DATA ANALYTICS TRAINING

FIDELITY INTERNATIONAL

HEART - SEIZURE PREDICTION

SUBMITTED BY -

SHUBHAM

Project Location

This project is stored on GitHub - an online project repository platform. This platform provides with many features that helps its user to keep their projects safely and securely. It also provides feature like version control of the file and allows others collaboration on projects.

https://github.com/Shubham14793/UCDPA_shubham

ABSTRACT

In this project I have tried to use and implement the learning that was provided by the trainer during the online course and data camp worksheets. During development of main project developed I realised that few topics that were discussed might not be directly utilised in project. Hence, I have created to project files-

1. DATA_IMPORT_&_ANALYSIS.IPYNB

This first file contains concepts like Importing a dataset using an api and then processing the response data. Creating function to reutilise the code and using regular expression to extract useful data from response of an api.

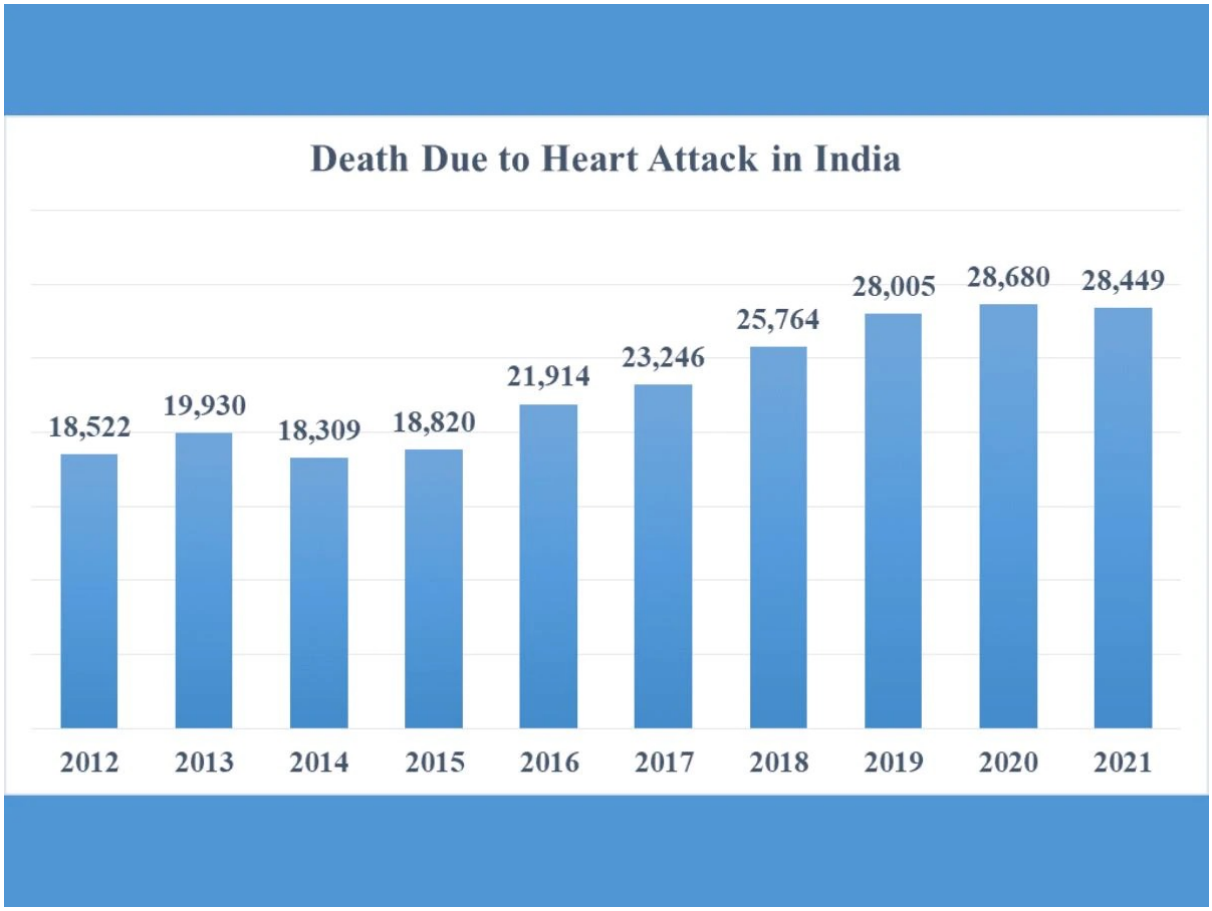
2. HEART_SEIZURE_DETECTION_MAIN.IPYNB

The second file contains the the main project. Heart-Seizure analysis is a predictive data analysis project which helps us identifying whether or not a person or subject is susceptible to Heart Seizure or Heart Attack. This project utilises different aspect of a patient or subject to predict its conclusion. We will be using a dataset available from highly reliable datasource. There are 2 separate datasets available of the patients. We will combine these dataset together and doing the analysis on the combined dataframe.

First we will be importing the data from the data source and storing it locally. We will understand the data through various methods available from python language and then perform certain operations that will provide us deep insights of the data. We will use python packages to do a comparative analysis of dataset through charts. We then will move on to using AI models to create a predictive model which will help us make decision on possibility of random person being diagnosed with Heart-Seizure.

INTRODUCTION

As the world is emerging out of COVID-19 pandemic, deteriorating heart conditions of the citizen has been another big concern for the countries. Developing countries like India have seen sudden increase in Heart-Attack cases where in most of the cases the victim resulted in loss of life. In the last 10 years, from 2012 to 2021, there had been 54% increase in heart attack deaths, according to NCRB.



According to an article publish by INDIA TODAY:

“Heart-related ailments among young Indians have become a concern for health experts. The issue is being debated why there has been a surge in heart attack cases in relatively younger age groups in India. Many people have associated it with Covid-19.”

This issue requires some immediate actions in order to understand and reduce down cardiovascular related cases. A healthy lifestyle has always been advised but there can be other sources as well which can contribute in addressing this long prevailing situation. Early detection of cardiac diseases and continuous supervision of clinicians can reduce the mortality rate. However, accurate detection of heart diseases in all cases and consultation of a patient for 24 hours by a doctor is not available since it requires more sapience, time and expertise.

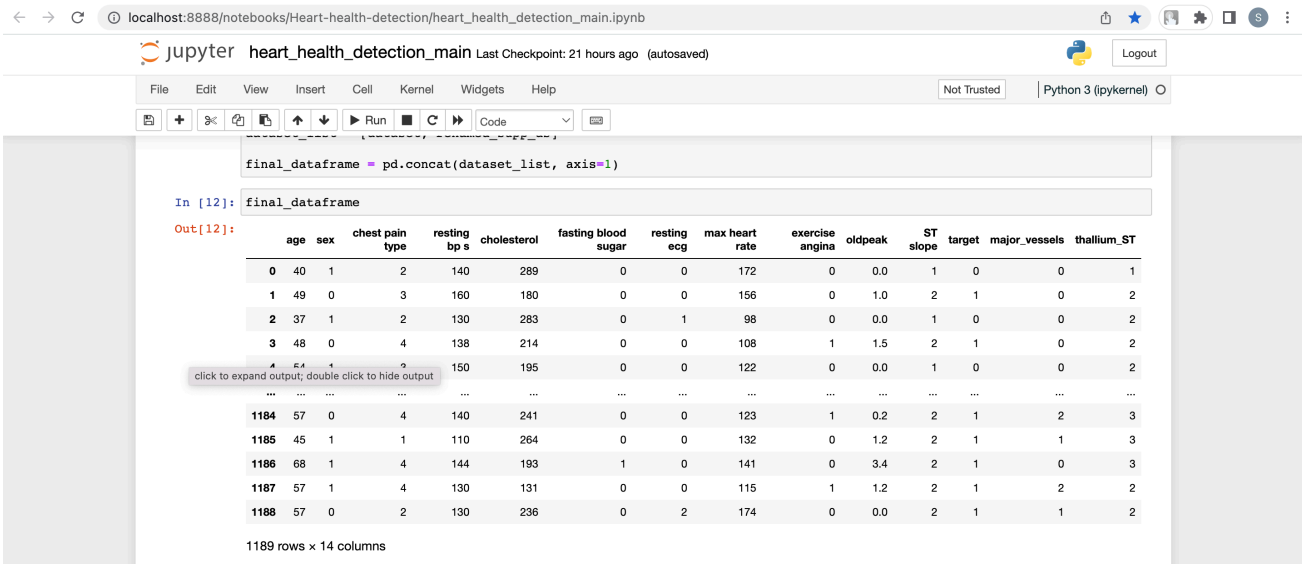
For the accurate detection of the heart disease we can utilise the advanced computer methodologies and techniques. Using Data Science and Artificial Intelligence for anomalies related to Heart condition can be a wonderful tool. In this project we are presenting a machine learning-based heart attack prediction method in which the analysis of different risk factors and prediction for heart attacks is done using ML approaches of Support Vector Machines, Logistic Regression. The data of heart disease symptoms has been collected dataset provided by IEEE.ORG and analysis has been performed on the data using ML methods. The focus has been on optimising the prediction on the basis of different parameters.

We observed that using ML approaches of Support Vector Machines, Logistic Regression helped AI model to predict possibility of Heart-Attack in a patient with more than 85% accuracy.

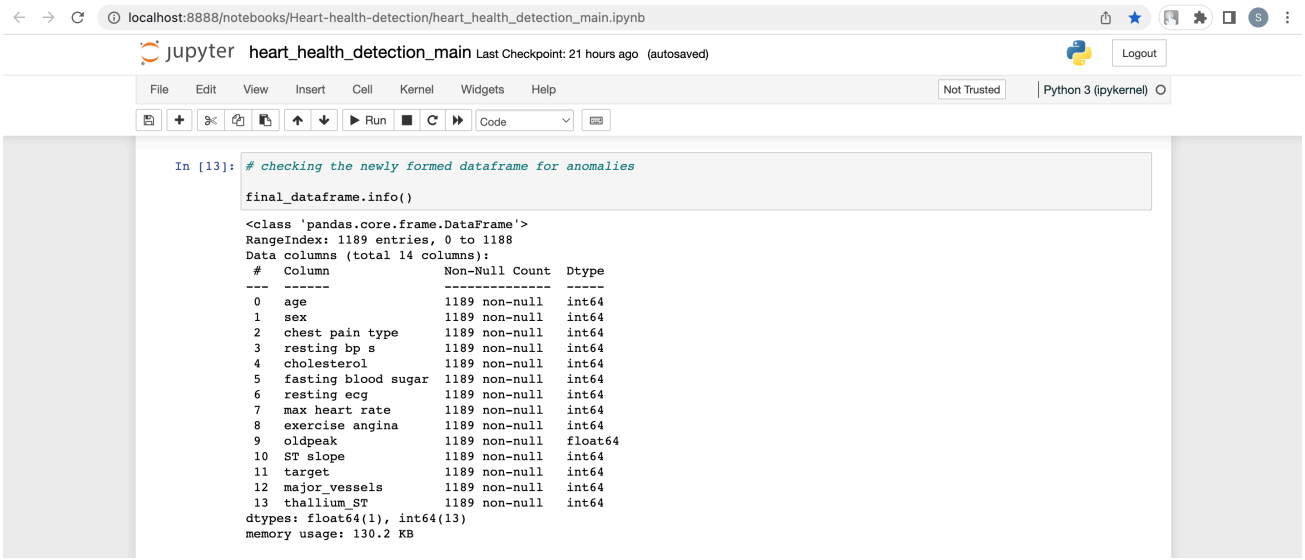
DATASET

The dataset that was utilised for this project is from ieee-dataport.org. The reason why I chose this dataset is that this heart disease dataset is created by combining 5 popular heart disease datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over multiple common features which makes it the largest heart disease dataset available so far for research purposes.

The dataset is consisting of 14 total columns and 1190 total rows. Each column is either of type int64 or float64.



	age	sex	chest pain type	resting bp s	cholesterol	fasting blood sugar	resting ecg	max heart rate	exercise angina	oldpeak	ST slope	target	major_vessels	thallium_ST
0	40	1	2	140	289	0	0	172	0	0.0	1	0	0	1
1	49	0	3	160	180	0	0	156	0	1.0	2	1	0	2
2	37	1	2	130	283	0	1	98	0	0.0	1	0	0	2
3	48	0	4	138	214	0	0	108	1	1.5	2	1	0	2
...	...	...	...	150	195	0	0	122	0	0.0	1	0	0	2
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1184	57	0	4	140	241	0	0	123	1	0.2	2	1	2	3
1185	45	1	1	110	264	0	0	132	0	1.2	2	1	1	3
1186	68	1	4	144	193	1	0	141	0	3.4	2	1	0	3
1187	57	1	4	130	131	0	0	115	1	1.2	2	1	2	2
1188	57	0	2	130	236	0	2	174	0	0.0	2	1	1	2



DATASET SOURCE

The dataset can be downloaded from below URL-

https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive

DATASET COLUMNS DESCRIPTION

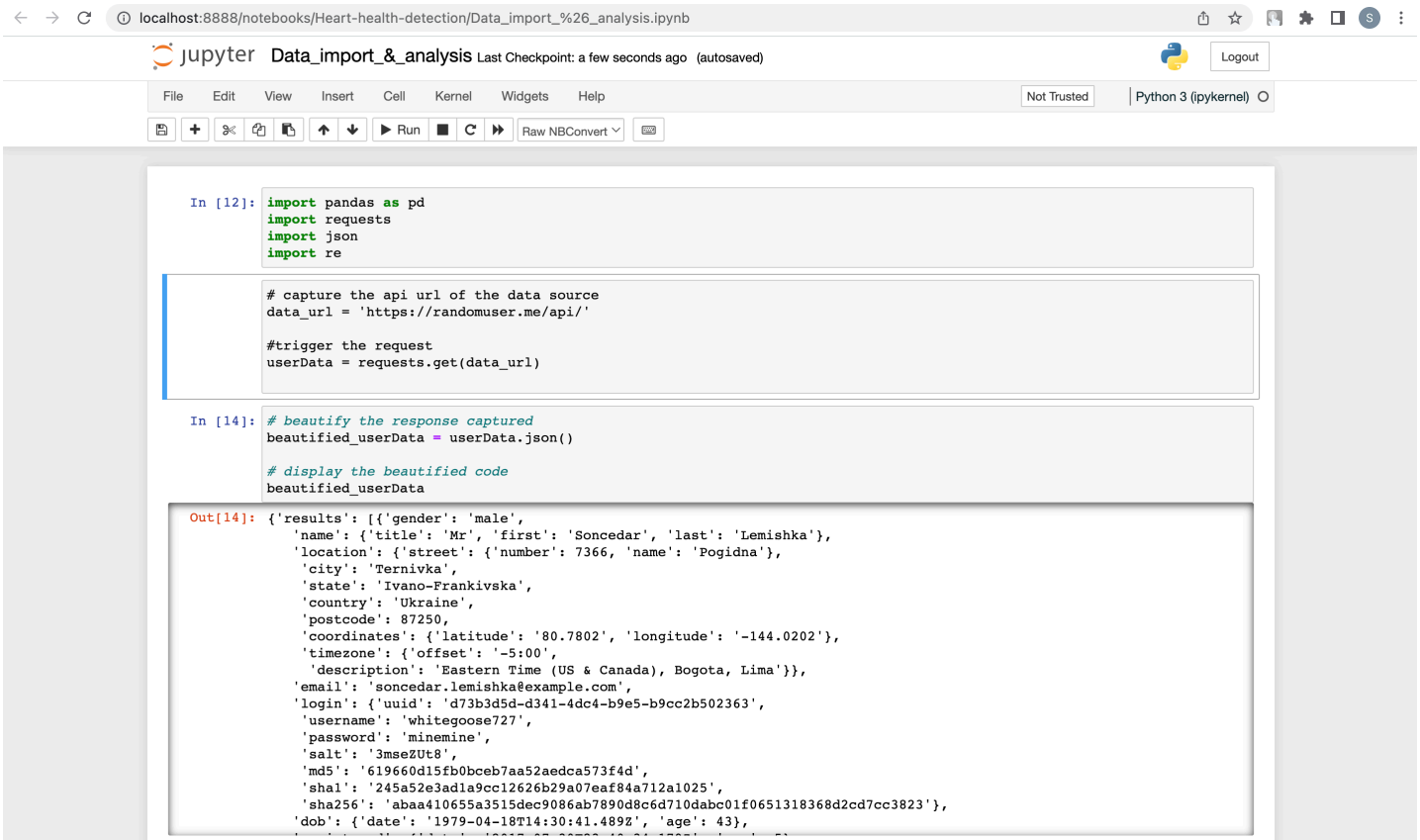
S.No.	Column Name	Column Description
1	Age	Age of the Subject
2	Sex	Gender of the Subject
3	Chest pain type	Categories - [0, 1, 2, 3] 0 -> Typical Pain 1 -> Atypical Pain 2 -> Non-specific chest pain 3 -> Asymptomatic
4	Resting BP S	Blood Pressure at resting state of Subject
5	Cholesterol	Cholesterol value measure by a device of Subject
6	Fasting blood sugar	High Blood Sugar during fasting (if > 120 then 1 otherwise 0)
7	Resting ECG	Electrocardiography during resting. Categories - [0, 1, 2] 0 -> Normal 1 -> ST-T wave normality () 2 -> Left ventricular hypertrophy (wall thickening of heart chamber)
8	Maximum Heart Rate	Maximum Heart Rate of the subject
9	Exercise Angina	Pain due to exercise [0 -> no, 1 -> yes]
10	Oldpeak	Previous Peak
11	ST Slope	Predictor of coronary artery disease
12	Target	Output variable
13	Major Vessel	Heart's Major Vessels
14	Thallium ST	Thallium stress test is an indicator of how well your blood flows to your heart.

IMPLEMENTATION PROCESS

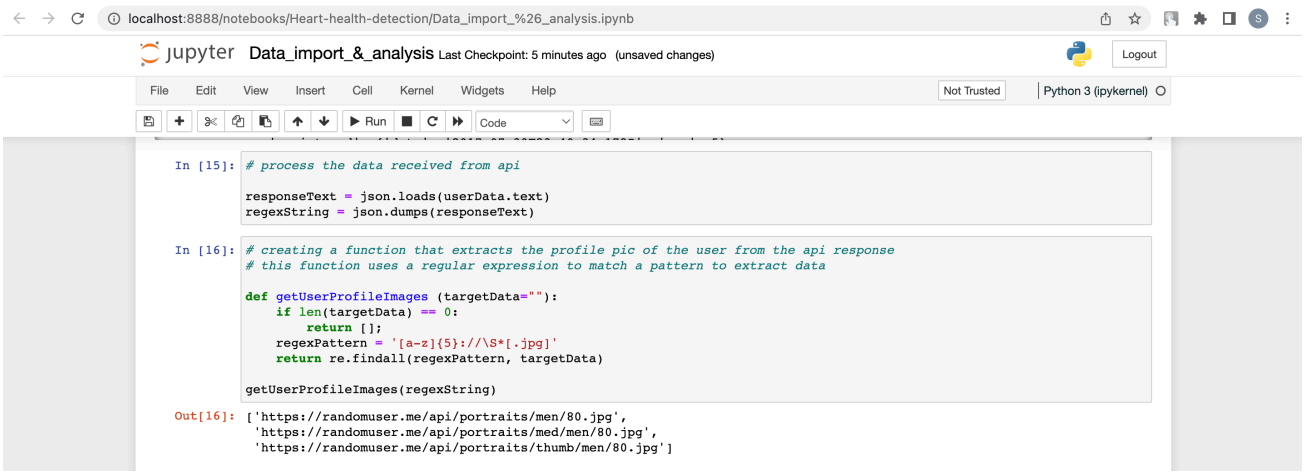
ADDITIONAL GRADING STEPS

In order to meet the grading requirement of the project, I have performed several steps additional to the main project. These additional steps include -

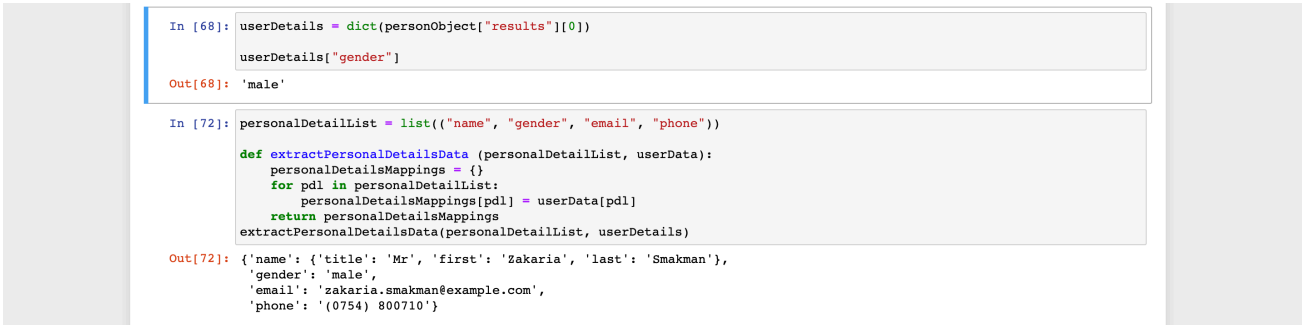
1. Importing Data using an API and processing the response.



2. Analysing Data using Regular expressions and iterating through the Data.



3. Implementing python language specific function to make reusable code and creating Lists or Dictionary data structure to store the values.

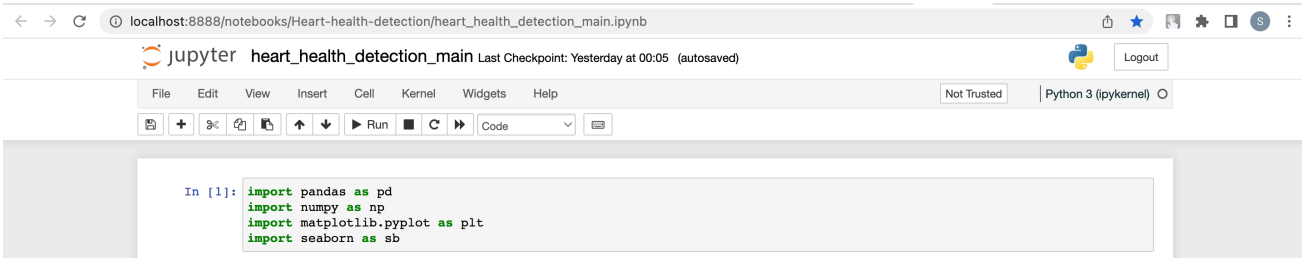


MAIN PROJECT

With goal of creating a model that is capable of predicting if a person is prone to a Heart Attack, I have performed series of operations on the dataset. The whole process of creating the model can be divided into several sub-sections.

1. PACKAGE IMPORT

In this section I imported all the project dependency. These dependencies are required to perform operations on the datasets that will be imported.



Pandas is a Python package **providing fast, flexible, and expressive data structures** designed to make working with “relational” or “labeled” data both easy and intuitive. It aims to be the fundamental high-level building block for doing practical, real-world data analysis in Python.

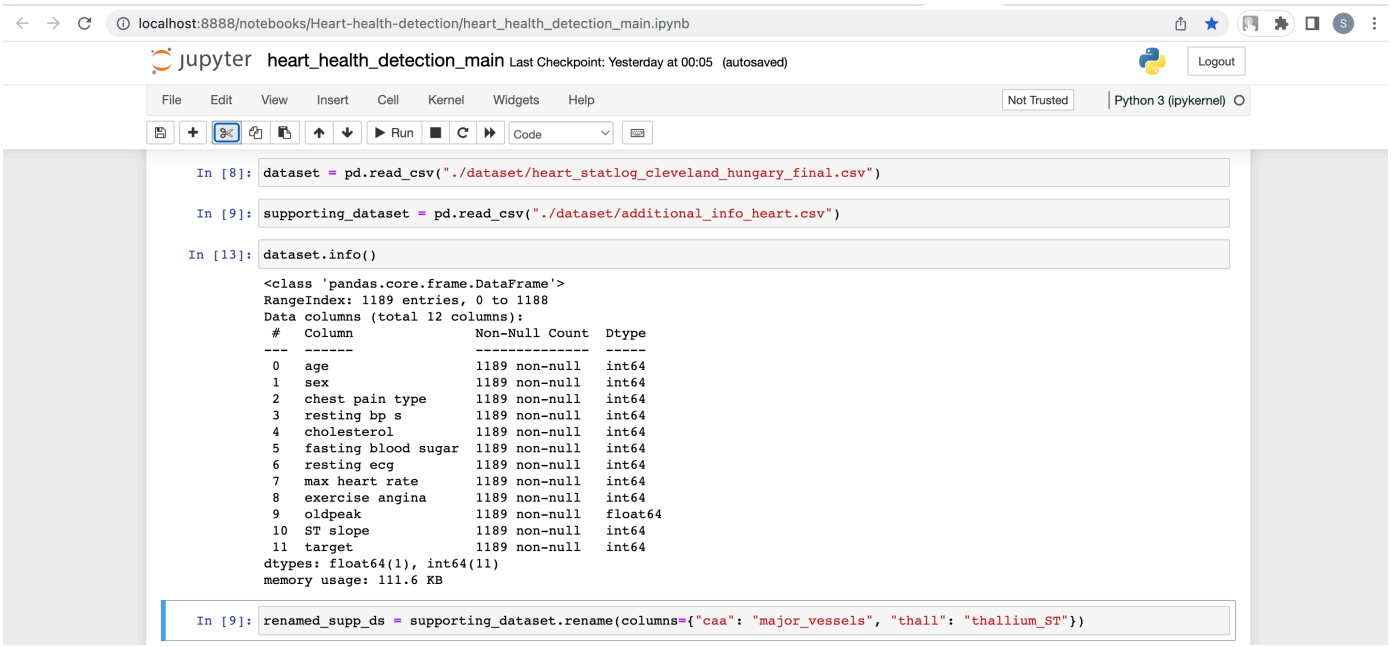
NumPy is **a Python library used for working with arrays**. It also has functions for working in domain of linear algebra, fourier transform, and matrices.

Matplotlib is **a plotting library for Python**. It is used along with NumPy to provide an environment that represents the data Graphically. This library has lot of functionalities that can be helpful in visual representation of the processed data

Seaborn is an amazing visualisation library for statistical graphics plotting in Python. It provides beautiful default styles and colour palettes to make statistical plots more attractive. It is built on the top of matplotlib library and also closely integrated to the data structures from pandas.

2. DATA IMPORT

In this section the dataset was imported after downloading it locally. Its benefit is that there is no dependency on internet connection. Using pandas based method can be utilised for importing the dataset.

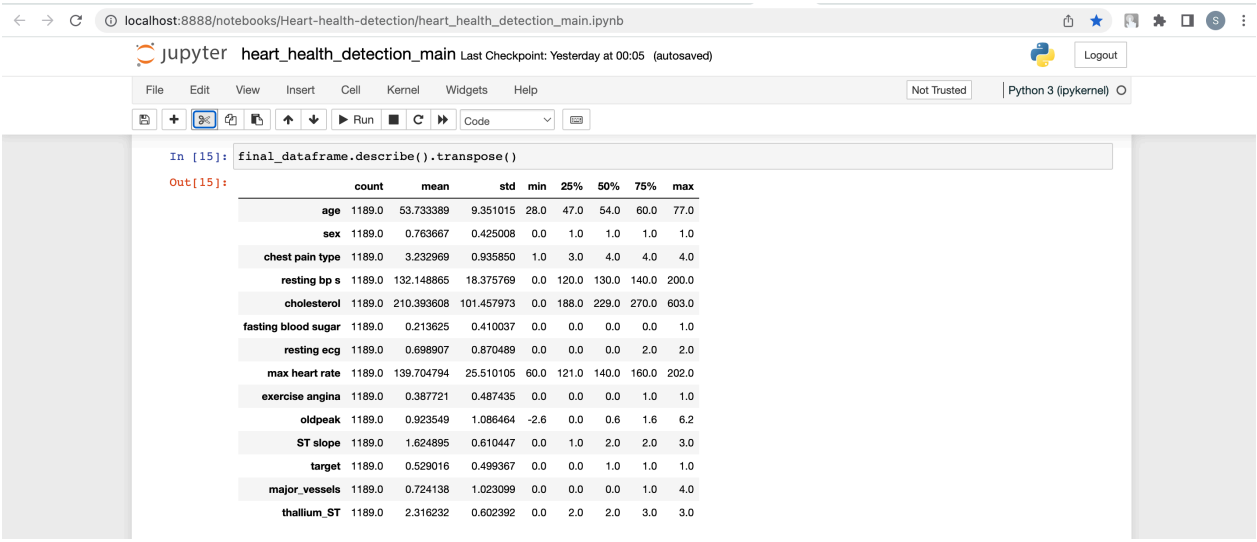


One the data is imported successfully. Information about the data can be checked using **info() method** provided by pandas package.

3. DATA CLEANING AND PROCESSING.

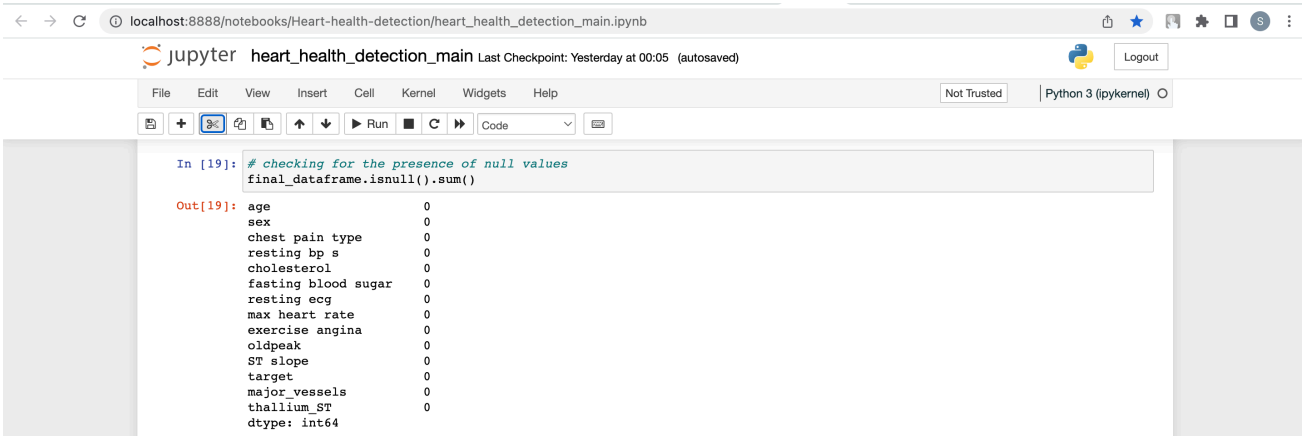
This part is one of the most important step of whole process. Effective data cleaning is a vital part of the data analytics process. If the data has inconsistencies or error, there will be very high chances of flawed results. Data cleaning activity provides several benefits like - **Organised data, avoids mistake, productivity improvement, cost reduction and improved mapping of the data**.

Pandas describe() method is used to view some basic statistical details like percentile, mean, std etc. of a data frame or a series of numeric values. When this method is applied to a series of string, it returns a different output which is shown below.



	count	mean	std	min	25%	50%	75%	max
age	1189.0	53.733389	9.351015	28.0	47.0	54.0	60.0	77.0
sex	1189.0	0.763667	0.425008	0.0	1.0	1.0	1.0	1.0
chest pain type	1189.0	3.232969	0.935850	1.0	3.0	4.0	4.0	4.0
resting bp s	1189.0	132.148865	18.375769	0.0	120.0	130.0	140.0	200.0
cholesterol	1189.0	210.393608	101.457973	0.0	188.0	229.0	270.0	603.0
fasting blood sugar	1189.0	0.213625	0.410037	0.0	0.0	0.0	0.0	1.0
resting ecg	1189.0	0.698907	0.870489	0.0	0.0	0.0	2.0	2.0
max heart rate	1189.0	139.704794	25.510105	60.0	121.0	140.0	160.0	202.0
exercise angina	1189.0	0.387721	0.487435	0.0	0.0	0.0	1.0	1.0
oldpeak	1189.0	0.923549	1.086464	-2.6	0.0	0.6	1.6	6.2
ST slope	1189.0	1.624895	0.610447	0.0	1.0	2.0	2.0	3.0
target	1189.0	0.529016	0.499367	0.0	0.0	1.0	1.0	1.0
major_vessels	1189.0	0.724138	1.023099	0.0	0.0	0.0	1.0	4.0
thallium_ST	1189.0	2.316232	0.602392	0.0	2.0	2.0	3.0	3.0

While making a Data Frame from a csv file, many blank columns are imported as null value into the Data Frame which later creates problems while operating that data frame. Pandas isnull() and notnull() methods are used to check and manage NULL values in a data frame



4. DATASET CATEGORISATION

The data columns provided by the Dataset are of different type i.e for few fields the value can be anything between a large range and it can be one of the limited options for other fields. Hence the dataset can be divided into categories-

1) Categorical columns 2) Continuous columns

But, why do we need to categorise the dataset ?

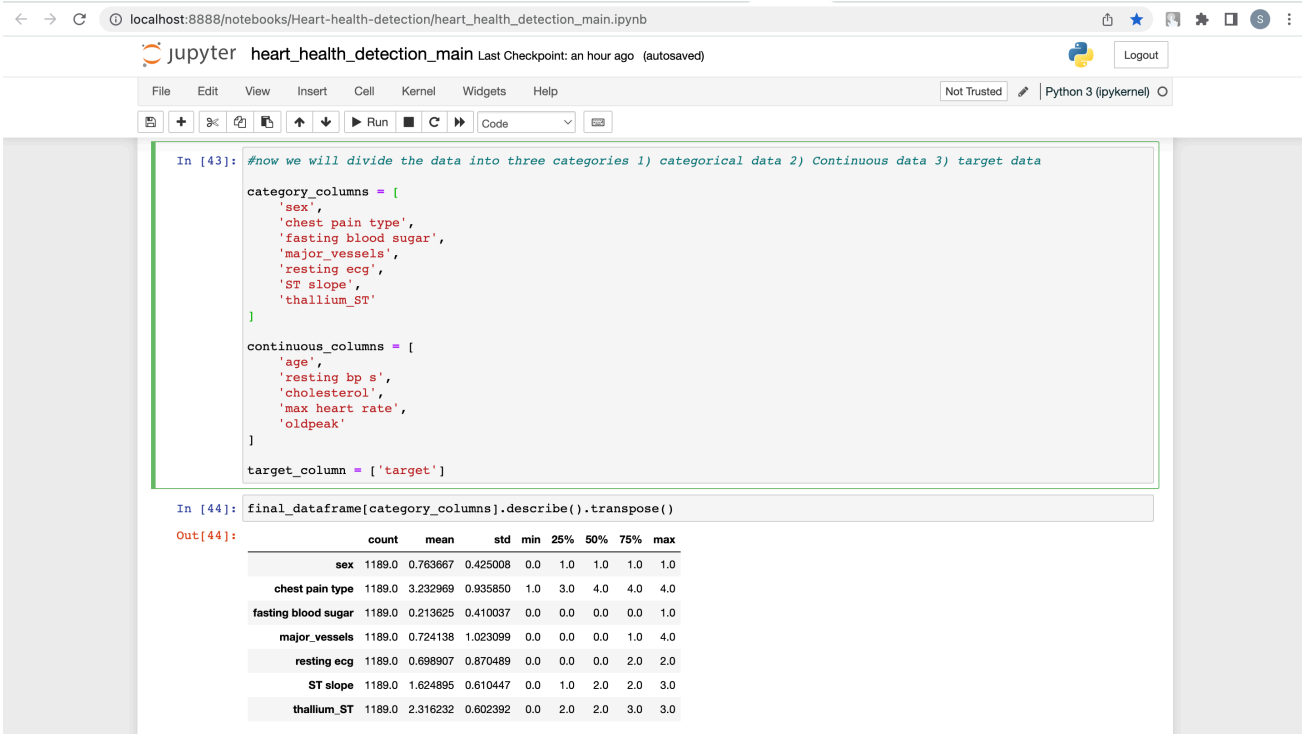
Well since we are doing analysis on the dataset, there are many techniques that Pandas library provide to do that. Using categories can bring some significant benefits:

— **Memory usage** for columns where there are many repeated values, categories can drastically reduce the amount of memory required to store the data in memory.

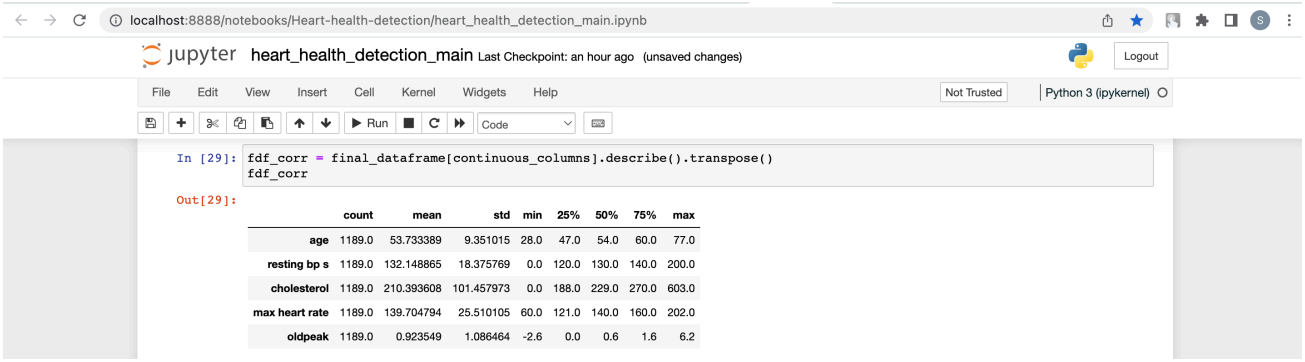
— **Runtime performance** are optimisations in place which can improve execution speed for certain operations.

— **ibrary integrations** have special functionality for categorical columns.

Here is an screenshot of categorisation that I did for the Dataset columns:



	count	mean	std	min	25%	50%	75%	max
sex	1189.0	0.763667	0.425008	0.0	1.0	1.0	1.0	1.0
chest pain type	1189.0	3.232969	0.935850	1.0	3.0	4.0	4.0	4.0
fasting blood sugar	1189.0	0.213625	0.410037	0.0	0.0	0.0	0.0	1.0
major_vessels	1189.0	0.724138	1.023099	0.0	0.0	0.0	1.0	4.0
resting ecg	1189.0	0.698907	0.870489	0.0	0.0	0.0	2.0	2.0
ST slope	1189.0	1.624895	0.610447	0.0	1.0	2.0	2.0	3.0
thallium_ST	1189.0	2.316232	0.602392	0.0	2.0	2.0	3.0	3.0



	count	mean	std	min	25%	50%	75%	max
age	1189.0	53.733389	9.351015	28.0	47.0	54.0	60.0	77.0
resting bp s	1189.0	132.148865	18.375769	0.0	120.0	130.0	140.0	200.0
cholesterol	1189.0	210.393608	101.457973	0.0	188.0	229.0	270.0	603.0
max heart rate	1189.0	139.704794	25.510105	60.0	121.0	140.0	160.0	202.0
oldpeak	1189.0	0.923549	1.086464	-2.6	0.0	0.6	1.6	6.2

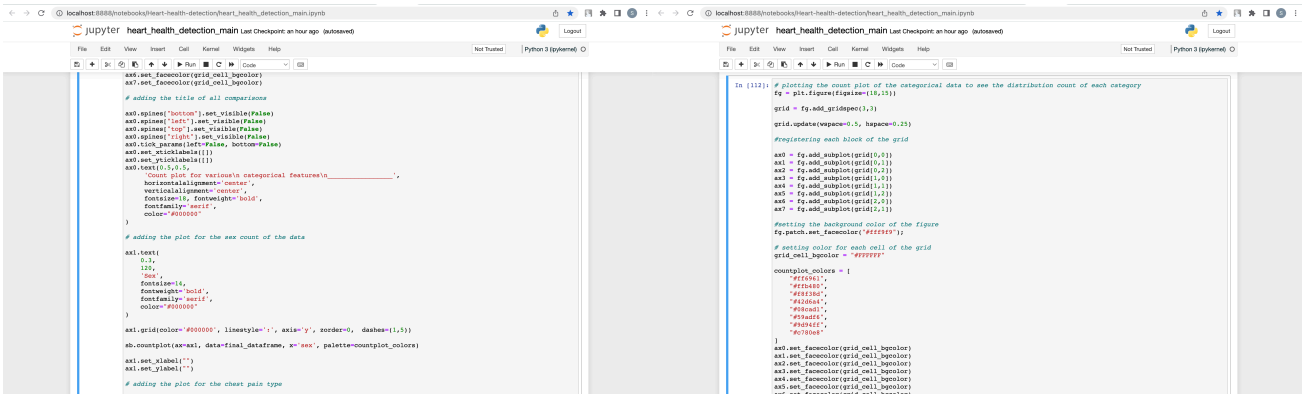
5. EXPLORATORY DATA ANALYSIS & VISUAL REPRESENTATION

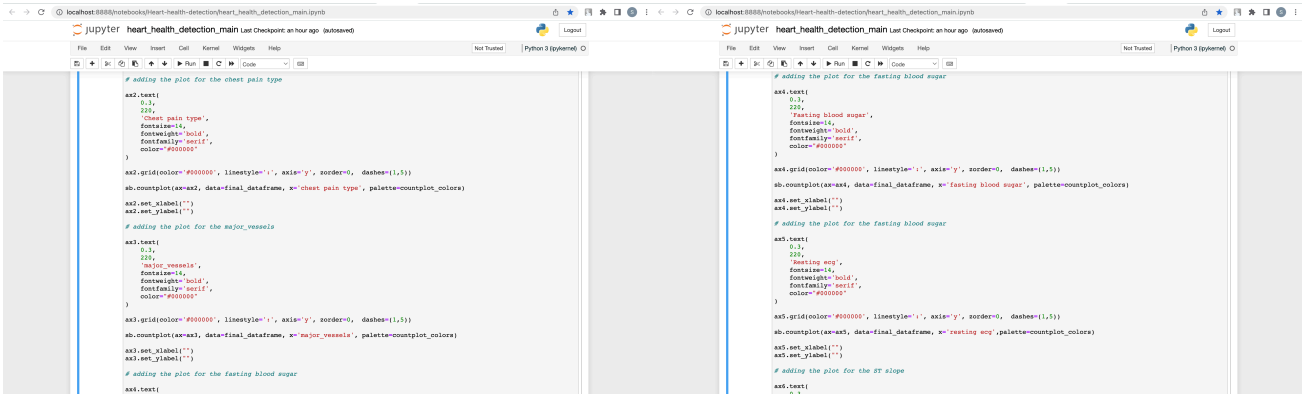
This section is responsible for the deep analysis of the dataset and is applied to investigate the dataset and summarise the key insights. This section provides the basic understanding of the data and its distribution. The exploration can be either done using graphs or python functions.

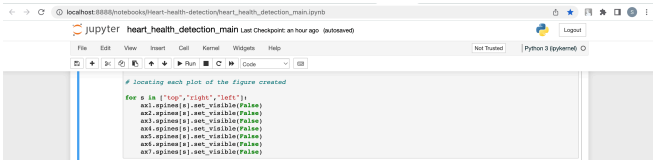
For this project I have used univariate analysis in which analysis on single attribute is done. This analysis is done for the continuous and categorical attributes of the dataset. For better and intuitive understanding of the data and its distribution I have utilised graphical feature provided by the ‘matplotlib’ and ‘seaborn’ libraries.

Count plotting on categorical data

countplot() method is used to Show the counts of observations in each categorical bin using graph bars. A count plot can be thought of as a histogram across a categorical, instead of quantitative, variable, so you can compare counts across nested variables.

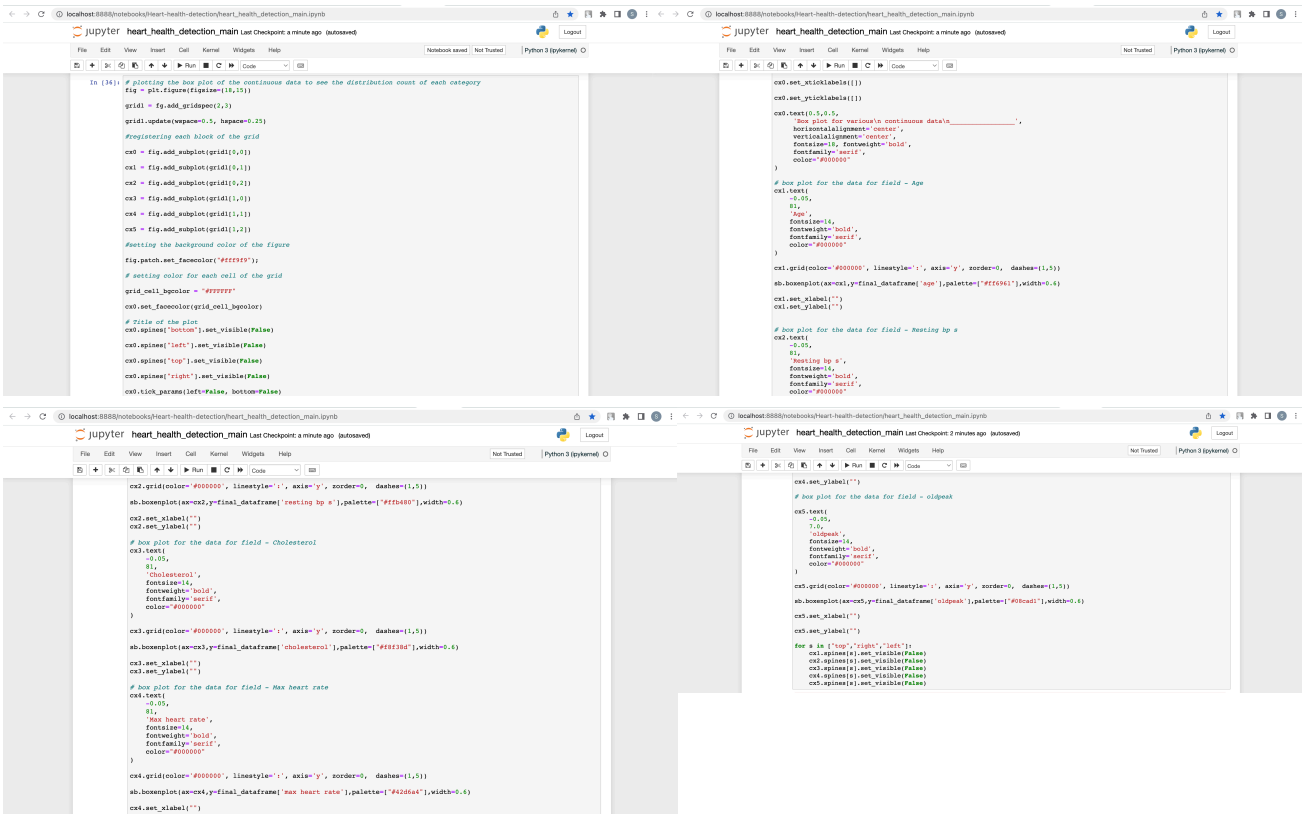






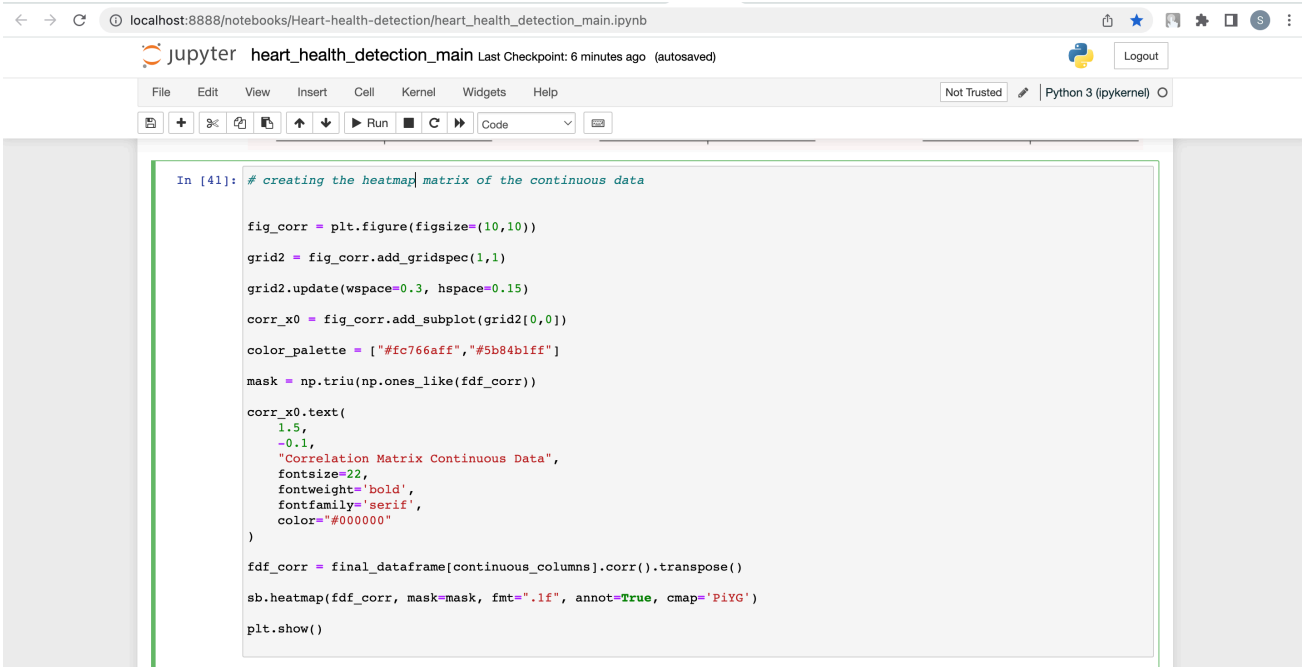
Box plotting on continuous data

A box plot, also known as a whisker plot, is used to display a summary of a set of data values with attributes such as minimum value, first quartile, median, third quartile, and maximum value. A box is created from the first quartile to the third quartile, where there is also a box with a vertical line through the median. Here the x-axis represents the data to be plotted and the y-axis represents the frequency distribution.



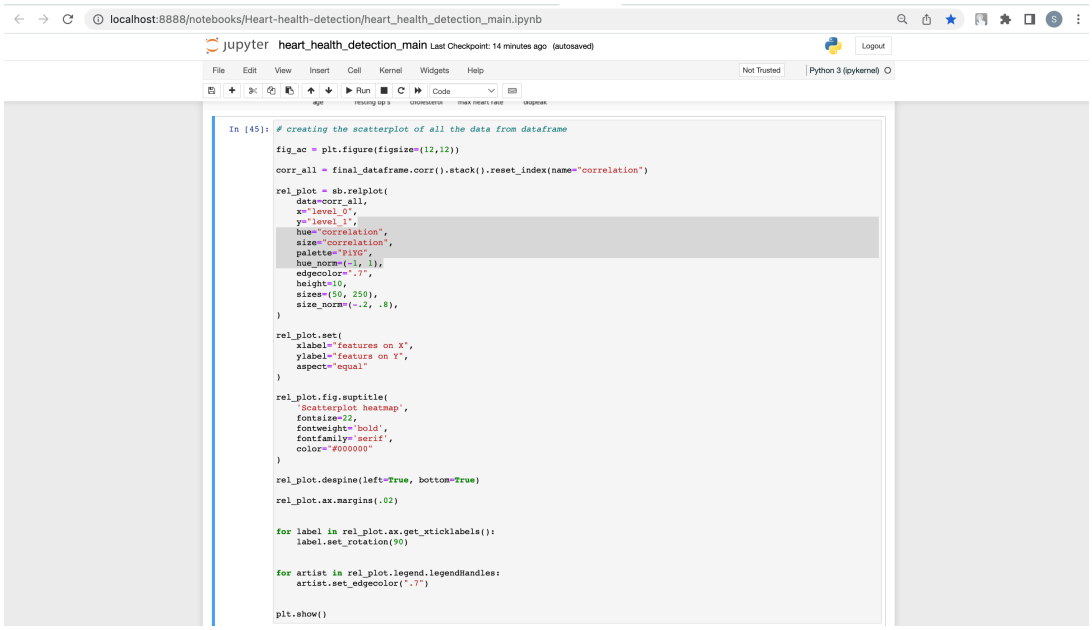
Heatmap matrix of continuous data

A heatmap is defined as a graphical representation of data using colours to visualise matrix values. In this case, to represent more common values or higher activity, a lighter colour is usually used, while to represent less common values or activity values, darker colours are preferred. Heatmaps are also defined by the name of the shading matrix. Heatmaps in Seaborn can be drawn using the seaborn.heatmap() function.



Scatterplot matrix of all data

Scatterplot is used with several semantic groupings which can help to understand well in a graph. They can plot two-dimensional graphics that can be



enhanced by mapping up to three additional variables while using the semantics of hue, size, and style parameters. All the parameter control visual semantic which are used to identify the different subsets. Using redundant semantics can be helpful for making graphics more accessible.

6. PACKAGES FOR AI MODEL

There are set of packages that we need in order to ready and train a predictive model. Here is the list of packages that will be needed-

RobustScaler- This technique of ML is used when there is a presence of outliers and the dataset.

train_test_split- This package is used to divide the dataset for training and testing suite of data.

Torch- this package provide high level features like - **Tensor computation (like NumPy) with strong GPU acceleration** & **Deep neural networks built on a tape-based autograd system**

Scikit Learn- It is a simple and efficient tools for predictive data analysis. It is built on NumPy, SciPy, and matplotlib. For regression related problems such as the one used in this project Scikit provides feature like Linear Regression learning models. Regression models a target prediction value based on independent variables. It is mostly used for finding out the relationship between variables and forecasting.

7. TRAINING MODEL WITH DATASET

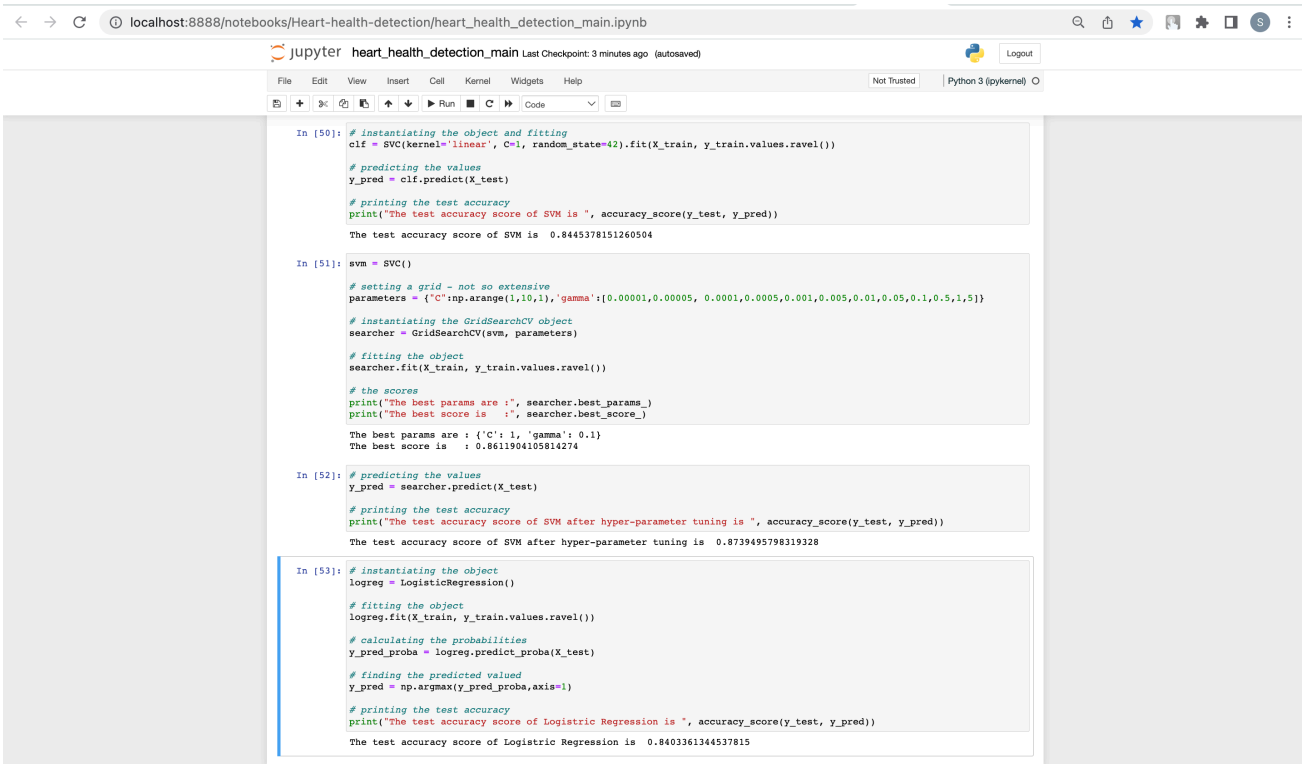
The process of training the model is quite simple. We first break the dataset into continuous and categorical sub-datasets. We then divide our whole dataset into two parts. One is used for training the model and other is used for testing the prediction accuracy. For splitting we use train_test_split() method and we keep 20% of our total data for testing purposes.

We then proceed to use method like fit_transform() for training the model. The method SVC() is responsible to creating instance of model that predicts the target variable for the testing suite. We can also check the efficiency of the model using accuracy_score method().

8. HYPER PARAMETER TUNING FOR IMPROVING PREDICTION.

By training a model with existing data, we are able to fit the model parameters. However, there is another kind of parameter, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.

For this project the methodology I have used is GridSearchCV(). In this methodology it searches for the best set of hyperparameters from a grid of hyperparameters values. GridSearchCV will go through all the intermediate combinations of hyperparameters which makes grid search computationally very expensive and will return the best possible combination.



We can also measure the accuracy of the model using method like accuracy_score() on the testing suite of the sub-dataset. We have package methods like logisticRegression() which builds a regression model to predict the probability that a given data entry belongs to the category numbered as “1”.

RESULTS

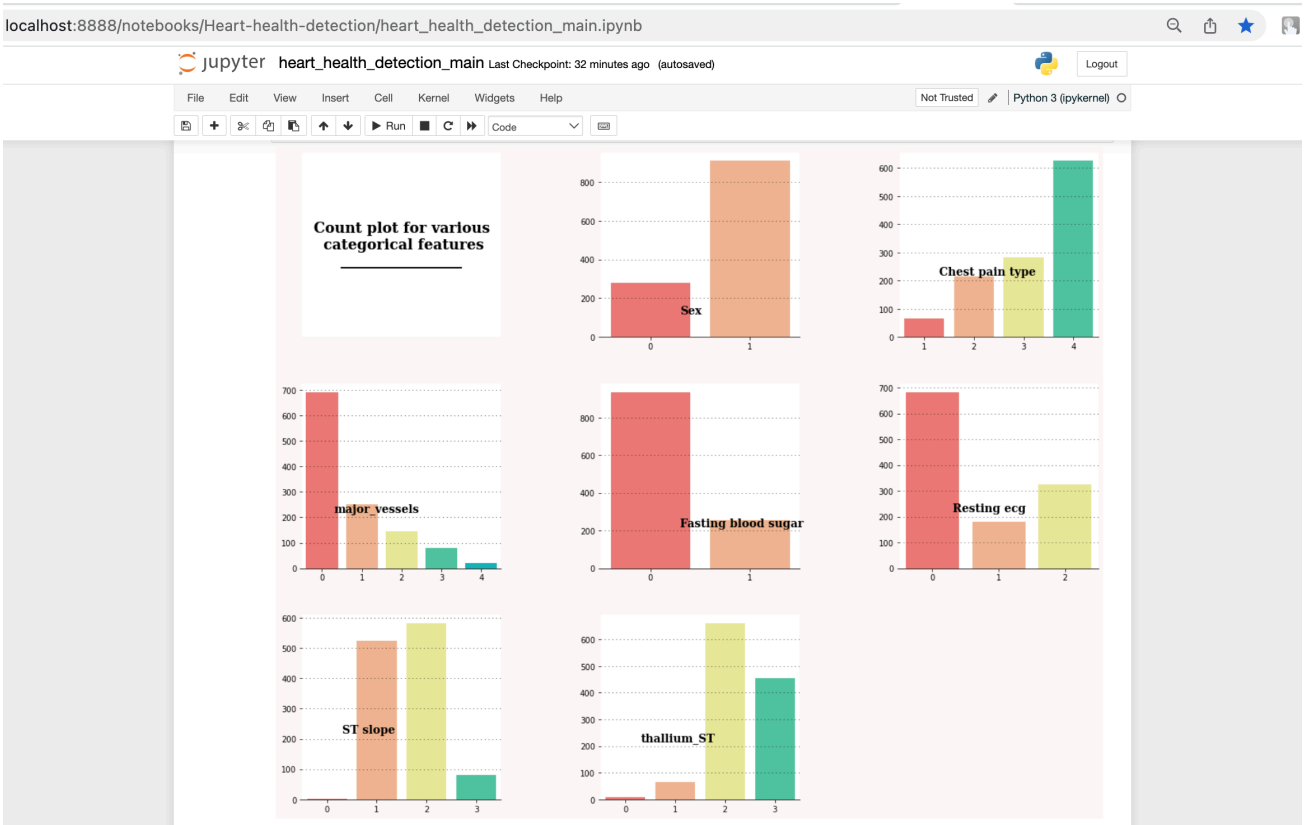
In this section of the project report we will discuss about the results of the operations that we performed on dataset. We used techniques that describes the data graphically and mathematically.

GRAPHICAL RESULTS ON DATASETS

COUNTPLOT OF THE CATEGORICAL COLUMNS

This graph represents the count plot of categorical columns from Dataset. It provides insight on the distribution of the dataset among the possible categories. e.g for Sex column there is only two possible categories, this graph informs us that around 300 subjects are of ‘Sex-0’ type and rest are of ‘Sex-1’ type.

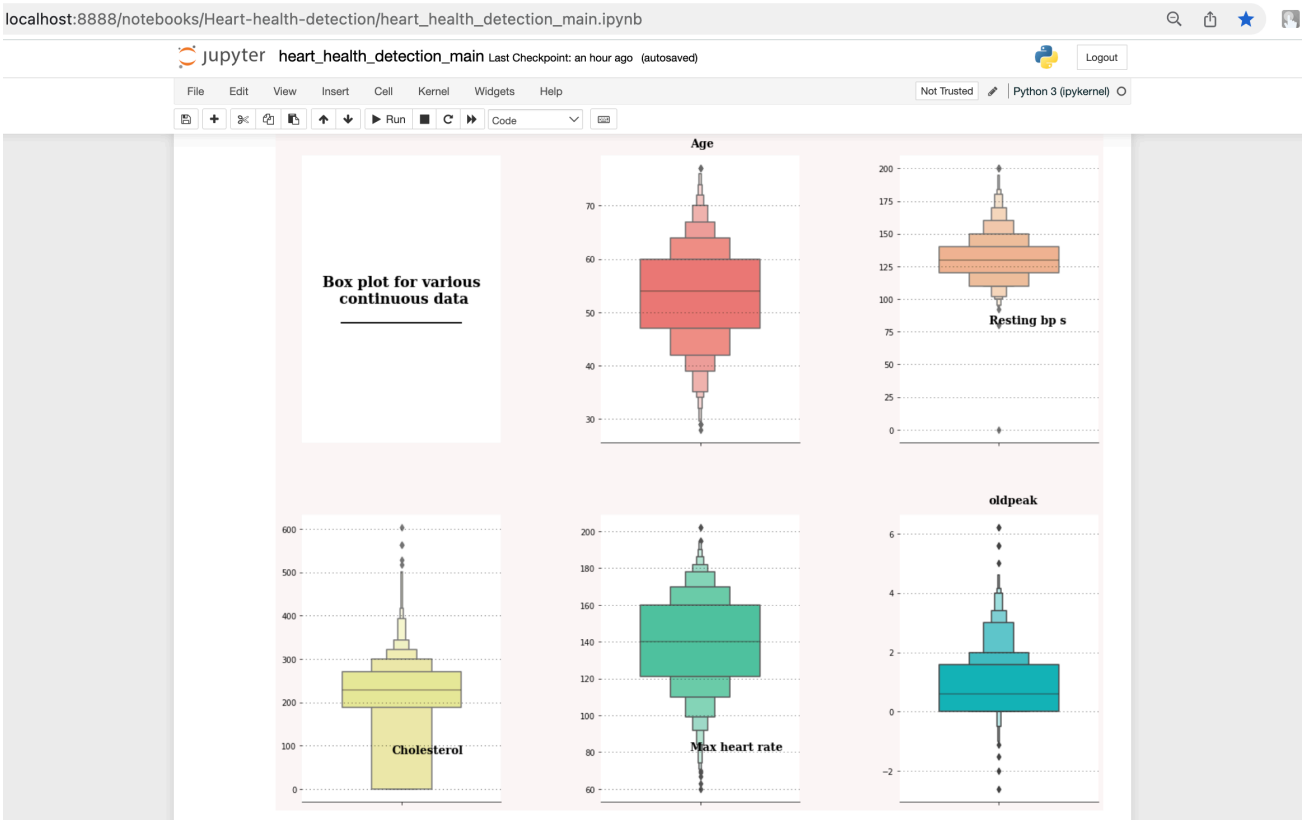
Similarly other columns categories can also be easily understood with the help of this graph.



BOXPLOT OF THE CONTINUOUS COLUMNS

This graph represents the boxplot of the continuous columns from Dataset. It provides an insight on minimum value, first quartile, median, third quartile, and maximum value of the column data. A box is created from first quartile to the third quartile and there is a median line from the box. Box plot also informs about the location of the outliers present in the column data.

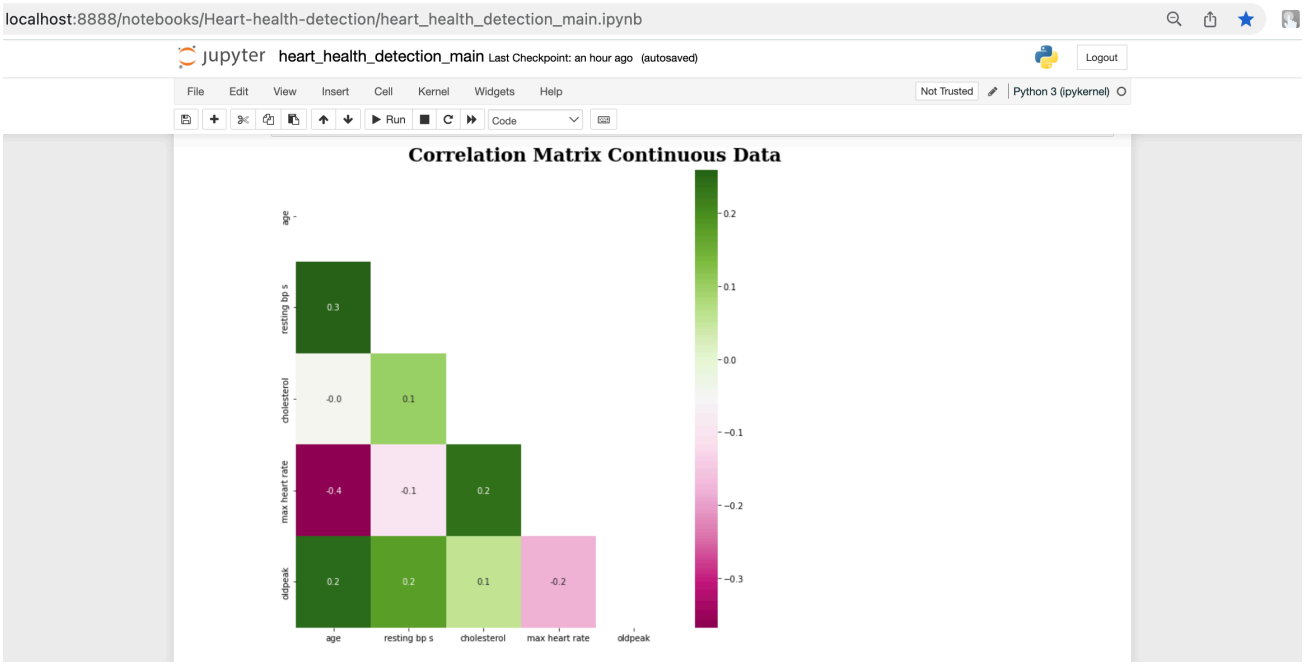
e.g. for the column ‘Resting Blood pressure’ there is one outlier which has the value 0. This value might not be useful for the prediction process and might lead to affect the prediction accuracy. Hence with help from this graphical data we can conclude that during Training of the AI model we should adopt a technique that removes the outliers from the columns data.



HEATMAP MATRIX OF THE CONTINUOUS COLUMNS

This graph represents the Heatmap of the continuous columns from Dataset. It provides an insight on the correlation between the two columns. The lighter colour represent the strong correlation between two columns and darker colour

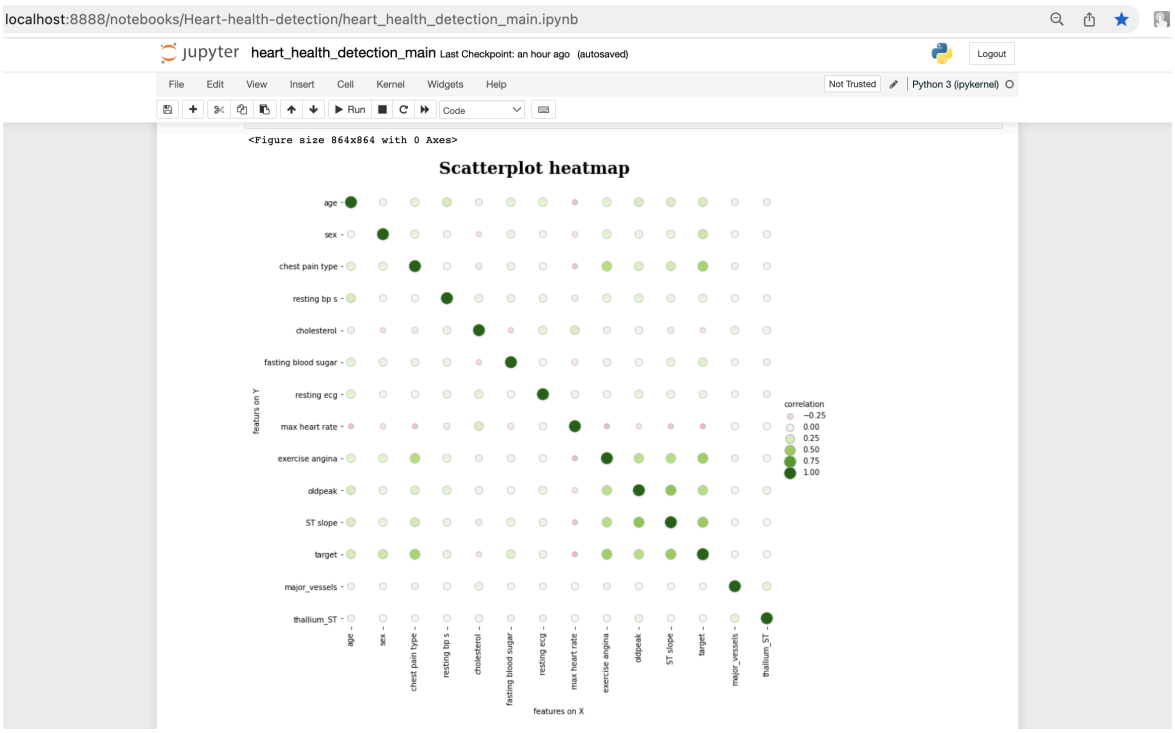
Represents weak correlation between two columns.



SCATTERPLOT MATRIX OF ALL COLUMNS

This graph represents the scatterplot of all the columns in dataset. It provides an insight on the relationship between the two columns. If there is a strong relation between two columns then they are presented with darker colour.

e.g It can be seen that there is a strong relationship between ‘chest type pain’ and ‘exercise angina’. Which can help conclude that a person who is prone to heart attack will experience pain in the chest during physical activities or workout sessions.


localhost:8888/notebooks/Heart-health-detection/heart_health_detection_main.ipynb
jupyter heart_health_detection_main Last Checkpoint: an hour ago (autosaved)
Logout
File Edit View Insert Cell Kernel Widgets Help
Not Trusted
Python 3 (ipykernel)
Run
Code
<Figure size 864x864 with 0 Axes>
Scatterplot heatmap
age
sex
chest pain type
resting bp s
cholesterol
fasting blood sugar
resting ecg
max heart rate
exercise angina
oldpeak
ST slope
target
major_vessels
thallium_ST
features on Y
features on X
correlation
−0.25
0.00
0.25
0.50
0.75
1.00

INSIGHTS

From the operation that I have performed and visual presentation of the dataset, here are the insights on the data

1. The dataset does not contain the null values and the data type of each column is of either INTEGER type of FLOAT type.

2. It can be seen that there are outliers present in continuous columns of Dataset.

3. There are less than twice subjects who have high Fasting blood sugar.

4. The scatterplot suggests that there is a strong relation between columns like ‘Chest Pain’, ‘Exercise Angina’ and ‘ST Slope’.

5. There is no strong relation between the two categories of columns evident from scatterplot of all the columns from dataset.

REFERENCES

1. https://news.abplive.com/science/world-heart-day-2022-70-of-heart-attack-deaths-last-year-occurred-in-30-60-age-group-1555818

2. https://www.indiatoday.in/health/story/why-there-is-a-sudden-surge-in-cardiac-arrest-cases-explain-experts-1998096-2022-09-08

3. https://www.datacamp.com/courses/introduction-to-data-visualization-with-matplotlib

4. https://www.youtube.com/watch?v=j9KpNo34Rp0

5. https://www.youtube.com/watch?v=CpWzLuQOg8g

6. www.youtube.com/watch?v=wetW1TFk1Sk

7. www.youtube.com/watch?v=ssNyZlc7_i0

8. https://practice.geeksforgeeks.org/courses/data-science-live?utm_source=GfG&utm_medium=gfg_submenu&utm_campaign=DS_Submenu/