Practical Machine Learning

Exploratory Data Analysis Report

Sanjeeb Humagain

April 9, 2024

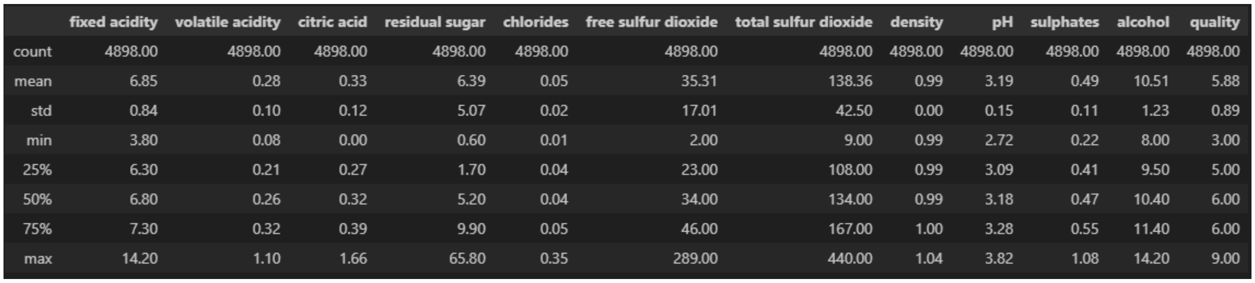
Introduction:

Exploratory data analysis (EDA) is used to check the pattern of data and prepare it before feeding it into any Machine Learning model to improve the performance of the model. Sometimes, it would be difficult to comprehend the nature of data until we visualize it. Analyzing the data would help to decide whether data cleaning is needed or not. Data preprocessing is crucial to transform the raw data into machine readable format. It involves steps such as, performing analysis, filtering the data, transformation, and encoding the data. Improper data preprocessing will degrade the quality and performance of the model. EDA is used to conduct an analysis on raw data to see if there are any missing values, visualize the data distribution, see the correlation between features and remove unnecessary data and features from the training data set. The goal is to make the data suitable for the ML model which will increase the model performance.

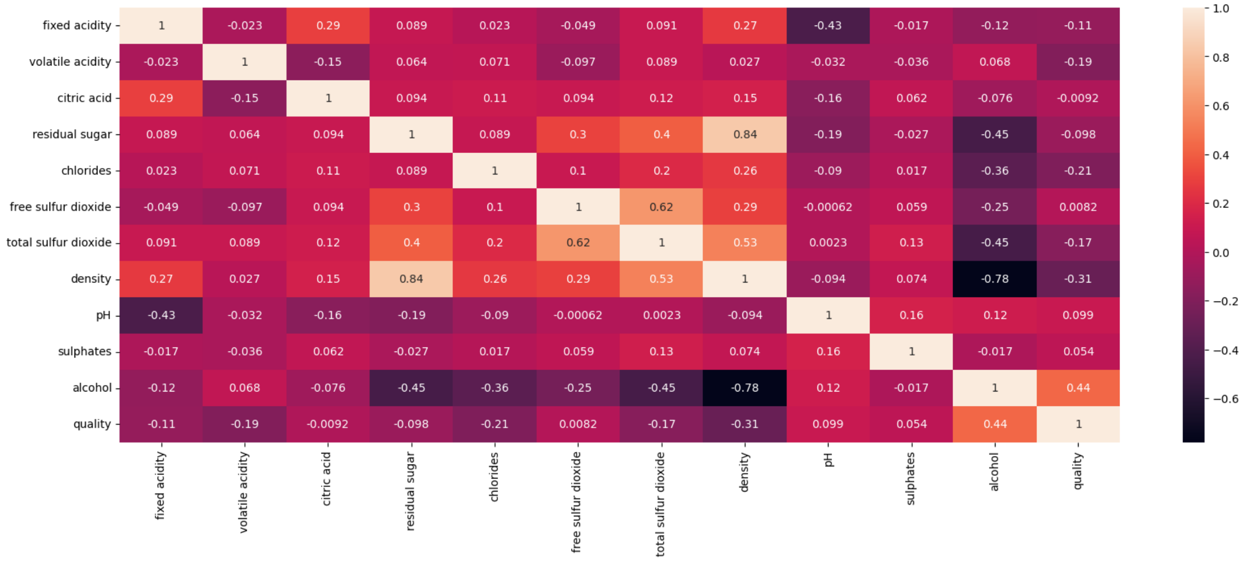
Results of the Analysis: White Wine

The Wine Quality data of white wine was used for this analysis which has 12 features and 4898 samples. Therefore, there will be 11 input features if we plan to use one as a target variable. The data does not have missing values which we could see by using info or describe function of pandas. Describe function gives the important statistical information of each column of the data set. Additionally, using the isnull() function helps to see the number of missing values in each features if they exist. This would help in data cleaning process. Additionally, the correlation of between the features was also visualized using heatmap function. This helped to get the understanding of how the features are correlated. We could remove some features if found uncorrelated and unnecessary. Also, the box plot helped to visualize the outlier which should be removed to make the performance of the ML model better. There were functions which we could use to remove some features, sort on the basis of some features that would help to understand the nature of features and compare them. Some of the images of the model is attached below.

info() function shows the number of non-null count of the features and its data types. describe() function shows the statistical measure of all the features. So, we could see the count, mean, standard deviation, minimum, first quartile, second quartile, third quartile, and maximum value of all the features as shown in the following table.



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
count	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00	4898.00
mean	6.85	0.28	0.33	6.39	0.05	35.31	138.36	0.99	3.19	0.49	10.51	5.88
std	0.84	0.10	0.12	5.07	0.02	17.01	42.50	0.00	0.15	0.11	1.23	0.89
min	3.80	0.08	0.00	0.60	0.01	2.00	9.00	0.99	2.72	0.22	8.00	3.00
25%	6.30	0.21	0.27	1.70	0.04	23.00	108.00	0.99	3.09	0.41	9.50	5.00
50%	6.80	0.26	0.32	5.20	0.04	34.00	134.00	0.99	3.18	0.47	10.40	6.00
75%	7.30	0.32	0.39	9.90	0.05	46.00	167.00	1.00	3.28	0.55	11.40	6.00
max	14.20	1.10	1.66	65.80	0.35	289.00	440.00	1.04	3.82	1.08	14.20	9.00



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.023	0.29	0.089	0.023	-0.049	0.091	0.27	-0.43	-0.017	-0.12	-0.11
volatile acidity	-0.023	1	-0.15	0.064	0.071	-0.097	0.089	0.027	-0.032	-0.036	0.068	-0.19
citric acid	0.29	-0.15	1	0.094	0.11	0.094	0.12	0.15	-0.16	0.062	-0.076	-0.0092
residual sugar	0.089	0.064	0.094	1	0.089	0.3	0.4	0.84	-0.19	-0.027	-0.45	-0.098
chlorides	0.023	0.071	0.11	0.089	1	0.1	0.2	0.26	-0.09	0.017	-0.36	-0.21
free sulfur dioxide	-0.049	-0.097	0.094	0.3	0.1	1	0.62	0.29	-0.00062	0.059	-0.25	0.0082
total sulfur dioxide	0.091	0.089	0.12	0.4	0.2	0.62	1	0.53	0.0023	0.13	-0.45	-0.17
density	0.27	0.027	0.15	0.84	0.26	0.29	0.53	1	-0.094	0.074	-0.78	-0.31
pH	-0.43	-0.032	-0.16	-0.19	-0.09	-0.00062	0.0023	-0.094	1	0.16	0.12	0.099
sulphates	-0.017	-0.036	0.062	-0.027	0.017	0.059	0.13	0.074	0.16	1	-0.017	0.054
alcohol	-0.12	0.068	-0.076	-0.45	-0.36	-0.25	-0.45	-0.78	0.12	-0.017	1	0.44
quality	-0.11	-0.19	-0.0092	-0.098	-0.21	0.0082	-0.17	-0.31	0.099	0.054	0.44	1

Fig (1) Correlation diagram of different features of white wine dataset

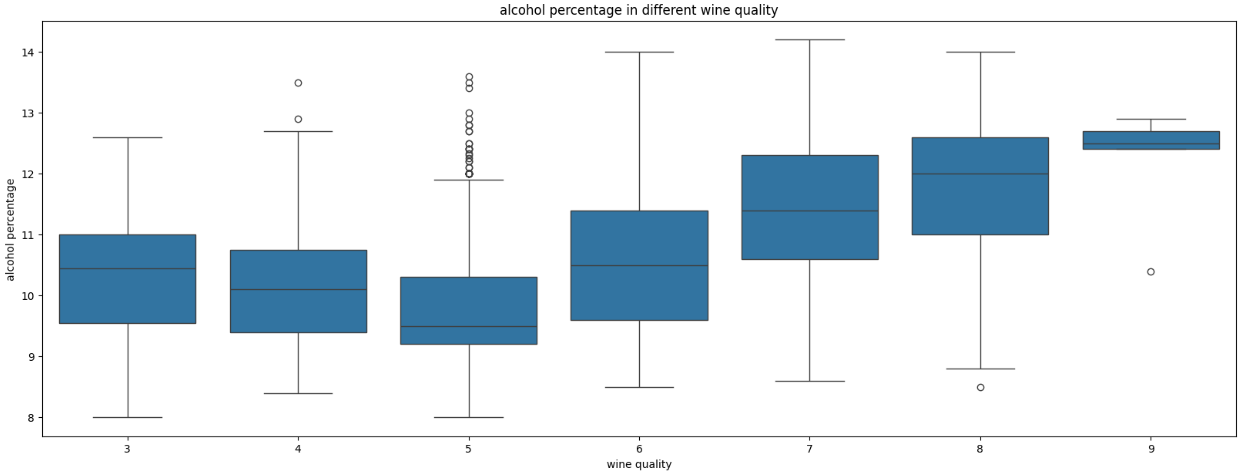


Fig (2) box plot of alcohol percentage with wine quality white wine dataset

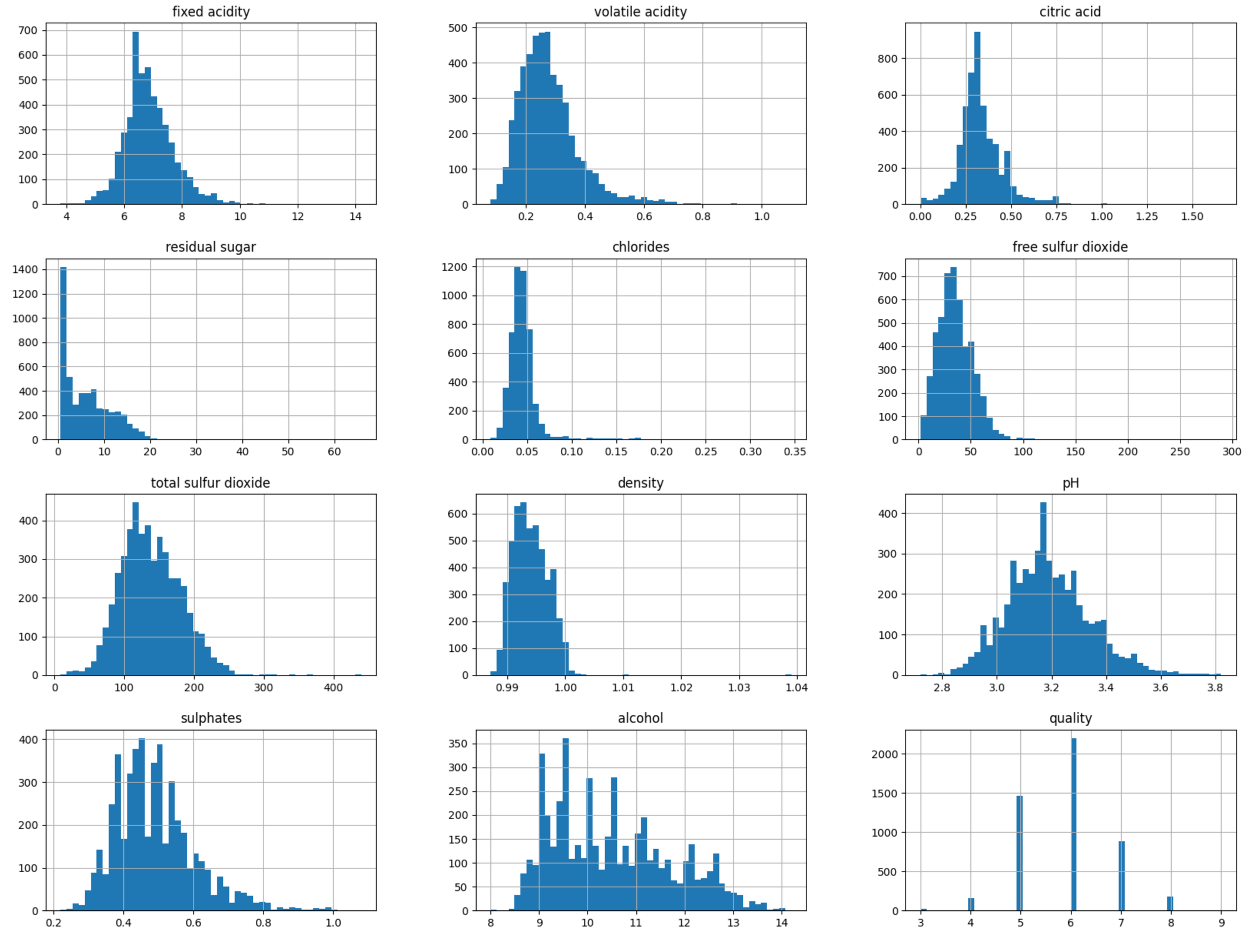


Fig (3) histogram of different features of white wine dataset

Figure 3 represents the histogram plot of different features and figure 4 below depicts the plot after normalization. Normalization is an important step which ensures that all the features contribute equally in any ML algorithm thereby improving the model performance. During this exercise, normalize function from sklearn library was used.

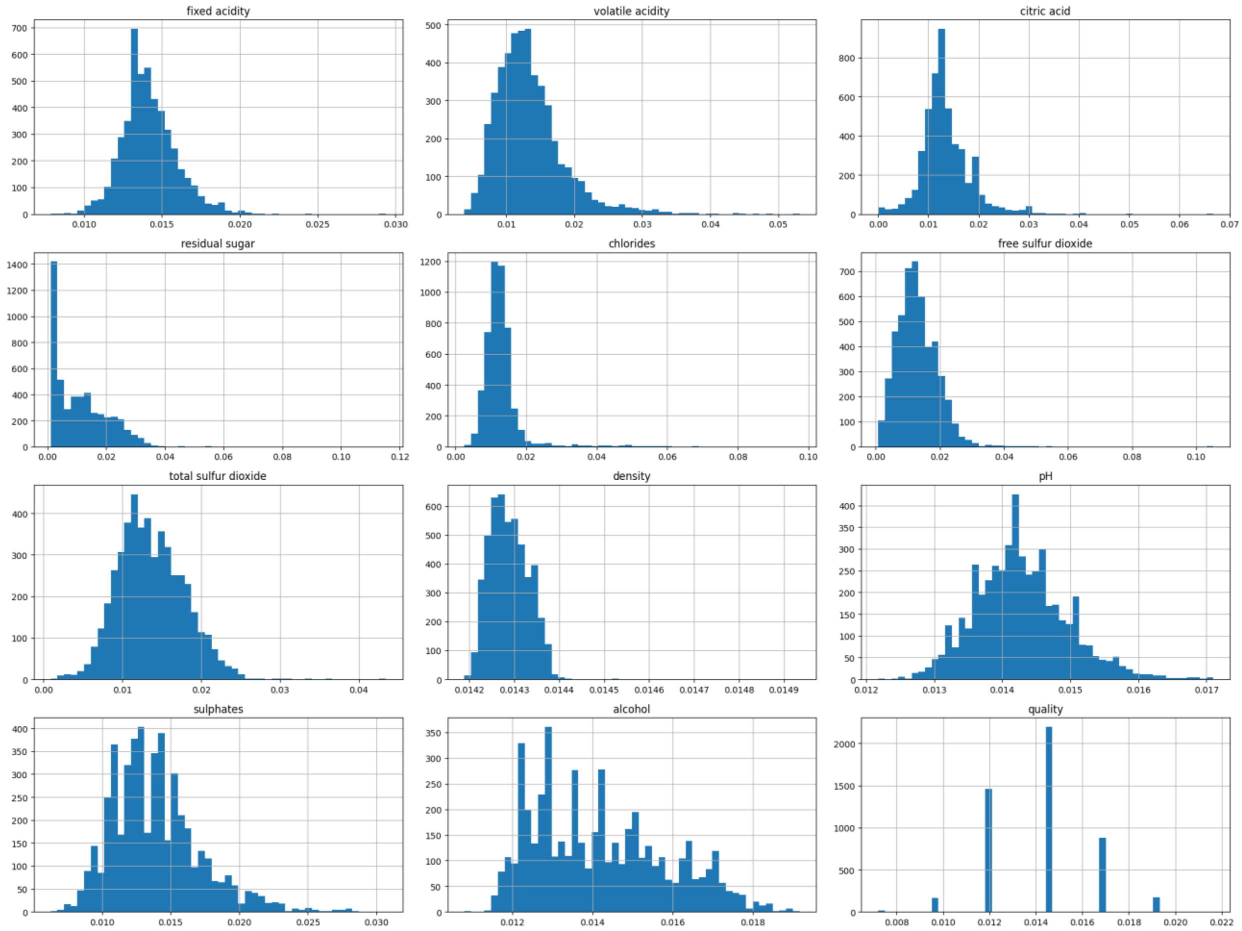


Fig (4): histogram of different features of white wine dataset after normalization

Results of the Analysis: Tumor Data

The data was originally in arff format which was converted to excel csv file using online converter. The data was found to have 339 rows and 19 columns. The data contains string values like yes, no etc. unlike the previous dataset that contains numerical values. I could see there are '?' sign in place of missing data but the isnull() function did not recognize the missing values. It was found that there was 1 value missing for sex, 67 values missing for histologic-type, 155 missing data for degree-of-diffe, 1 missing for skin and 1 missing for axillar. After removing the missing value using dropna() function. The dataset will have 132 rows of data.

Therefore, data preprocessing is one of the most important step in any ML task as data will not always be in ready to use format. In order to do further tasks, we could replace the string values with integer and perform the evaluation. In this way finally we can apply ML in this kind of datasets. One of the differences between wine data and tumor data is the former did not have missing data. Additionally, the tumor data has string value rather than numerical values in wine data. This makes tumor data to undergo a lot of preprocessing compared to the wine data.

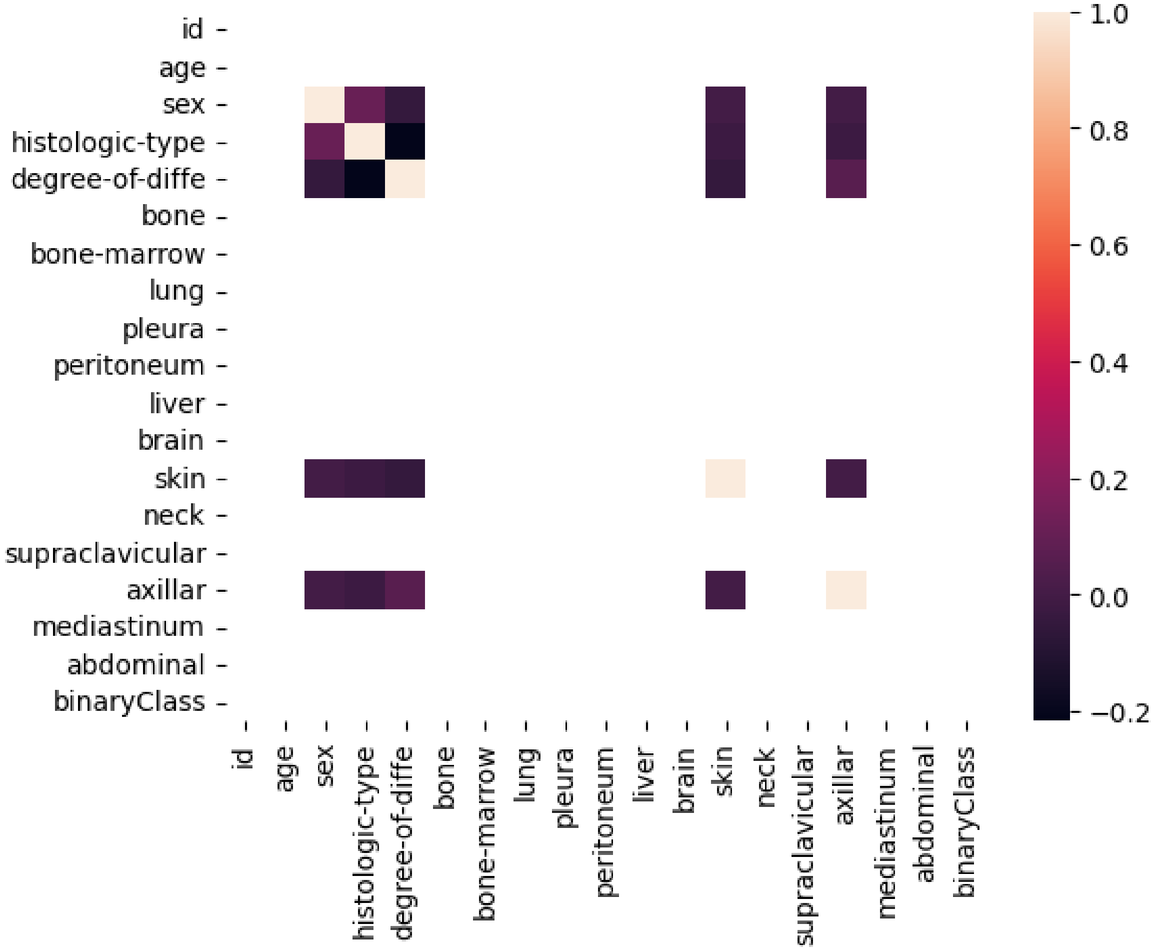


Fig (5): Nullity correlation heatmap for tumor data

The Nullity correlation heatmap show above is a visualization tool used to understand the relationship between missing values in different variable of the tumor dataset.

The frequency plot for age and gender is as shown in the following plot.

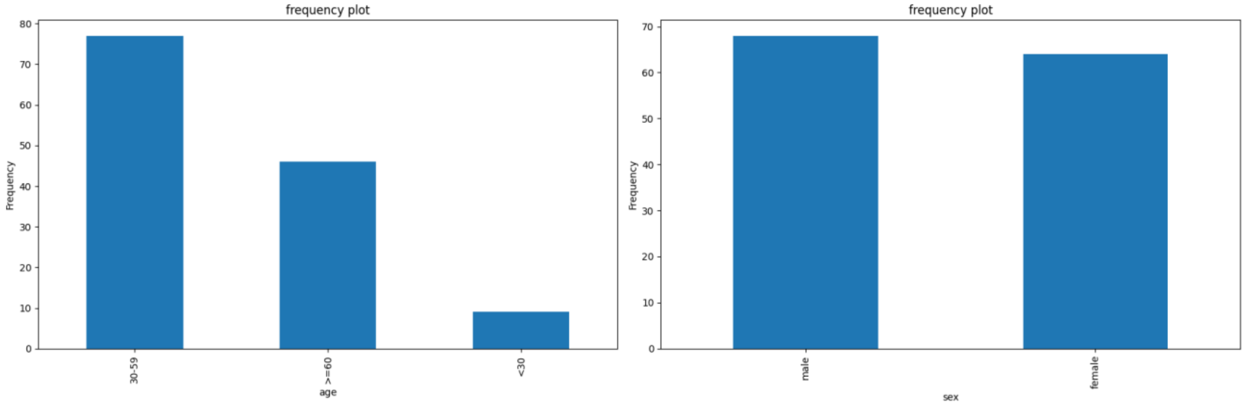


Fig (6): Frequency plot of tumor data