



Heart Failure Prediction

Investigating Advanced Strategies to Mitigate the possibility of heart failure and inferring possible causes

Abhik Rana

Bachelor's in Statistics

Statistical Methods - III

Kolkata, November 2024



Heart Failure Prediction

Investigating Advanced Strategies to Mitigate the possibility of heart failure and inferring possible causes

Abhik Rana

Supervisor: Prof. Ayan Basu

Professor, ISRU, Indian Statistical Institute, Kolkata

Bachelor's in Statistics

Statistical Methods - III

Project

Kolkata, November 2024

DECLARATION OF AUTHORSHIP

Has undersigned, hereby it his declared that this work entitled “Heart Failure Prediction” is the original work and that it has not previously in its entirety or in part been submitted at any university or higher education institution for the award of any degree, diploma, or other qualifications. It is also hereby declared that to the best of the knowledge, this work contains no material previously published or written by another person, except where due reference, acknowledgement, and citation is made.

Kolkata, November 2024

Abhik Rana

ABSTRACT

With a plethora of medical data available and the rise of Data Science, medical professionals are taking up the challenge of attempting to create indicators for the foreseen diseases that might be contracted, Cardiovascular diseases (CVDs) are the number 1 cause of death globally, taking an estimated 17.9 million lives each year, which accounts for 31% of all deaths worldwide. Heart failure is a common event caused by CVDs. People with cardiovascular disease or who are at high cardiovascular risk (due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease) need early detection and management wherein a statistical model for analysis of data collected can be of great help.

CONTENTS

Contents	vii
List of Figures	xi
1 Exploration of the data	1
1.1 Aim	1
1.2 Source of Data	1
1.3 Dataset Attributes	1
2 Analysis of the data	3
3 Critical Exploration of the Data	6
3.1 Distribution of the categorical features	7
3.2 Distribution of all the numerical features	8
3.3 Target visualization of heart diseases	9
3.4 Comparison of Categorical Features against heart failures	9
3.5 Categorical Features against Positive Heart Disease Cases	10
3.6 Numerical Features against Heart Diseases	11
3.7 Numerical features against Categorical features w.r.t Target variable . .	13
3.7.1 How does sex influence Numerical Features?	13
3.7.2 How does Chest Pain Type compare against Numerical features? .	14
3.7.3 How does FastingBS influence the numerical features?	14
3.7.4 How does RestingECG influence the numerical features?	15
3.7.5 How does ExcerciseAngina influence the numerical features? . .	16
3.7.6 How does ST_Slope influence the numerical features?	17

3.8	Influence of numerical features by the other numerical features w.r.t target variables	18
3.9	Summary of the observations	19
4	Inference	20
4.1	Correlation matrix	20
4.2	χ^2 -test for categorical features	21
4.3	ANOVA test for numerical features	22
Appendices		
A	Appendix A	28

LIST OF FIGURES

2.1	Head of the dataset.	3
2.2	Mean value of all the features present.	5
3.1	Distribution of all the categorical features.	7
3.2	Distribution of heart diseases according to gender	7
3.3	Distribution of all numerical features.	8
3.4	Distribution of oldpeak.	8
3.5	Visualization of percentage of diseased vs non-diseased candidates in the dataset.	9
3.6	Comparison of categorical features against diseased candidates	9
3.7	Comparison of categorical features against the positive cases of heart diseases.	10
3.8	Comparison of numerical features against the positive cases of heart diseases.	11
3.9	Comparison of numerical features against average positive cases of heart diseases.	12
3.10	Comparison of numerical features against gender.	13
3.11	Comparison of numerical features against gender.	13
3.12	Comparison of numerical features against chest pain type.	14
3.13	Comparison of numerical features against chest pain type.	14
3.14	Comparison of numerical features against FastingBS.	14
3.15	Comparison of numerical features against FastingBS.	15
3.16	Comparison of numerical features against RestingECG.	15
3.17	Comparison of numerical features against RestingECG.	15

3.18	Comparison of numerical features against ExerciseAngina.	16
3.19	Comparison of numerical features against ExerciseAngina.	16
3.20	Comparison of numerical features against ST_Slope.	17
3.21	Comparison of numerical features against ST_Slope.	17
3.22	Comparison of numerical features against other numerical features. . .	18
4.1	The correlation matrix between various features.	20
4.2	Correlation between features and heart disease.	21
4.3	χ^2 -scores for the categorical features.	21
4.4	ANOVA-scores for numerical features	22

EXPLORATION OF THE DATA

1.1 Aim

1. To classify / predict whether a patient is prone to heart failure depending on multiple attributes.
2. It is a **binary classification** with multiple numerical and categorical features.

1.2 Source of Data

This dataset was created by combining different datasets already available independently but not combined before. In this dataset, 5 heart datasets are combined over 11 common features which makes it the largest heart disease dataset available so far for research purposes. 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

Every dataset used can be found under the Index of heart disease datasets from UCI Machine Learning Repository.

1.3 Dataset Attributes

1. **Age** : age of the patient [years]
2. **Sex** : sex of the patient [M: Male, F: Female]

3. **ChestPainType** : chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. **RestingBP** : resting blood pressure [mm Hg]
5. **Cholesterol** : serum cholesterol [mm/dl]
6. **FastingBS** : fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. **RestingECG** : resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. **MaxHR** : maximum heart rate achieved [Numeric value between 60 and 202]
9. **ExerciseAngina** : exercise-induced angina [Y: Yes, N: No]
10. **Oldpeak** : oldpeak = ST [Numeric value measured in depression]
11. **ST_Slope** : the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. **HeartDisease** : output class [1: heart disease, 0: Normal]

ANALYSIS OF THE DATA

A general idea of the data may be obtained:

```
1 data = pd.read_csv('heart.csv')
2 data.head()
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	M	ATA	140	289	0	Normal	172	N	0.00	Up	0
1	49	F	NAP	160	180	0	Normal	156	N	1.00	Flat	1
2	37	M	ATA	130	283	0	ST	98	N	0.00	Up	0
3	48	F	ASY	138	214	0	Normal	108	Y	1.50	Flat	1
4	54	M	NAP	150	195	0	Normal	122	N	0.00	Up	0

Figure 2.1: *Head of the dataset.*

The data has a dimension of 918 rows with 12 columns. The column names are respectively,

```
1 data.columns
```

```
1 Index(['Age', 'Sex', 'ChestPainType', 'RestingBP', 'Cholesterol', 'FastingBS',
2       'RestingECG', 'MaxHR', 'ExerciseAngina', 'Oldpeak', 'ST_Slope',
3       'HeartDisease'],
4       dtype='object')
```

Now we check for possible gap in the data if there is any,

```
1 data.info()
```

```

1 <class 'pandas.core.frame.DataFrame'>
2 RangeIndex: 918 entries, 0 to 917
3 Data columns (total 12 columns):
4  #   Column                Non-Null Count  Dtype
5  ---  ---
6  0   Age                   918 non-null    int64
7  1   Sex                   918 non-null    object
8  2   ChestPainType         918 non-null    object
9  3   RestingBP             918 non-null    int64
10  4   Cholesterol            918 non-null    int64
11  5   FastingBS             918 non-null    int64
12  6   RestingECG            918 non-null    object
13  7   MaxHR                 918 non-null    int64
14  8   ExerciseAngina        918 non-null    object
15  9   Oldpeak               918 non-null    float64
16  10  ST_Slope              918 non-null    object
17  11  HeartDisease          918 non-null    int64
18 dtypes: float64(1), int64(6), object(5)
19 memory usage: 86.2+ KB

```

Looks like there is no null elements in the dataset. Now we look at the summary of the data in the given dataset.

```

1 data.describe().T

```

	count	mean	std	min	25%	50%	75%	max
Age	918.00	53.51	9.43	28.00	47.00	54.00	60.00	77.00
RestingBP	918.00	132.40	18.51	0.00	120.00	130.00	140.00	200.00
Cholesterol	918.00	198.80	109.38	0.00	173.25	223.00	267.00	603.00
FastingBS	918.00	0.23	0.42	0.00	0.00	0.00	0.00	1.00
MaxHR	918.00	136.81	25.46	60.00	120.00	138.00	156.00	202.00
Oldpeak	918.00	0.89	1.07	-2.60	0.00	0.60	1.50	6.20
HeartDisease	918.00	0.55	0.50	0.00	0.00	1.00	1.00	1.00

For experimental reason we look at the mean value of all the features in the cases of heart disease and non-heart disease.

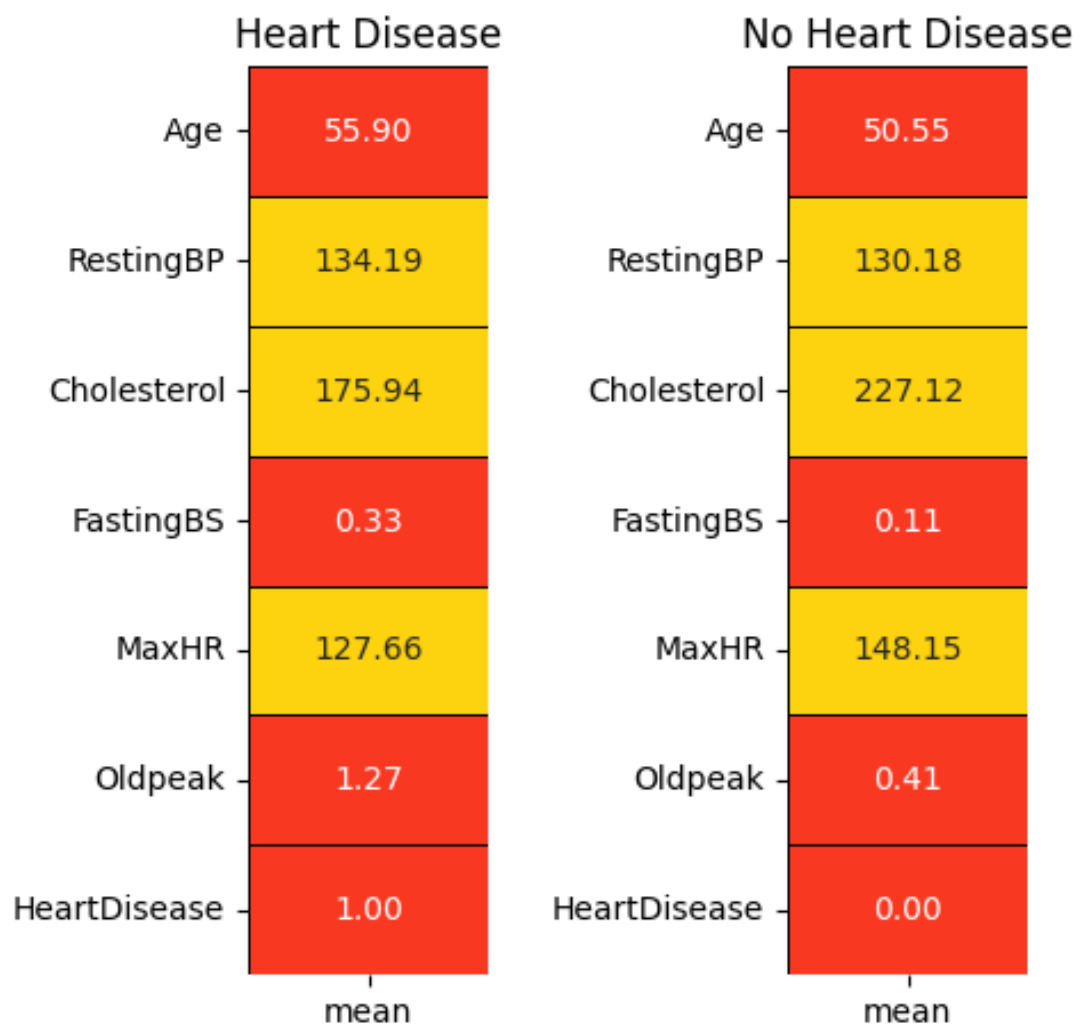


Figure 2.2: Mean value of all the features present.

CRITICAL EXPLORATION OF THE DATA

For exploring the data properly, we divide the data into two groups: one being categorical and the other being numerical.

- Here, categorical features are defined if the attribute has less than 6 unique elements else it is a numerical feature.
- Typical approach for this division of features can also be based on the datatypes of the elements of the respective attribute.

Eg : datatype = integer, attribute = numerical feature ; datatype = string, attribute = categorical feature

- For this dataset, as the number of features are less, we can manually check the dataset as well.

After analyzing the data-types completely, we come to the conclusion:

-
- 1 Categorical Features : Sex ChestPainType FastingBS RestingECG ExerciseAngina ST_Slope
↔ HeartDisease
 - 2 Numerical Features : Age RestingBP Cholesterol MaxHR Oldpeak
-

3.1 Distribution of the categorical features

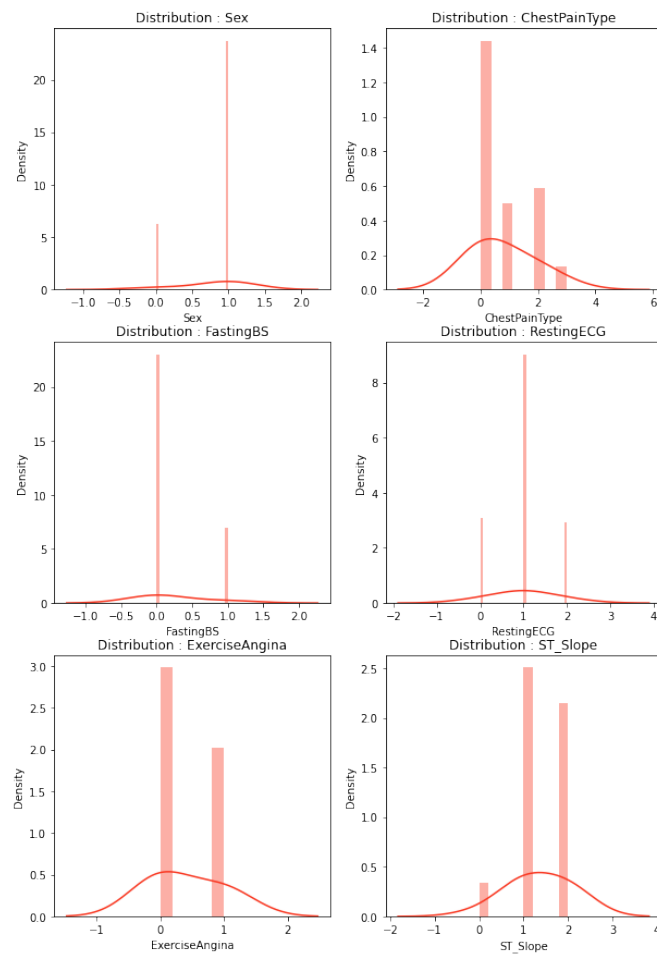


Figure 3.1: Distribution of all the categorical features.

And the distribution of heart diseases according to gender is:

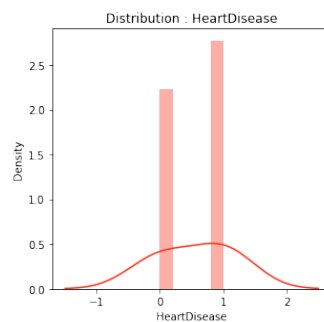


Figure 3.2: Distribution of heart diseases according to gender

We deduce that all the categorical features are nearly **normally distributed**.

3.2 Distribution of all the numerical features

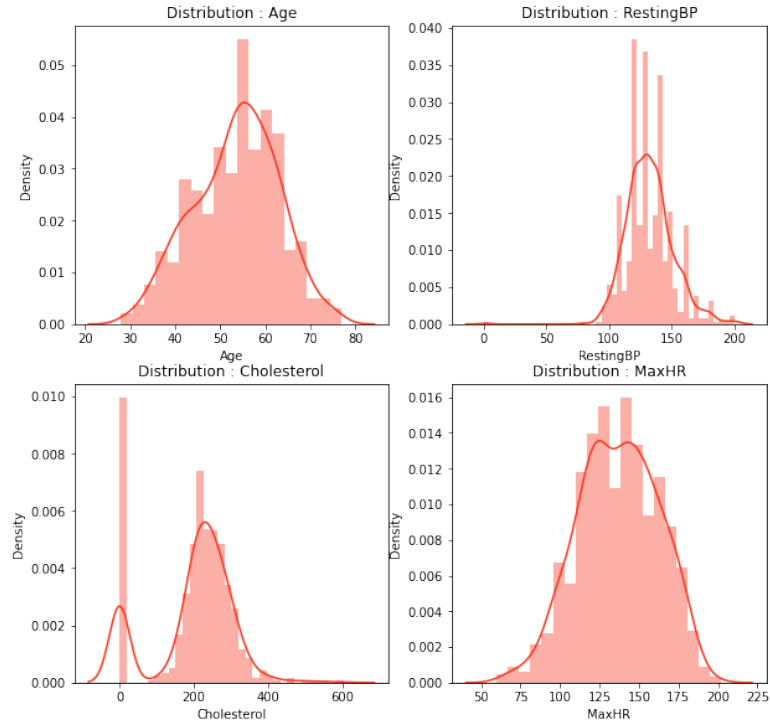


Figure 3.3: *Distribution of all numerical features.*

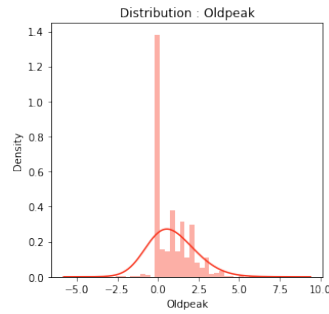


Figure 3.4: *Distribution of oldpeak.*

We deduce that

- **Oldpeak's** data distribution is rightly skewed.
- **Cholestrol** has a bi-modal data distribution.

3.3 Target visualization of heart diseases

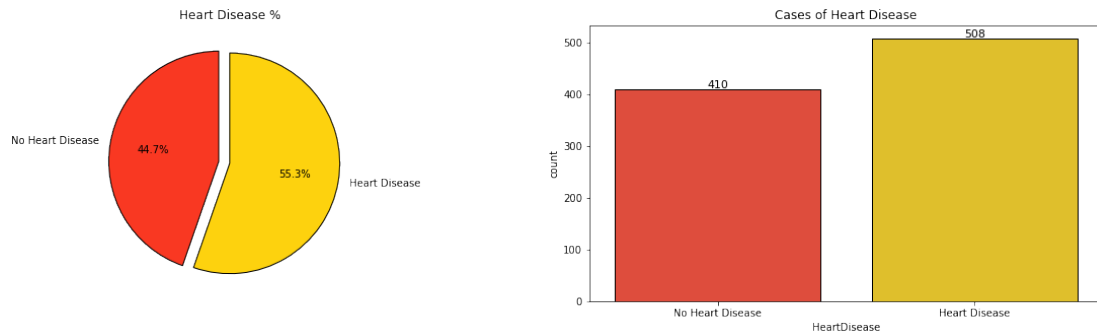


Figure 3.5: Visualization of percentage of diseased vs non-diseased candidates in the dataset.

We find that the data-set is pretty much **evenly balanced**.

3.4 Comparison of Categorical Features against heart failures

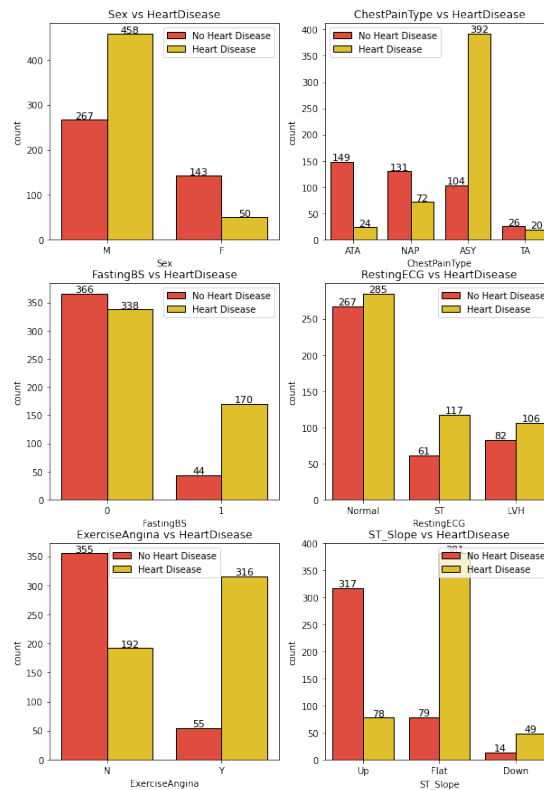


Figure 3.6: Comparison of categorical features against diseased candidates

Conclusions

- **Male** population has more heart disease patients than no heart disease patients. In the case of **Female** population, heart disease patients are less than no heart disease patients.
- **ASY** type of chest pain boldly points towards major chances of heart disease.
- **Fasting Blood Sugar** is tricky! Patients diagnosed with Fasting Blood Sugar and no Fasting Blood Sugar have significant heart disease patients.
- **RestingECG** does not present with a clear cut category that highlights heart disease patients. All the 3 values consist of high number of heart disease patients.
- **Exercise Induced Angina** definitely bumps the probability of being diagnosed with heart diseases.
- With the **ST_Slope** values, **flat** slope displays a very high probability of being diagnosed with heart disease. **Down** also shows the same output but in very few data points.

3.5 Categorical Features against Positive Heart Disease Cases

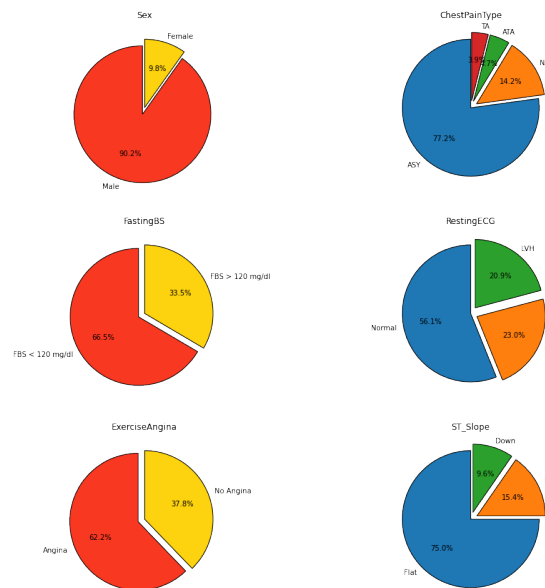


Figure 3.7: Comparison of categorical features against the positive cases of heart diseases.

Conclusion

- Out of all the heart disease patients, a staggering 90% patients are **male**.
- When it comes to the type of chest pain, **ASY** type holds the majority with 77% that lead to heart diseases.
- **Fasting Blood Sugar** level < 120 mg/dl displays high chances of heart diseases.
- For **RestingECG**, **Normal** level accounts for 56% chances of heart diseases than **LVH** and **ST** levels.
- Detection of **Exercise Induced Angina** also points towards heart diseases.
- When it comes to **ST_Slope** readings, **Flat** level holds a massive chunk with 75% that may assist in detecting underlying heart problems.

3.6 Numerical Features against Heart Diseases

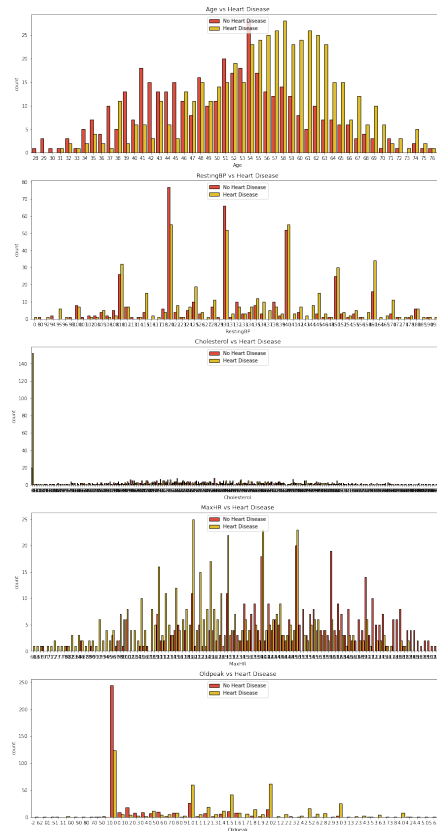


Figure 3.8: Comparison of numerical features against the positive cases of heart diseases.

Because of too many unique data points in the above features, it is difficult to gain any type of insight. Thus, we will convert these numerical features, except age, into

categorical features for understandable visualization and gaining insights purposes. We scale the individual values of these features. This brings the varied data points to a constant value that represents a range of values. Here, we divide the data points of the numerical features by 5 or 10 and assign its quotient value as the representative constant for that data point. The scaling constants of 5 and 10 are decided by looking into the data and intuition.

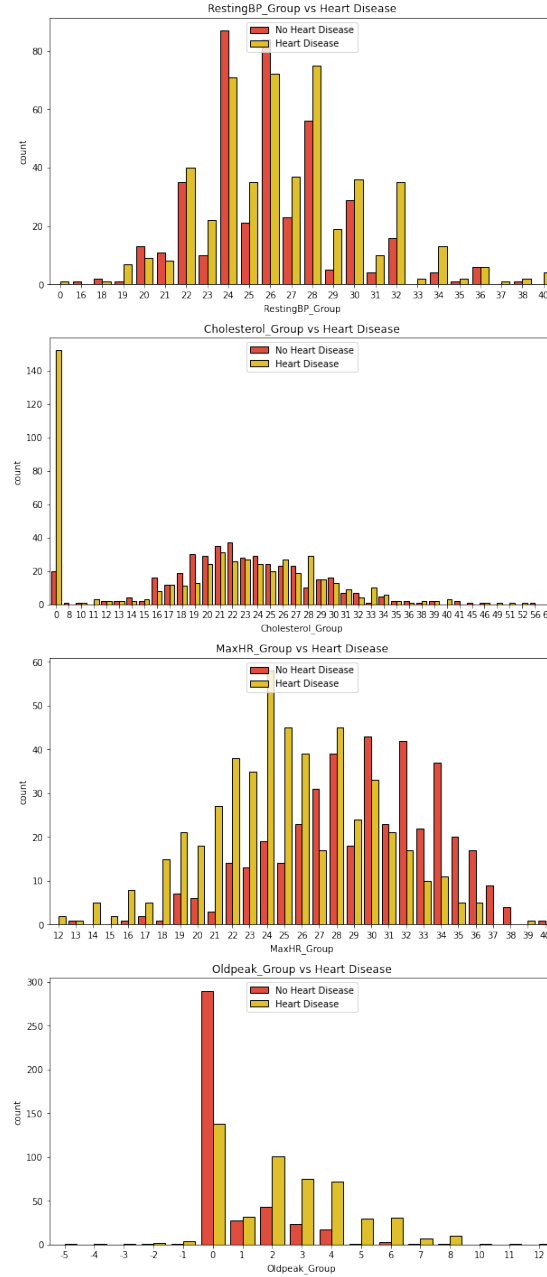


Figure 3.9: Comparison of numerical features against average positive cases of heart diseases.

Observations

- From the **RestingBP** group data, **95** (19x5) - **170** (34x5) readings are most prone to be detected with heart diseases.
- **Cholesterol** levels between **160** (16x10) - **340** (34x10) are highly susceptible to heart diseases.
- For the **MaxHR** readings, heart diseases are found throughout the data but **70** (14x5) - **180** (36x5) values has detected many cases.
- **Oldpeak** values also display heart diseases throughout. **0** (0x5/10) - **4** (8x5/10) slope values display high probability to be diagnosed with heart diseases.

3.7 Numerical features against Categorical features w.r.t Target variable

3.7.1 How does sex influence Numerical Features?



Figure 3.10: Comparison of numerical features against gender.

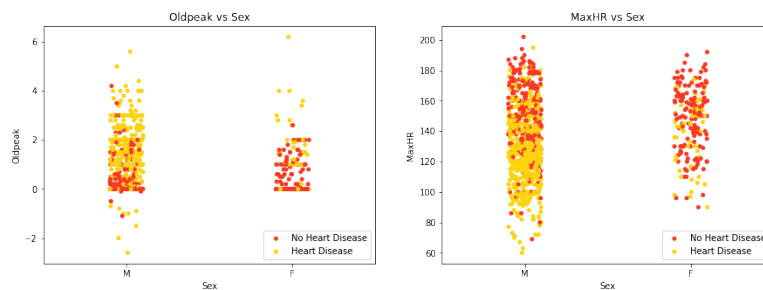


Figure 3.11: Comparison of numerical features against gender.

- **Male** population displays heart diseases at near about all the values of the numerical features. Above the age of 50, positive old peak values and maximum

heart rate below 140, heart diseases in male population become dense.

- **Female** population data points are very less as compared to **male** population data points. Hence, we cannot point to specific ranges or values that display cases of heart diseases.

3.7.2 How does Chest Pain Type compare against Numerical features?

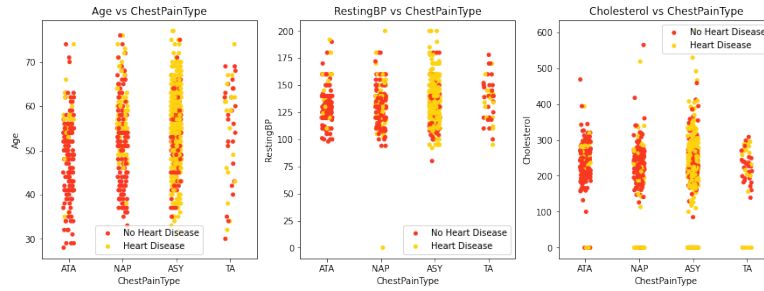


Figure 3.12: Comparison of numerical features against chest pain type.

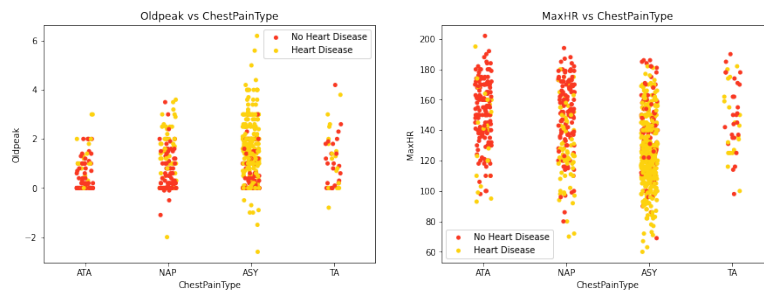


Figure 3.13: Comparison of numerical features against chest pain type.

ASY type of chest pain dominates other types of chest pain in all the numerical features by a lot.

3.7.3 How does FastingBS influence the numerical features?

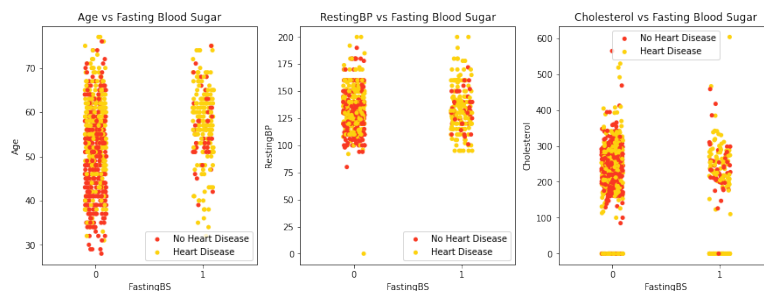


Figure 3.14: Comparison of numerical features against FastingBS.

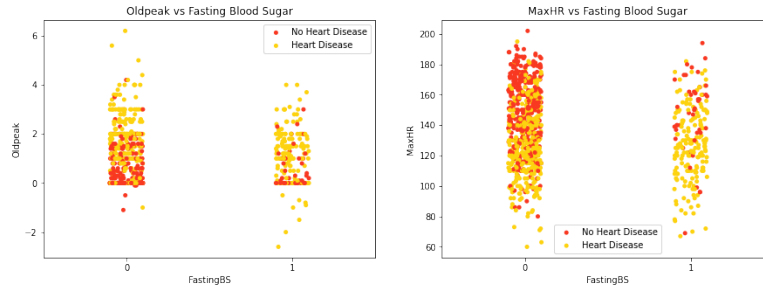


Figure 3.15: Comparison of numerical features against FastingBS.

- Above the **age 50**, heart diseases are found throughout the data irrespective of the patient being diagnosed with Fasting Blood Sugar or not.
- **Fasting Blood Sugar** with **Resting BP** over 100 has displayed more cases of heart diseases than patients with no fasting blood sugar.
- **Cholesterol** with **Fasting Blood Sugar** does not seem to have an effect in understanding reason behind heart diseases.
- Patients that have not been found positive with **Fasting Blood Sugar** but have maximum heart rate below 130 are more prone to heart diseases.

3.7.4 How does RestingECG influence the numerical features?

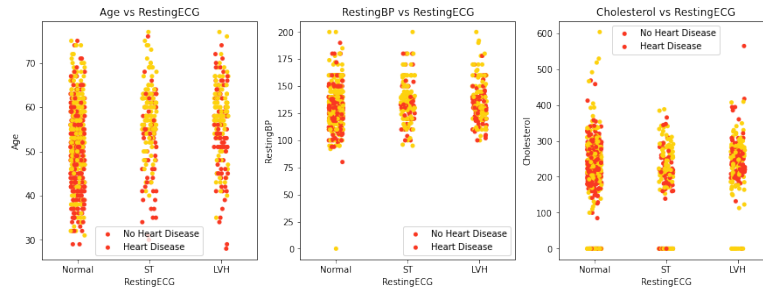


Figure 3.16: Comparison of numerical features against RestingECG.

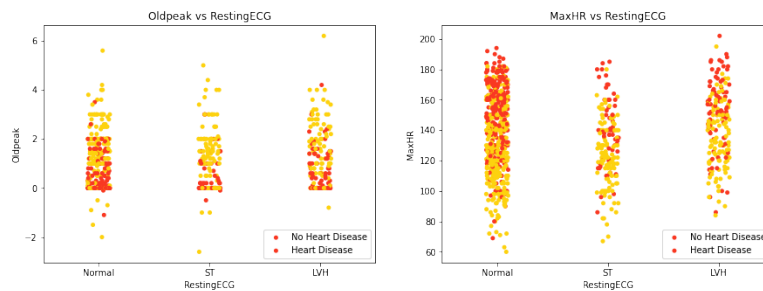


Figure 3.17: Comparison of numerical features against RestingECG.

- Heart diseases with **RestingECG** values of **Normal**, **ST** and **LVH** are detected starting from 30,40 and 40 respectively. Patients above the age of 50 are more prone than any other ages irrespective of **RestingECG** values.
- Heart diseases are found consistently throughout any values of **RestingBP** and **RestingECG**.
- **Cholesterol** values between 200 - 300 coupled with **ST** value of **RestingECG** display a patch of patients suffering from heart diseases.
- For **maximum Heart Rate** values, heart diseases are detected in dense below 140 points and **Normal** RestingECG. **ST** and **LVH** throughout the maximum heart rate values display heart disease cases.

3.7.5 How does ExerciseAngina influence the numerical features?

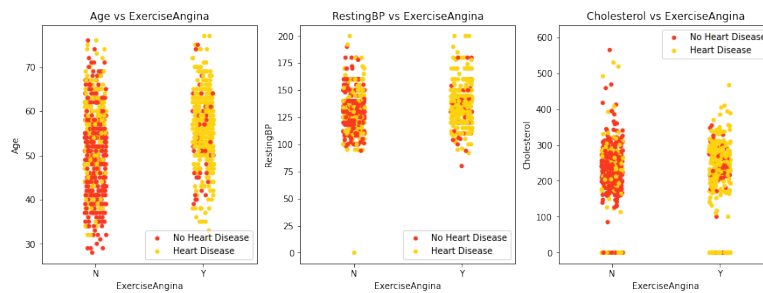


Figure 3.18: Comparison of numerical features against ExerciseAngina.

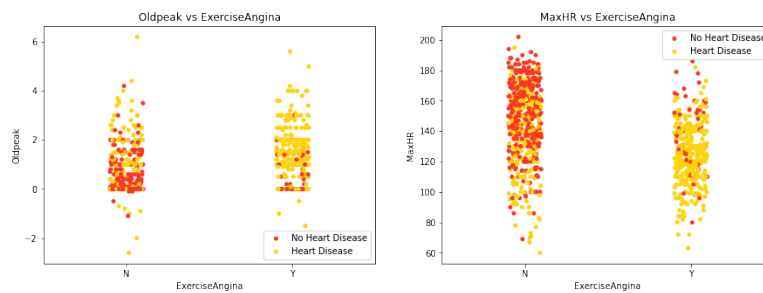


Figure 3.19: Comparison of numerical features against ExerciseAngina.

A crystal clear observation can be made about the relationship between **heart disease** case and **Exercise induced Angina**. A positive correlation between the 2 features can be concluded throughout all the numerical features.

3.7.6 How does ST_Slope influence the numerical features?

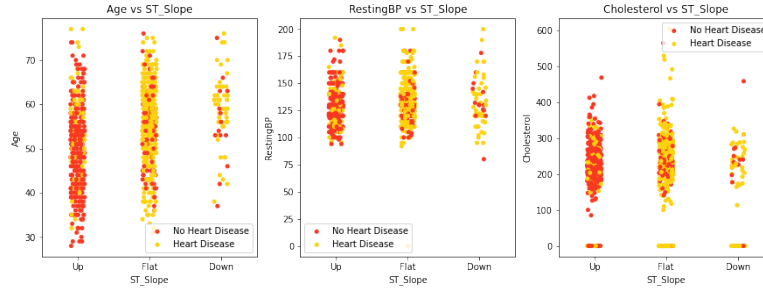


Figure 3.20: Comparison of numerical features against ST_Slope.

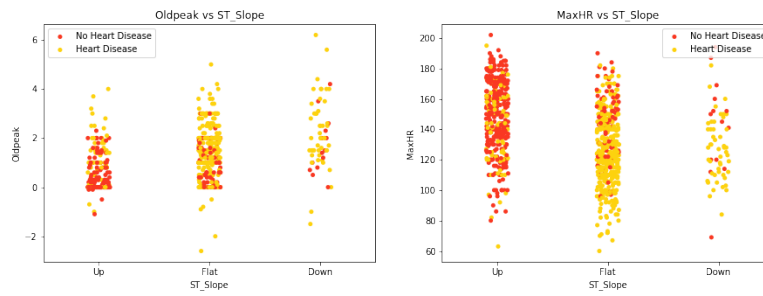


Figure 3.21: Comparison of numerical features against ST_Slope.

- Another crystal clear positive observation can be made about the positive correlation between **ST_Slope** value and **Heart Disease** cases.
- **Flat, Down** and **Up** in that order display high, middle and low probability of being diagnosed with heart diseases respectively.

3.8 Influence of numerical features by the other numerical features w.r.t target variables

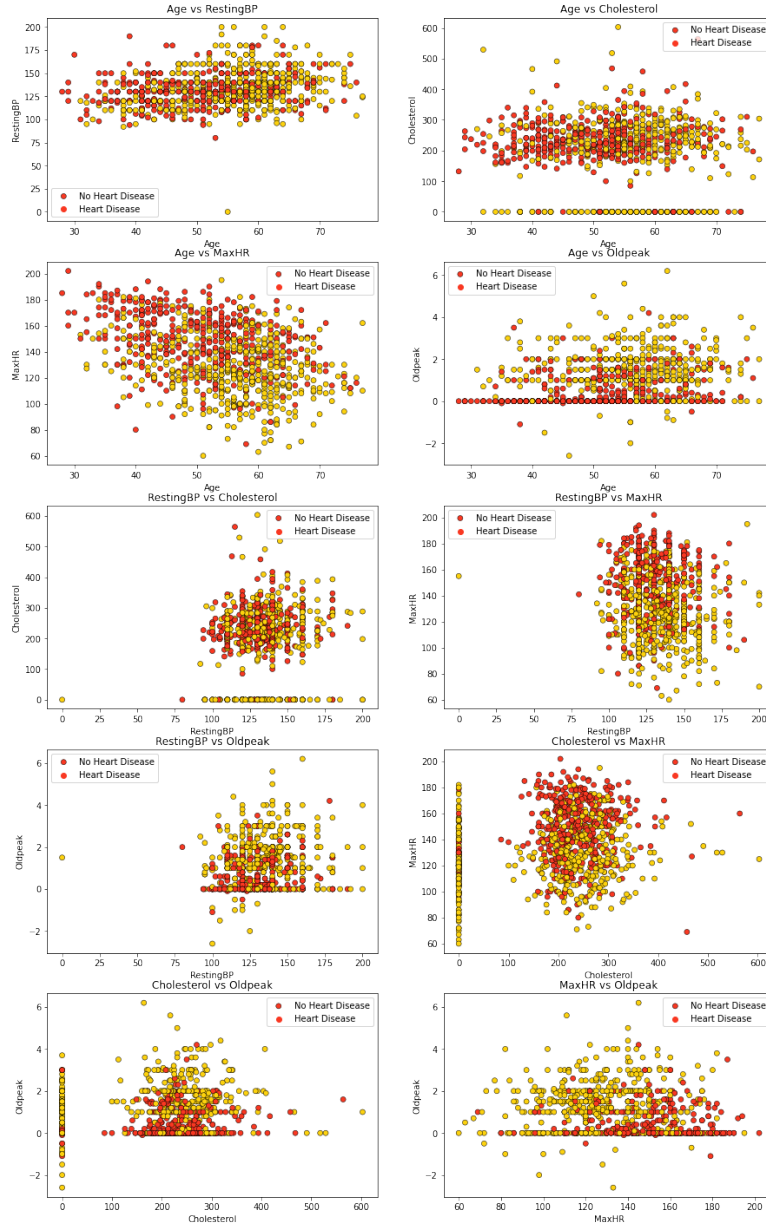


Figure 3.22: Comparison of numerical features against other numerical features.

Conclusions

- - For age 50+, RestingBP* between 100 - 175, Cholesterol level of 200 - 300, Max Heart Rate* below 160 and positive oldpeak values displays high cases of heart disease.

- For **RestingBP** values 100 - 175, highlights too many heart disease patients for all the features.
- **Cholesterol** values 200 - 300 dominates the heart disease cases.
- Similarly, **Max Heart Rate** values below 140 has high probability of being diagnosed with heart diseases.

3.9 Summary of the observations

Now with the help of the above observations we may find the order of increasing risk of heart failures due to the studied features:

- **Categorical features:**
 1. Sex : Male > Female
 2. ChestPainType : ASY > NAP > ATA > TA
 3. FastingBS : (FBS < 120 mg/dl) > (FBS > 120 mg/dl)
 4. RestingECG : Normal > ST > LVH
 5. ExerciseAngina : Angina > No Angina
 6. ST_Slope : Flat > Up > Down
- **Numerical features:**
 1. Age : 50+
 2. RestingBP : 95 - 170
 3. Cholesterol : 160 - 340
 4. MaxHR : 70 - 180
 5. Oldpeak : 0 - 4

Now that we have understood the typical values of the features, we may move on to the next step where we select the appropriate features for modeling.

INFERENCE

Firstly we want to study the correlation between various features we have studied above. But before that we need to do data scaling. It may be carried out in 2 options : 1) **Normalization** 2) **Standardization**. As most of the algorithms assume the data to be normally (Gaussian) distributed, **Normalization** is done for features whose data does not display normal distribution and **standardization** is carried out for features that are normally distributed where their values are huge or very small as compared to other features.

- **Oldpeak** feature is normalized as it had displayed a right skewed data distribution.
- **Age, RestingBP, Cholesterol** and **MaxHR** features are scaled down because these features are normally distributed.

4.1 Correlation matrix



Figure 4.1: The correlation matrix between various features.

As we may observe, it is a huge matrix with too many features. We only want to look at correlation against heart diseases.

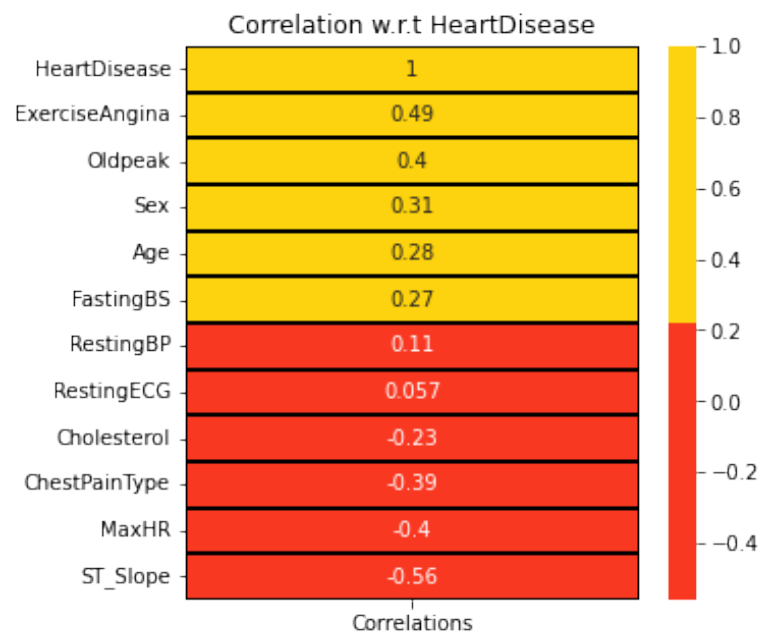


Figure 4.2: Correlation between features and heart disease.

Except for **RestingBP** and **RestingECG**, everyone displays a positive or negative relationship with **HeartDisease**.

4.2 χ^2 -test for categorical features

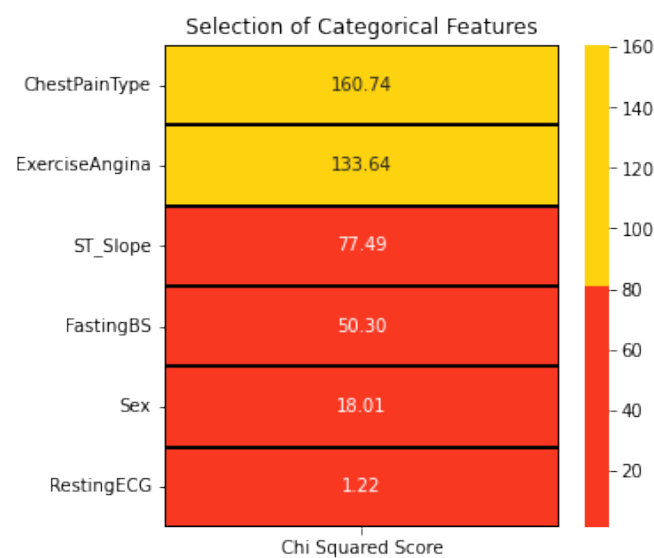


Figure 4.3: χ^2 -scores for the categorical features.

Except **RestingECG**, all the remaining categorical features are pretty important for predicting heart diseases.

4.3 ANOVA test for numerical features

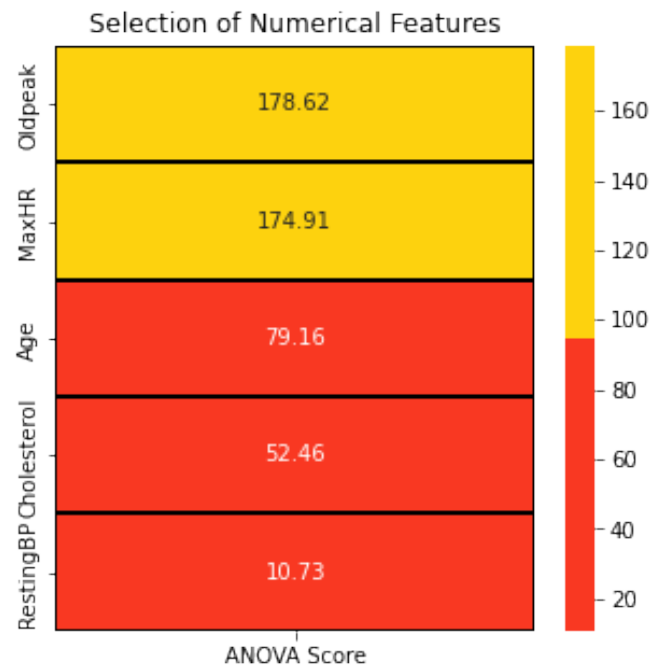



Figure 4.4: ANOVA-scores for numerical features

RestingBP doesn't seem as influential as the remaining features.

APPENDICES

APPENDIX A

All the codes used in the text can be found in my GitHub repository named heart-failure-analysis .

