Report

Data exploration

1. Pandas profiler (please find the result as output.html in repo) - This tool automatically explores the dataframe and creates a detailed report of the dataset.

2. Capturing unclean data - noticed that there were ‘?’ values in a couple of features (ca and thal).

3. Cleaning -since there were only a small number of unclean data, I decided to remove the rows which contained those values

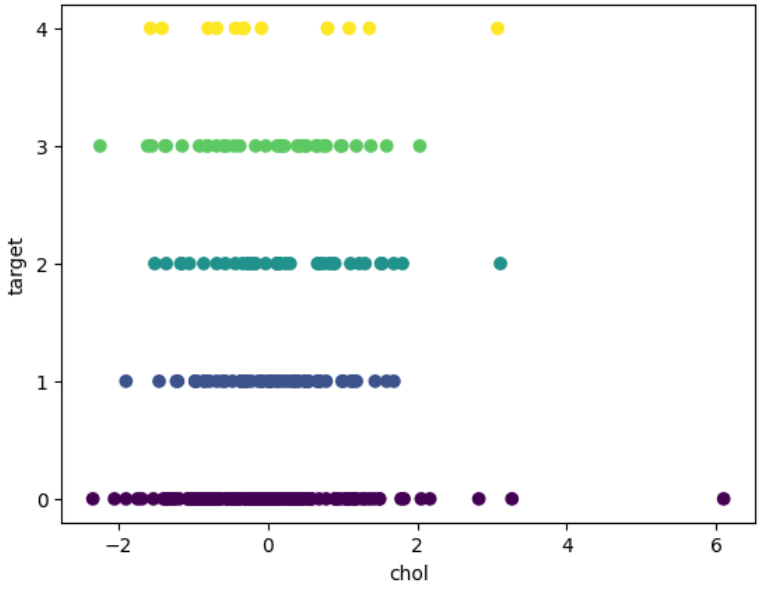
4. Changing data types - changed the data type of categorical columns from int to object type to avoid confusion.

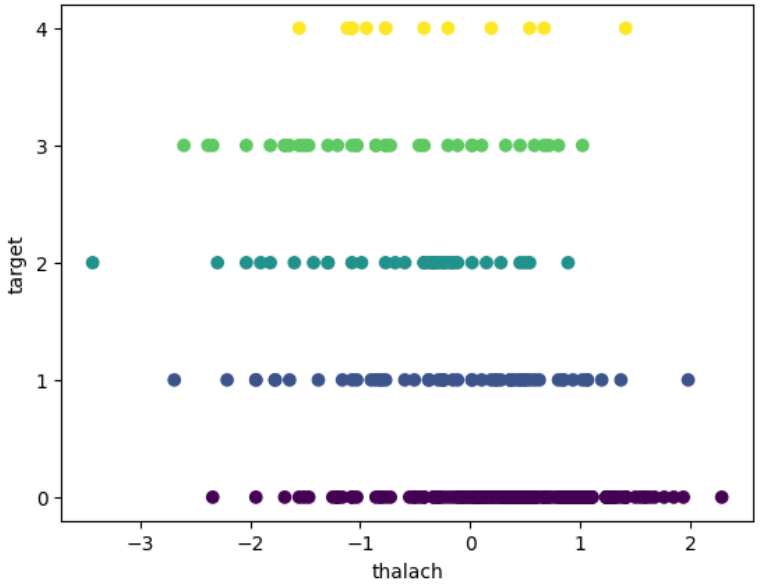
5. Scaled the numerical columns using the standard scaler provided by sci kit learn.

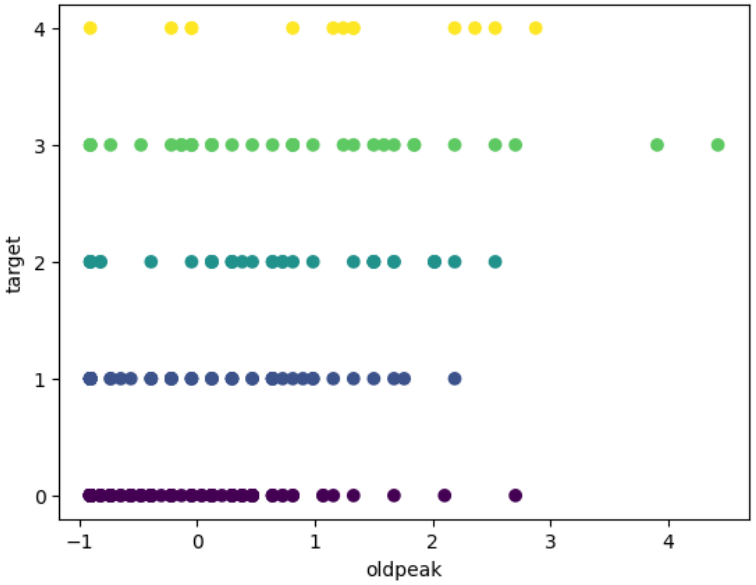
6. Feature selection - since there are both categorical and numerical values in the features and the target is a categorical variable, I decided to use mutual information as my criteria for feature selection.
    a. What is mutual information - Mutual information measures the entropy drop of one random variable given other
    b. Why mutual information? -
        i. Mutual information can capture non linear relationships between variables
        ii. Can be used for both numerical and categorical variables

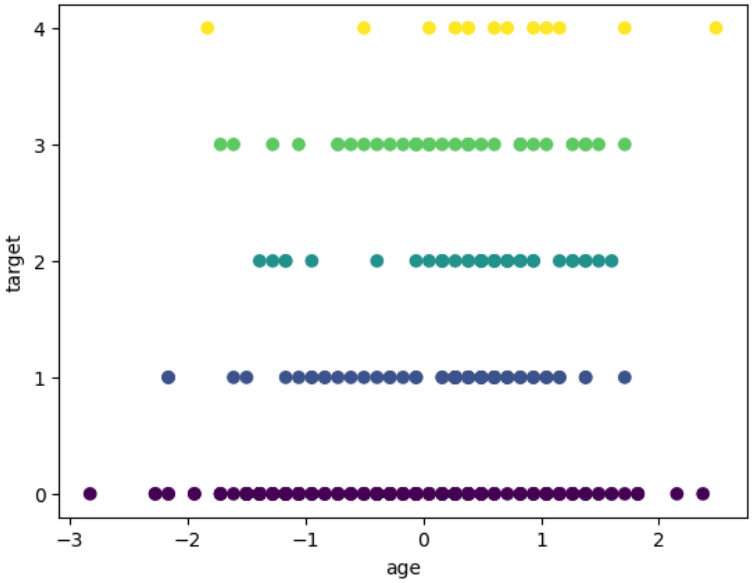
Data Analysis

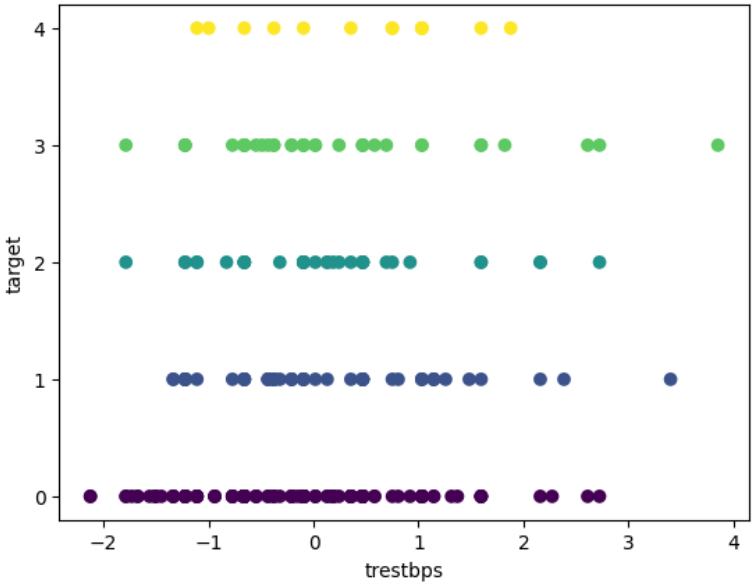
1. Scatter plots - to visualise the distribution of values of numerical features with respect to the target



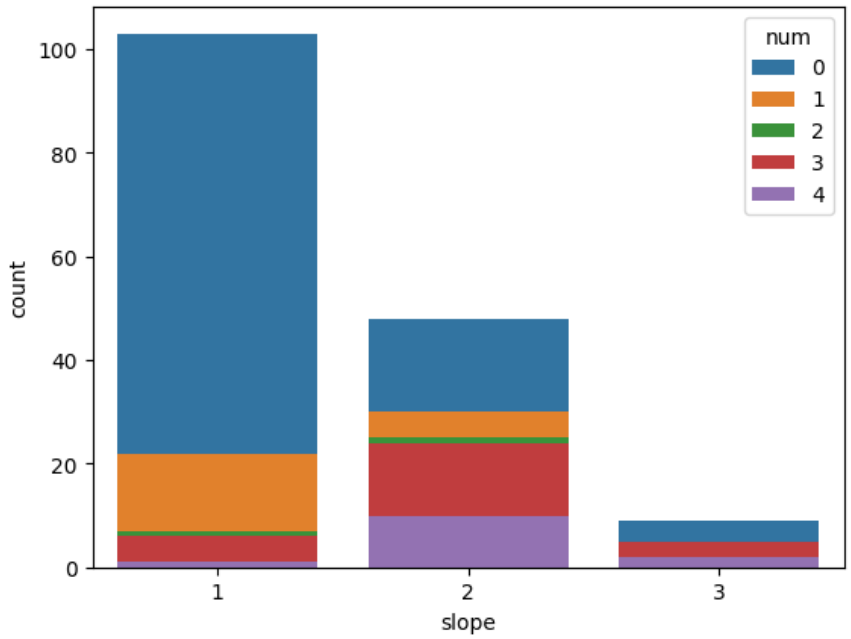


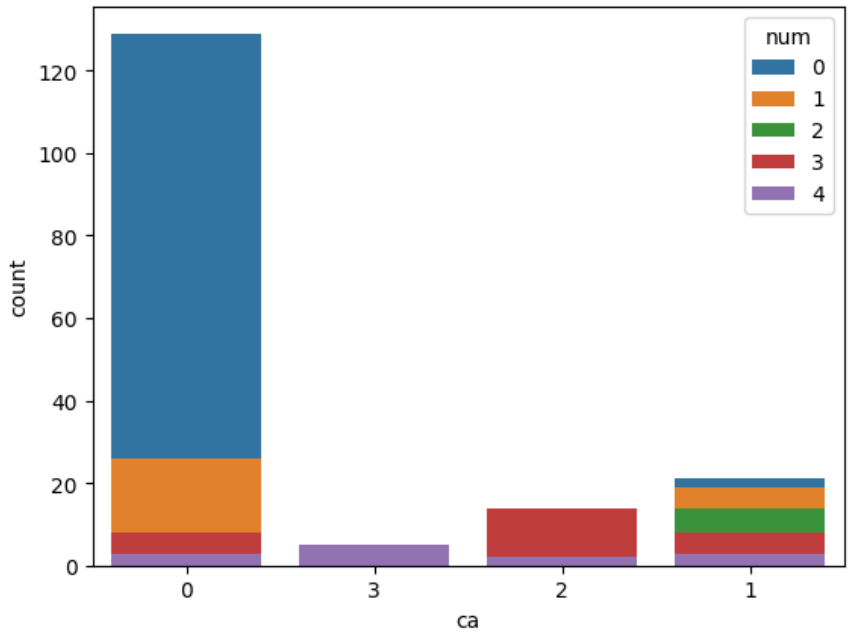

target
oldpeak
4
3
2
1
0
−1
0
1
2
3
4

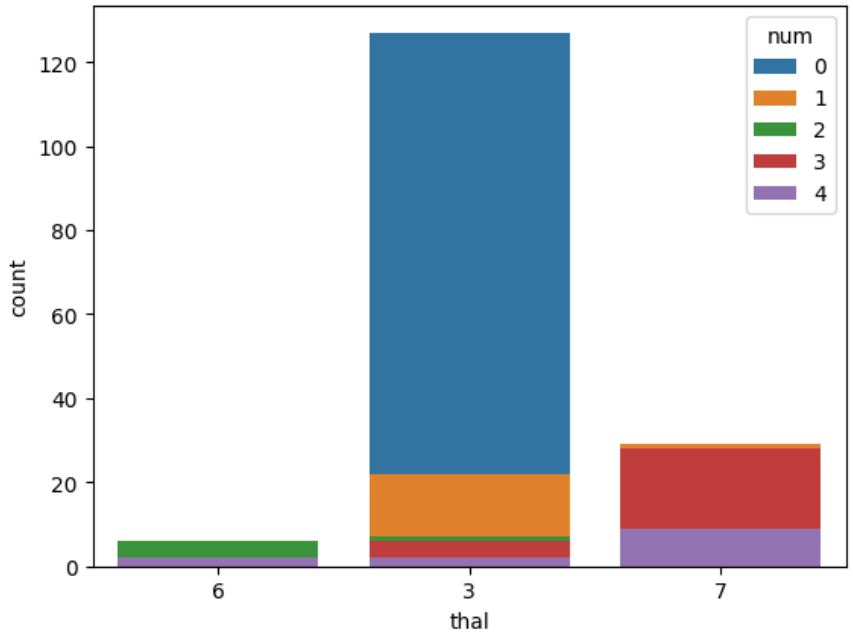

target
age
4
3
2
1
0
−3
−2
−1
0
1
2

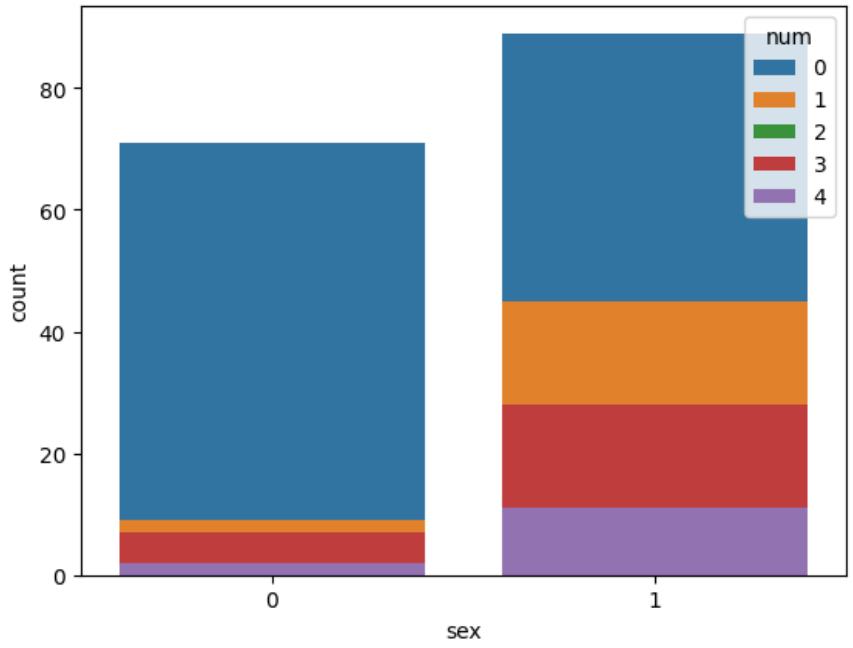

target
trestbps
4
3
2
1
0
−2
−1
0
1
2
3
4

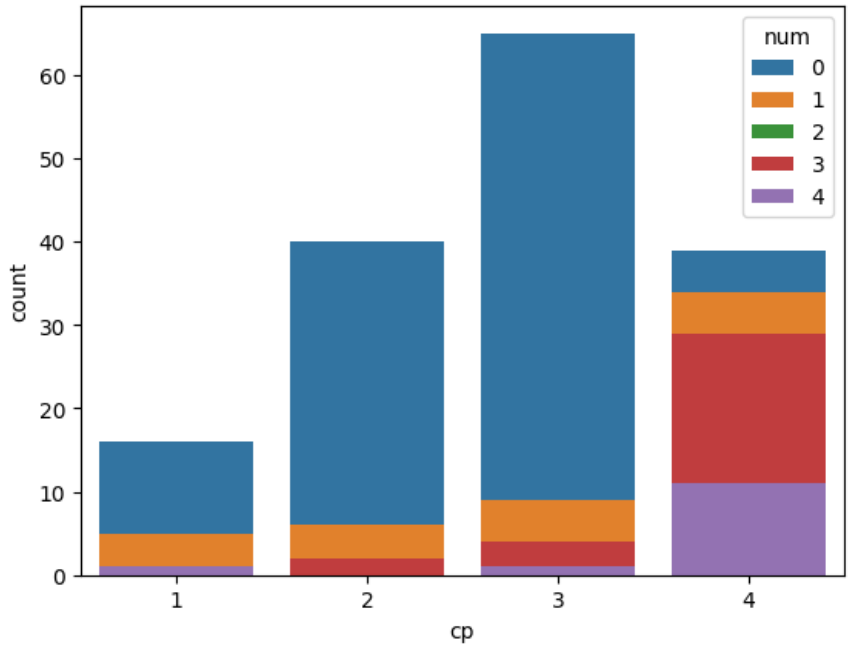
2. Barplots - to visualise the distribution of categories of categorical features with respect to the target variable

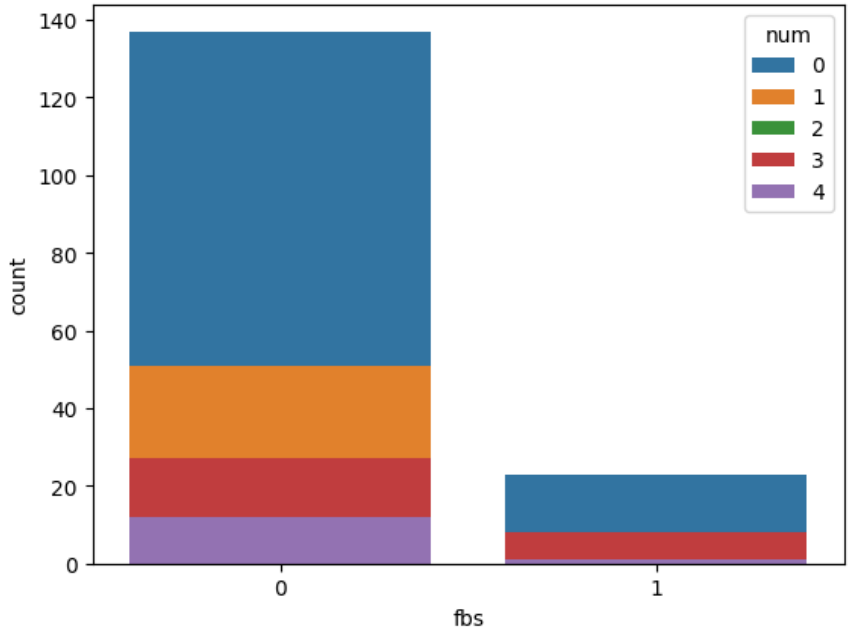


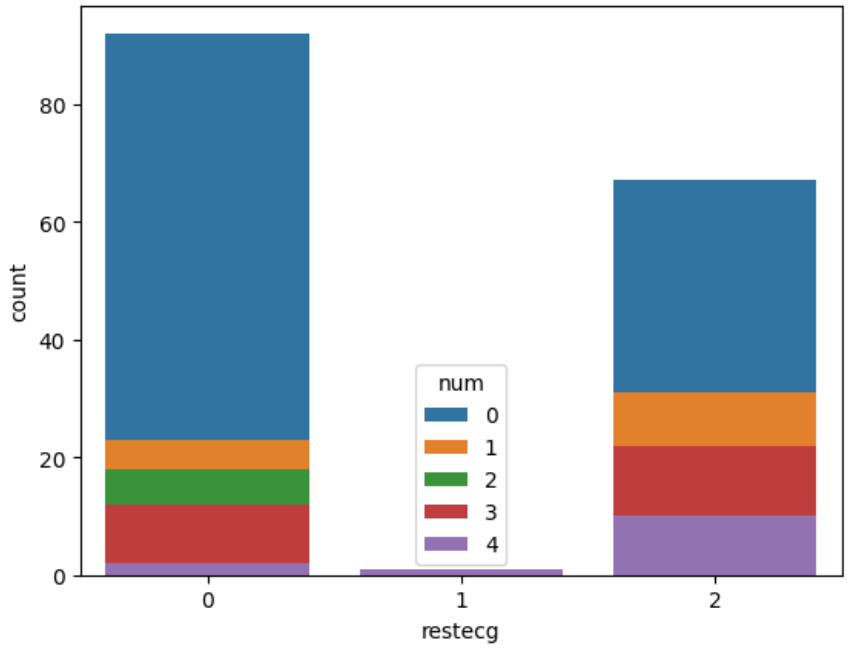


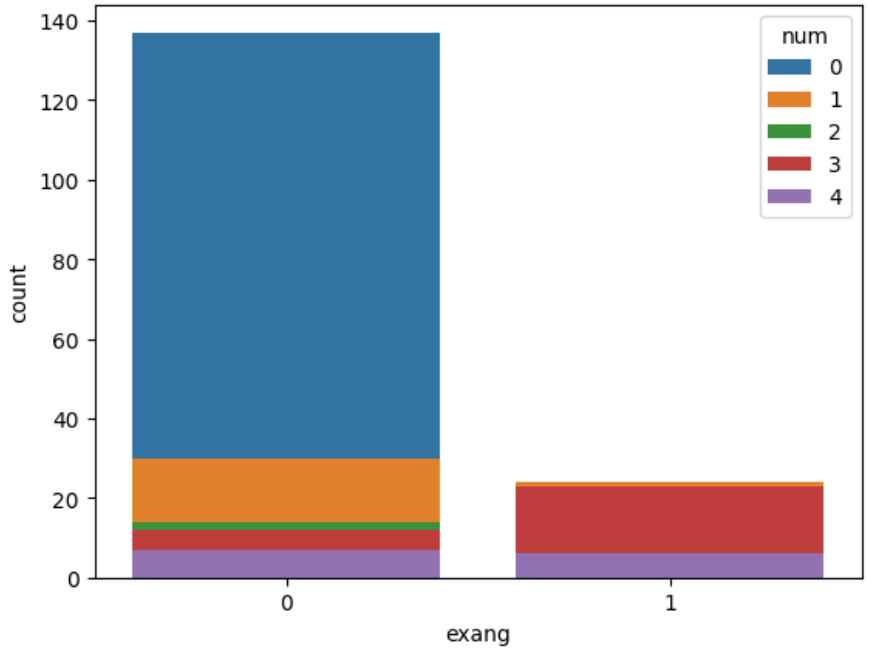

num
0
1
2
3
4
120
100
80
60
40
20
0
count
6
3
7
thal


num
0
1
2
3
4
80
60
40
20
0
count
0
1
sex


num
0
1
2
3
4
count
60
50
40
30
20
10
0
1
2
3
4
cp


num
0
1
2
3
4
count
140
120
100
80
60
40
20
0
0
1
fbs





Observations -

1. When the number of major vessels (ca) is 0, is it likely that the person does not have heart disease
2. When exercise induced angina is present(exang) , it is likely that the person does not have heart disease
3. When the results of a nuclear stress test (thal) is 3, it is likely that the person does not have heart disease.

Comment on logistic regression -

I do not think we can use logistic regression for this case. It is usually used to predict the probability of a binary event occurring. We should be able to predict probabilities of multiple classes. Therefore we need to use an algorithm which supports multiclass classification.