Data Science Python Project

–

on Heart Decease

PROJECT BY-NANDINI DESAI

SUPERVISED BY-MR.REZA MOOSAVI

AGENDA

- Objective
- Data Exploring
- Data Preprocessing
- Data Modelling
- Model Comparison
- Model Evaluation and Deployment
- Conclusion

Mission Statement and Goals

- The primary objective of this machine learning project would be to design and implement a predictive model that can assess the likelihood of an individual having heart disease or not

DATA INTRODUCTION-922 rows and 12 columns

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
152	40	M	ATA	130	275.00	0	Normal	150	N	0.00	Up	0
454	58	M	ASY	136	203.00	1	Normal	123	Y	1.20	Flat	1
315	74	M	ATA	145	0.00	1	ST	123	N	1.30	Up	1
338	63	M	ASY	140	0.00	1	LVH	149	N	2.00	Up	1
723	59	M	ASY	140	177.00	0	Normal	162	Y	0.00	Up	1
337	63	M	ASY	150	0.00	1	ST	154	N	3.70	Up	1
245	54	M	TA	120	171.00	0	Normal	137	N	2.00	Up	0
828	43	F	NAP	122	213.00	0	Normal	165	N	0.20	Flat	0
96	43	M	ATA	142	207.00	0	Normal	138	N	0.00	Up	0
509	58	M	ASY	110	198.00	0	Normal	110	N	0.00	Flat	1
751	67	F	NAP	152	277.00	0	Normal	172	N	0.00	Up	0
541	76	M	NAP	104	113.00	0	LVH	120	N	3.50	Down	1
171	40	M	NAP	140	235.00	0	Normal	188	N	0.00	Up	0
865	60	M	ASY	145	282.00	0	LVH	142	Y	2.80	Flat	1
447	77	M	ASY	124	171.00	0	ST	110	Y	2.00	Up	1

DATA INSIGHTS

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 922 entries, 0 to 921
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   Age             922 non-null    int64  
 1   Sex             922 non-null    object 
 2   ChestPainType   922 non-null    object 
 3   RestingBP       922 non-null    int64  
 4   Cholesterol     920 non-null    float64
 5   FastingBS       922 non-null    int64  
 6   RestingECG      922 non-null    object 
 7   MaxHR           922 non-null    int64  
 8   ExerciseAngina  920 non-null    object 
 9   Oldpeak         922 non-null    float64
 10  ST_Slope        920 non-null    object 
 11  HeartDisease    922 non-null    int64  
dtypes: float64(2), int64(5), object(5)
memory usage: 86.6+ KB
```

```
heart.duplicated().sum()
```

```
4
```

```
heart.isna().sum()
```

```
Age               0
Sex               0
ChestPainType     0
RestingBP         0
Cholesterol       2
FastingBS         0
RestingECG        0
MaxHR             0
ExerciseAngina    2
Oldpeak           0
ST_Slope          2
HeartDisease      0
dtype: int64
```

UNIVARIATE ANALYSIS

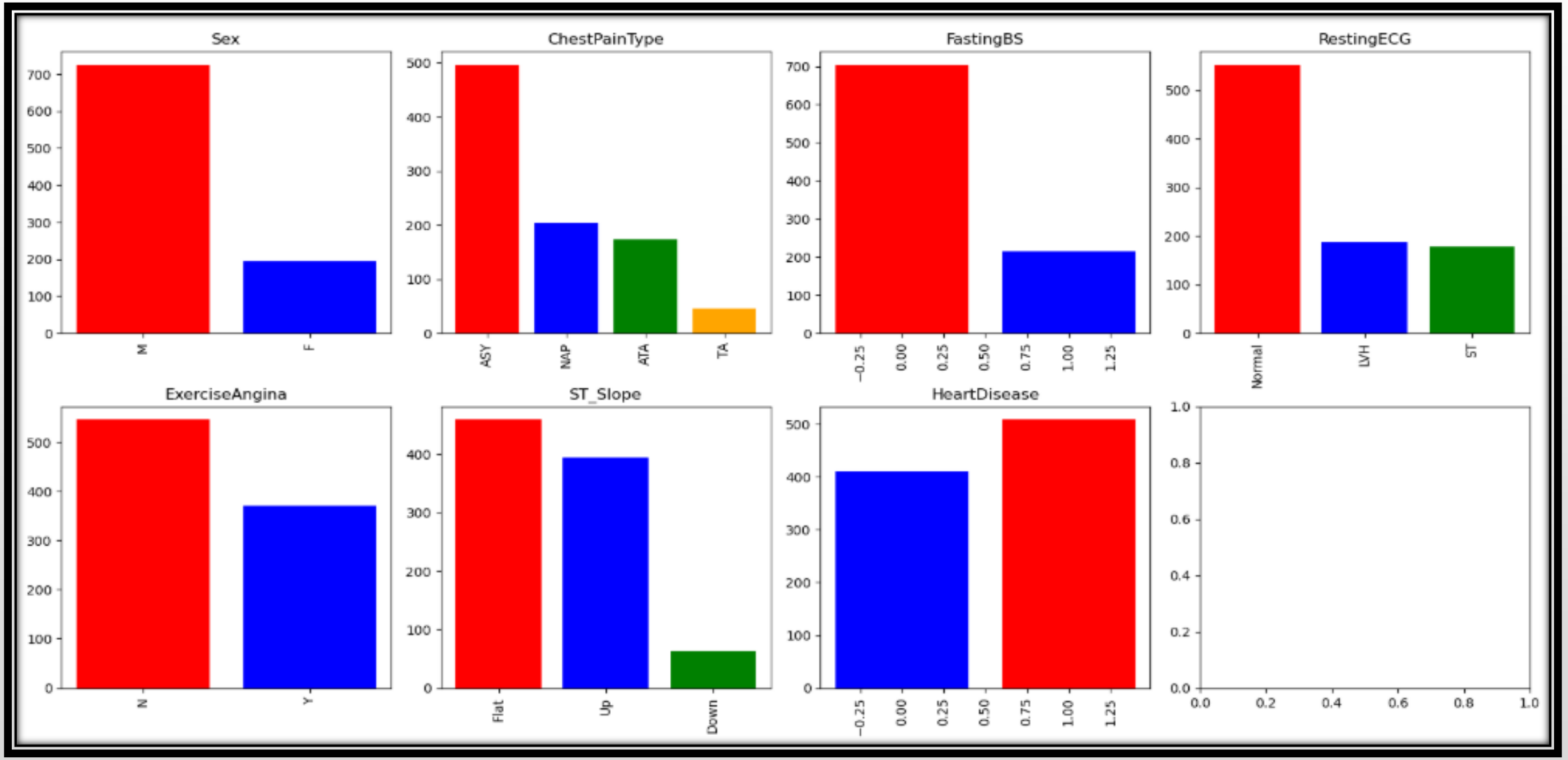
Numerical Variables

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.00	918.00	916.00	918.00	918.00	918.00	918.00
mean	53.51	132.40	200.73	0.23	138.43	0.89	0.55
std	9.43	18.51	124.00	0.42	49.70	1.07	0.50
min	28.00	0.00	0.00	0.00	-175.00	-2.60	0.00
25%	47.00	120.00	173.00	0.00	120.00	0.00	0.00
50%	54.00	130.00	223.00	0.00	138.00	0.60	1.00
75%	60.00	140.00	267.00	0.00	156.00	1.50	1.00
max	77.00	200.00	1960.00	1.00	1120.00	6.20	1.00

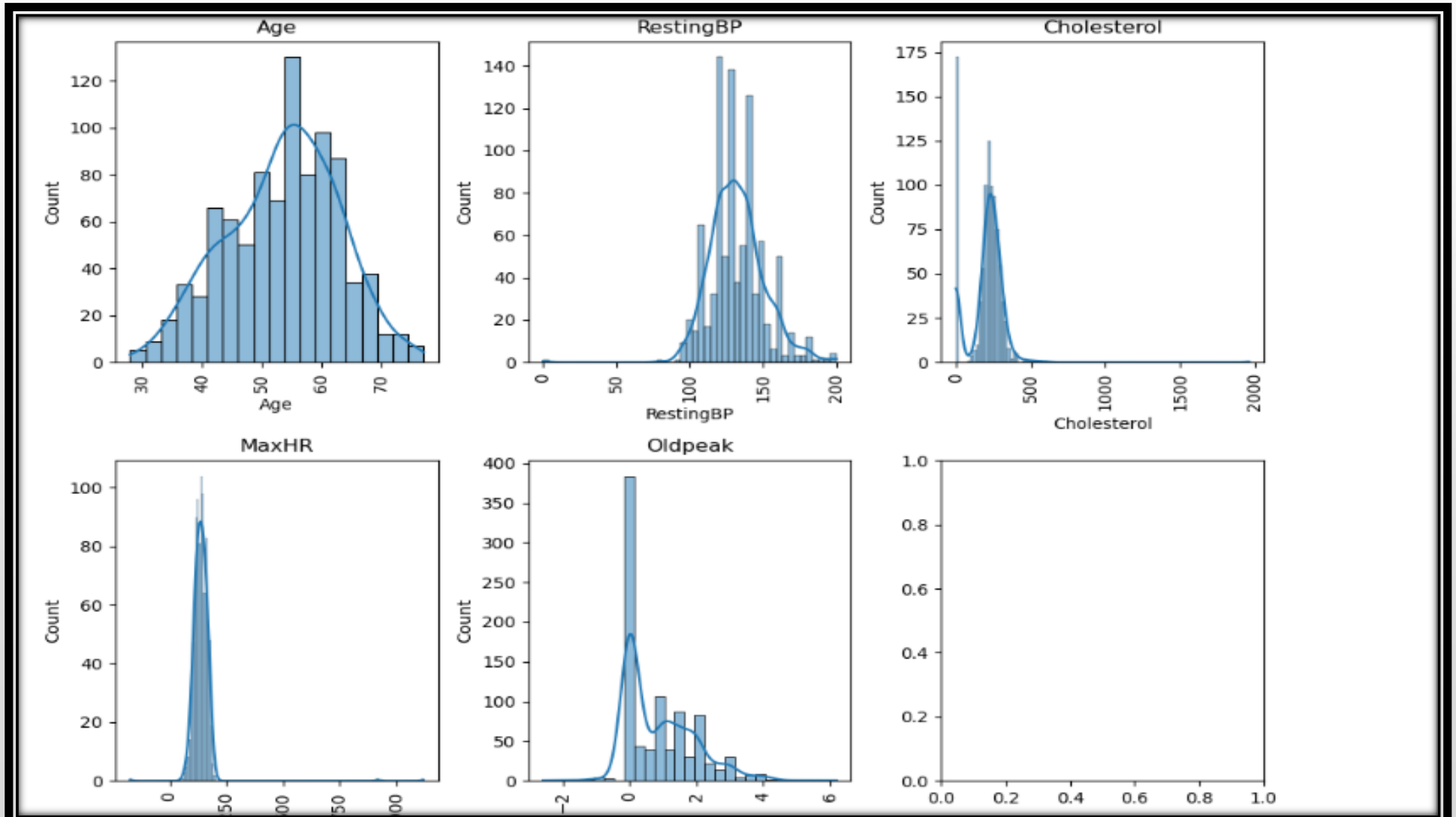
Categorical Variables

	Sex	ChestPainType	RestingECG	ExerciseAngina	ST_Slope
count	918	918	918	916	916
unique	2	4	3	2	3
top	M	ASY	Normal	N	Flat
freq	725	496	552	545	459

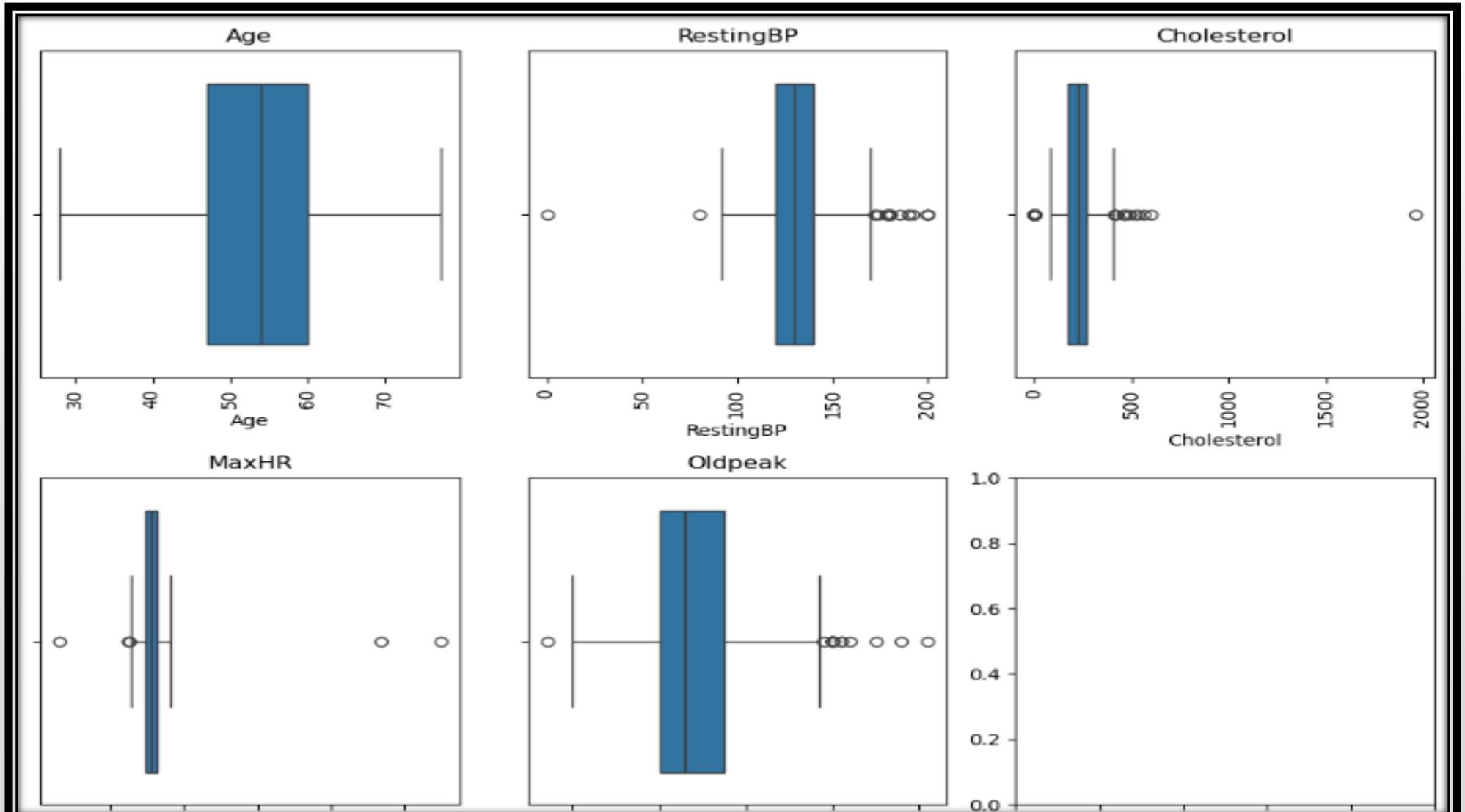
UNIVARIATE ANALYSIS-CATEGORICAL VARIABLES



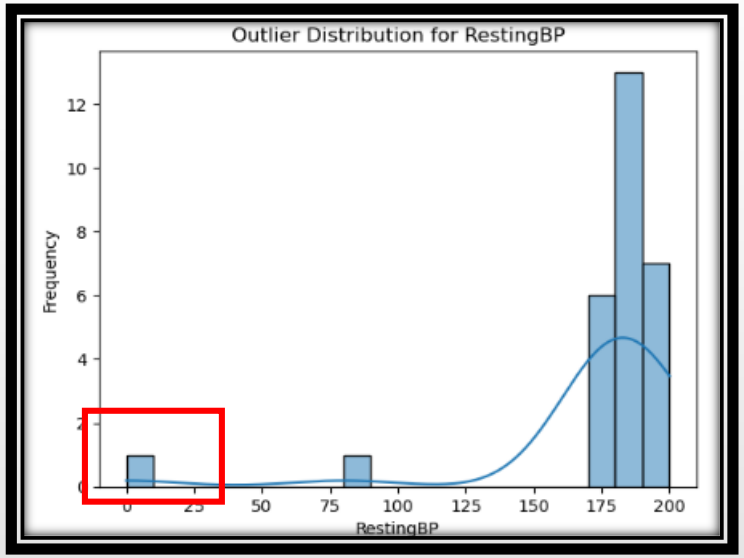
UNIVARIATE ANALYSIS-NUMERICAL VARIABLES

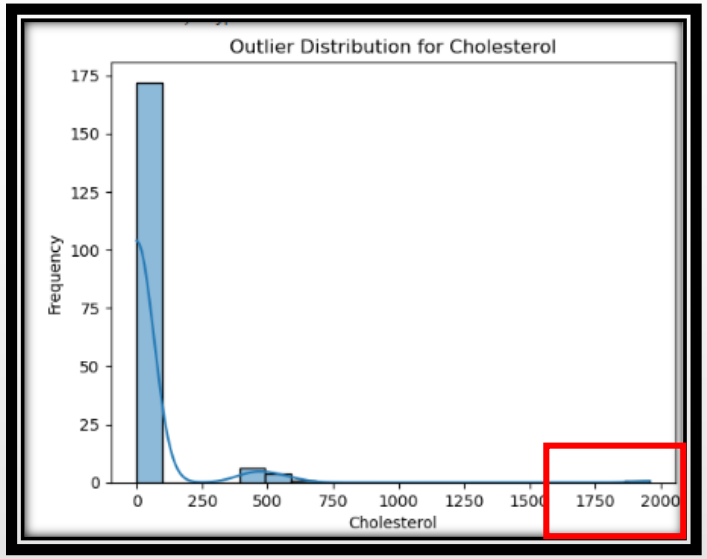


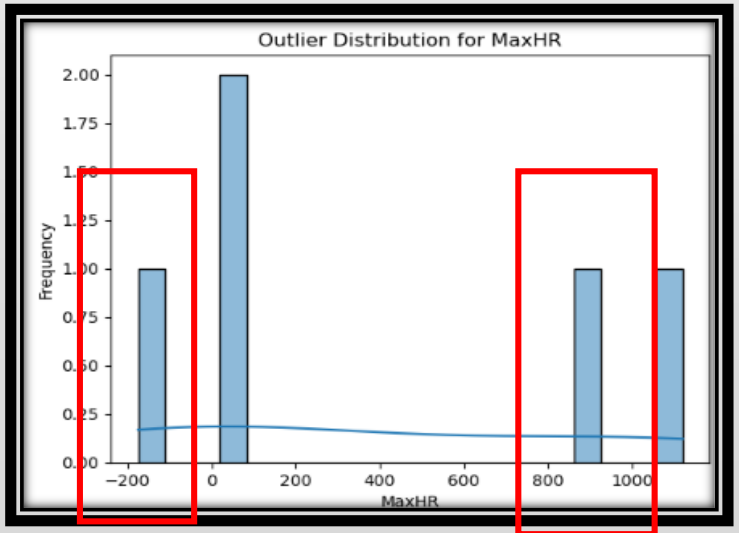
OUTLIER DETECTION

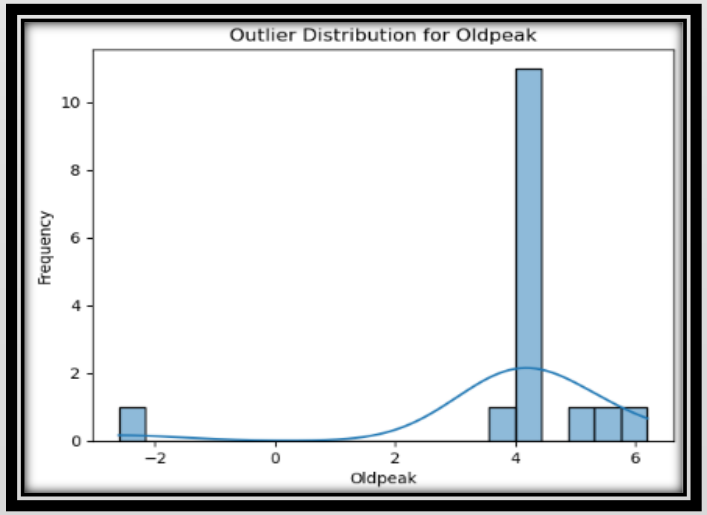


OUTLIER DETECTION



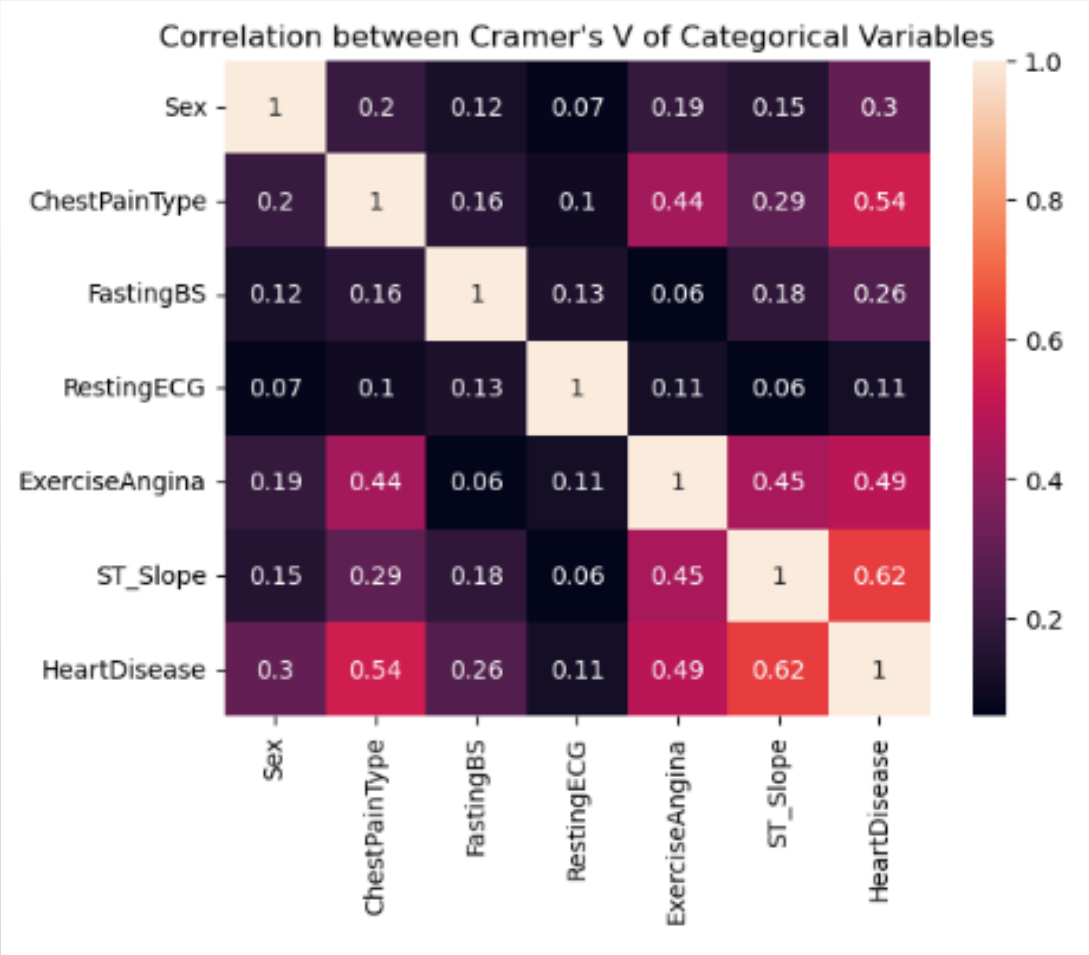




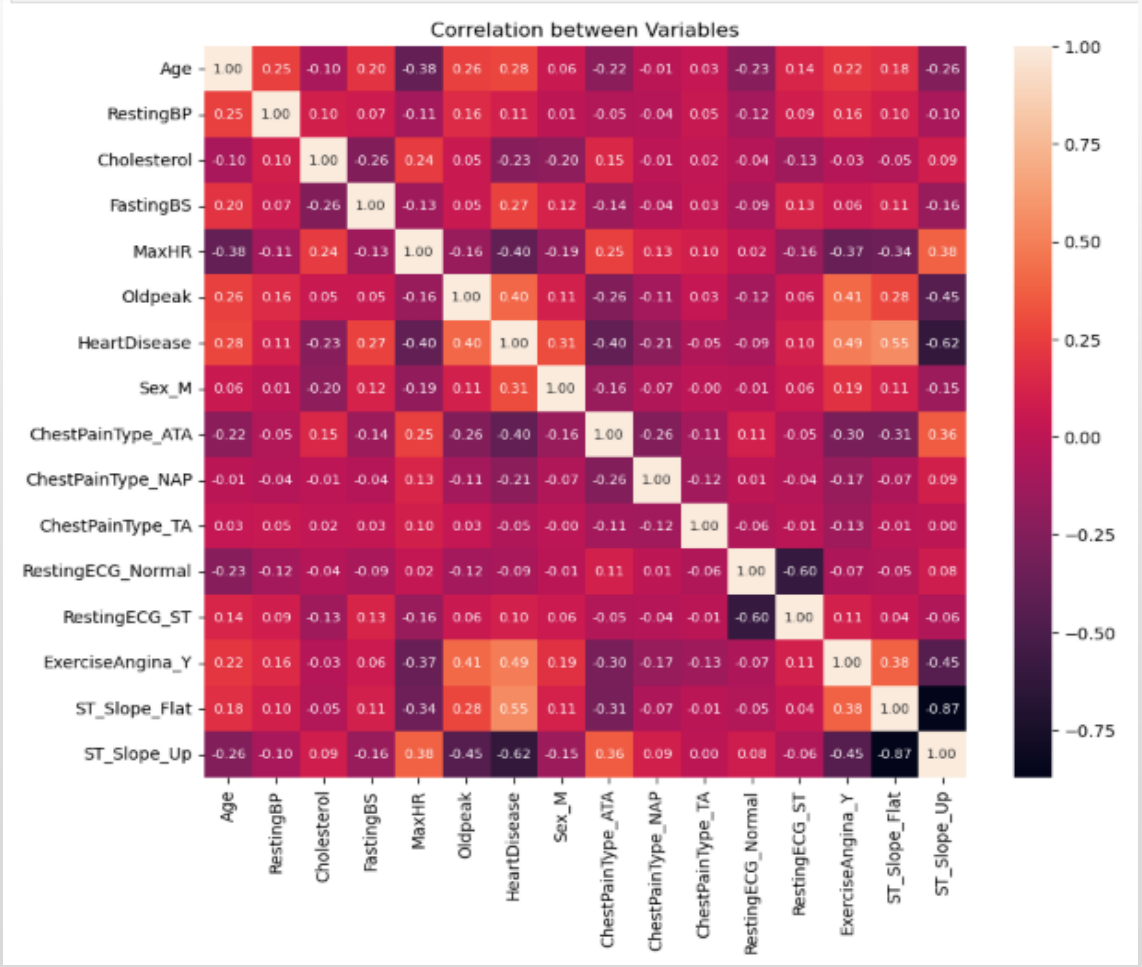


BIVARIATE ANALYSIS

Categorical

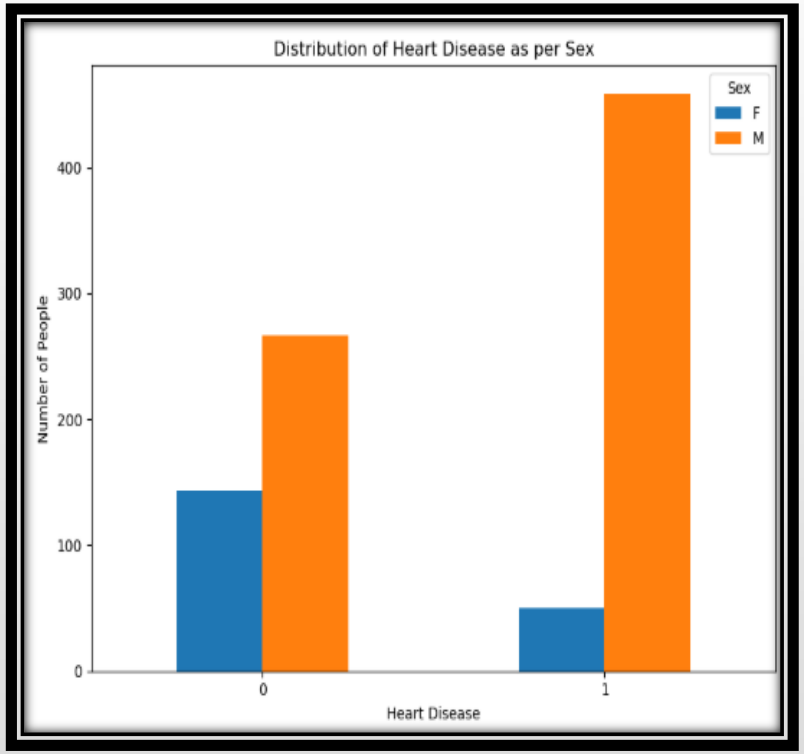


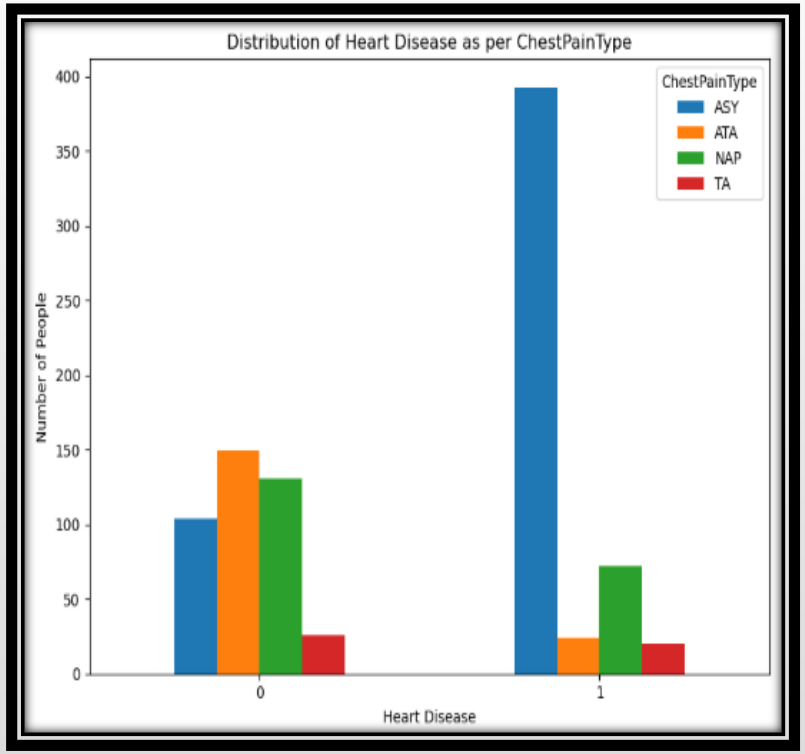
Numerical

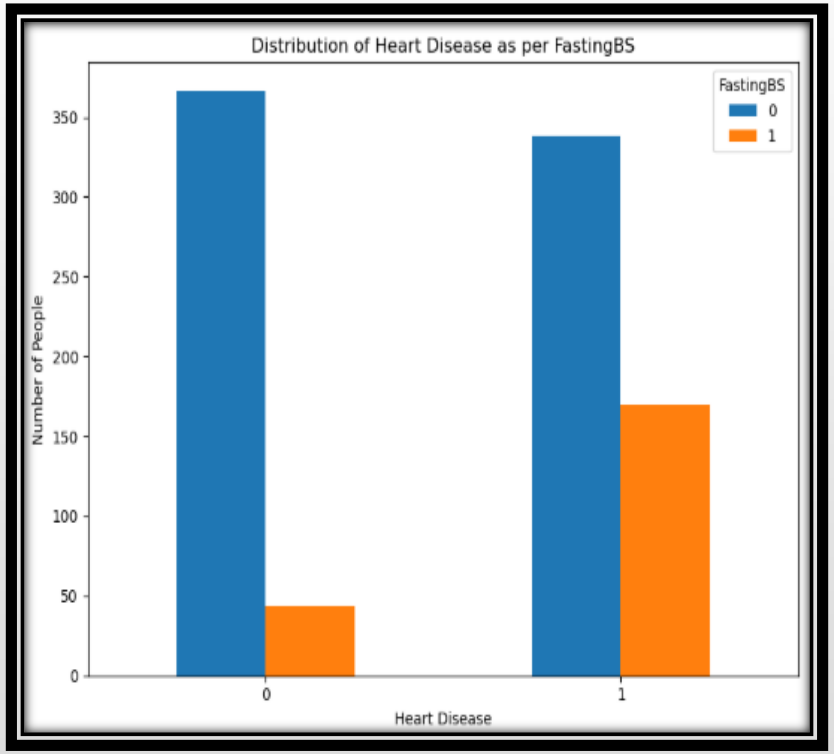


BIVARIATE ANALYSIS

Statistically Significant association

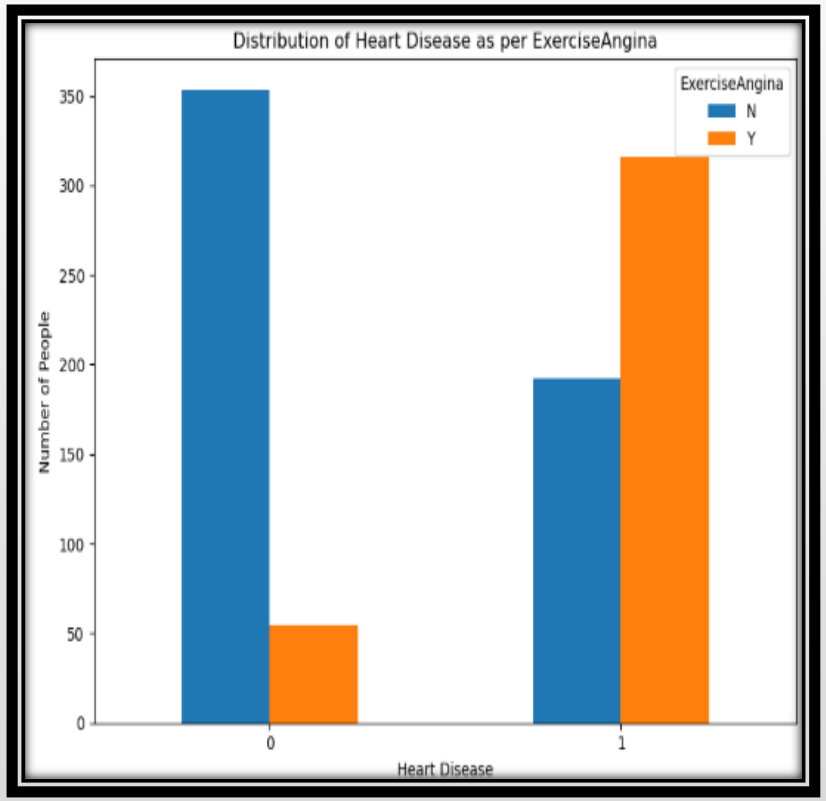


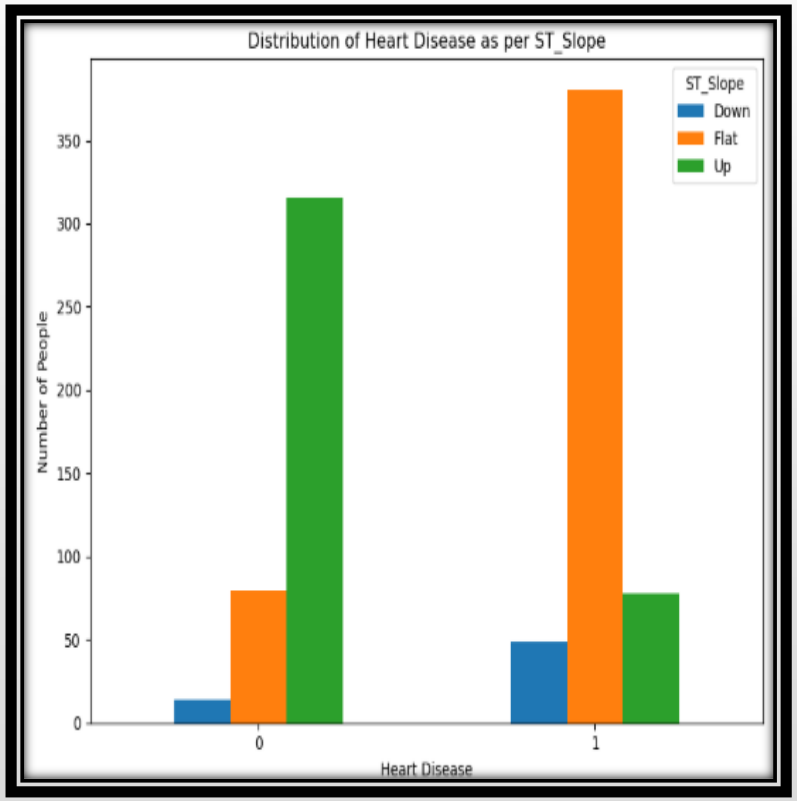


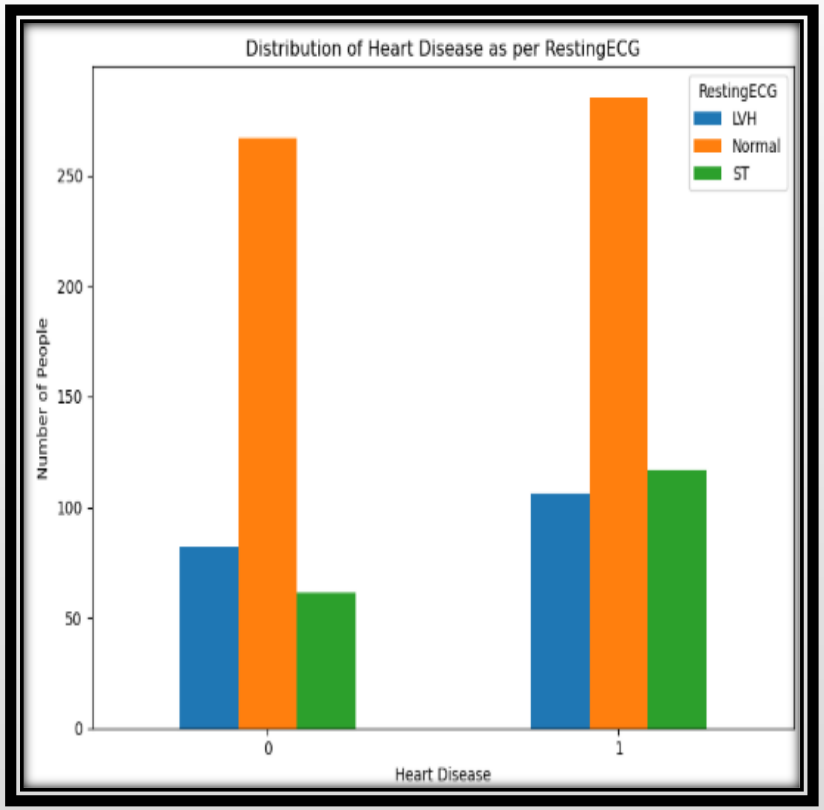


BIVARIATE ANALYSIS

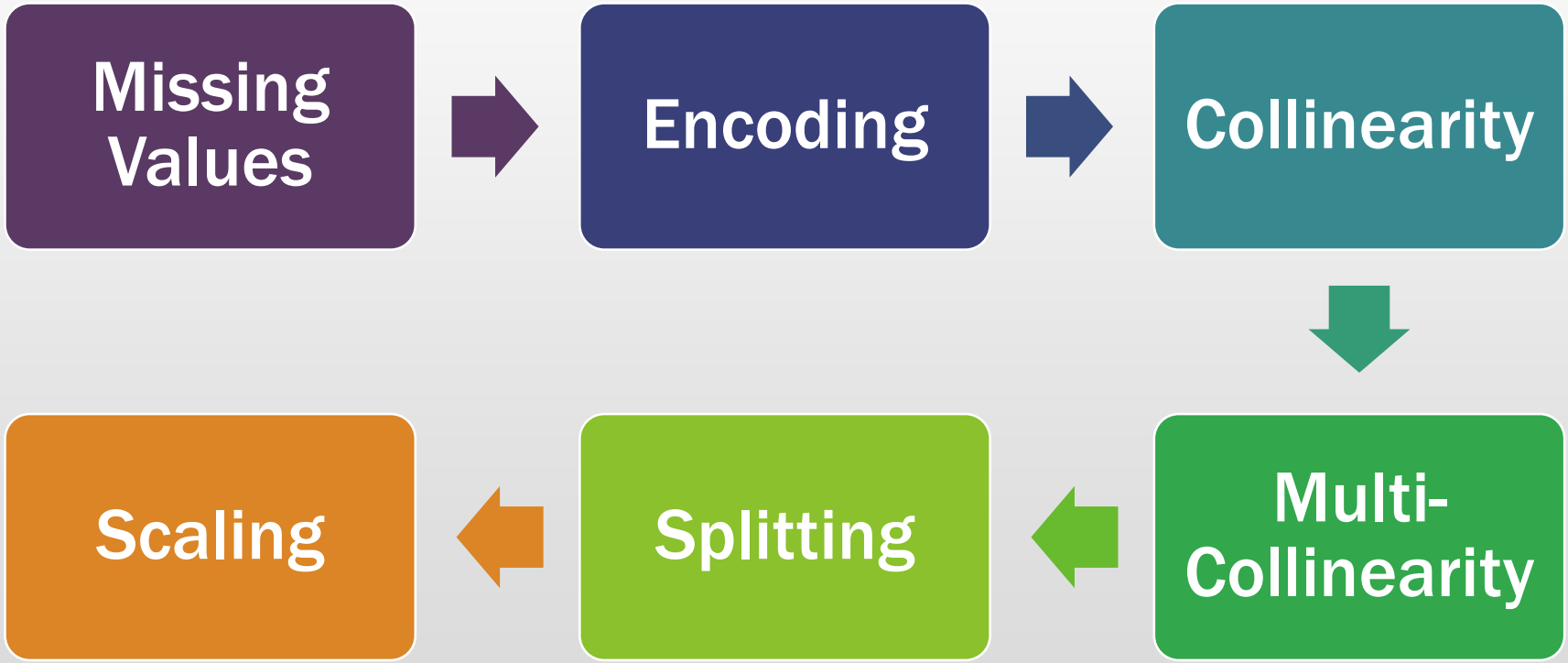
Statistically Significant association







DATA PREPROCESSING



DATA MODELLING CLASSIFICATION

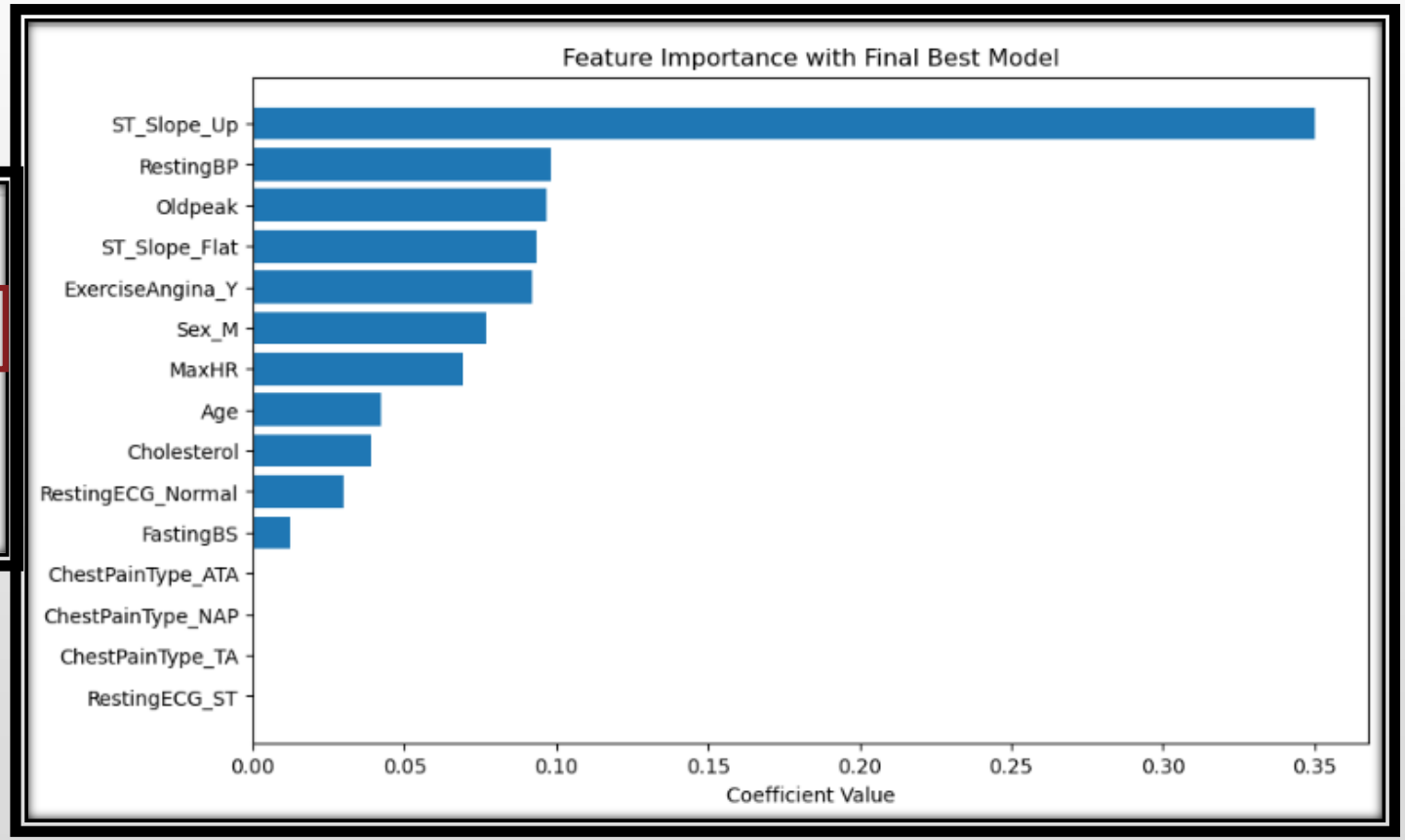
- MODEL1- Random Forest with Uncleaned Data and Default Parameters
- MODEL2- Random Forest with Cleaned Data and Default Parameters
- MODEL3- Random Forest with Uncleaned Data and Hyper parameter Tuning
- MODEL4- Random Forest with Cleaned Data and Hyper parameter Tuning
- MODEL5- XGBoost with Uncleaned Data and Default Parameters
- MODEL6- XGBoost with Cleaned Data and Default Parameters
- MODEL7- XGBoost with Uncleaned Data and Hyper parameter Tuning
- MODEL8- XGBoost with Cleaned Data and Hyper parameter Tuning
- MODEL9- Logistic Regression with Default Parameters
- MODEL10- Logistic Regression with Hyper parameter Tuning
- MODEL11- Decision Tree Classifier with Default Parameters
- MODEL12- Decision Tree Classifier with Hyper parameter Tuning

MODEL COMPARISON

	Train_CV	Train_Accuracy	Test_Accuracy	Train f1_score	Test f1_score	Train Recall_score	Test Recall_score	Train roc_auc_score	Test roc_auc_score	Type2_error
RF_Unclean_Default	0.86	1.00	0.87	1.00	0.88	1.00	0.91	1.00	0.96	0.09
RF_clean_Default	0.87	1.00	0.88	1.00	0.89	1.00	0.91	1.00	0.95	0.09
RF_Unclean_Best	0.88	1.00	0.87	1.00	0.89	1.00	0.93	1.00	0.96	0.07
RF_clean_Best	0.88	1.00	0.90	1.00	0.92	1.00	0.96	1.00	0.95	0.04
XGB_Unclean_Default	0.87	1.00	0.88	1.00	0.89	1.00	0.89	1.00	0.93	0.11
XGB_clean_Default	0.86	1.00	0.84	1.00	0.86	1.00	0.87	1.00	0.92	0.13
XGB_Unclean_Best	0.87	0.92	0.92	0.93	0.93	0.94	0.98	0.97	0.97	0.02
XGB_clean_Best	0.88	0.92	0.89	0.93	0.91	0.95	0.96	0.98	0.96	0.04
LogReg_Default	0.87	0.88	0.87	0.89	0.88	0.91	0.91	0.93	0.96	0.09
LogReg_Best	0.87	0.87	0.88	0.89	0.90	0.90	0.93	0.93	0.96	0.07
DecisonTree_Default	0.79	0.99	0.83	1.00	0.84	0.99	0.83	1.00	0.83	0.17
DecisionTree_Best	0.83	0.87	0.90	0.89	0.92	0.92	0.98	0.93	0.93	0.02

MODEL EVALUATION ON UNSEEN DATA (Validation Set)

	Accuracy	f1_score	Test Recall_score	Test roc_auc_score
XGBoost	0.87	0.88	0.90	1.00
Decision Tree	0.87	0.88	0.90	0.92
Logistic Regression	0.87	0.88	0.90	0.91



SUMMARY

- XGBoost with Hyper Parameter Tuning is overall best model for given Heart Dataset
- Decision Tree Classifier and Logistic Regression with Hyper Parameter Tuning take 2nd and 3rd position with overall scores
- ST_Slope_Up (Slope of the peak excecise ST Segment)- The ST segment rises, which could indicate some mild ischemia or stress response.
- Other Important Features-
  - RestingBP- Resting Blood Pressure- Normal Range= 120/80 mmhg
  - OldPeak- Depression of the ST segment in an ECG. The amount of ST segment depression measured during a stress test.
  - ST_Slope_Flat- ST segment remains horizontal or flat, indicating that there is no significant upward or downward deviation in the ST segment during the stress test.
  - Exercise_Angina- A type of chest pain or discomfort that occurs during physical exertion or exercise.

THANK YOU

STAY HEALTHY STAY FIT