2023

2023

Exploratory Data Analysis on UCI Heart Disease Dataset

Assignment - 1

Tharindu Gunasekera

Table of Contents

Introduction

Objectives

The objective of the assignment is to explore, analyze and model a heart-disease dataset. Through data exploration, analysis and modelling we aim to gain insights into the data set, identify trends and patterns as well as determine how suitable logistic regressions is as a predictive model for heart disease. We aim to gain deeper understanding on how to handle data exploration and analysis while showing the potential of this process for predicting heart disease.

About the Dataset

The data set is used is one of four available data sets in the UCI Machine Learning Repository. The dataset used is specifically the processed version of the Cleveland Clinic Foundation's Heart Disease dataset. This dataset is used as it is the most commonly used out of the available four, owing to the completeness of its data. The processed version of the data is different from the original in the following ways.

- Only 14 out of the 76 variables present were selected for analysis
- The processed data replaces the -9 value in the data set which was used to represent missing data, with '?'
- The processed data has all 303 records available without corruption. Some part of the data in the original is corrupted, this leads to a lesser number of records being easily identifiable in it.

Data Exploration

Dealing with Invalid Values

After initial observations of the data. It was noticed that two categorical variables 'ca' and 'thal' which were supposed to have 4 and 3 categories respectively, had 5 and 4 categories. This was due to the presence of invalid values denoted by '?'. Further there was the potential that invalid values could also exist in fields of numerical variables.

To deal with this, we detected all records which had a missing value and eliminated them from the data set. This resulted in the categorical variables getting the appropriate possible unique values and decreased the number of records from the initial 302 to 296.

Applying Types

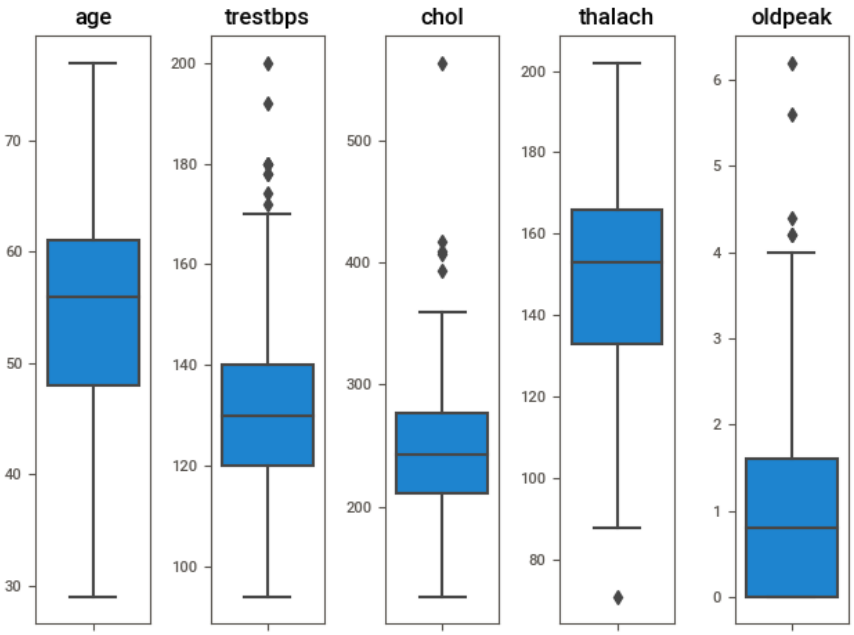
After invalid values were eliminated, appropriate typing was applied to each column of the data set. In here an additional consideration was taken due to the objective of this assignment. As the requirement was to detect the presence of heart disease but not the specifics regarding its scale, the categorical variable 'num' which was used to show the presence of heart disease was made to take binary values between 0 (present) and 1 (not present). Initially it had values 0-4 with increase signifying the severity of heart disease.

Transformation and Normalization

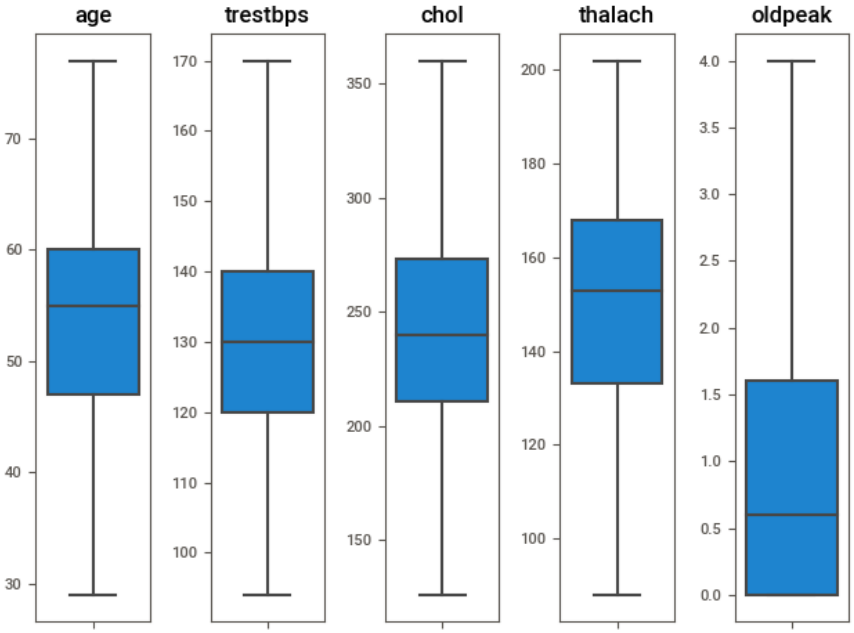
As the purpose of this analysis was not to create a machine learning model but rather to analyze the data and come to conclusions regarding, normalizations and transformations were avoided. These would distort the data so that relationships between specific features may not be initially clear as their values

would change. Further this would also have the effect of affecting the distribution of data and therefore the inferences we gain from analyzing it.

Dealing with Outliers



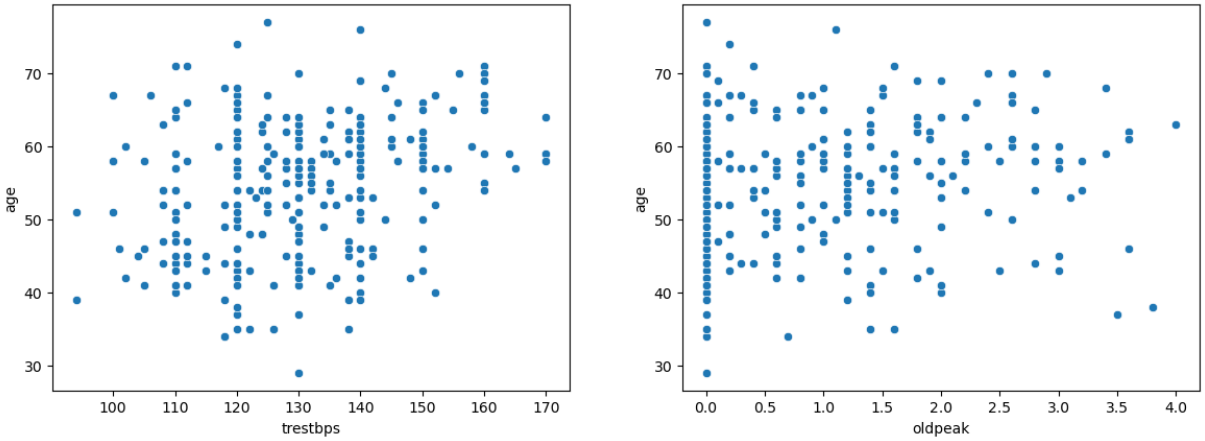
Outliers were identified on a per column basis by analyzing box plots made for each individual feature. These outliers were removed as they would affect the distribution of the data unequally. Though there are other ways to deal with outliers, we decided to remove them as other methods may affect the distribution of data. Multivariate outliers however, were ignored. This process reduces the observations in our data set from 296 to 277.



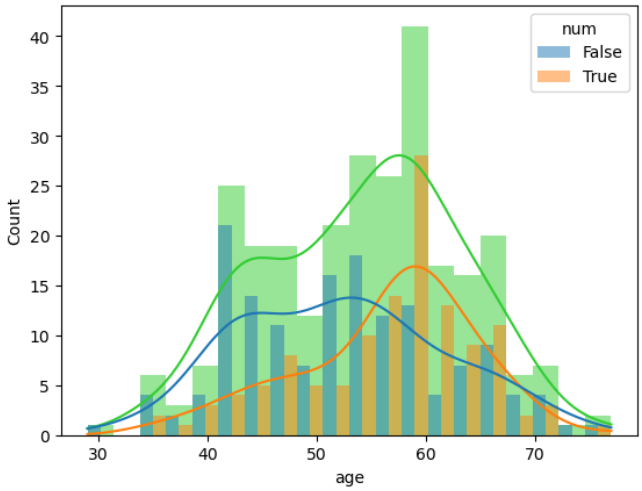
Data Analysis

Observations on continuous variables

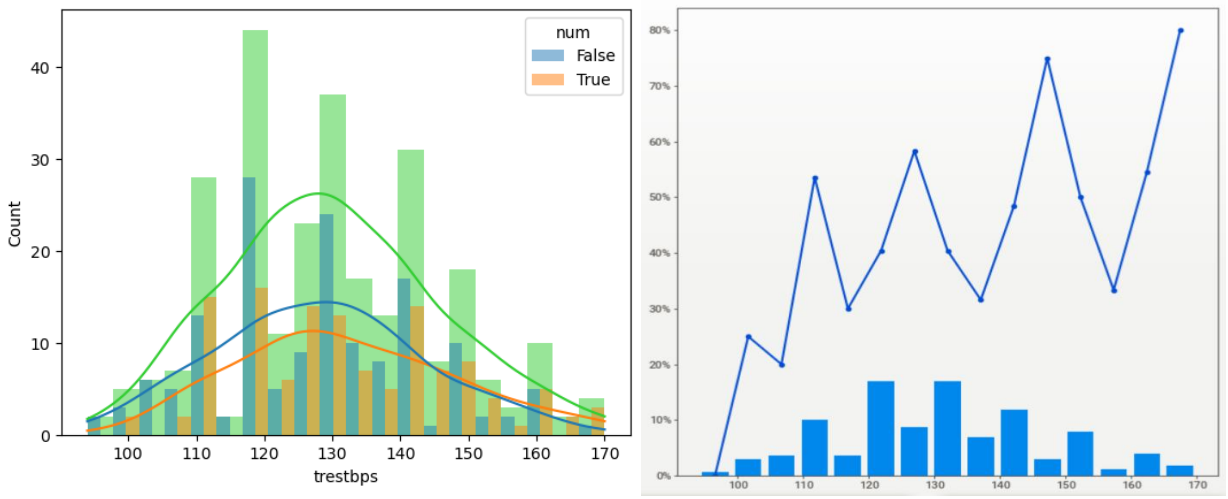
There were no significant observations to be made by drawing scatter plots for independent continuous variables. No relationships or extreme outliers to be noted. Following are two sample scatter plots. Showing.



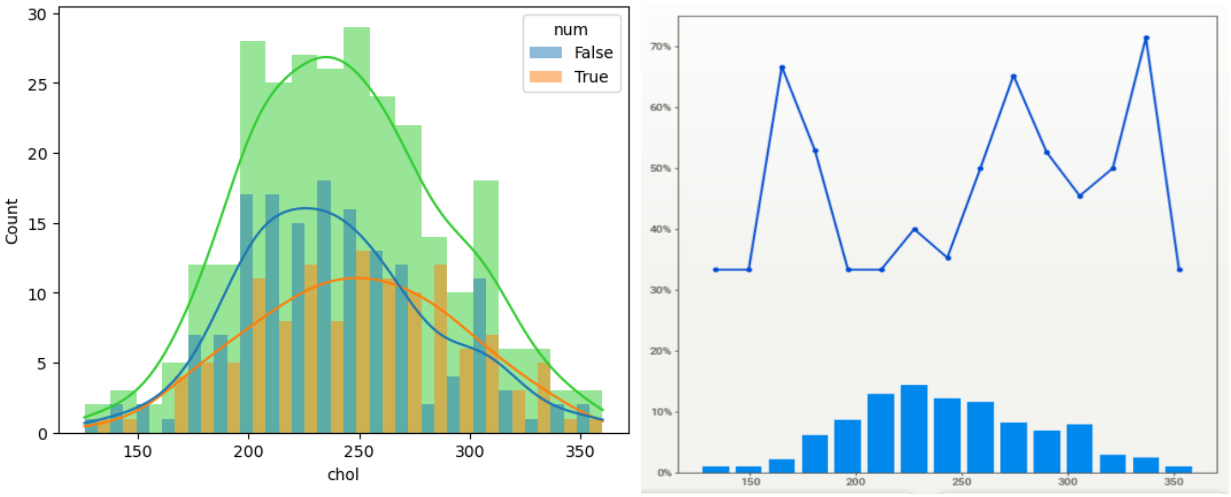
When histograms are observed, age seems to have a bimodal distribution with a slight negative skew. Further there is an increase in heart disease for patients in their 60s.



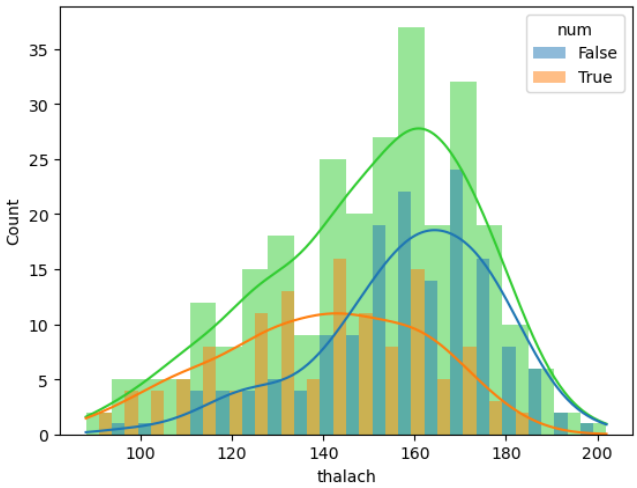
There seems to be a relationship between increase trestbps and increasing heart disease. Though there are dips in probability with a pattern at values such as 120, 140. This means that increasing blood pressure was cause to suspect presence of heart disease.



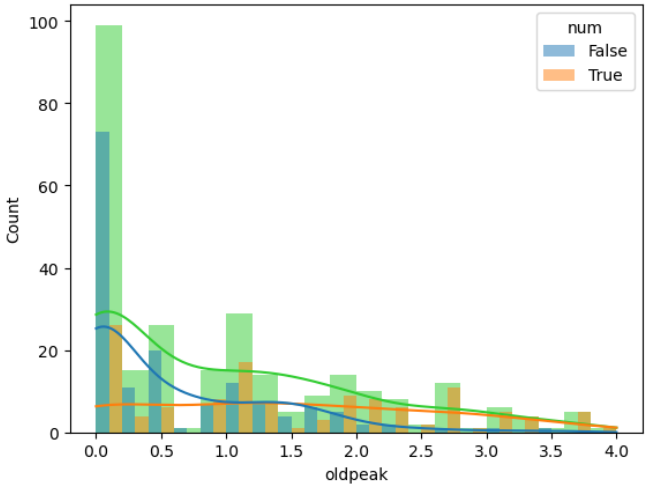
Increasing chol is generally cause for concern. Notably there are spikes in presence of heart disease when chol is higher than 250 or in the range 150-200. Higher cholesterol values have are frequently linked to presence of heart disease and this dataset seems to follow the same pattern.

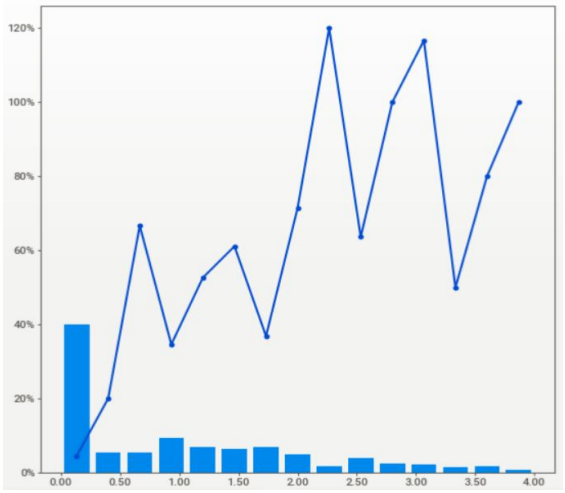


Counterintuitively, there seems to be a strongly visible relationship between lower thalach values and the presence of heart disease. This means that lower resting heart rates can be linked to the presence of heart disease, a point that may be deserving of further inspection.



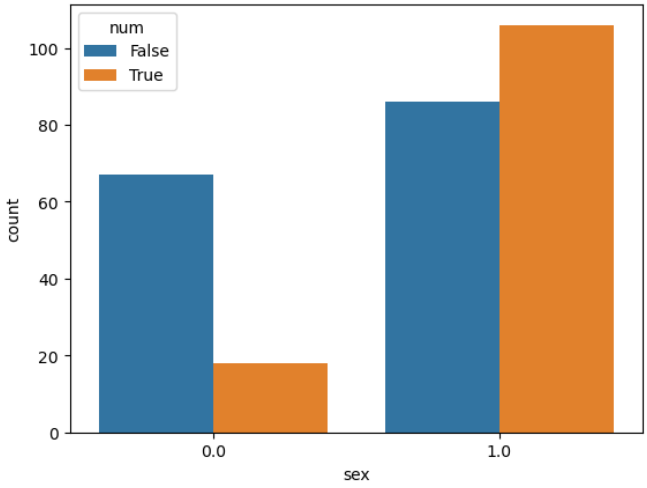
oldpeak seem to have a strongly positively skewed distribution. Further, higher oldpeak values seem to indicate presence of heart disease with probability being lowest when oldpeak is 0. Higher ST depression correlates with presence of heart disease.



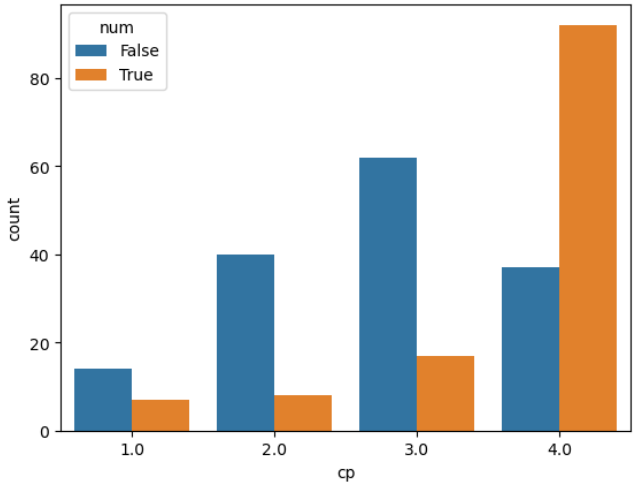


Observations on categorical variables

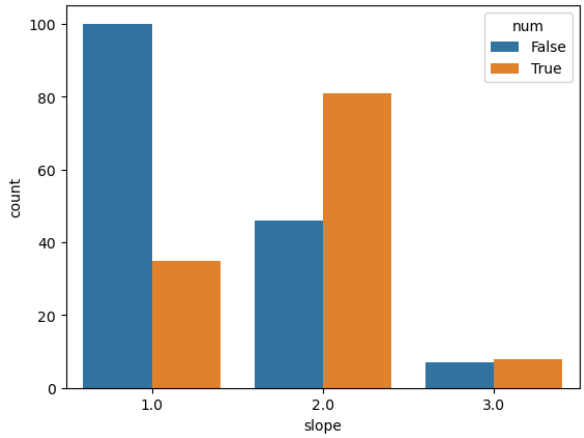
Higher likelihood over heart disease overall in males over females. Total dataset also contains higher number of males over females observed at 192 and 85 respectively.

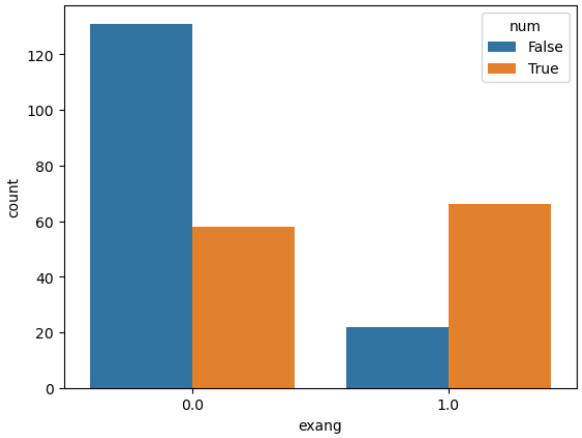


There seems to be a spike in heart disease in cp is 1 i.e., patient is experiencing typical anginal pain. There is an even higher spike when cp is 4 i.e., patient is asymptomatic. Risk of heart disease seems to be lowest when angina is atypical in nature.

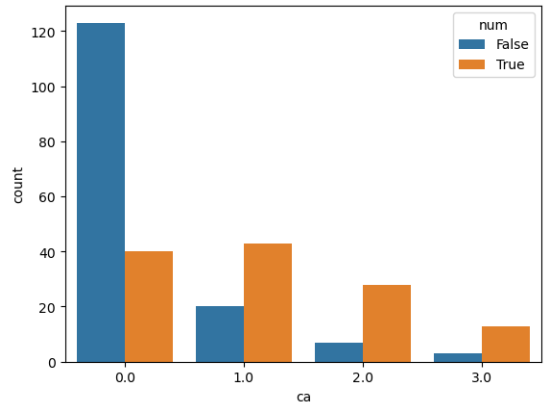


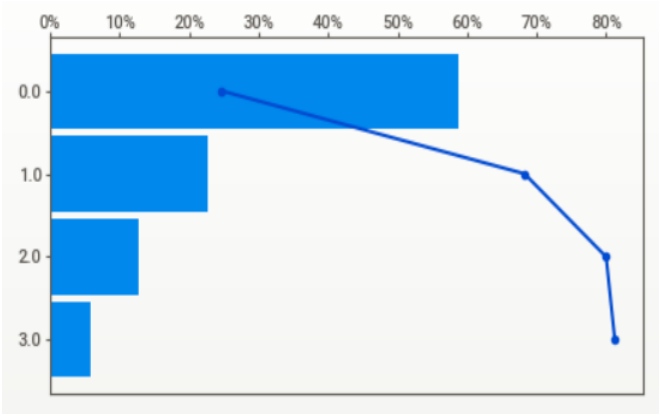
Patients with an exang value of 1 are notably at risk for heart disease. This is true for patients with a slope value of 2 as well. This means that exercise induced angina as well as downsloping of the peak exercise ST segment are both cause for concern.



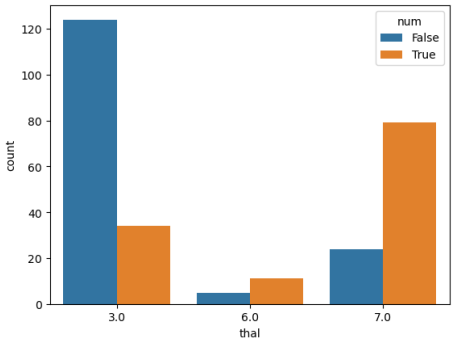


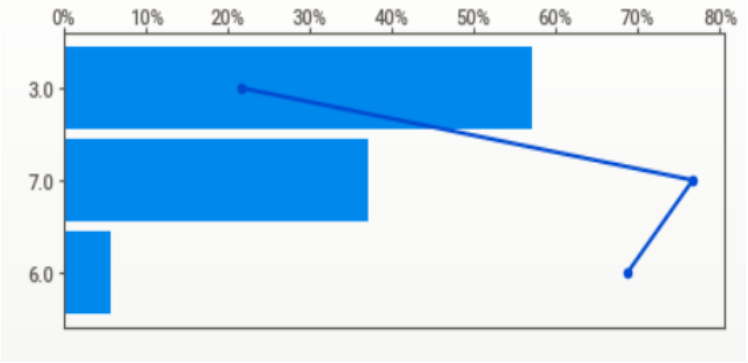
There seems to be a general trend of increasing presence of heart disease with increase in ca value. That is a higher number of vessels colored by fluoroscopy correlates with presence of heart disease.





Similarly there is a general trend for increasing thal values signifying increase in heart disease.





Data Modelling

Suitability for Logistics Regression

When determining whether logistics regression is a suitable model to predict heart disease a few considerations and assumptions need to be made. These assumptions are;

- Outcomes are of appropriate type, they should not be continuous values
- Linearity of independent variables
- No strongly influential outliers present in the data set
- Absence of multi-collinearity across considered independent variables
- Independence of each individual observation or record
- Analyzed data set is of a sufficiently large data set.

The dependent variable in this circumstance is 'num'. Initially num was an ordinal discrete value which could takes values 0-4. At this point, ordinal logistics regressions might have been suitable to predict heart disease data. As the variables was changed to only signify presence or absence of heart disease, binary logistics regression should be used. Therefore, num is suitable as a target variable in logistics regression.

Linearity of independent variables when related to their specific log can be found by analyzing the below figure. These are the results of a Box-Tidwell test.

Generalized Linear Model Regression Results

Dep. Variable:	num	No. Observations:	277
Model:	GLM	Df Residuals:	268
Model Family:	Binomial	Df Model:	8
Link Function:	Logit	Scale:	1.0000
Method:	IRLS	Log-Likelihood:	-158.36
Date:	Fri, 02 Jun 2023	Deviance:	316.72
Time:	09:26:02	Pearson chi2:	277.
No. Iterations:	5	Pseudo R-squ. (CS):	0.2070
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
age	1.0824	0.836	1.294	0.196	-0.557	2.722
trestbps	0.5139	0.684	0.751	0.453	-0.827	1.855
chol	-0.0020	0.164	-0.012	0.990	-0.324	0.320
thalach	0.4286	0.469	0.914	0.361	-0.491	1.348
age:Log_age	-0.2157	0.168	-1.287	0.198	-0.544	0.113
trestbps:Log_trestbps	-0.0852	0.116	-0.732	0.464	-0.313	0.143
chol:Log_chol	0.0010	0.025	0.040	0.968	-0.049	0.051
thalach:Log_thalach	-0.0794	0.079	-1.007	0.314	-0.234	0.075
const	-30.2689	21.892	-1.383	0.167	-73.176	12.638

It we consider p larger than 0.05 to be statistically significant, all values pass and therefore seem to be linearly related. 'oldpeak' was intentionally excluded from this test despite being a continuous variable due being able to have 0 values. Therefore, this assumption is valid.

During data exploration we found and removed outliers on the basis of their visibility on box plots. Assuming that the removal of outliers marked in the box-plots is enough, there should be no strongly influential outliers present in the data set. To verify this further we can find Cook's distance for each individual data point.

Collinearity across independent variables can be analyzed by considering the Variance Inflation Factor for each continuous variable in our data set. Below are the values that are found when this is done.

	variables	VIF
0	age	37.670899
1	trestbps	64.962379
2	chol	29.609890
3	thalach	30.784962
4	oldpeak	2.104720

When these results are analyzed, while oldpeak seems to have a moderate VIF value. All 4 other variables have a VIF value larger than 10, showing very high correlation. Due to this observation, this assumption is not satisfied.

Independence of each record is automatically assumed based on how the data was collected. As the data is assumed to have been collected regarding distinct patients in a manner that was independent of each other. This assumption is satisfied.

There are a few ways in which we can analyze whether dataset is of sufficiently large size. The first method is to see whether the least frequent outcome for a given independent variables appears at least 10 times. In this data set, while this is true for categorical variables, it is not true for the continuous variables. Another method is to simply say that there should be a minimum of 500 observations for a data set to be sufficiently large, this also fails as the data set only has 277 observations in its current state. Therefore, this assumption fails.