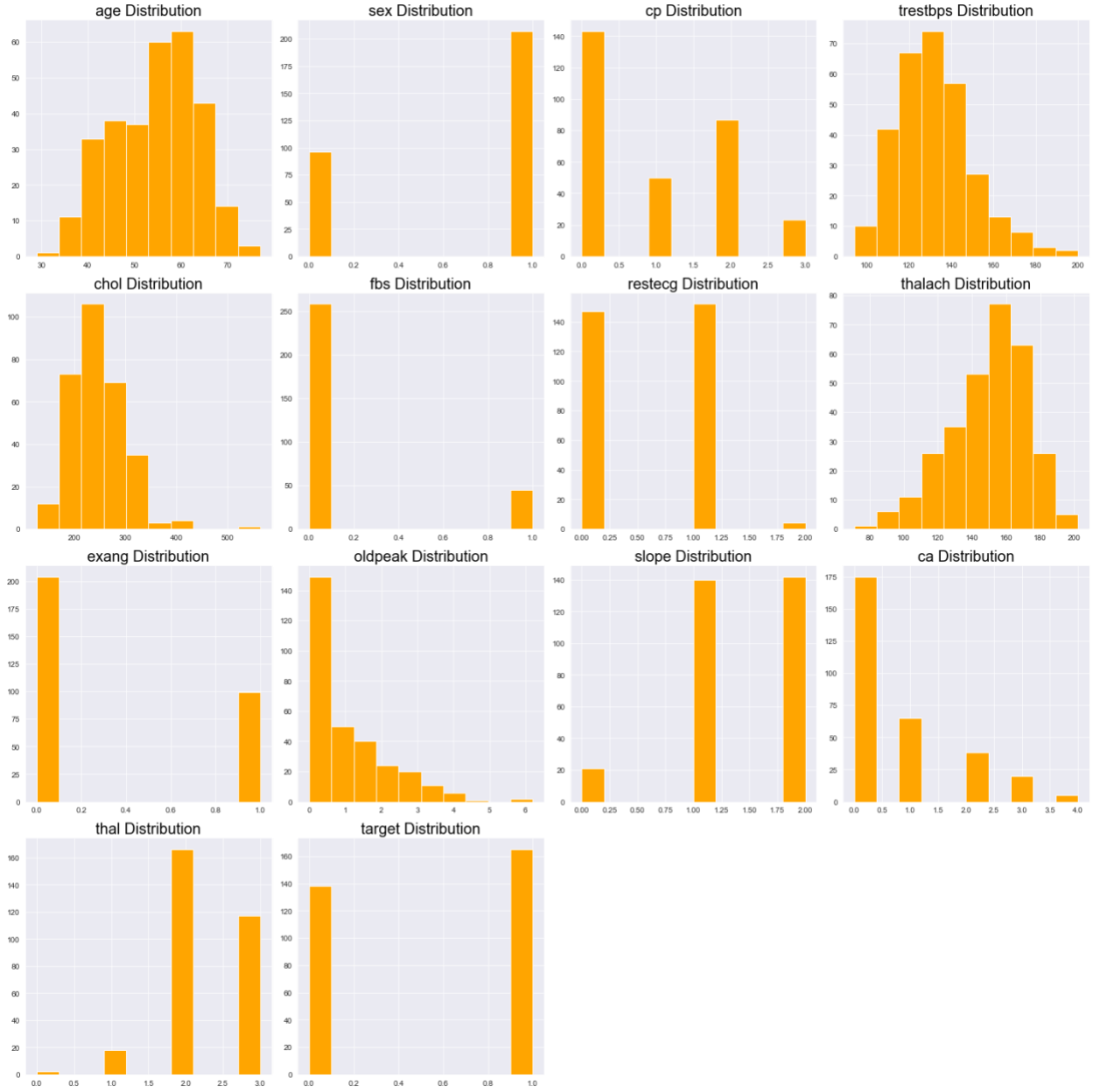


It is evident from the heatmap that there is no single feature that is very highly correlated with our target value. Some of the features have a positive and some have negative correlation with the target values.

The Hist-plots for the dataset are:

```
> def draw_histograms(dataframe, features, rows, cols):
>     fig = plt.figure(figsize = (20,20))
>     for i, feature in enumerate(features):
>         ax = fig.add_subplot(rows, cols, i+1)
>         dataframe[feature].hist(bins = 10, ax = ax, facecolor = 'orange')
>         ax.set_title(feature + " Distribution", color = 'black', fontsize
  = 20)
```

```
>       fig.tight_layout()
>       plt.show()
>   draw_histograms(df, df.columns, 4, 4)
>   sns.set_style('darkgrid')
```



The histogram plots show how each feature and label is distributed along different ranges. This tells us that the data needs to be scaled. A categorical variable will have a discrete bar. Before applying machine learning, we will have to take care of the categorical variables by using dummy variables or any other method suitable.