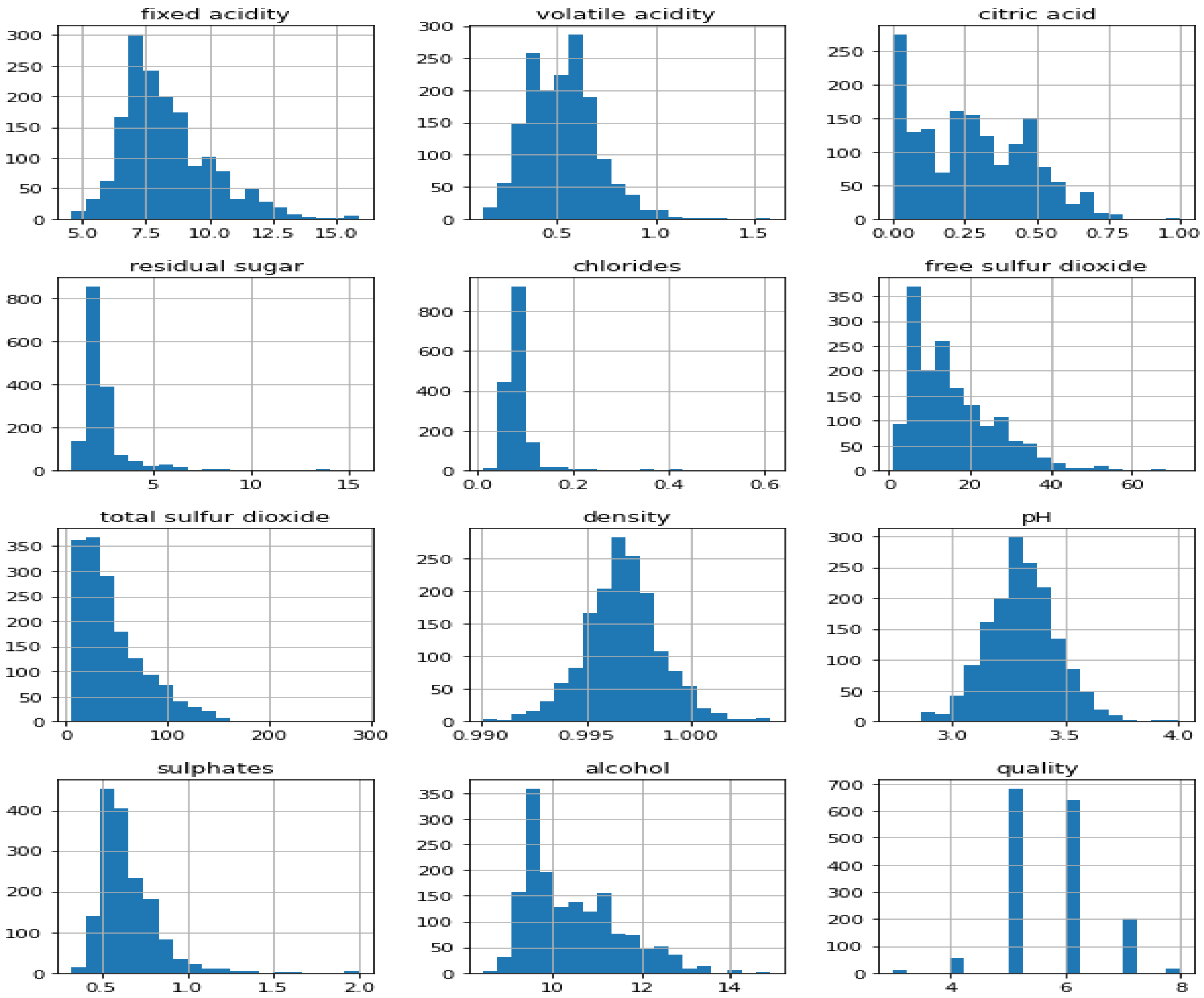
CAPSTONE PROJECT

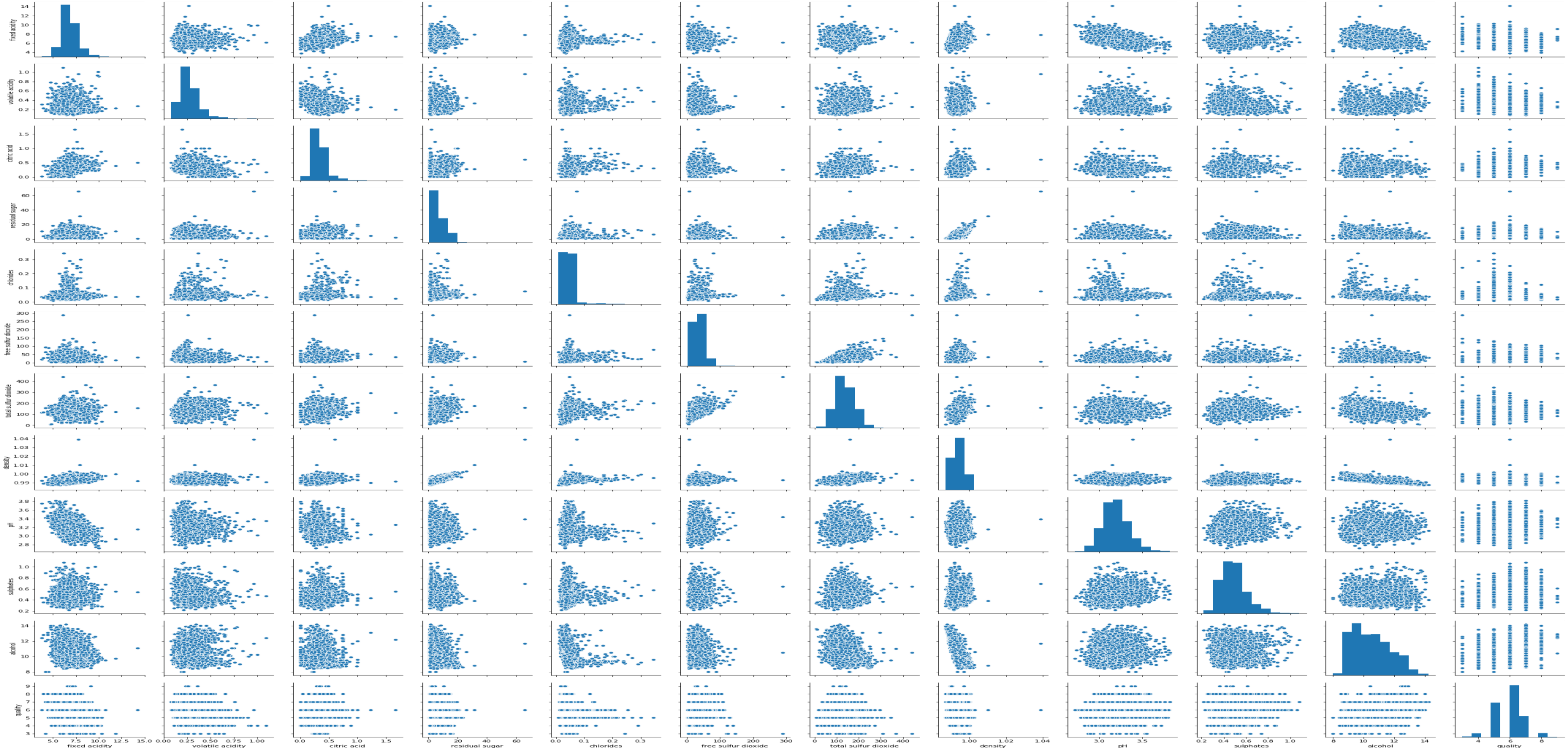
By:- Aditya Aryan

Batch:- 24th March WKDY-2022

The below plot shows the distribution of the dataset. We get to know that most of the wine is of average quality. There are fewer wines that are of very high quality and great tasting, and very few wines that aren't so good.

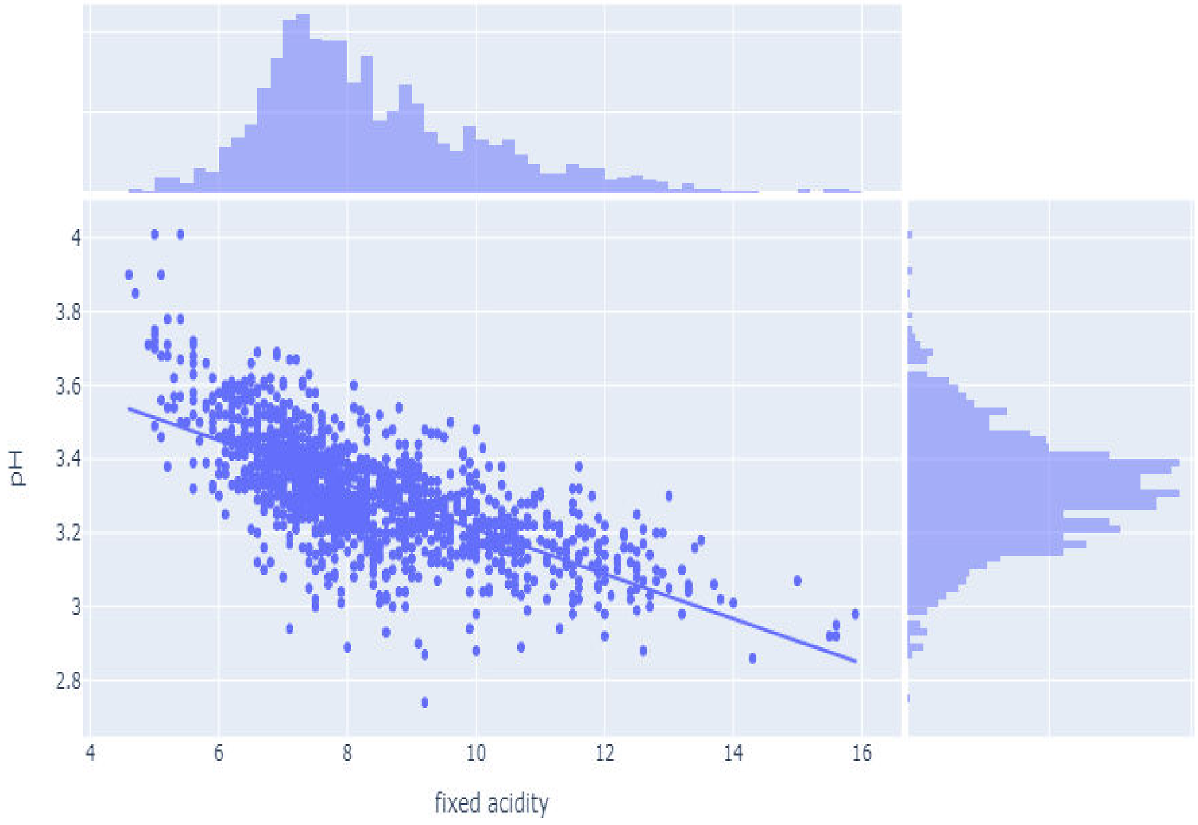


Lets check which of the other columns are highly correlated to Quality. We will first display a pair plot to visualize the correlation between variables. From the above scatterplot we can get some interesting details. For some of the features, the distribution appears to be fairly linear. For some others, the distribution appears to be negatively skewed. So this confirms our initial suspicions — there are indeed some interesting co-dependencies between some of the features.



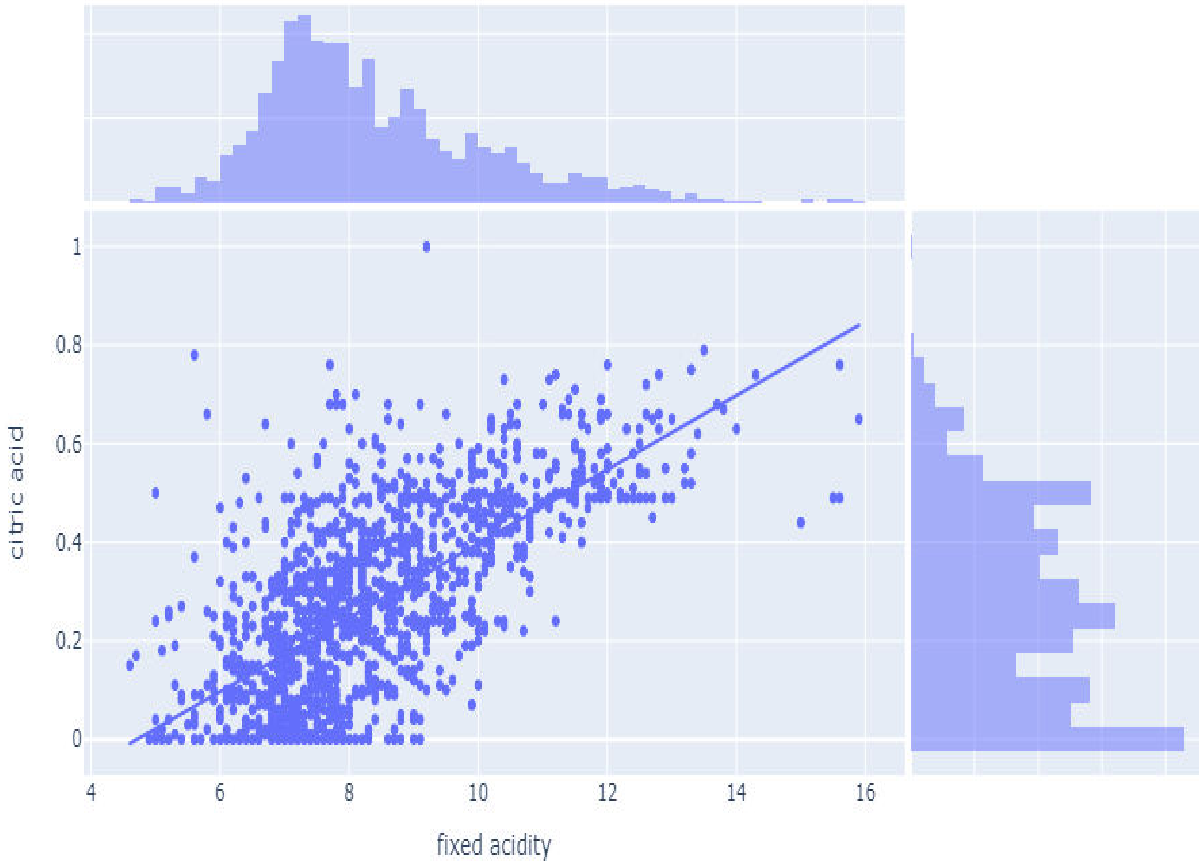
Relation between pH and Fixed Acidity

Below is a scatter plot to visualize the relation between pH and Fixed Acidity with marginal histograms to visualize their distribution. A trendline is created for better understanding of the plot. We can see that pH and Fixed Acidity has a negative correlation between each other. As the Fixed Acidity increases the pH level decreases, a lower pH indicates more acidic wine and vice-versa



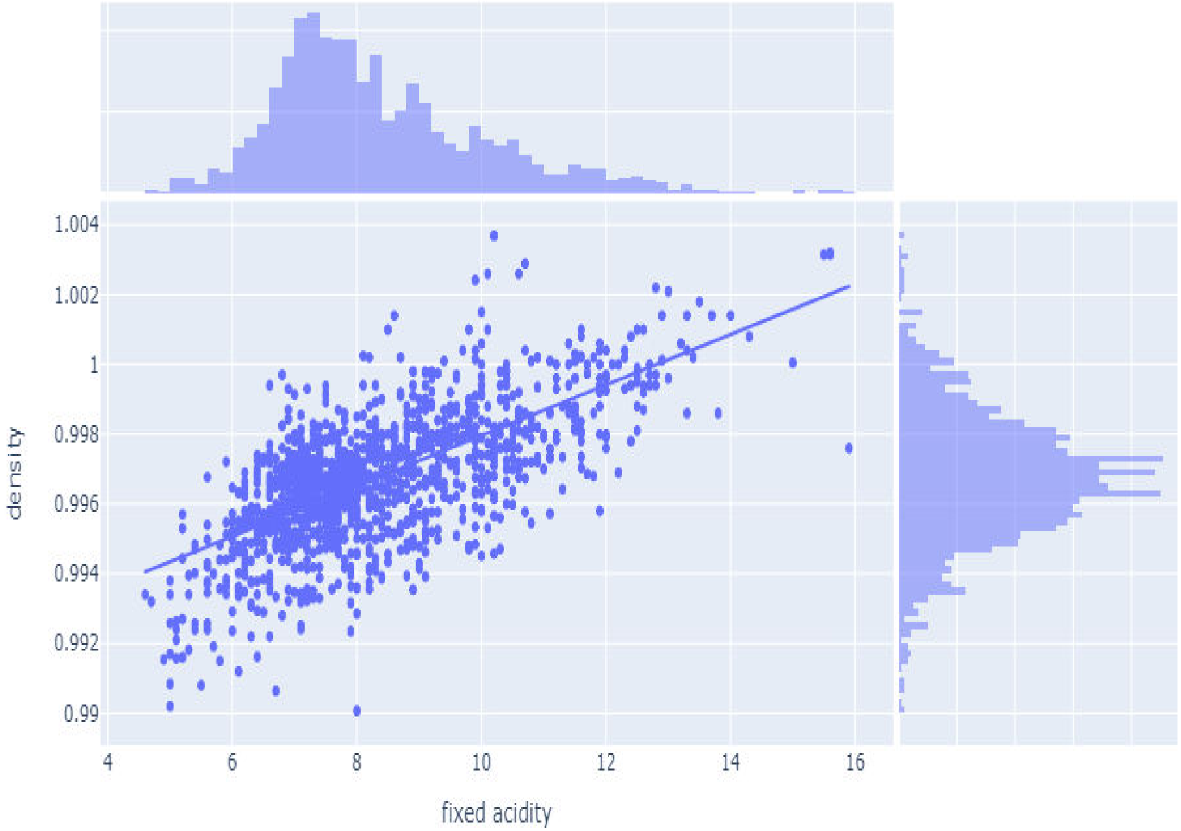
Relation between Fixed Acidity and Citric Acid

Citric acid has a minor presence in wine, but a noticeable one nonetheless. The quantity of citric acid in wine is about 1/20th that of tartaric acid. It’s mostly added to wines after fermentation due to yeast’s tendency to convert citric acid to acetic acid. It has an aggressive acidic taste, is often added by winemakers to increase a wine’s total acidity, and should be added very cautiously. From the below plot we



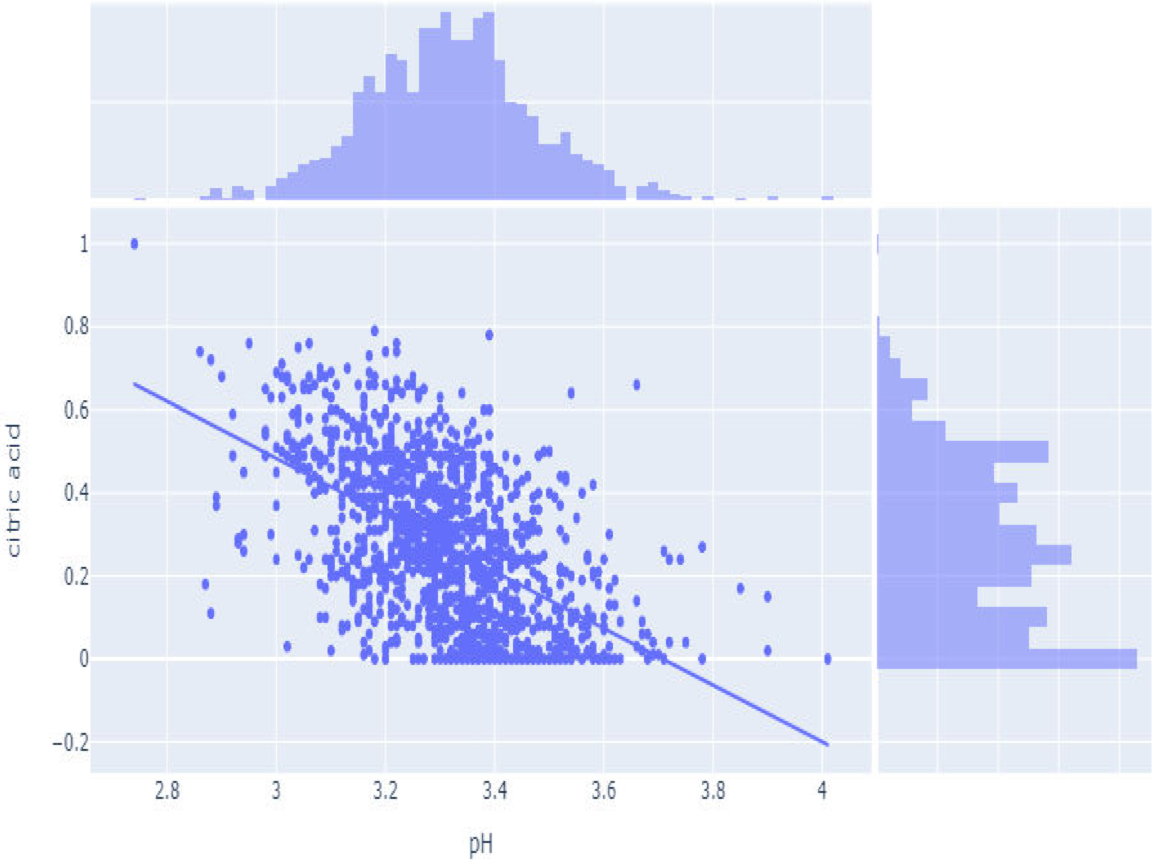
Relation between Fixed Acidity and density

Tartaric acid plays a key role in the stability of wines and influences the taste, color and odor of the final product. Below scatterplot shows that density increases with a increase in fixed acidity. It tells that wine tasters like wine which are more dense in fixed acidity.

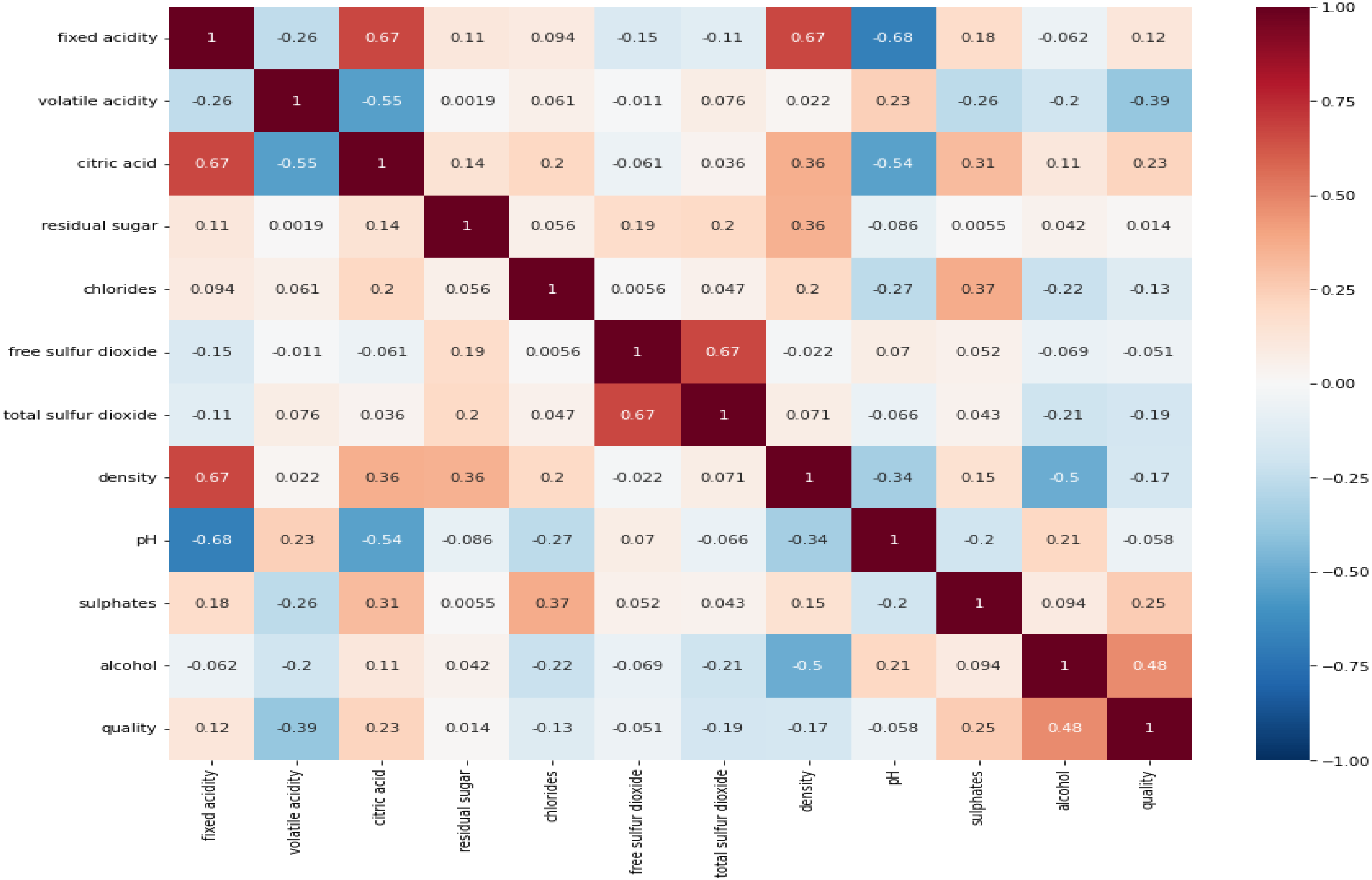


Relation between pH and citric acid

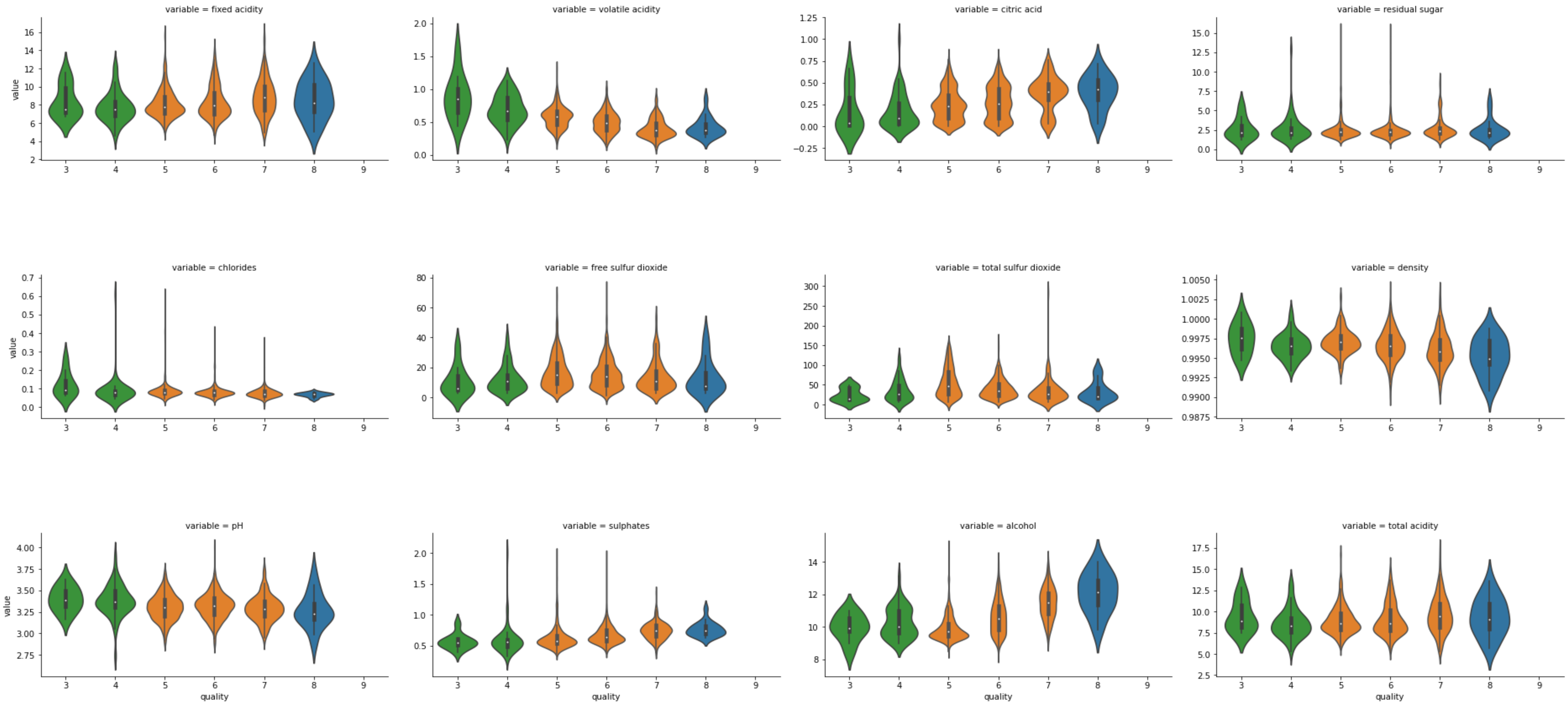
Below scatterplot shows that pH level increases with the increase in citric acid.



We can plot a heatmap of co-relations between features, which will help us get more insights. Free Sulphur dioxide and Total Sulphur dioxide have some positive relation to Residual Sugar. On further inspection, I found that the quantity of Sulphur dioxide is dependent on sugar content. We see some variables like sulphates , alcohol and citric acid contents seem to be correlated to quality.



We see some variables like sulphates , alcohol and citric acid contents seem to be correlated to quality. We also observe that some variables seem to have a weak correlation with quality, like chlorides and residual sugars content. Below we can visualize FacetGrid violin plot for different variables vs quality.

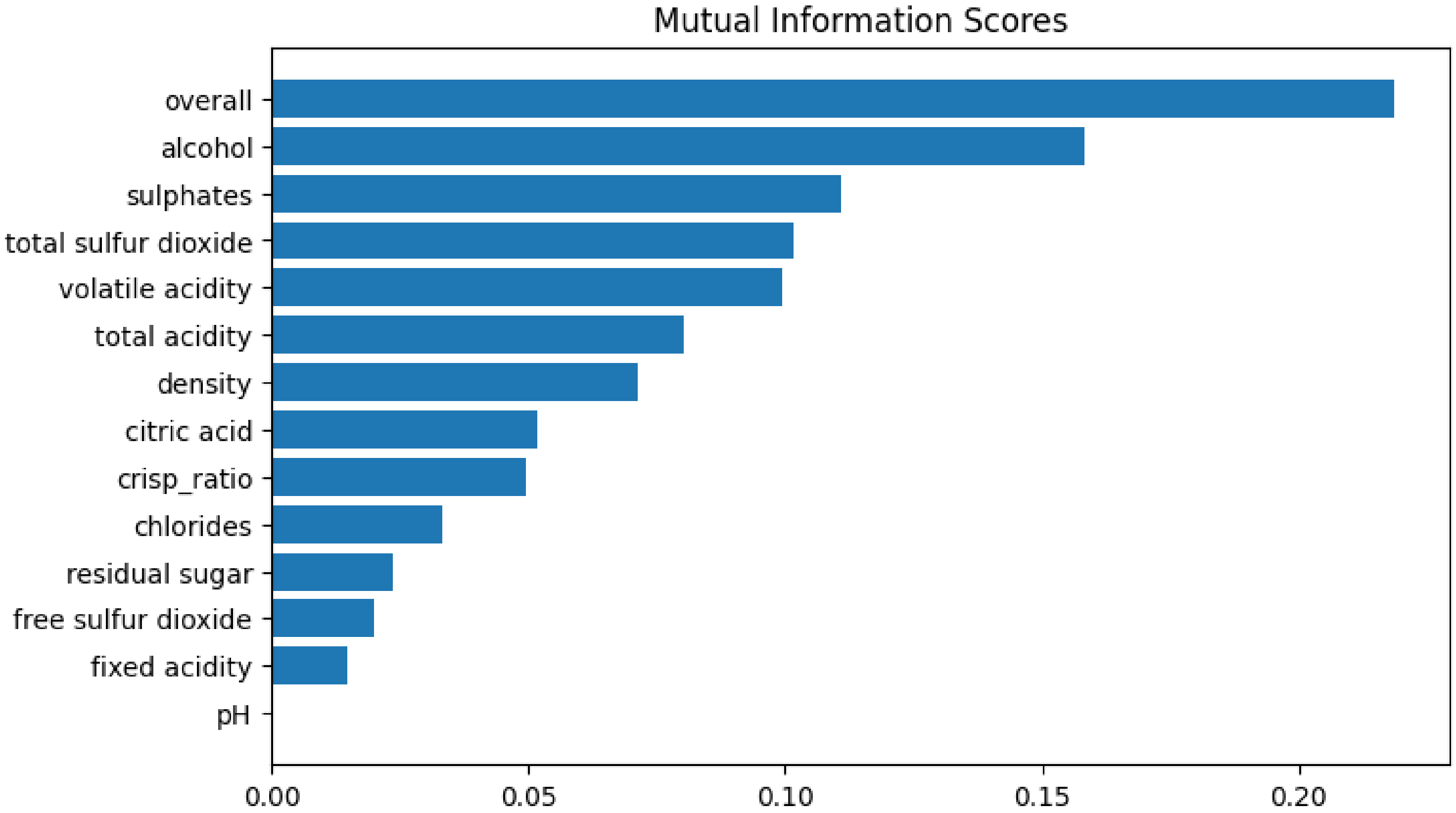


Feature Creation

Created three new columns called first is total_acidity which is the sum of fixed acidity, citric acid, and volatile acidity, the second is crisp_ratio which is the ratio of total acidity upon residual sugar and the third is overall which consider all wines with ratings 7 and above to be of very good quality, wines with 5 and 6 to be of average quality, and wines less than 5 to be of insipid quality which is the binned value of quality, it tells us if the wine's quality is good, bad or medium

Feature Selection

For feature selection I tried mutual information since it is a univariate metric so it can't detect interaction between univariate metrics. Therefore, I picked all the features shows correlation among each other.



MODEL BUILDING AND SELECTION

STEPS:

I tried Logistic Regression, Support Vector Machine, Random Forest Classifier and XGBoost Classifier. I used GridSearchCV for hyperparameter tuning and found out that XGBoost outperforms other algorithms. In the preprocessing step MinMaxScaler was used to standardize the dataset in order to prepare it for modelling. Sklearn's train_test_split was used for splitting the data into training and testing set. I used all the preprocessing and model building steps inside a pipeline in order to make the workflow much easier to read and understand and to enforce the implementation and order of steps in your project.

DIFFICULTIES FACED:

The difficulties I faced during the project was in increasing my models accuracy, which at first was only around 60-65%, after applying some feature engineering I was able to achieve better results. Apart from feature engineering a particular problem I faced was during class imbalance. I overcame the difficulty by tuning in the parameters for XGBoost and in case of Random Forest I used sklearn's compute_class_weight function to calculate the weight and pass it as a dictionary in weights parameter. Target variable was mapped using map() function of pandas. Another problem occurred while using GridSearchCV for parameter tuning, the score obtained from cross-validating were less compared to setting parameters manually. The GridSearchCV process not only attempts all of the values that you have in the, but also performs some data manipulation: the "folds" of the data. This is resampling data multiple times so as to help make the final classifier as robust to new data as possible.