PROJECT PRESENTATION DS102/DS104

Evan Sim Yih Shan

INTRODUCTION OF DATASET

According to WHO fact sheets, the world's biggest killer is ischemic heart disease that responsible for 16% of the world's total deaths. Since 2000, the largest increase in deaths has been rising by more than 2 million to 8.9 million deaths in 2019. This dataset chosen is to predict possible heart disease for early detection.

This dataset was created by combining different datasets. In this dataset, 5 heart datasets are combined over 11 common features. The five datasets used for its curation are:

Cleveland: 303 observations

Hungarian: 294 observations

Switzerland: 123 observations

Long Beach VA: 200 observations

Stalog (Heart) Data Set: 270 observations

Total: 1190 observations

Duplicated: 272 observations

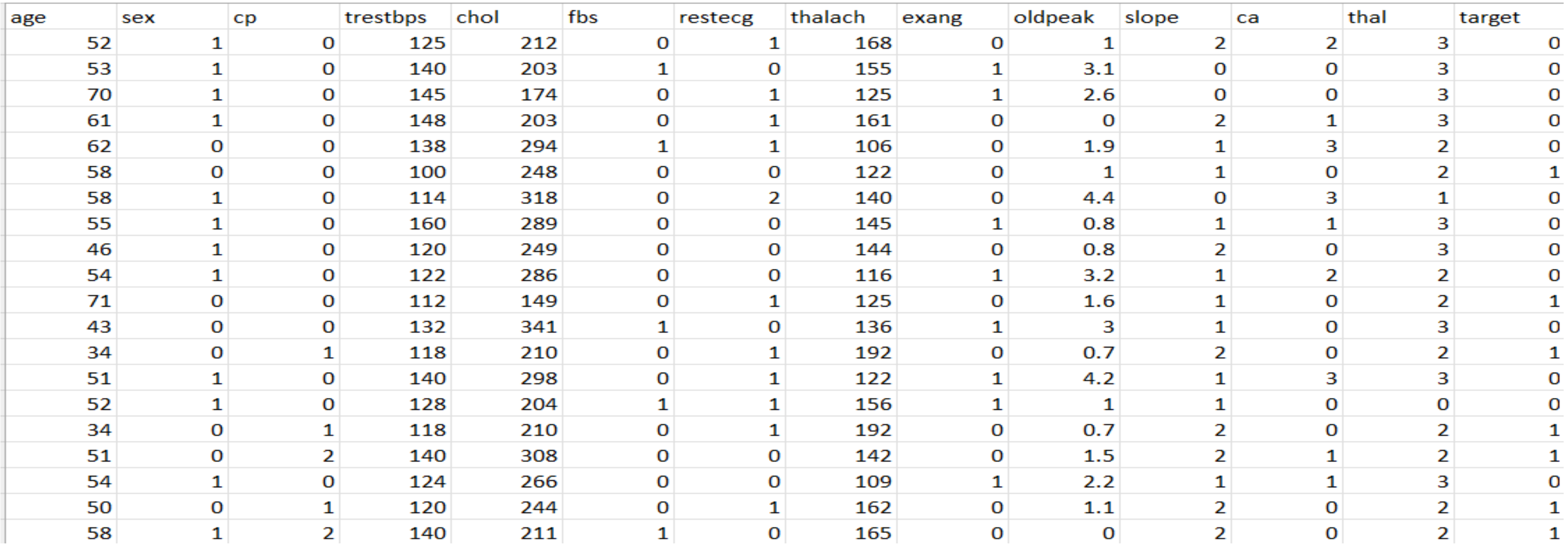
Final dataset: 918 observations

DETAILS OF DATASET

Size of the data: 918 row

Origin of Dataset: https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction

Attributes: Age, Sex, Chest Pain type, Resting BP, Cholesterol, Fasting Blood Sugar, Resting ECG, Max Heart Rate, Exercise Angina, Old Peak, ST segment slope and Output of heart disease.



age	sex	cp	trestbps	chol	fbs	restecg	thalach	exang	oldpeak	slope	ca	thal	target
52	1	0	125	212	0	1	168	0	1	2	2	3	0
53	1	0	140	203	1	0	155	1	3.1	0	0	3	0
70	1	0	145	174	0	1	125	1	2.6	0	0	3	0
61	1	0	148	203	0	1	161	0	0	2	1	3	0
62	0	0	138	294	1	1	106	0	1.9	1	3	2	0
58	0	0	100	248	0	0	122	0	1	1	0	2	1
58	1	0	114	318	0	2	140	0	4.4	0	3	1	0
55	1	0	160	289	0	0	145	1	0.8	1	1	3	0
46	1	0	120	249	0	0	144	0	0.8	2	0	3	0
54	1	0	122	286	0	0	116	1	3.2	1	2	2	0
71	0	0	112	149	0	1	125	0	1.6	1	0	2	1
43	0	0	132	341	1	0	136	1	3	1	0	3	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	1	0	140	298	0	1	122	1	4.2	1	3	3	0
52	1	0	128	204	1	1	156	1	1	1	0	0	0
34	0	1	118	210	0	1	192	0	0.7	2	0	2	1
51	0	2	140	308	0	0	142	0	1.5	2	1	2	1
54	1	0	124	266	0	0	109	1	2.2	1	1	3	0
50	0	1	120	244	0	1	162	0	1.1	2	0	2	1
58	1	2	140	211	1	0	165	0	0	2	0	2	1

Microsoft Excel
ma Separated Valı

CHALLENGERS

1. *Data Readability – Certain attributes need to figure it out what is the relationship to the heart disease*

2. Data Accuracy – as this database is combined by 5 different institutions (Cleveland, Hungary, Switzerland, Long Beach VA & Stalog (Heart) Data set which is sample size average 238 may not get more accurate conclusion.

3. Data Types- Certain attributes is object datatype might need to convert certain categories during analysis. Some data might need to do a benchmark in blood pressure, cholesterol , maximum heart rate as different practice in other countries different benchmark.

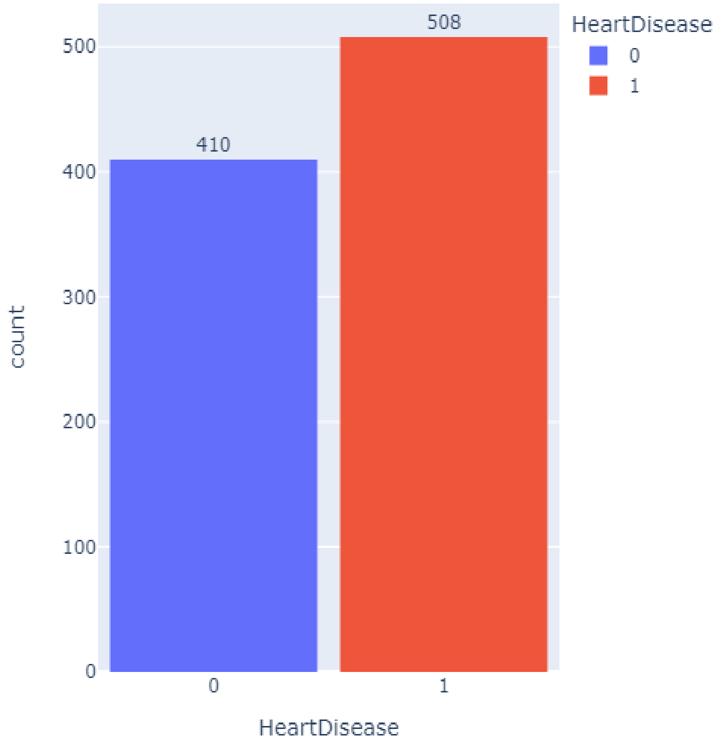
QUESTIONS FOR ANALYSIS

1. What’s the population for this data sets? How many person who had heart disease? What’s the percentage in this population in these 5 combined data sets?
2. Do Male gender tend to easily to get heart disease compared to female?
3. There are claims that high cholesterol, hypertension and diabetes are the leading cause of getting heart disease. Is that higher in each 3 group can contribute to get heart disease?
4. Is higher Max HR can lead to heart disease?
5. Does Age go higher will lead to get heart disease? What is the odds of getting it when one year older?
6. Can type of chest pain, type of resting ECG, presence of angina induced during exercise, type of ST slope after exercise detect person to have heart disease?

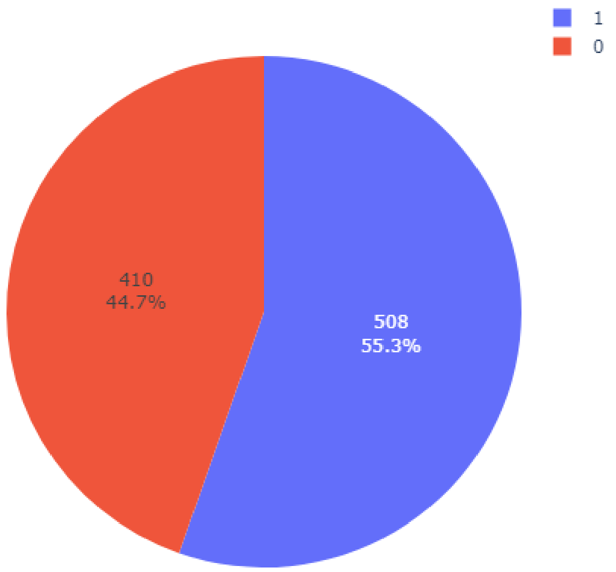
QUESTION NO1

What’s the population for this data sets? How many person who had heart disease? What’s the percentage in this population in these 5 combined data sets?

Total Number of Heart Disease



Total Number of Heart Disease in Percentage

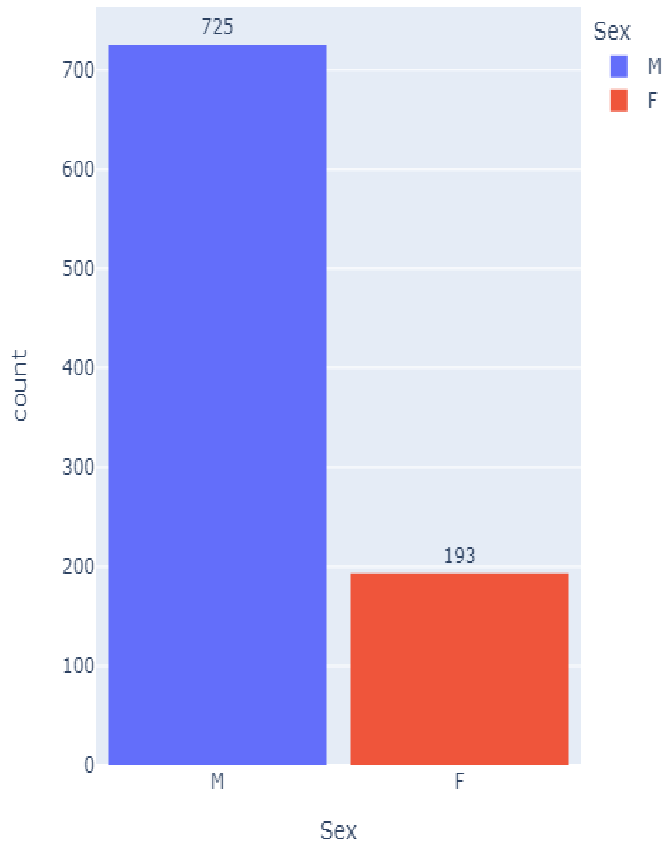


Total Population: 918

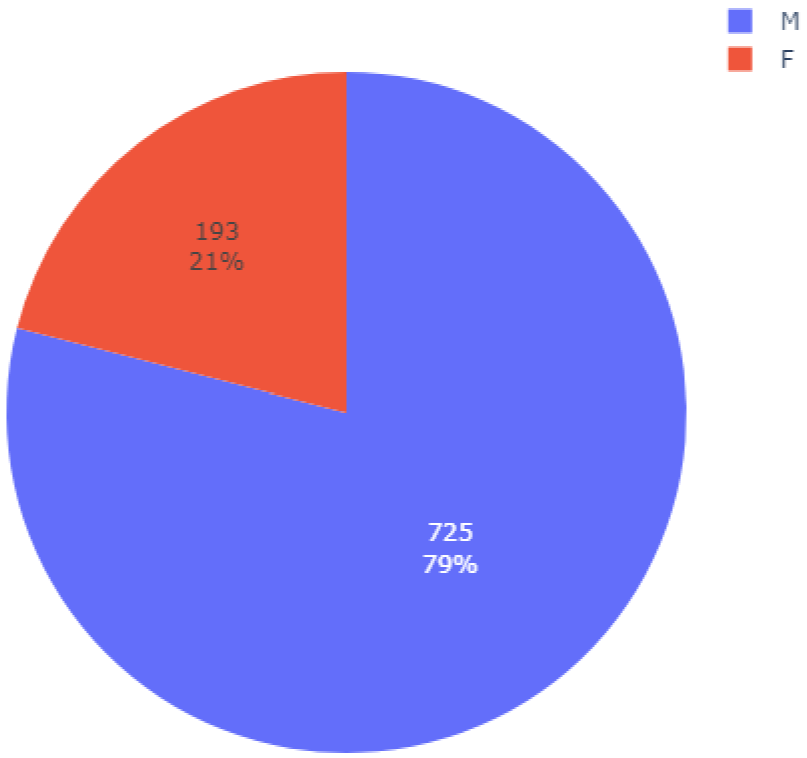
QUESTION NO2

Do Male gender tend to easily to get heart disease compared to female?

Total Male & Female Population

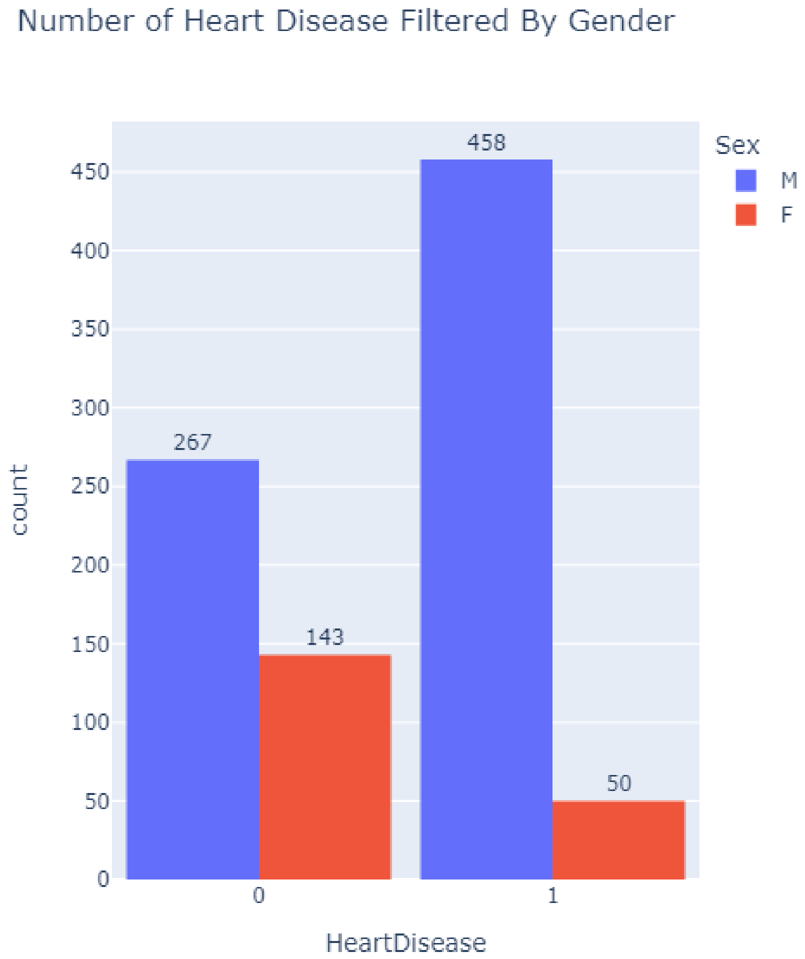


Total Male & Femele in population in %



QUESTION NO2

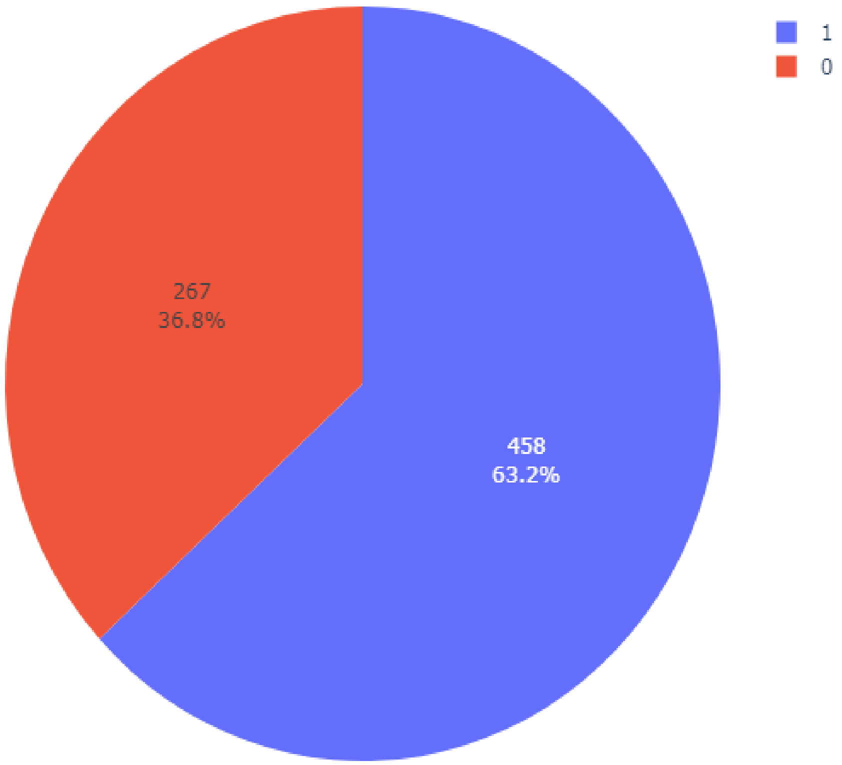
Do Male gender tend to easily to get heart disease compared to female?



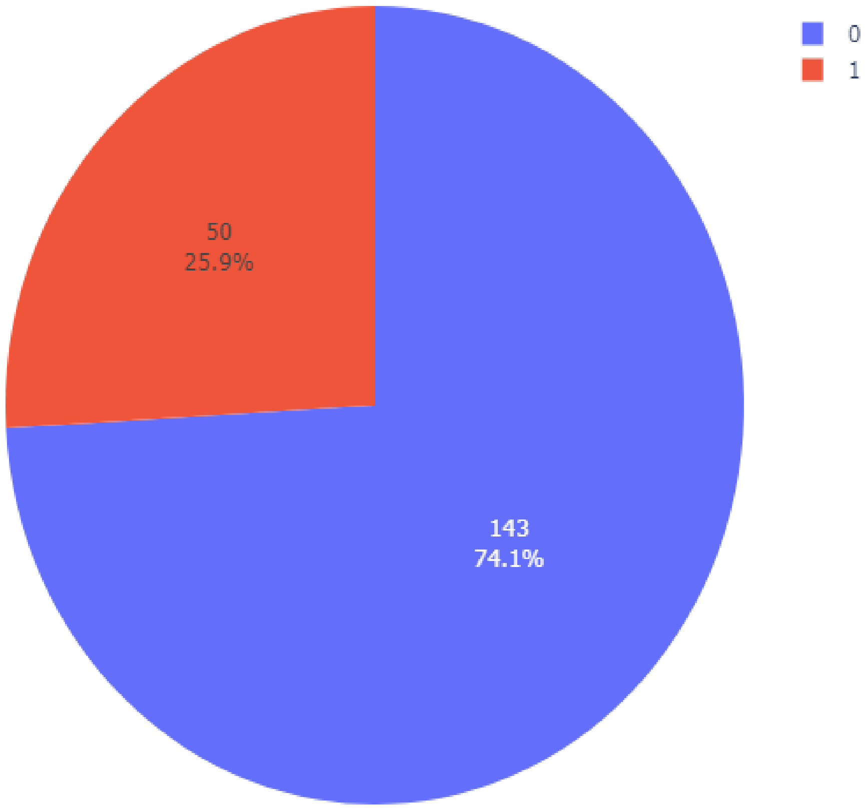
QUESTION NO2

Do Male gender tend to easily to get heart disease compared to female?

Number of Heart Disease in Male Category in Percentage



Number of Heart Disease in Female Category in Percentage

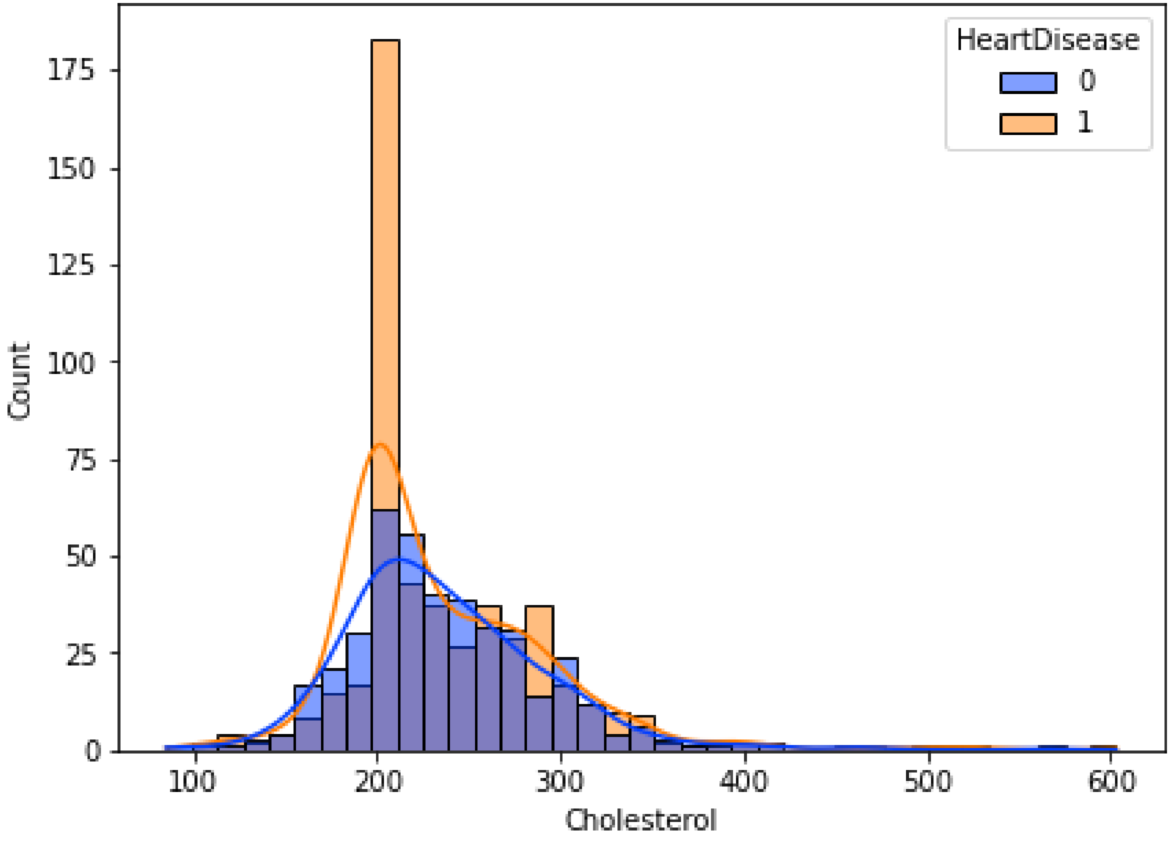


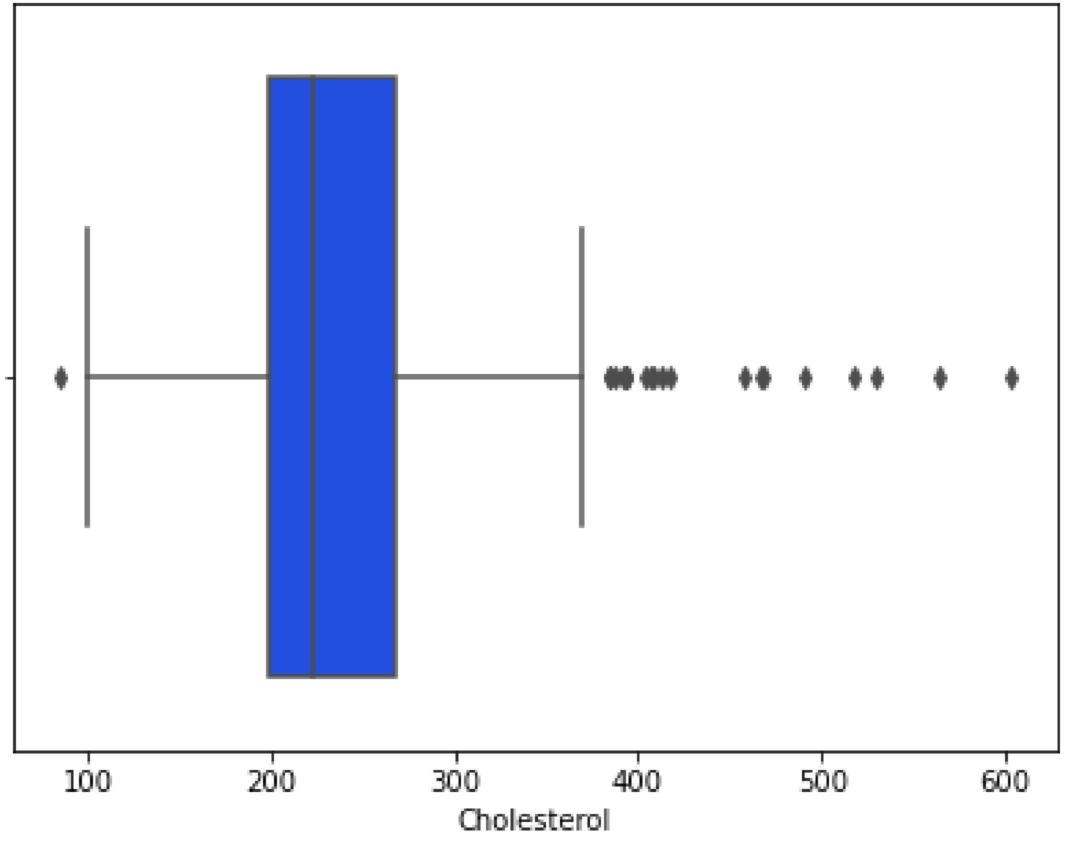
QUESTION NO3

There are claims that high cholesterol, hypertension and diabetes are the leading cause of getting heart disease. Is that higher in each 3 group can contribute to get heart disease.

QUESTION NO3

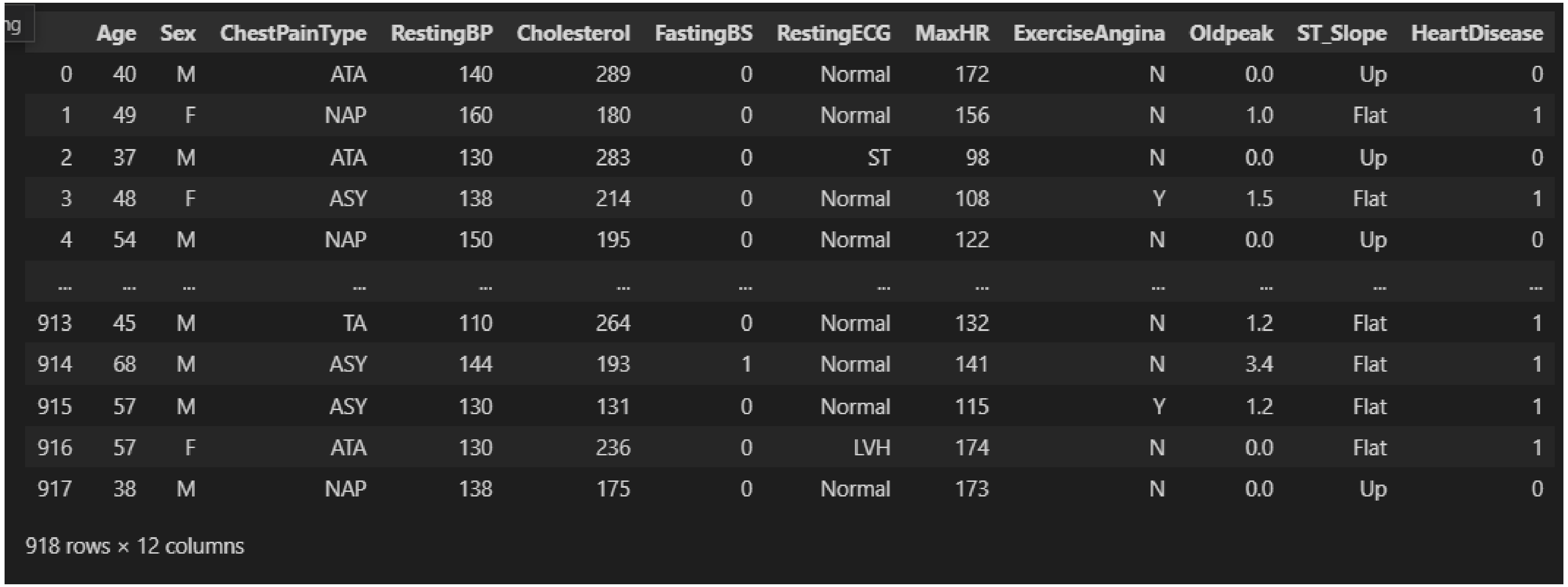
Cholesterol – Comparison between cholesterol and heart disease





QUESTION NO3

Cholesterol Data Frame



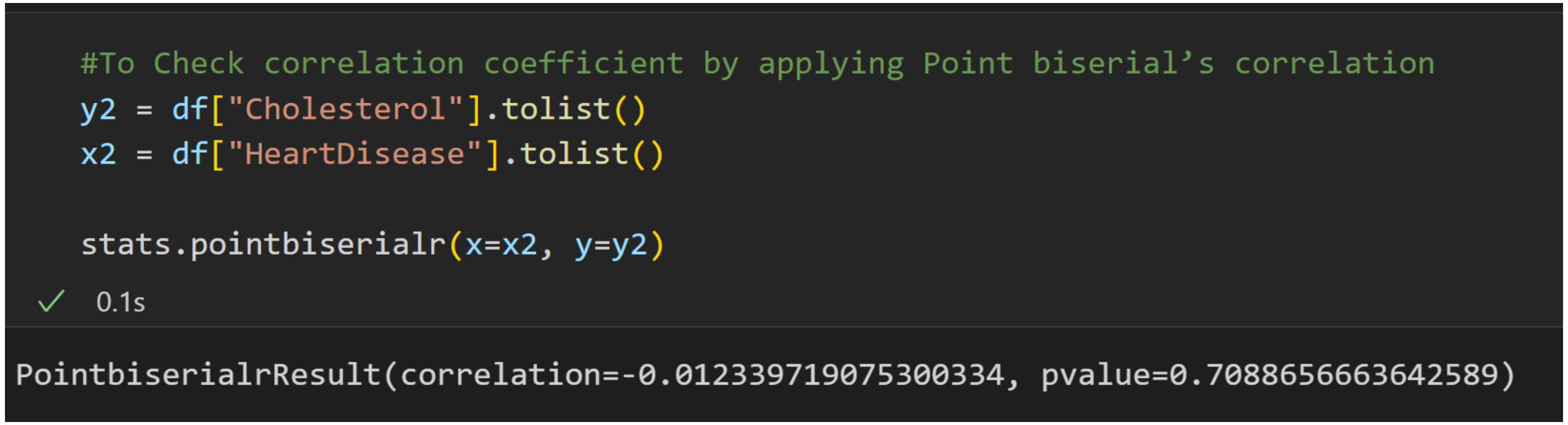
	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

918 rows × 12 columns

Cholesterol Data Type: Continuous Data
Heart Disease: Binary Data
Check Correlation Coefficient – apply point biserial

QUESTION NO3

Cholesterol Data Frame



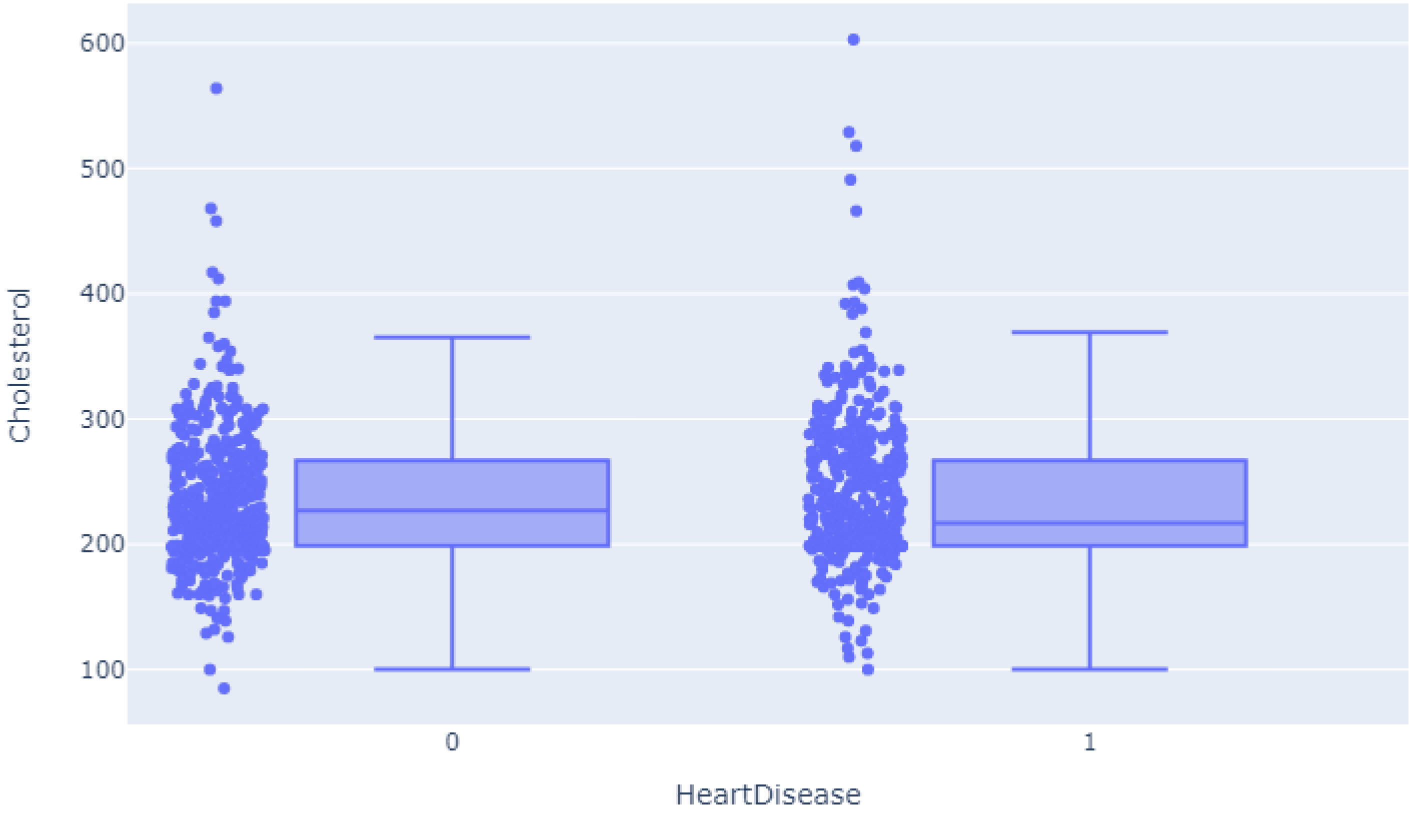
```
#To Check correlation coefficient by applying Point biserial’s correlation
y2 = df["Cholesterol"].tolist()
x2 = df["HeartDisease"].tolist()

stats.pointbiserialr(x=x2, y=y2)
```

✓ 0.1s

```
PointbiserialrResult(correlation=-0.012339719075300334, pvalue=0.7088656663642589)
```

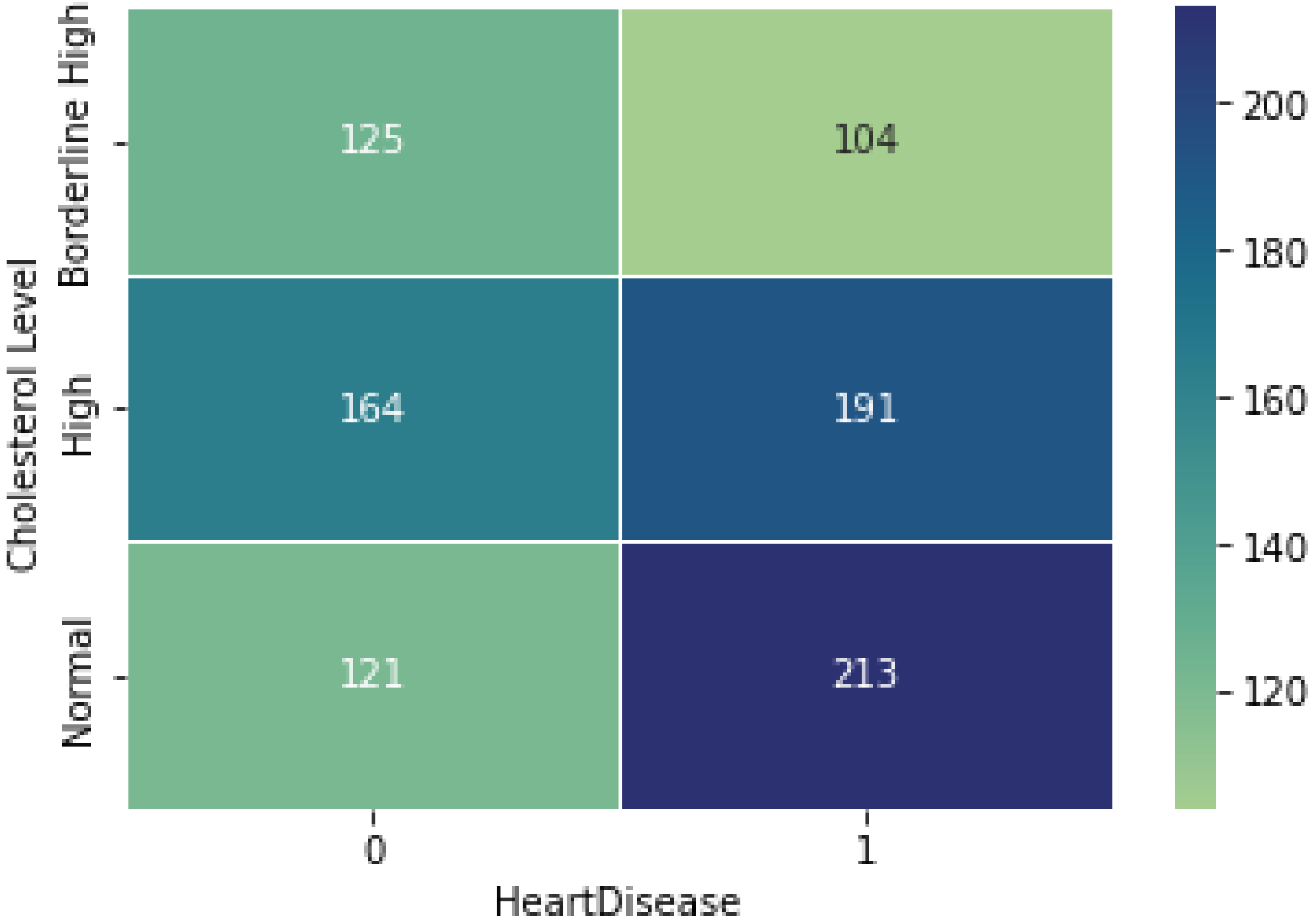
QUESTION NO3

Cholesterol Group



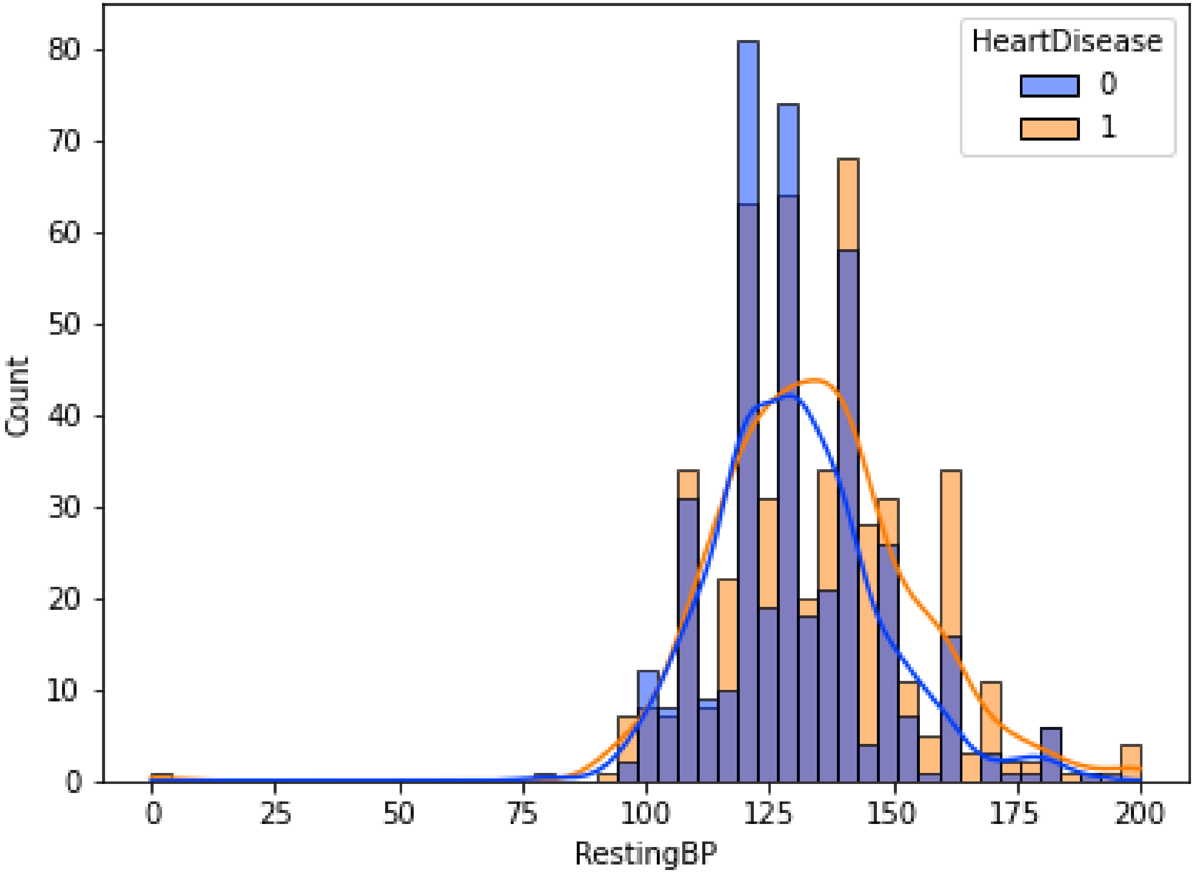
QUESTION NO3

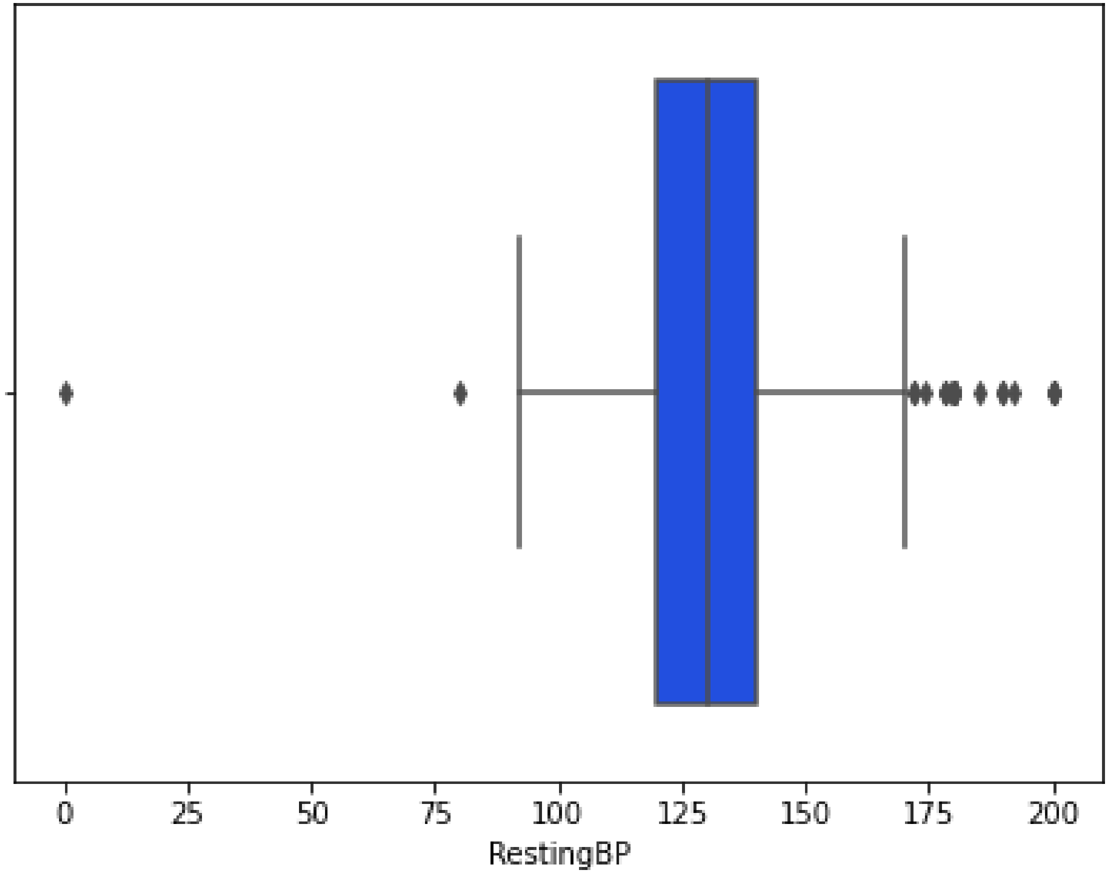
Cholesterol Group



QUESTION NO3

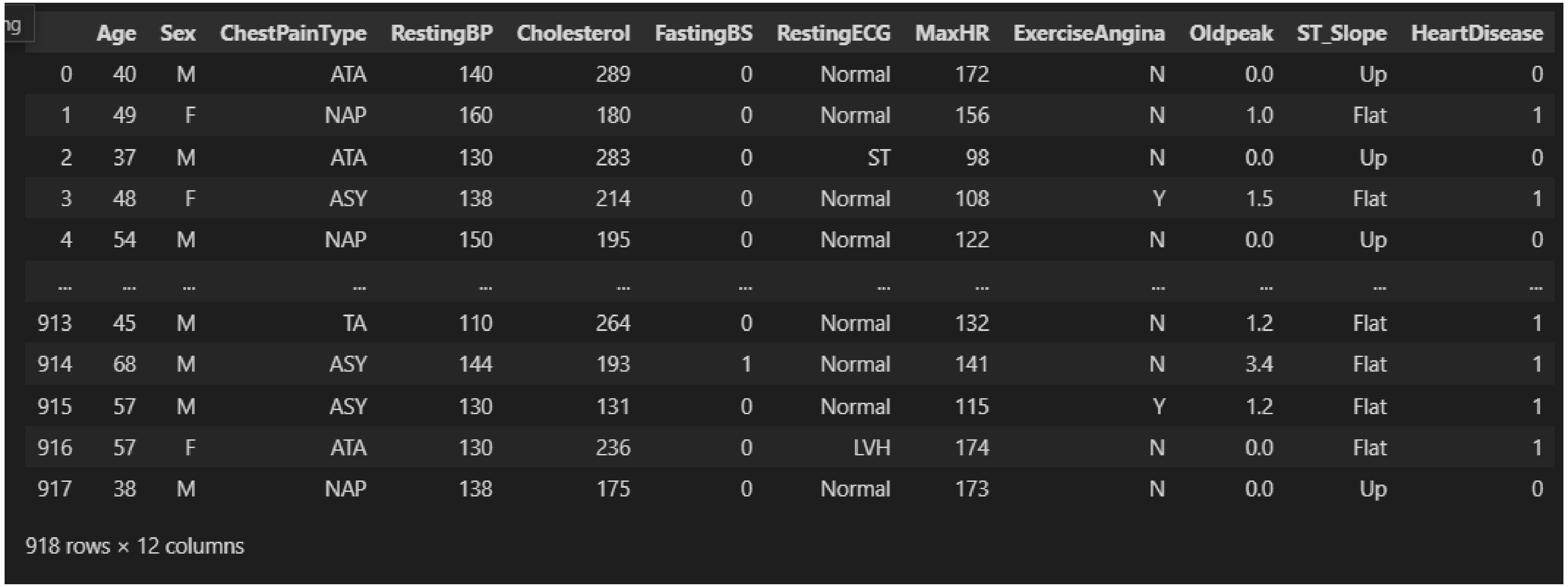
Blood Pressure – Comparison between blood pressure and heart disease





QUESTION NO3

Blood Pressure Data Frame



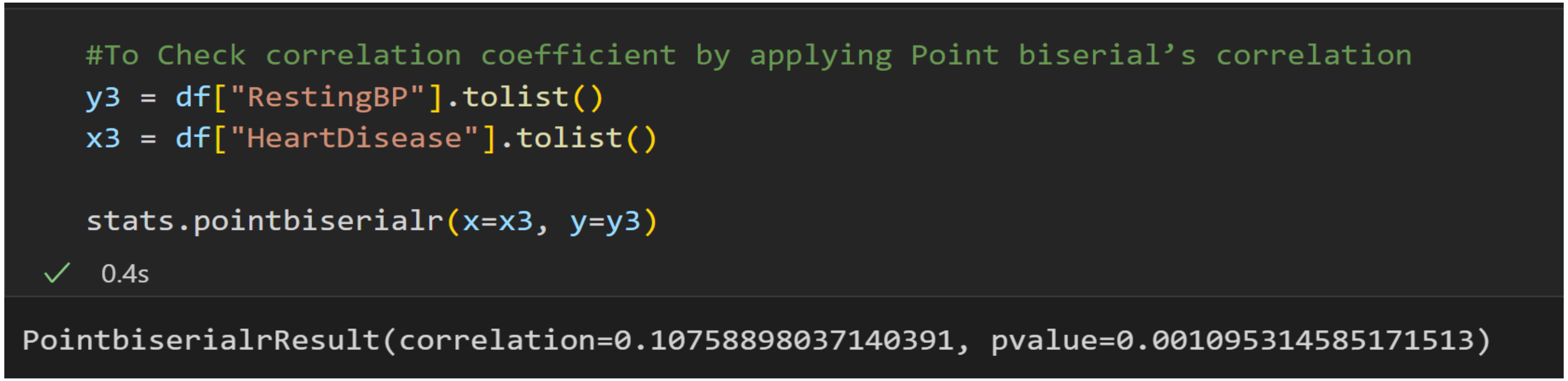
	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

918 rows × 12 columns

Blood Pressure Data Type: Continuous Data
Heart Disease: Binary Data
Check Correlation Coefficient – apply point biserial

QUESTION NO3

Blood Pressure Data Frame



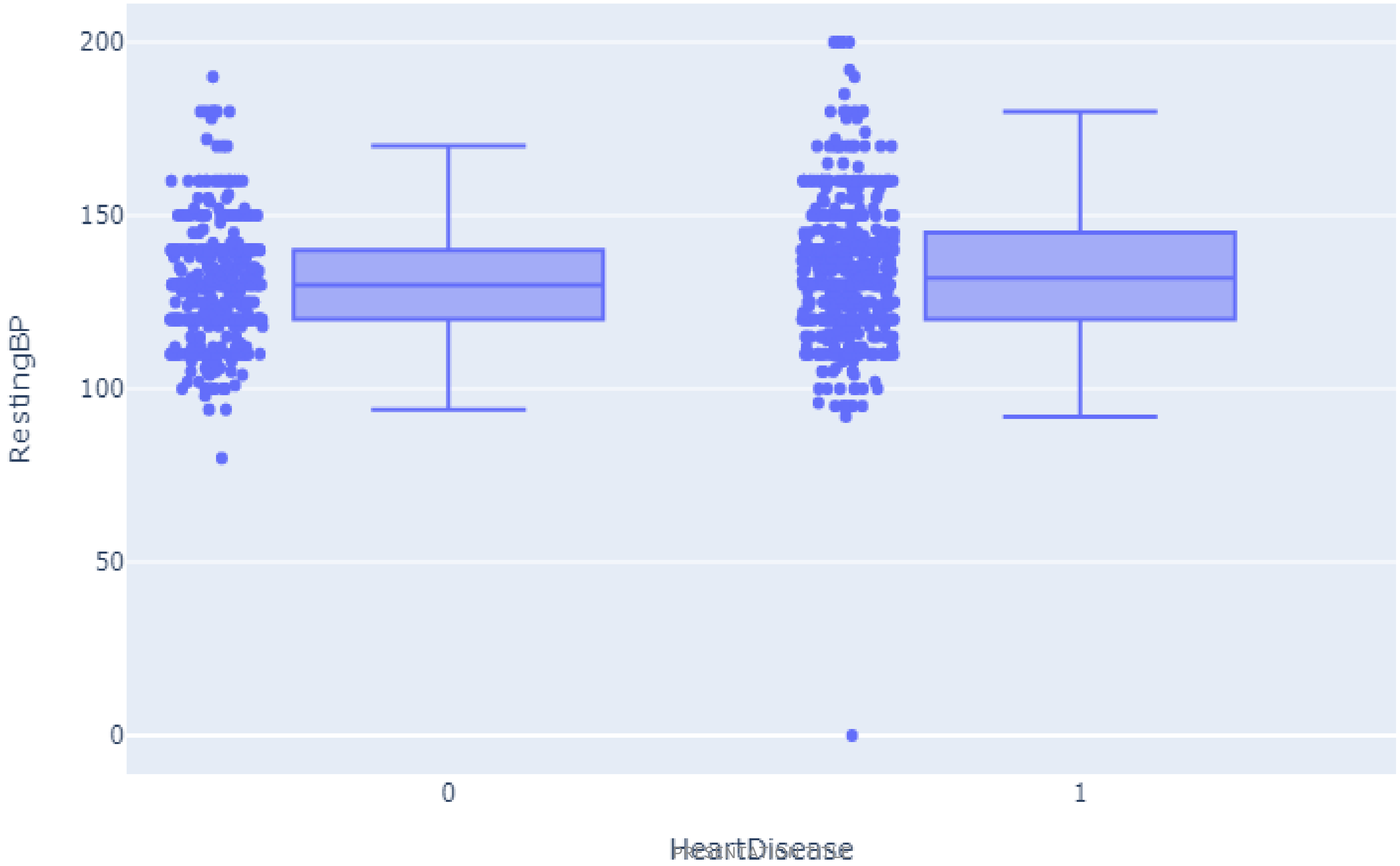
```
#To Check correlation coefficient by applying Point biserial’s correlation
y3 = df["RestingBP"].tolist()
x3 = df["HeartDisease"].tolist()

stats.pointbiserialr(x=x3, y=y3)
```

✓ 0.4s

```
PointbiserialrResult(correlation=0.10758898037140391, pvalue=0.001095314585171513)
```

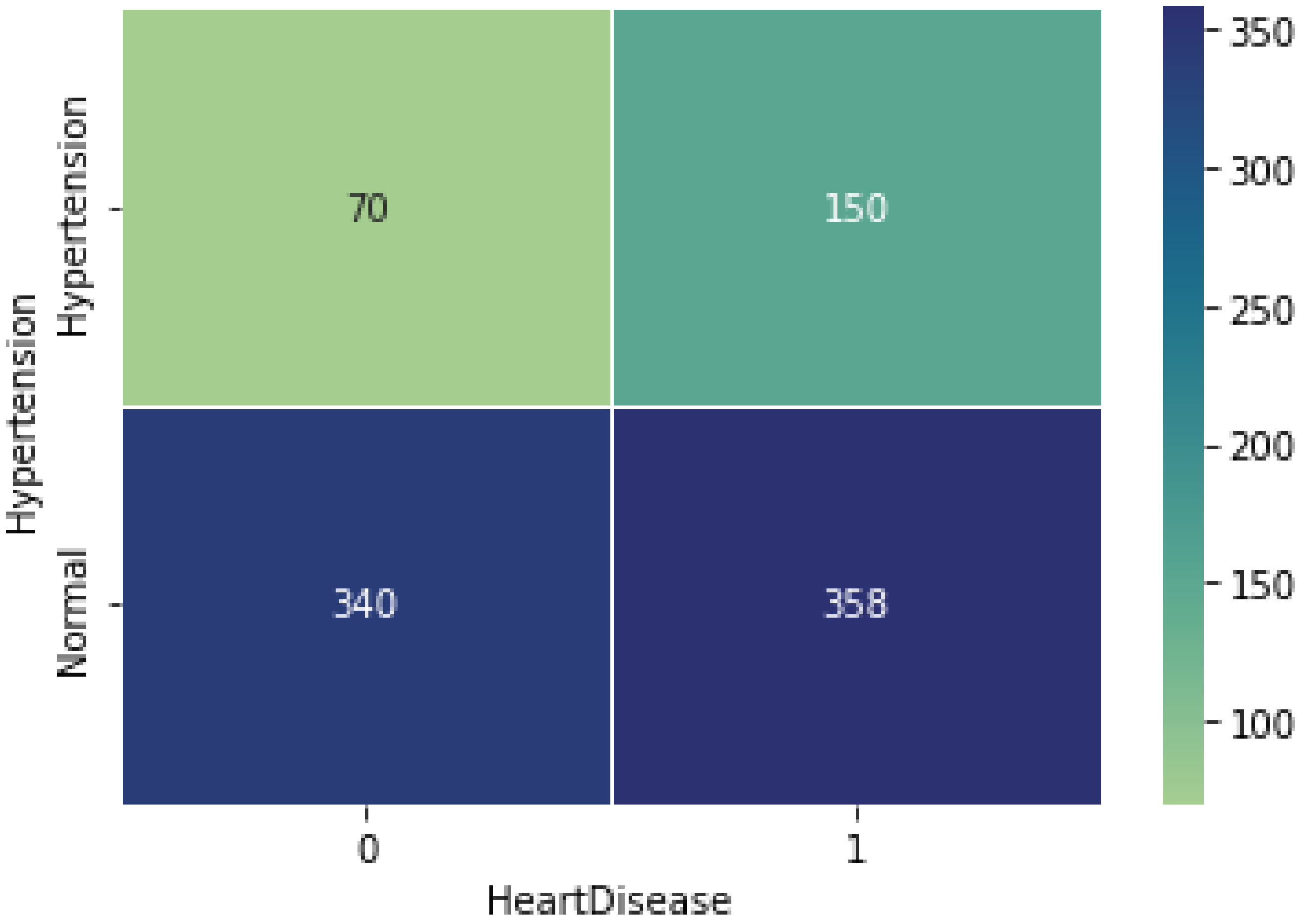
QUESTION NO3

Blood Pressure Group



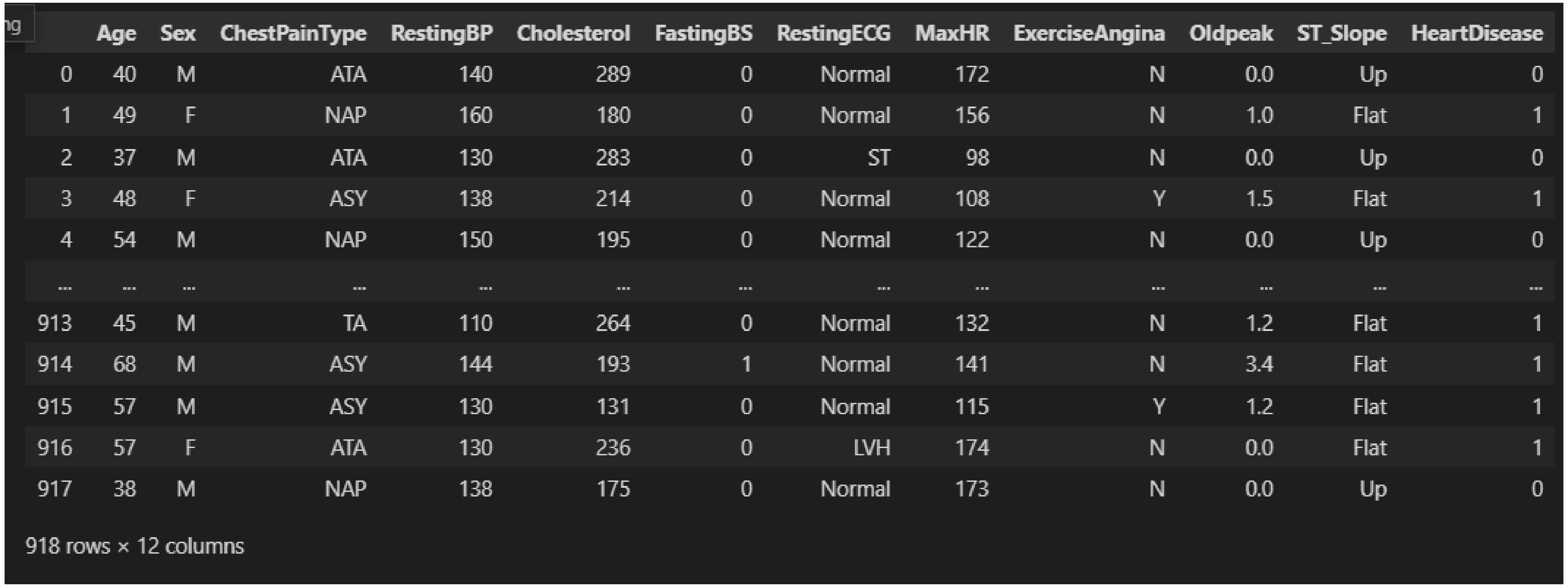
QUESTION NO3

Blood Pressure Group



QUESTION NO3

Diabetes Data Frame



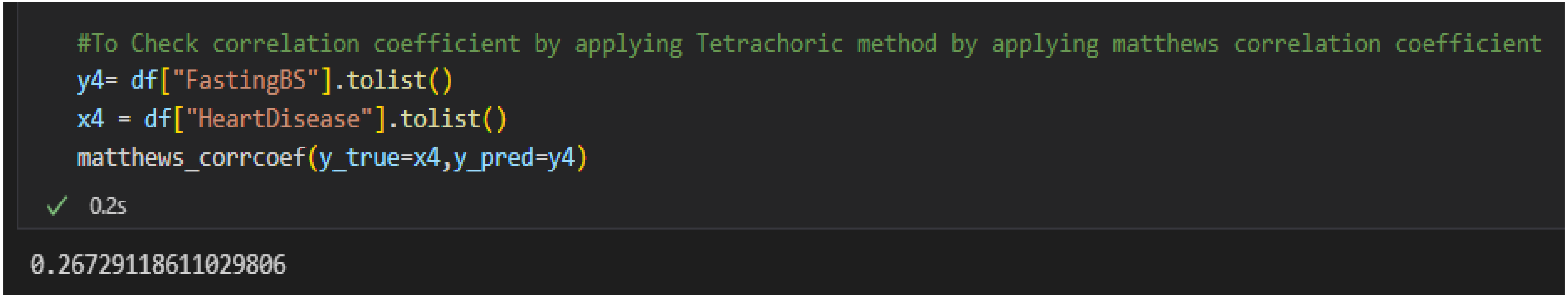
	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

918 rows × 12 columns

Diabetes Data Type: Binary Data
Heart Disease: Binary Data
Check Correlation Coefficient – apply tetrachoric method by finding Matthews Correlation

QUESTION NO3

Diabetes Group



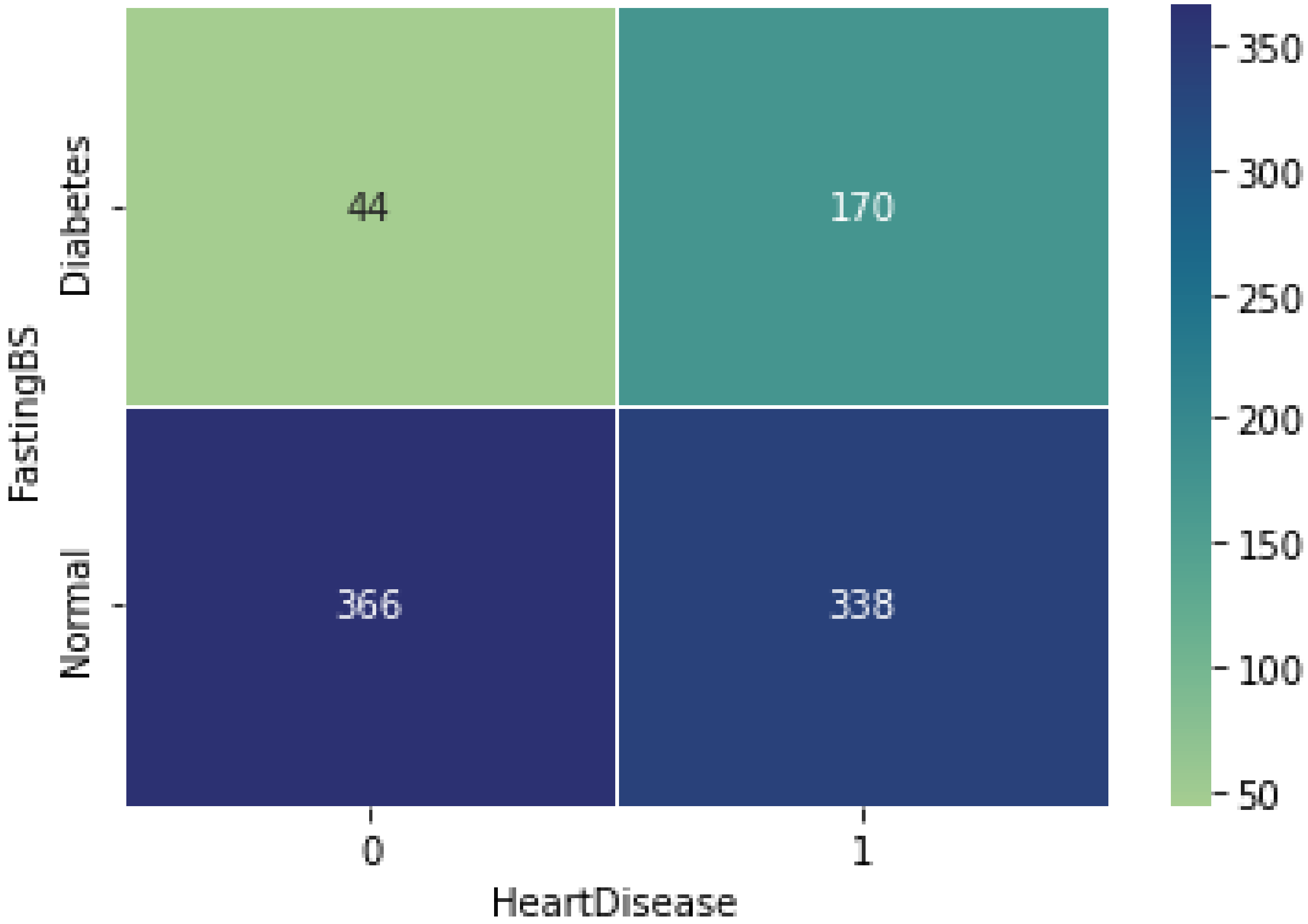
```
#To Check correlation coefficient by applying Tetrachoric method by applying matthews correlation coefficient
y4= df["FastingBS"].tolist()
x4 = df["HeartDisease"].tolist()
matthews_corrcoef(y_true=x4,y_pred=y4)
```

✓ 0.2s

```
0.26729118611029806
```

QUESTION NO3

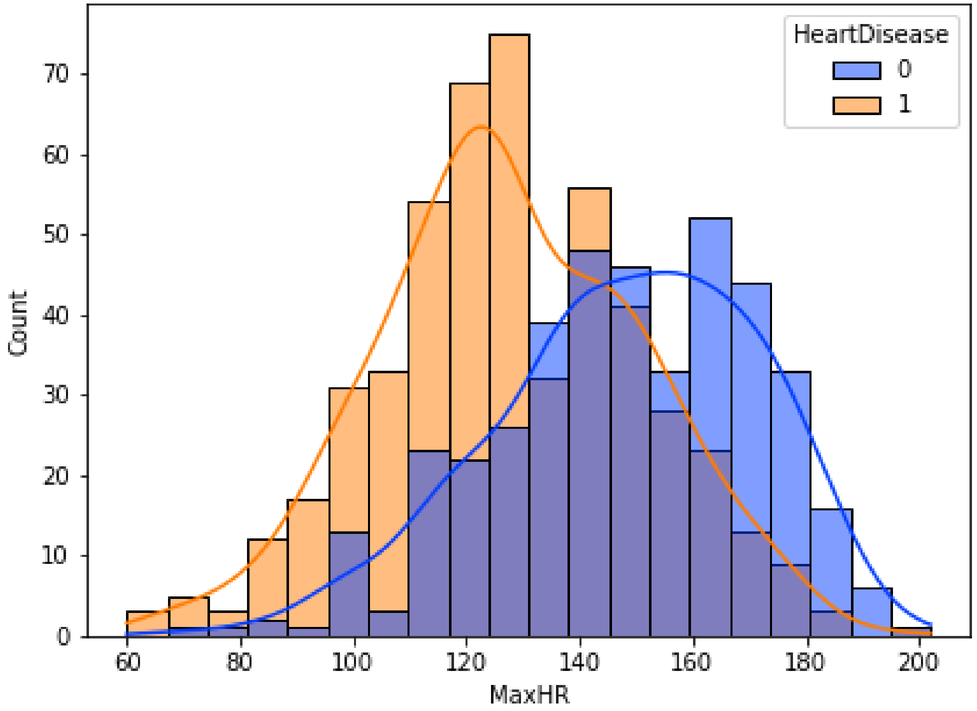
Diabetes Group

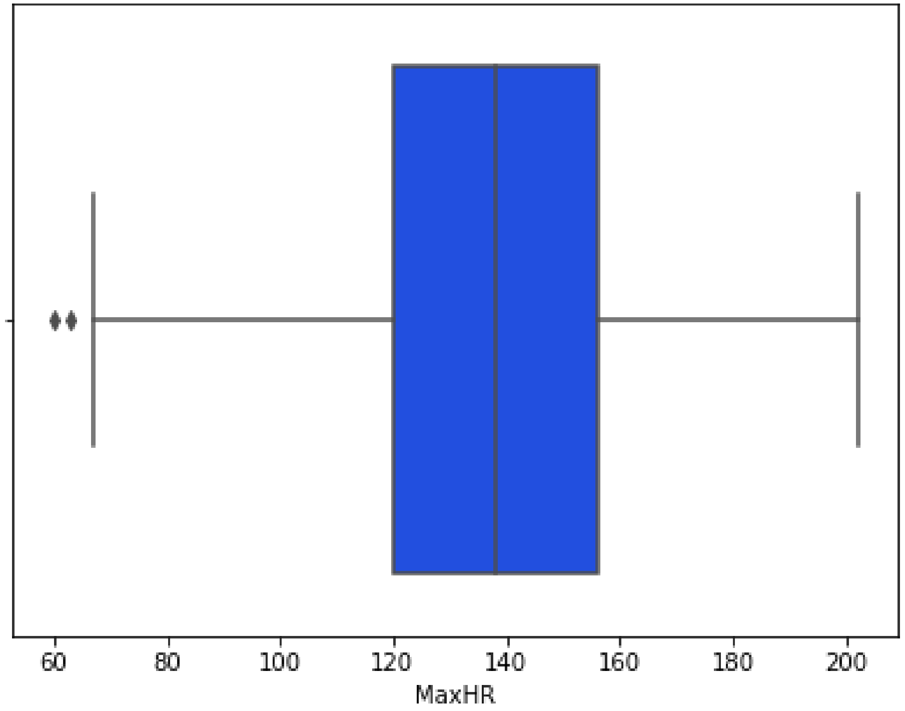


QUESTION NO4

Is higher Max HR can lead to heart disease?

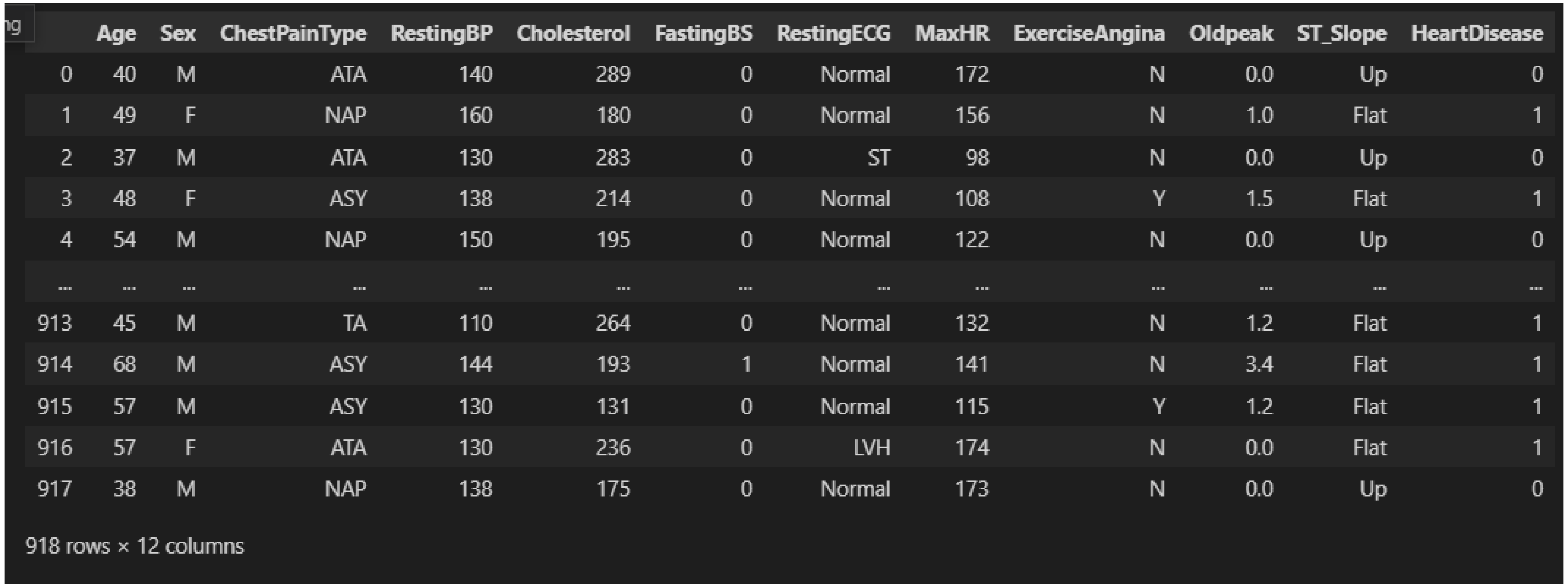
Comparison between Max HR and heart disease





QUESTION NO4

Max HR Data Frame



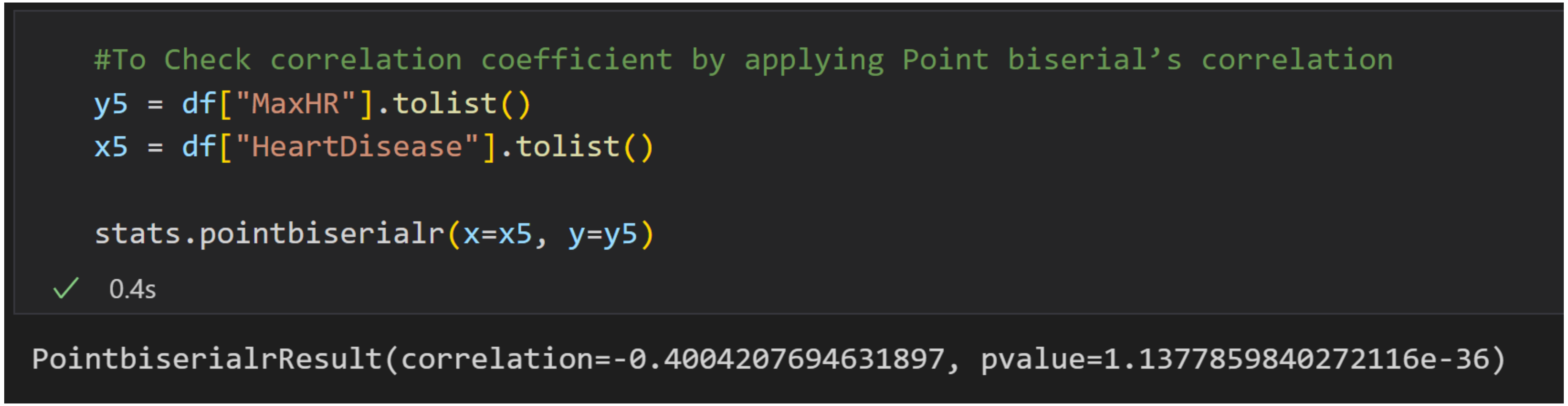
	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

918 rows × 12 columns

Max HR Data Type: Continuous Data
Heart Disease: Binary Data
Check Correlation Coefficient – apply point biserial

QUESTION NO4

Max HR Data Frame



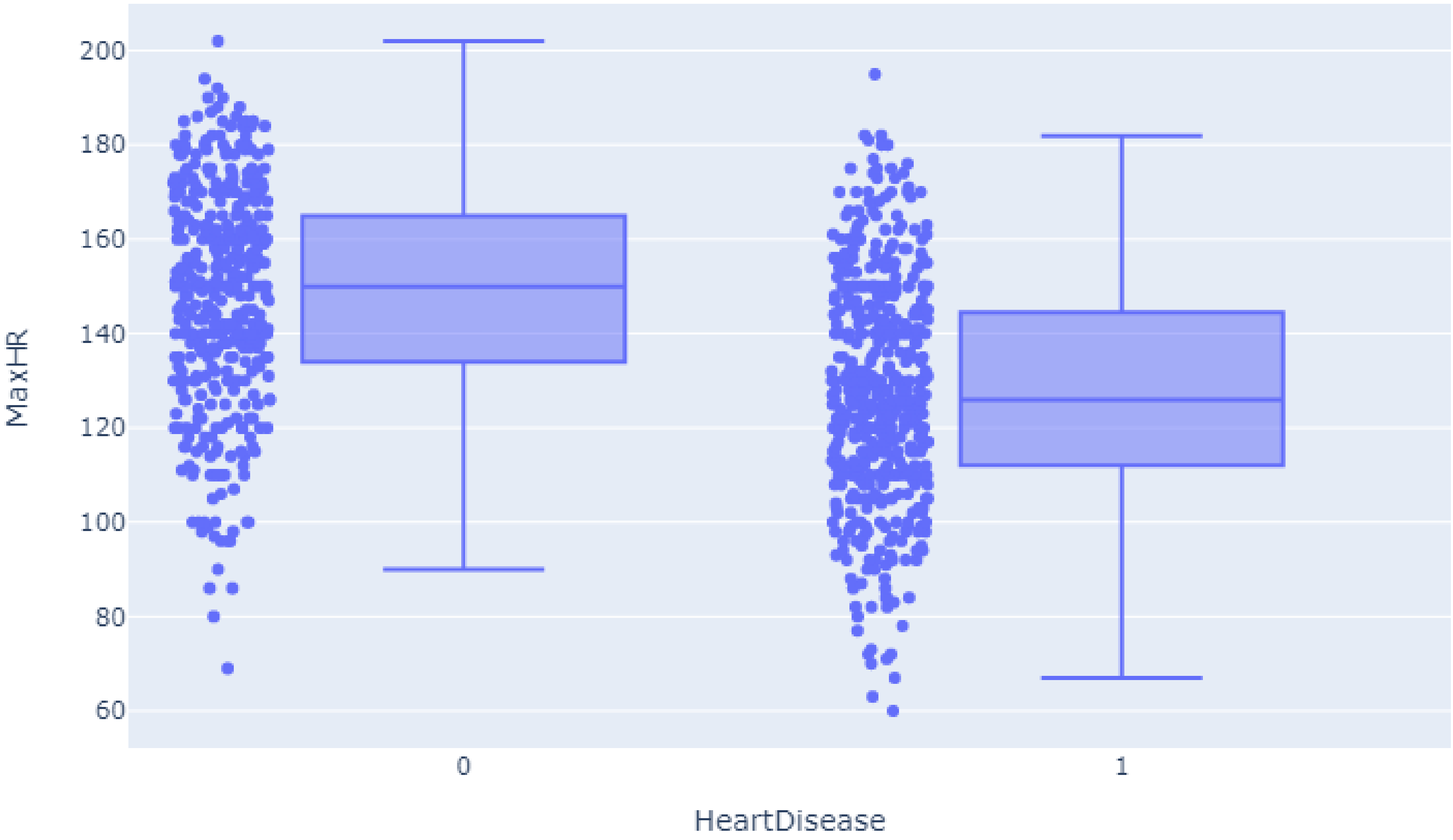
```
#To Check correlation coefficient by applying Point biserial’s correlation
y5 = df["MaxHR"].tolist()
x5 = df["HeartDisease"].tolist()

stats.pointbiserialr(x=x5, y=y5)
```

✓ 0.4s

```
PointbiserialrResult(correlation=-0.4004207694631897, pvalue=1.1377859840272116e-36)
```

QUESTION NO4

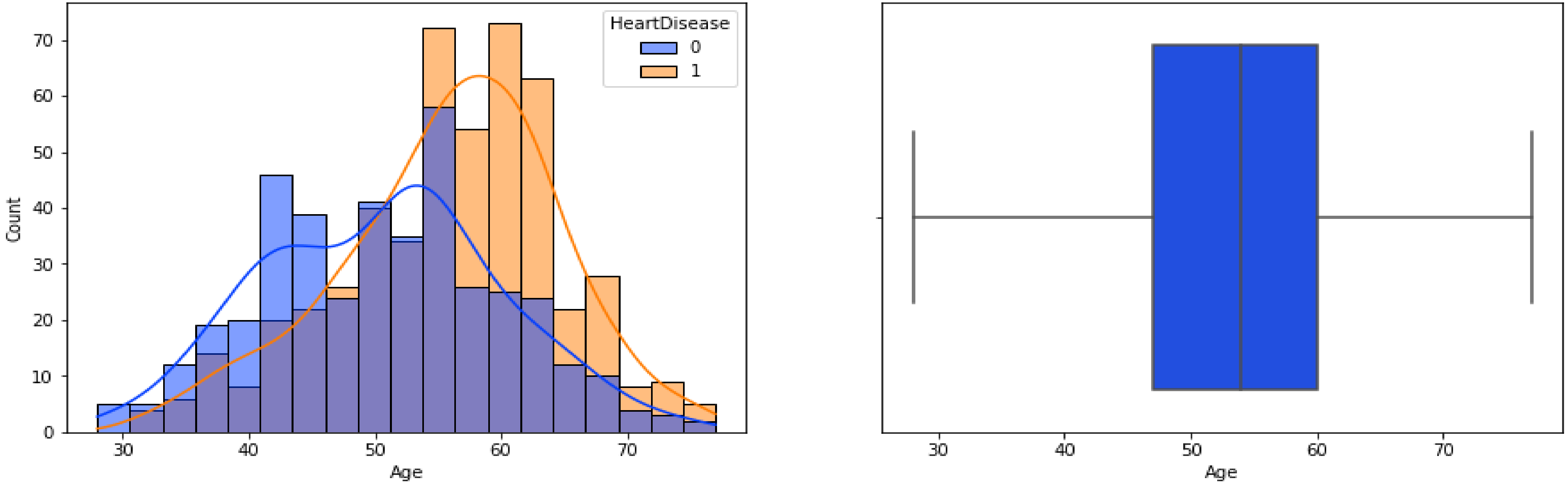
Max HR Group



QUESTION NO5

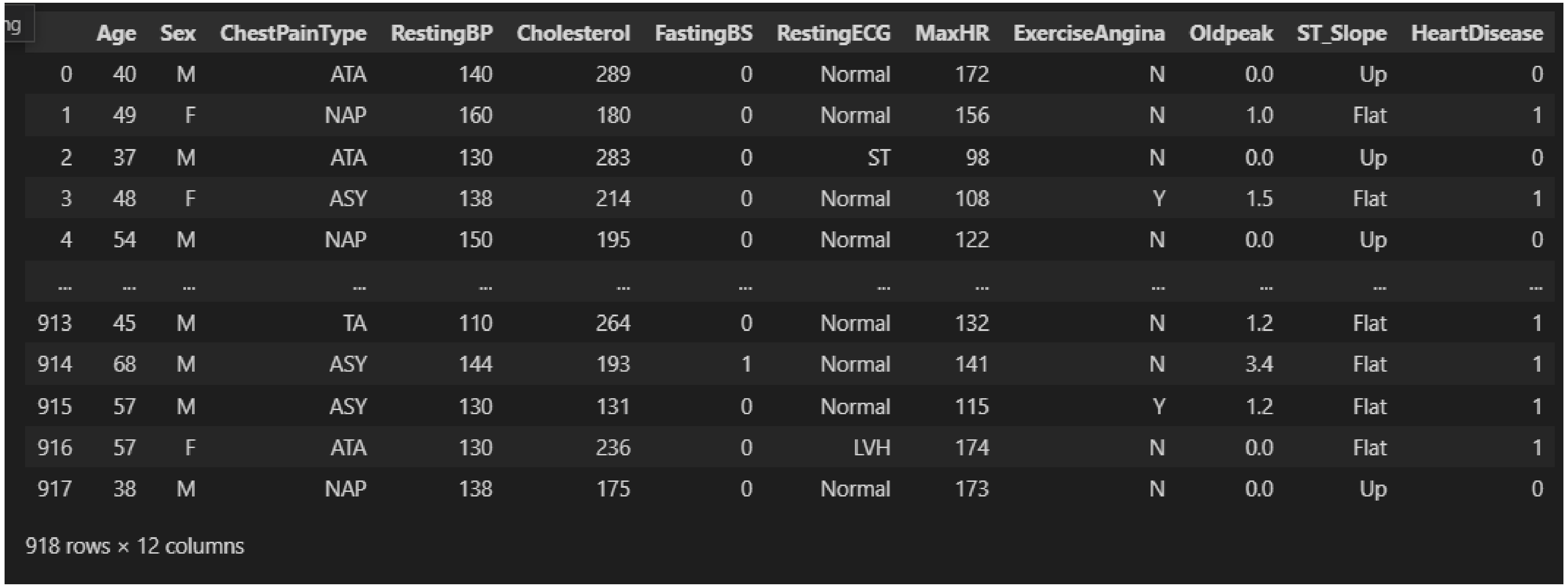
Does Age go higher will lead to get heart disease? What is the odds of getting it when one year older?

Comparison between age and heart disease



QUESTION NO5

Age Data Frame



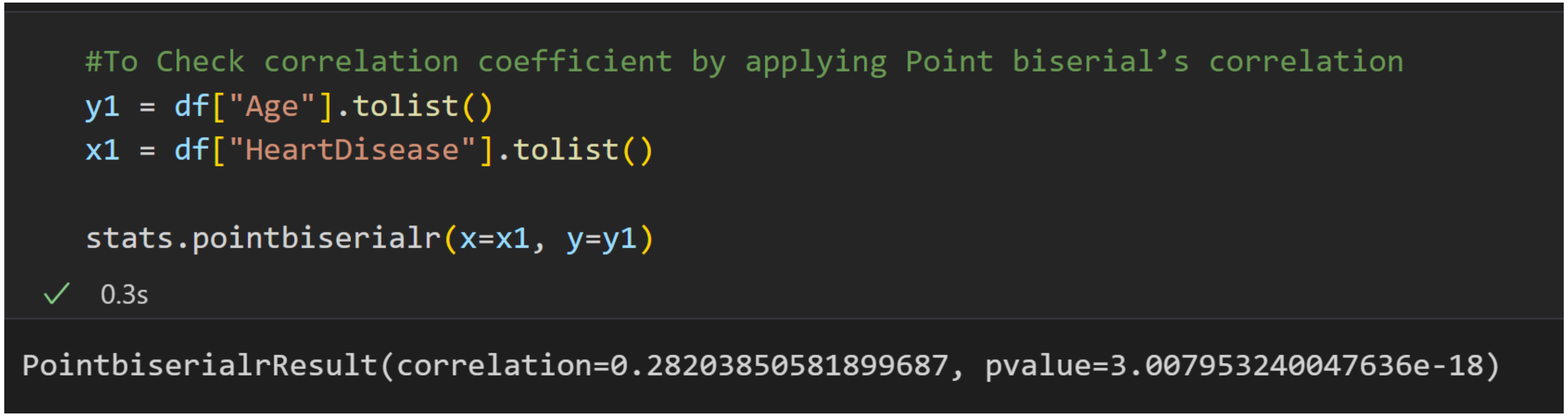
	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.0	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.0	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.0	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.5	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.0	Up	0
...	...	...	...	...	...	...	...	...	...	...	...	...
913	45	M	TA	110	264	0	Normal	132	N	1.2	Flat	1
914	68	M	ASY	144	193	1	Normal	141	N	3.4	Flat	1
915	57	M	ASY	130	131	0	Normal	115	Y	1.2	Flat	1
916	57	F	ATA	130	236	0	LVH	174	N	0.0	Flat	1
917	38	M	NAP	138	175	0	Normal	173	N	0.0	Up	0

918 rows × 12 columns

Age Data Type: Continuous Data
Heart Disease: Binary Data
Check Correlation Coefficient – apply point biserial

QUESTION NO5

Age Data Frame



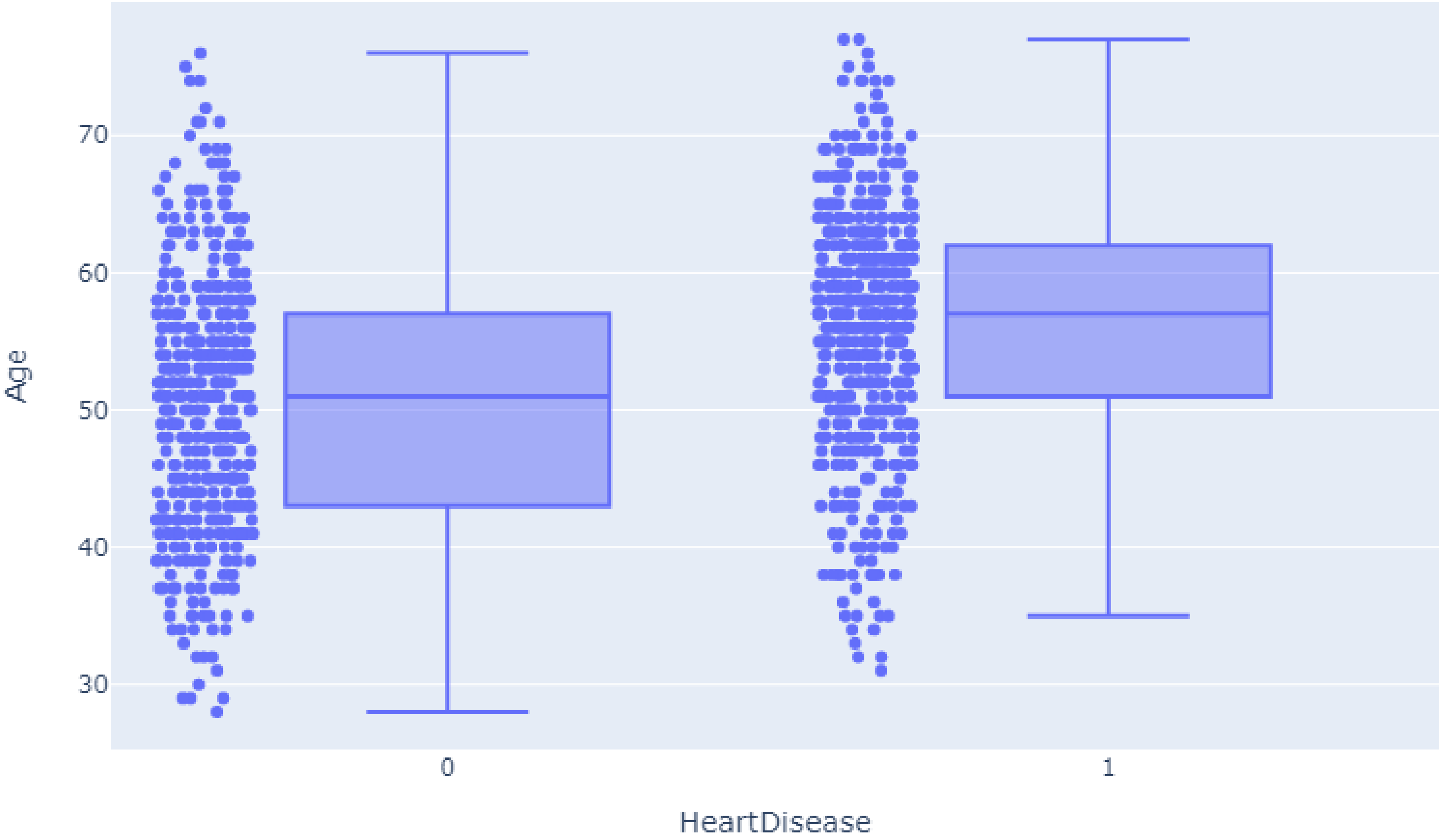
```
#To Check correlation coefficient by applying Point biserial’s correlation
y1 = df["Age"].tolist()
x1 = df["HeartDisease"].tolist()

stats.pointbiserialr(x=x1, y=y1)
```

✓ 0.3s

```
PointbiserialrResult(correlation=0.28203850581899687, pvalue=3.007953240047636e-18)
```

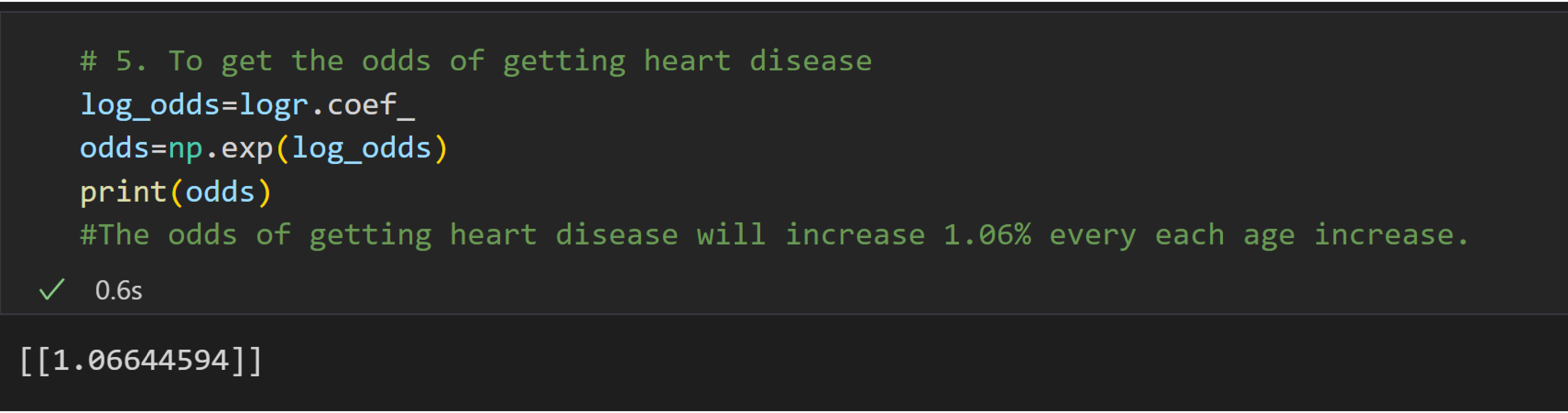
QUESTION NO5

Age Group



QUESTION NO 5

Age Group



```
# 5. To get the odds of getting heart disease
log_odds=logr.coef_
odds=np.exp(log_odds)
print(odds)
#The odds of getting heart disease will increase 1.06% every each age increase.
```

✓ 0.6s

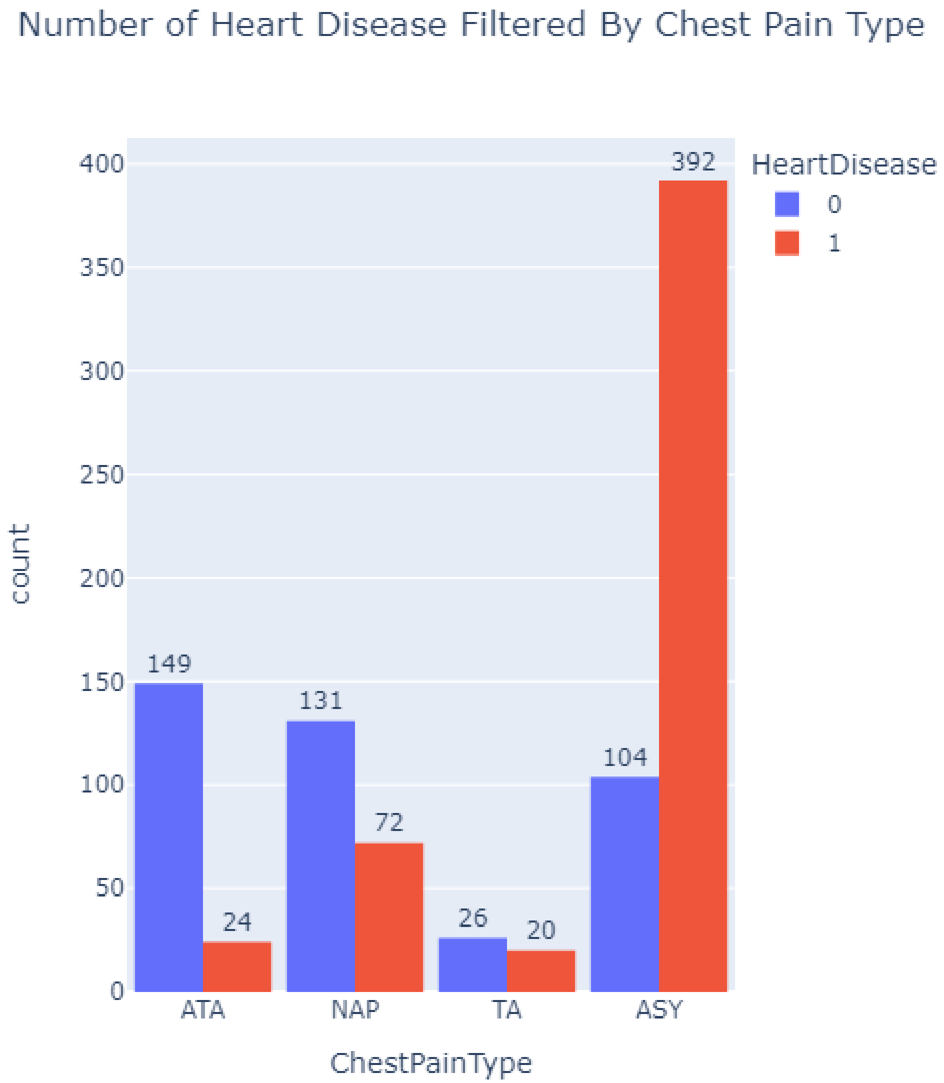
```
[[1.06644594]]
```

QUESTION NO6

Can type of chest pain, type of resting ECG, presence of angina induced during exercise, type of ST slope after exercise detect person to have heart disease?

QUESTION NO6

Chest Pain Type



ATA - Atypical Angina

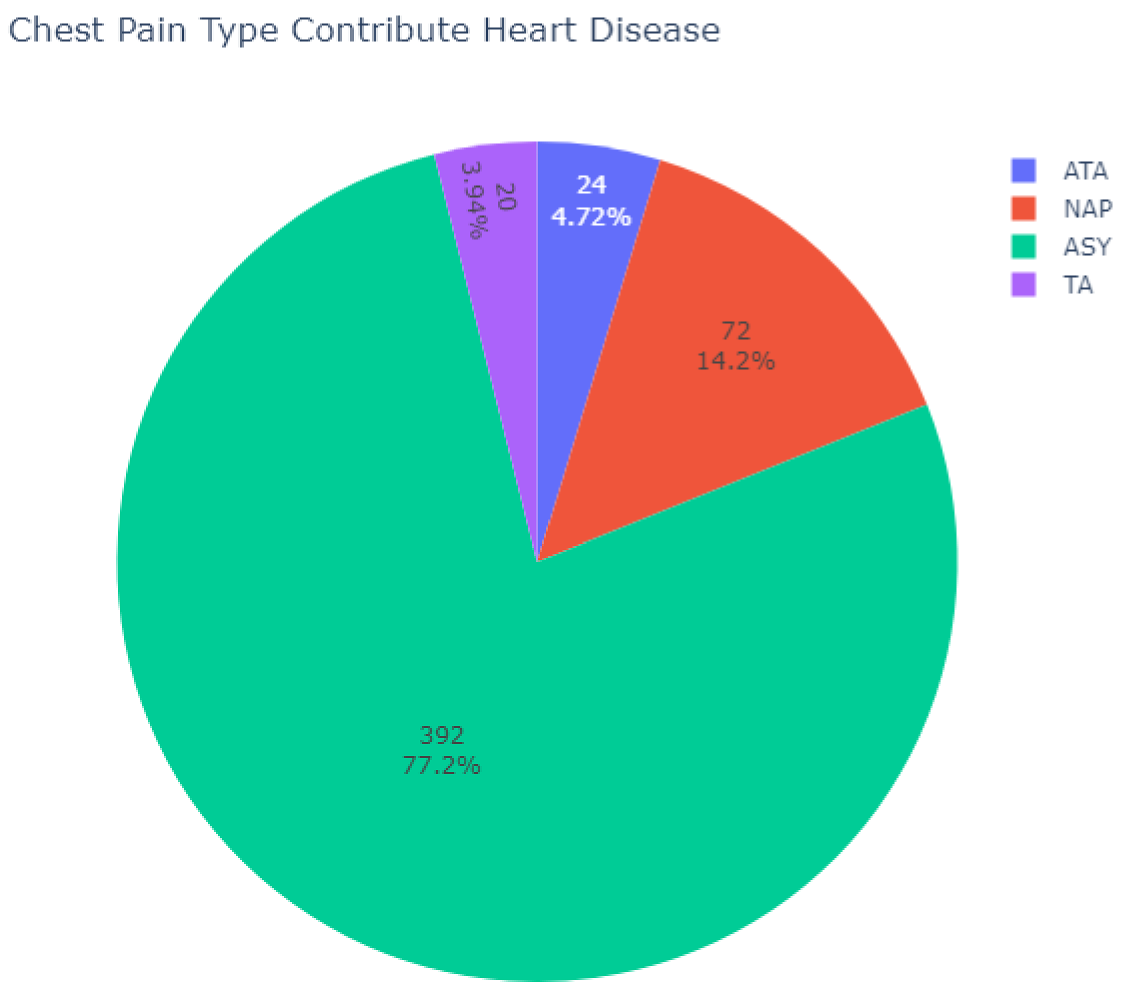
NAP - Non Angina Pain

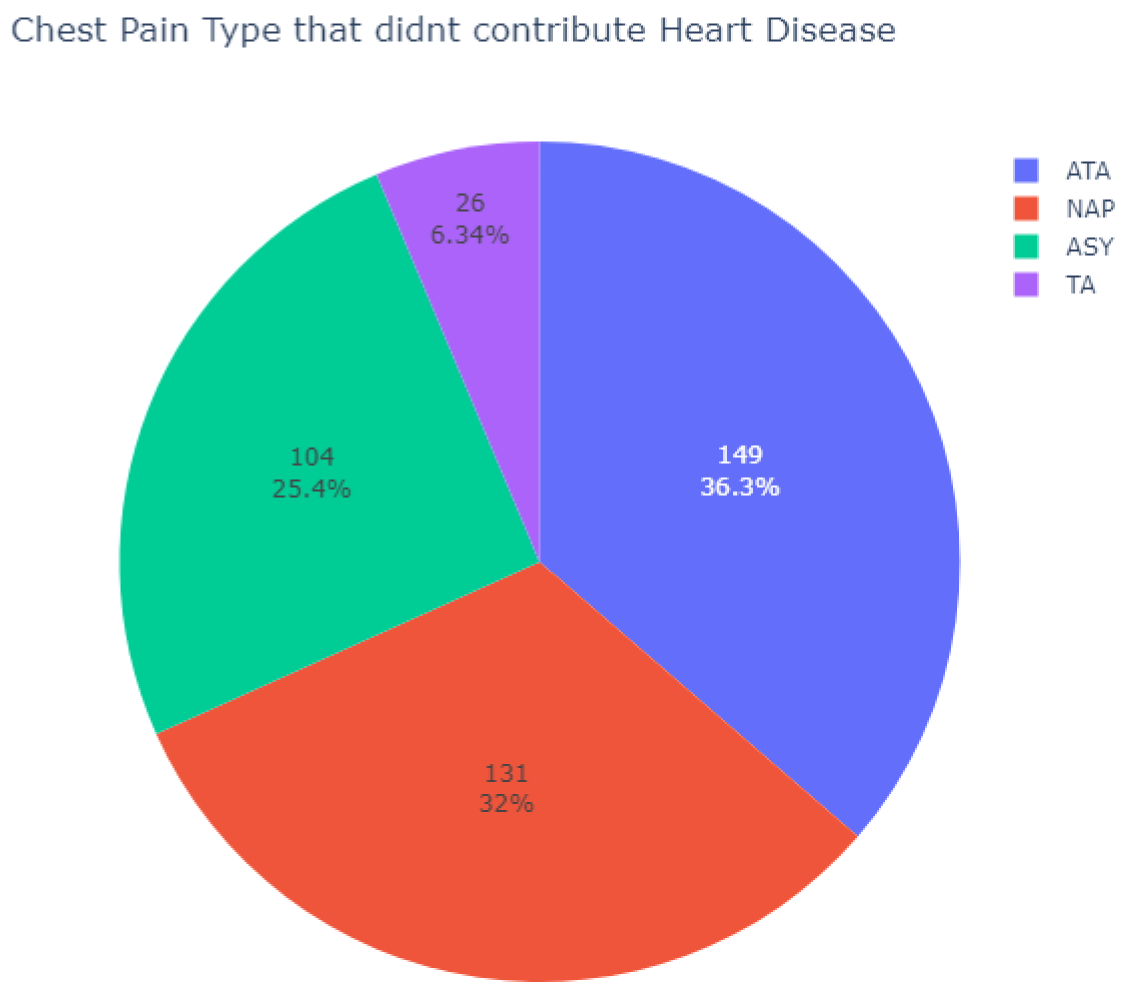
ASY – Asymptomatic

TA - Typical Angina

QUESTION NO6

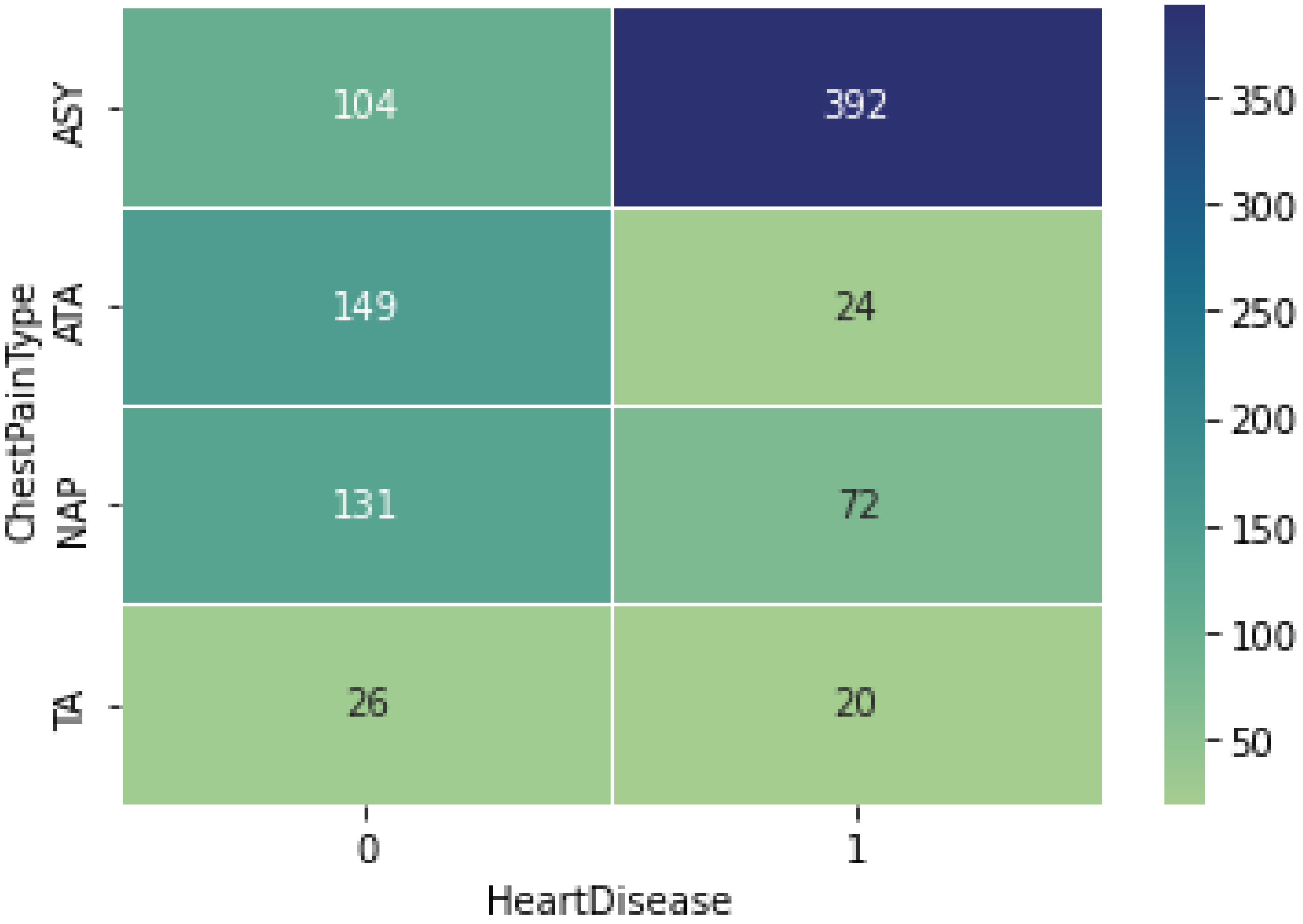
Chest Pain Type


Chest Pain Type Contribute Heart Disease


Chest Pain Type that didnt contribute Heart Disease

QUESTION NO6

Chest Pain Type

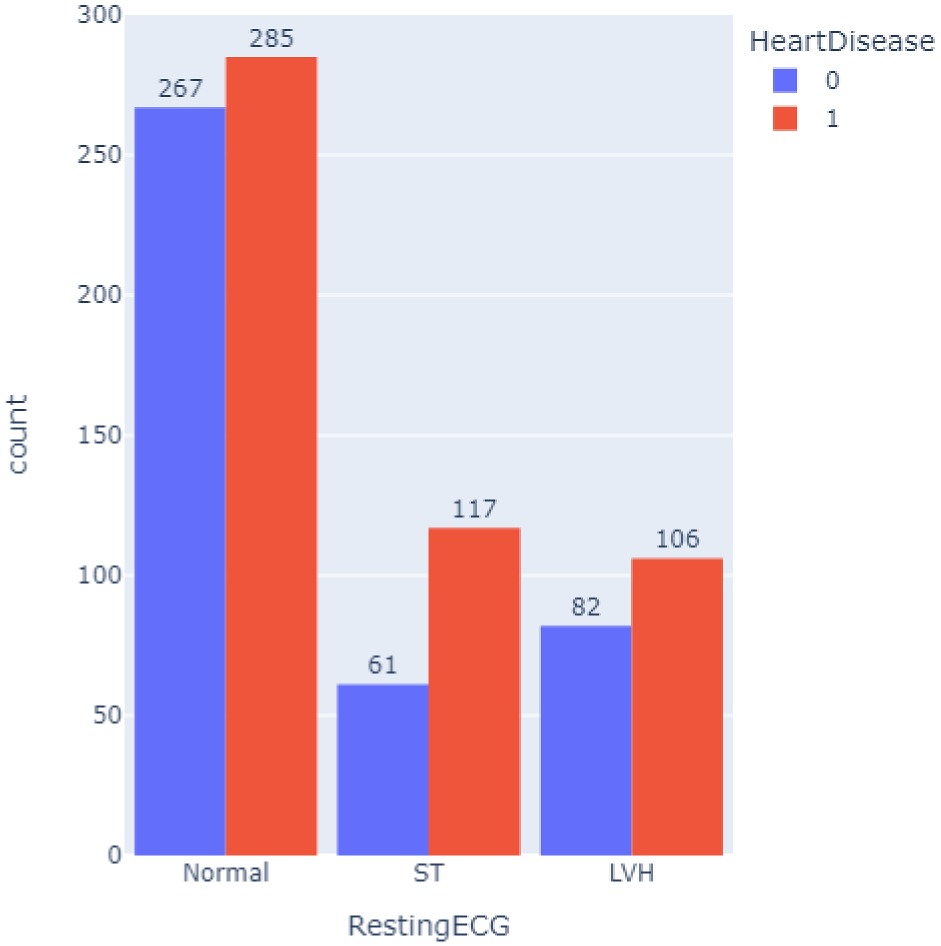


ATA - Atypical Angina
NAP - Non Angina Pain
ASY – Asymptomatic
TA - Typical Angina

QUESTION NO6

Resting ECG Type

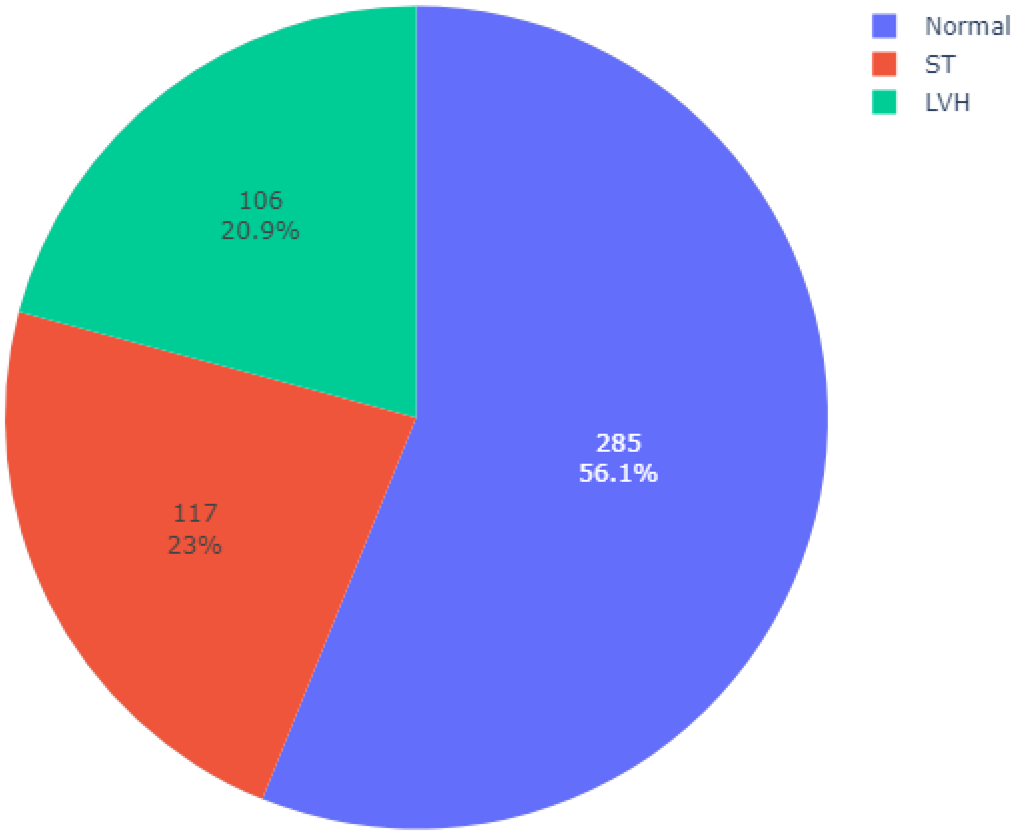
Number of Heart Disease Filtered By Resting ECG Type

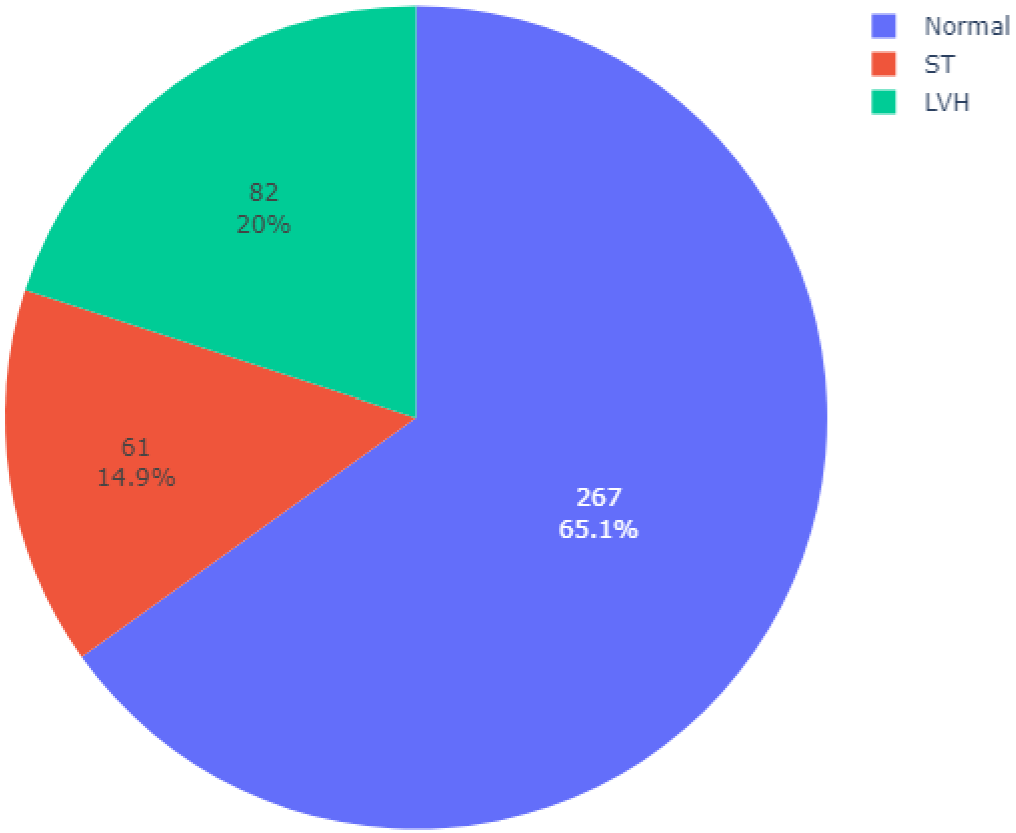


Normal – normal resting ECG
ST – having ST or T wave abnormality
LVH – left ventricle hypertrophy

QUESTION NO6

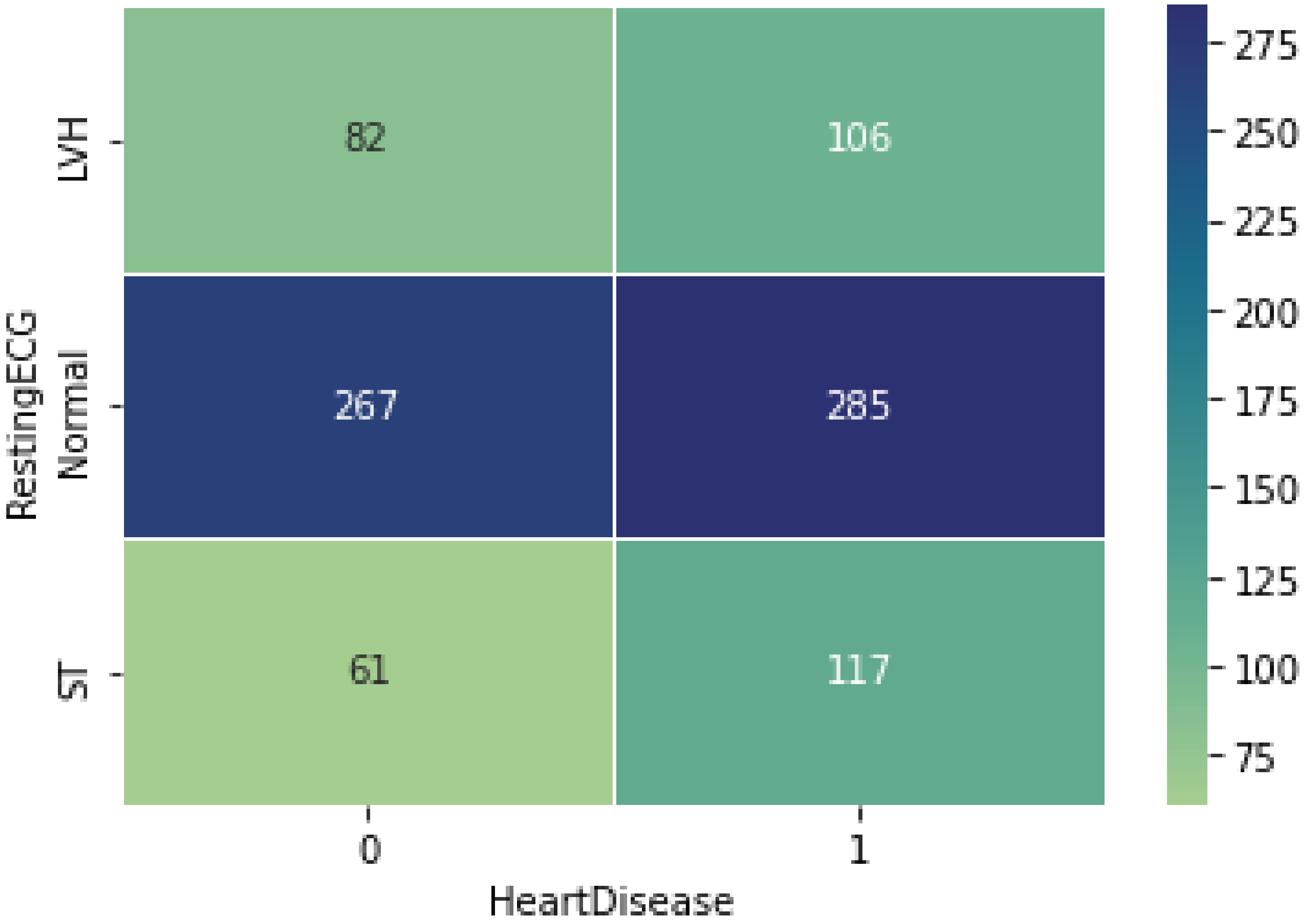
Resting ECG Type


Resting ECG Type Contribute Heart Disease


Resting ECG Type that didnt contribute Heart Disease

QUESTION NO6

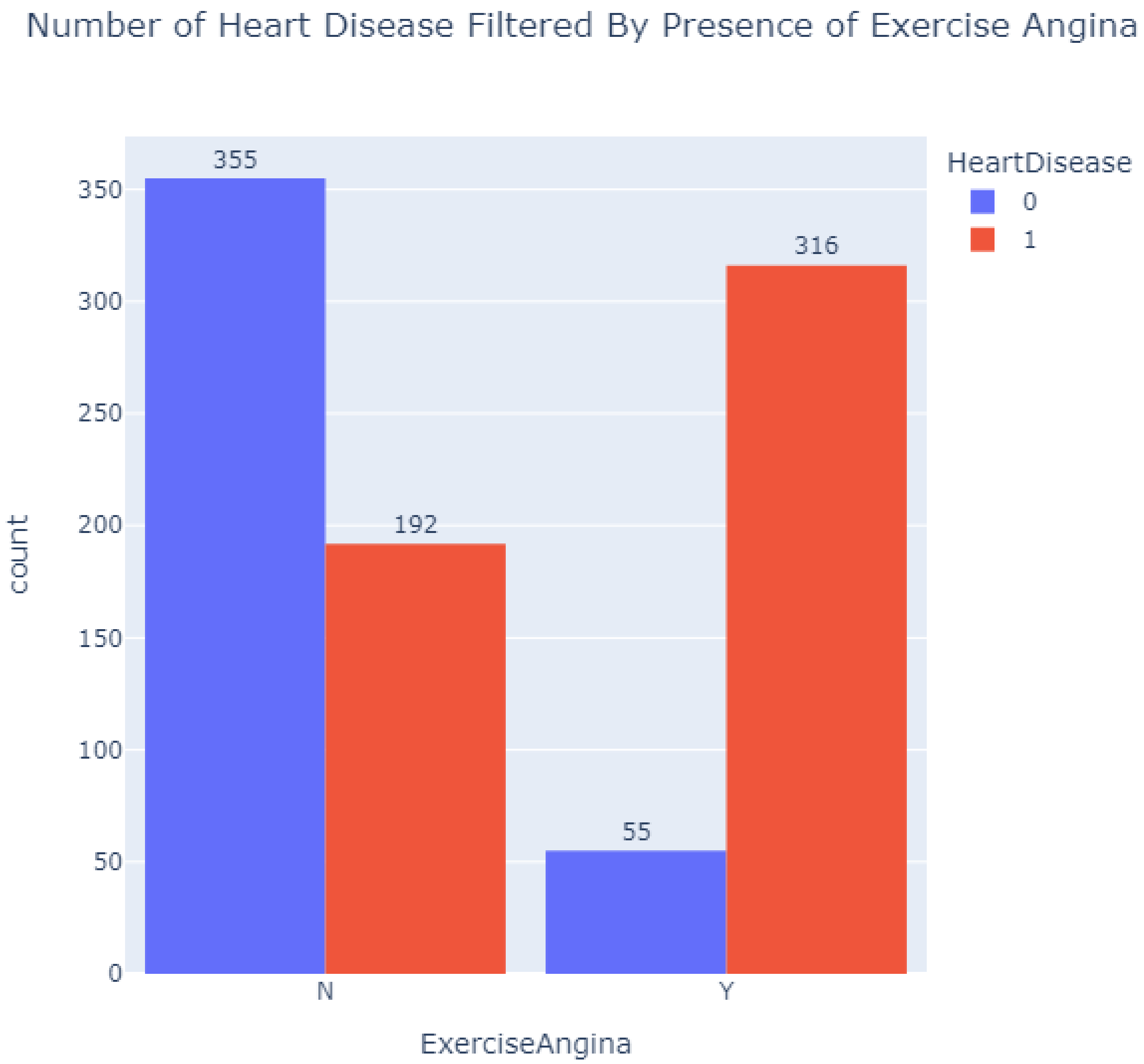
Resting ECG Type



Normal – normal resting ECG
ST – having ST or T wave abnormality
LVH – left ventricle hypertrophy

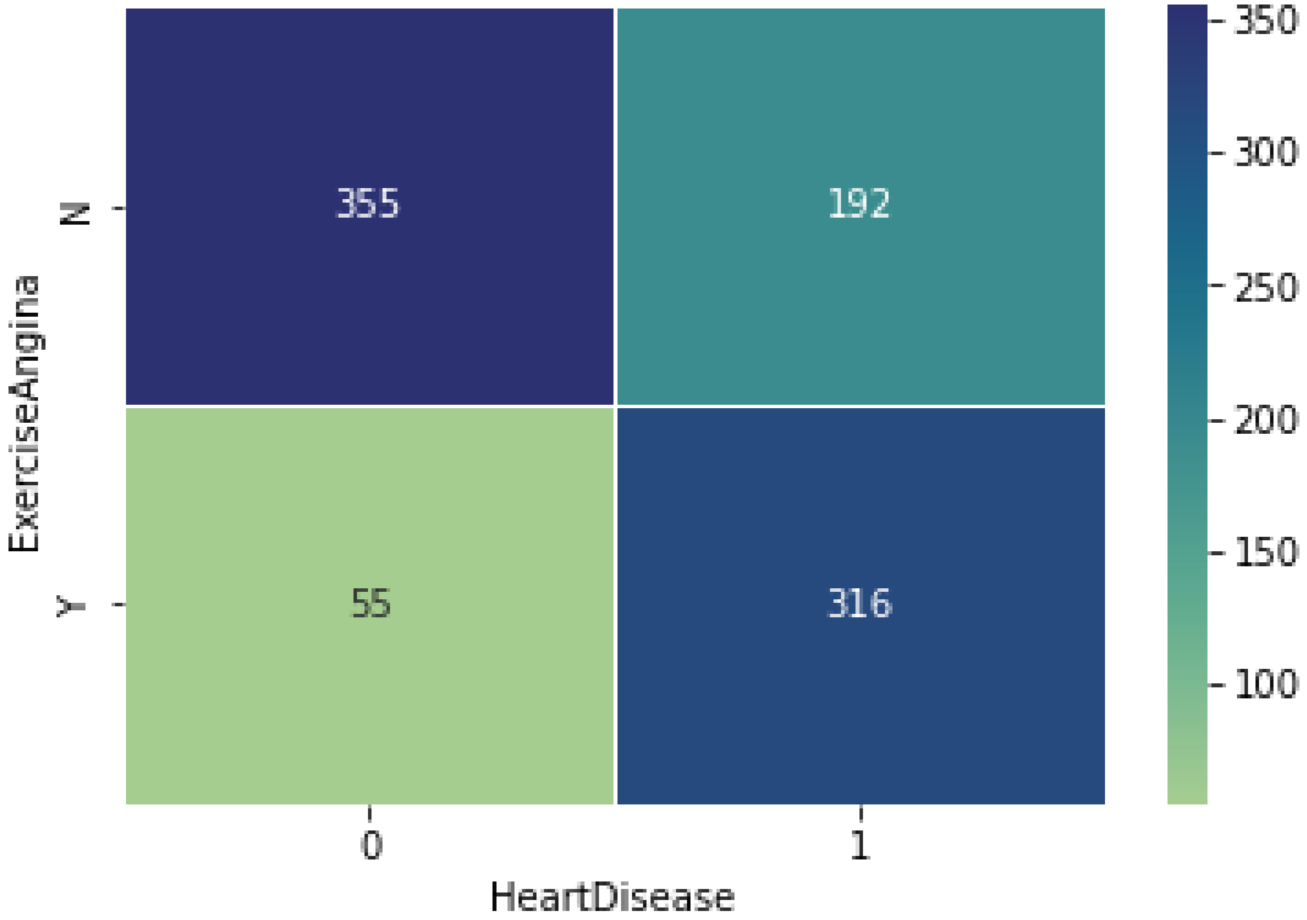
QUESTION NO6

Presence of exercise Angina



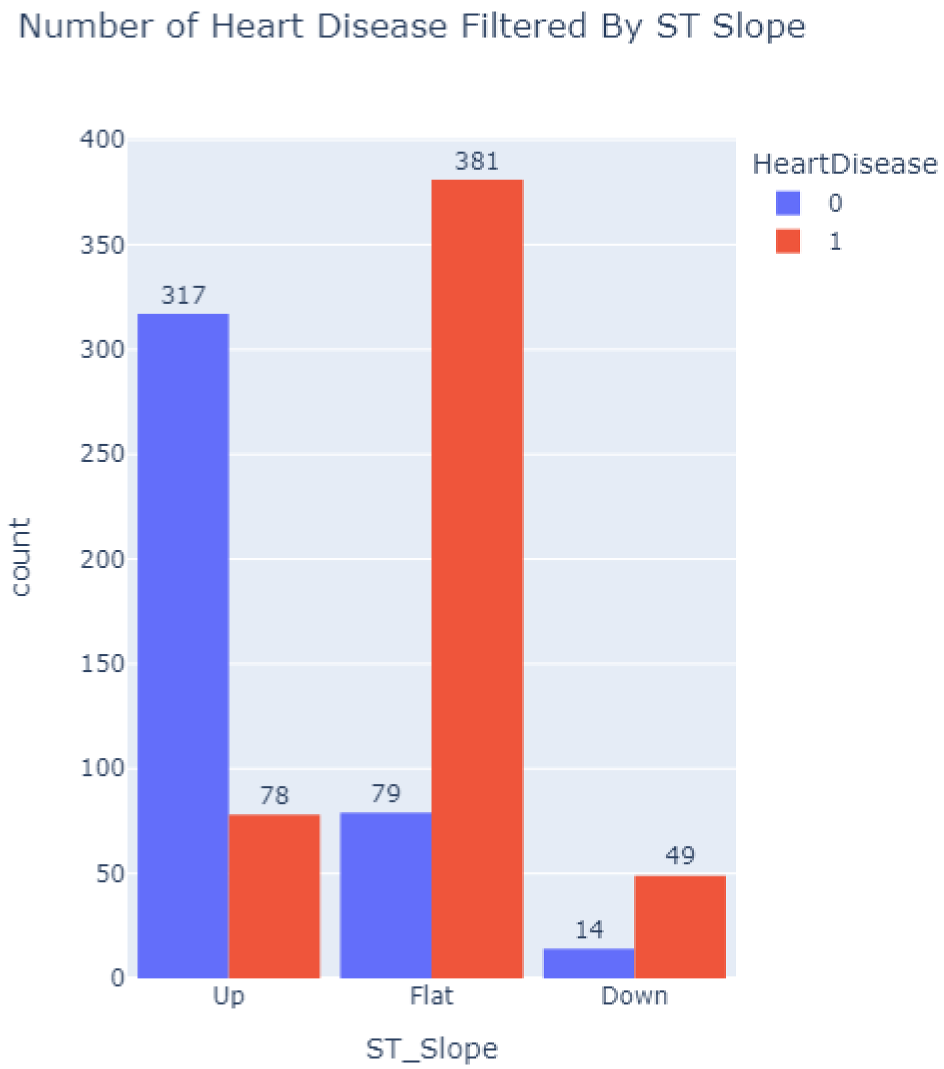
QUESTION NO6

Presence of exercise Angina



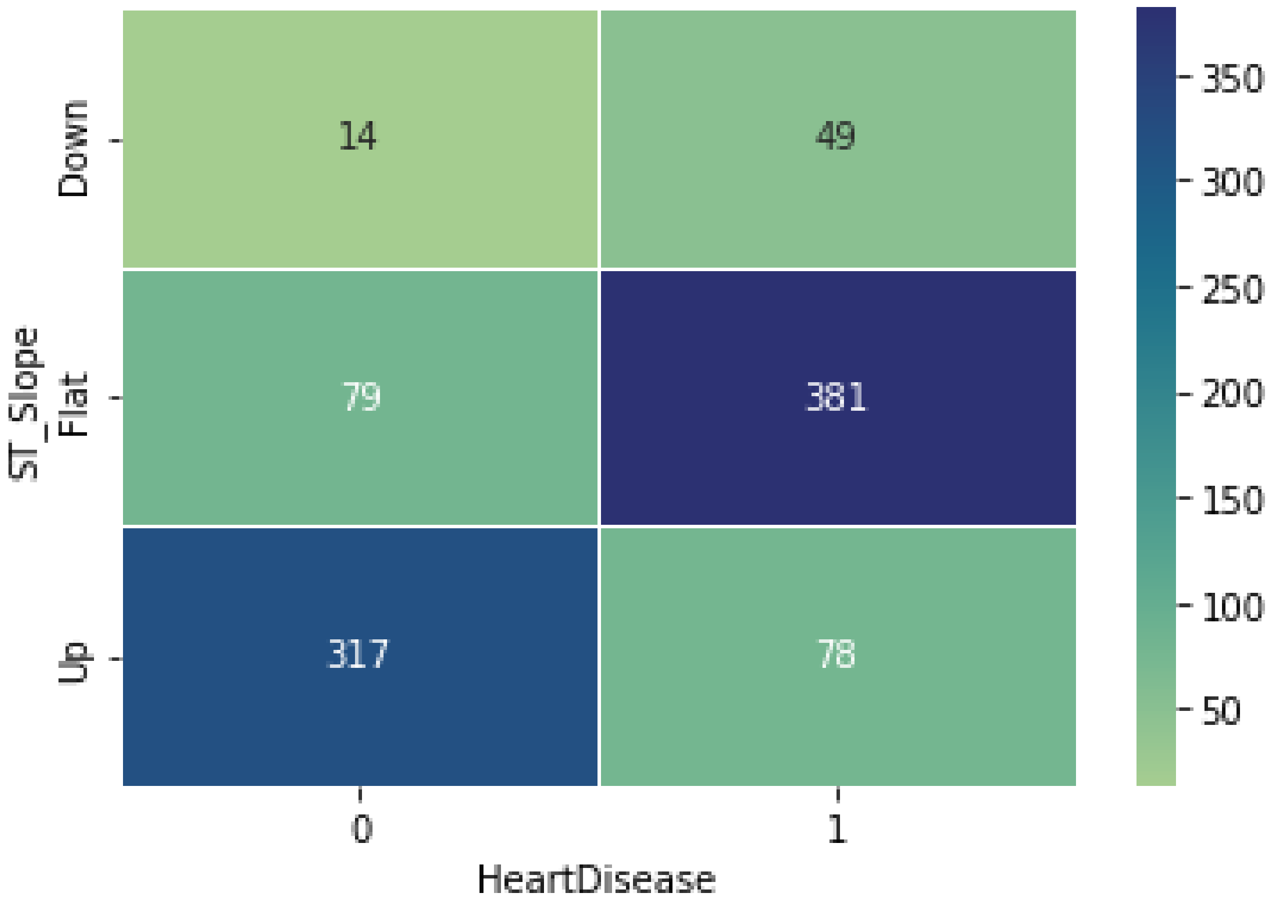
QUESTION NO6

ST Slope after exercise



QUESTION NO6

Presence of exercise Angina



CONCLUSION

- Sample size 918. 55.3% of the population had heart disease.
- Male tend to easy get heart disease compared to female.
- Although cholesterol, hypertension and diabetes claimed to be leading cause of getting heart disease, there is no correlation to get heart disease if one of these 3 group in high risk.
- Higher Max HR wont led to get heart disease.
- Sample shows that when age go higher will lead to get heart disease. The odds of getting it increase by 1.06% with every increase of age.
- Type of chest pain, type of resting ECG, presence of angina induced during exercise and type of ST stop after exercise is a good tool to predict person whether to have heart disease.

THANK YOU