Statistical Analysis and Regression Project

**Student - David Ispiryan**

**Instructor - Dr. Ourania Mangira**

1. **The difference in linear and logistic regressions.**

For Linear Regression:

* Is used to predict the continuous dependent variable
* Is used to find the best fitted line, so output can be predicted
* The output is a continuous/integer value
* Is used least square method for accuracy estimation
* The relationship between X’s and Y shall be linear
* Collinearity between X independent variables

For Logistic Regression:

* Is used to predict the categorical dependent variable
* Is used to solve classification problems
* The output is categorical value, (0,1) or (no, yes)
* Is used maximum likelihood method
* The relationship between X’s and Y is not required to be linear
* No collinearity between X independent variables

**Key assumptions for linear and logistic regressions.**

For Linear Regression:

* Linearity of the data - it is used residuals vs fitted values plot which will show a line. If the line is horizontal that means a good linear relationship.
* Normality of residuals – it is used QQ-plot to understand if the residuals are normally distributed. If points are aligned with the line than it is normally distributed.
* Spread-Location – it is used with Homogeneity of residuals variance. If points are spread equally on the horizontal line, it means a good homoscedasticity, if not it is called heteroscedasticity.
* Residuals vs Leverage (Independence of residuals error terms) – it is used to find out influential cases and extreme values that can have a huge influence on the data.

For Logistic Regression:

* It requires to be no or little multicollinearity between X variables.
* It requires for X variables to be linearly related to log odds and not on dependent variable. So, it is used QQ plot again.
* It requires Shapiro’s or Kurtosis test to test the normality.
* It requires Levene’s test to understand the homogeneity of variance.

**Applications over any domains:**

Linear regression:

* Is used in finance and economics for modeling relationships such as income and expenditure or stock price.
* Is used in Sport analysis to calculate the likelihood of winning or calculating the average goals a team will score.
* Is used in marketing analysis to understand which part of business need to be improved like the relationship in advertising and price.
* Is used in healthcare and pharmaceutics for modeling patients’ outcomes and medical parameters.
* Is used in environmental health to find relationship between water or oil or gas.

https://sg.indeed.com/career-advice/career-development/linear-regression

Logistic regression:

* Is used in finance to find the fraud detections.
* Is used in healthcare to predict the probability of a disease based on other factors.
* Is used in marketing to divide customers into segments.
* Is used in IT for example to find out whether the email is a spam or not.

**Measuring the performance:**

Linear regression:

* Mean Absolute error which measures the mean difference between predicted and actual values.
* Mean Squared error is the mean of squared differences between predicted and actual values.
* Root mean Squared error is the square root of MSE.
* R-squared which measures the proportion of variance in X’s and Y.

Logistic regression:

* ROC curve which is visualization of true positive/negative and false positive/negative values.
* Precision is the correctly predicted positive instance and Recall is the actual positive instances like (True negative/positive and False negative/positive).
* F1 score is the mean of precision and it is used to measure model’s accuracy.

**Alterations and improvements:**

For linear regression we can add X variables for the model to be improved. Hence R squared will be higher. There are few methods to build regression models for improvement and Alterations:

* You can use all independent variables together to use in linear regression and then remove one by one to see how the regression assumptions change.
* You can use only one independent variable and then add new X variables to see the improvements in assumptions.
* You can add polynomial term in any variables that we need to see the change.
* You can add logarithmic expression (mostly used in dependent variable) to transform skewed variable into normalized one.

For logistic regression we can change the threshold of recall and precision if the model has too many False predictions.

I set rate\_classifier to identify from 0 up to 0.6 as 0 and anything above as 1. So, we can change that in order to have a better accuracy of the model.

* You can set the threshold for recall to 0.3 instead of 0.5 by changing our rate\_classifier, if you have a lot of False positives.
* You can set the threshold for precision up to 0.7 by changing our rate\_classifier if you have more False Positives.

**Challenges:**

One of the main challenges we may face in regression models:

* The outliers as it can be very sensitive to the data.
* Underfitting as it means that the model is accurate but the predictions were not.
* Overfitting as it means that the model is not accurate but the predictions were. (Training and Test).
* Ratio of recall and precision should be high.

1. Given the wine\_quality.csv data, we can see that there are no missing values, so we do not have to remove any data or replace missing or NA values with mean or median of any variable.

All our variables are numeric so there are no categorical variables in the data but we might create one in order to divide into specific groups for a particular variable like the quality of wine.

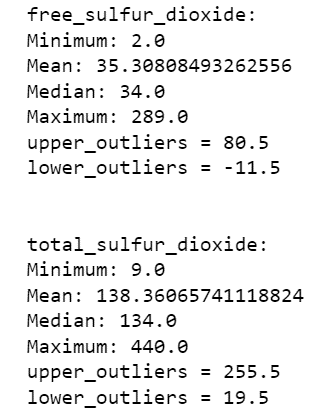
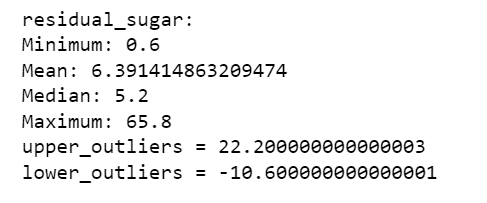
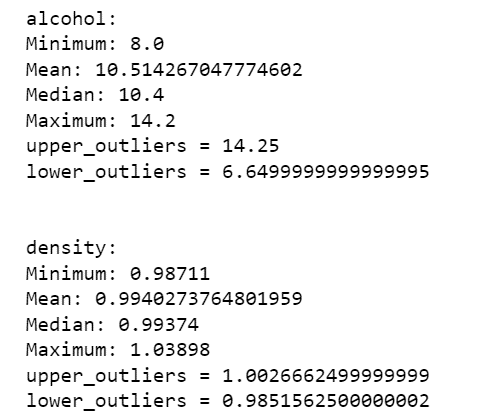
The first thing is to understand the correlations of variables in data. After reviewing a python code to find a correlation higher than ±0.7 as it is considered a strong positive correlation, we see that there is a strong linear relationship between following variables – residual sugar and density (r = 0.84). We create a visualization with seaborn with these two numeric variables using scatterplot or regplot. It is obvious that we have an outlier as it is too far and we check that with calculating the IQR and finding which points are outside minimum and maximum values with these formula Q3 + 1.5\*IQR and Q1 – 1.5\*IQR.

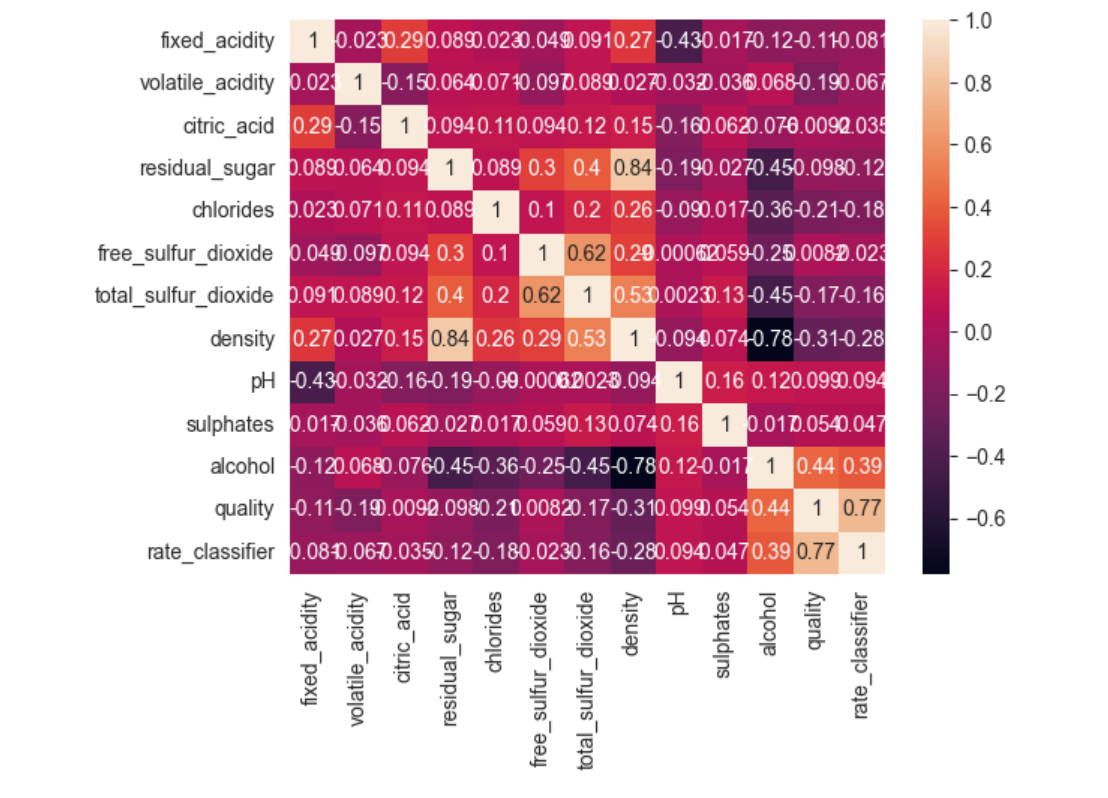
I have also checked the variance and the standard deviation of these variable – residual sugar, alcohol. The variation for both variables are quite high ( >1 ).

As for Standard Deviations they have a high STD which means that it has a large impact on the data.

Regarding the skewness of these variable we can see that the median and the mean are almost the same number so our data is more normally distributed.

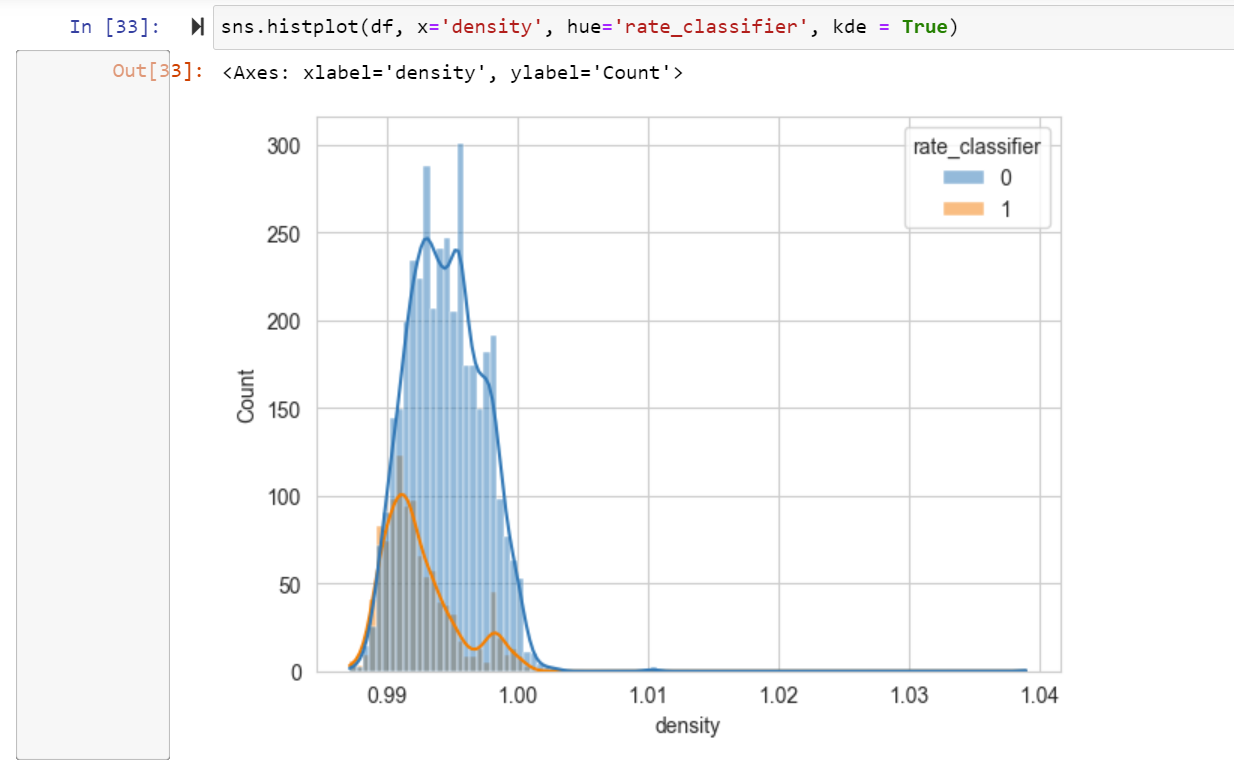
Here we can see the visualization of the correlation matrix. Cells which have a bright color have a strong correlation.





After removing the residual sugar outlier values which are bigger than 22.2 and removing density values bigger than 1.00266, we can see the visualization much clear and by removing the outliers this will help us in building regression models with a better assumption results and with better visualizations in plotting assumptions.

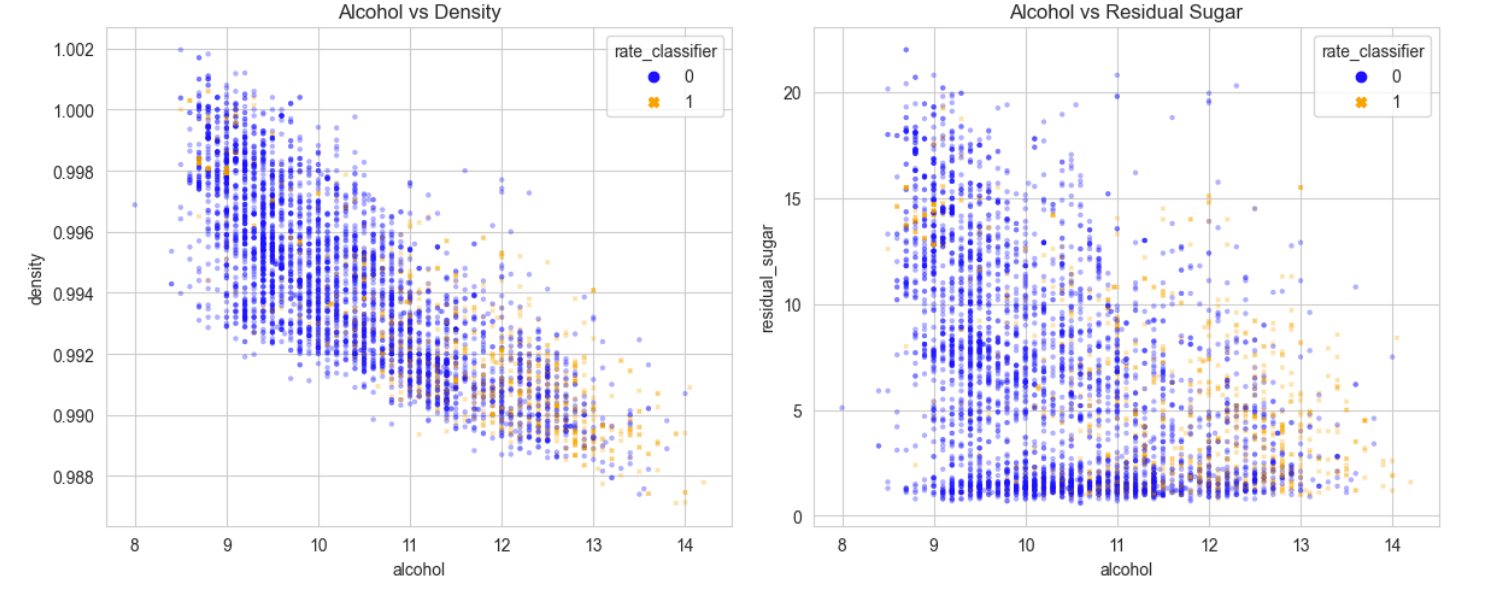
With histogram plot of density, we can see that most of data is lied on 0.99-1.00 interval



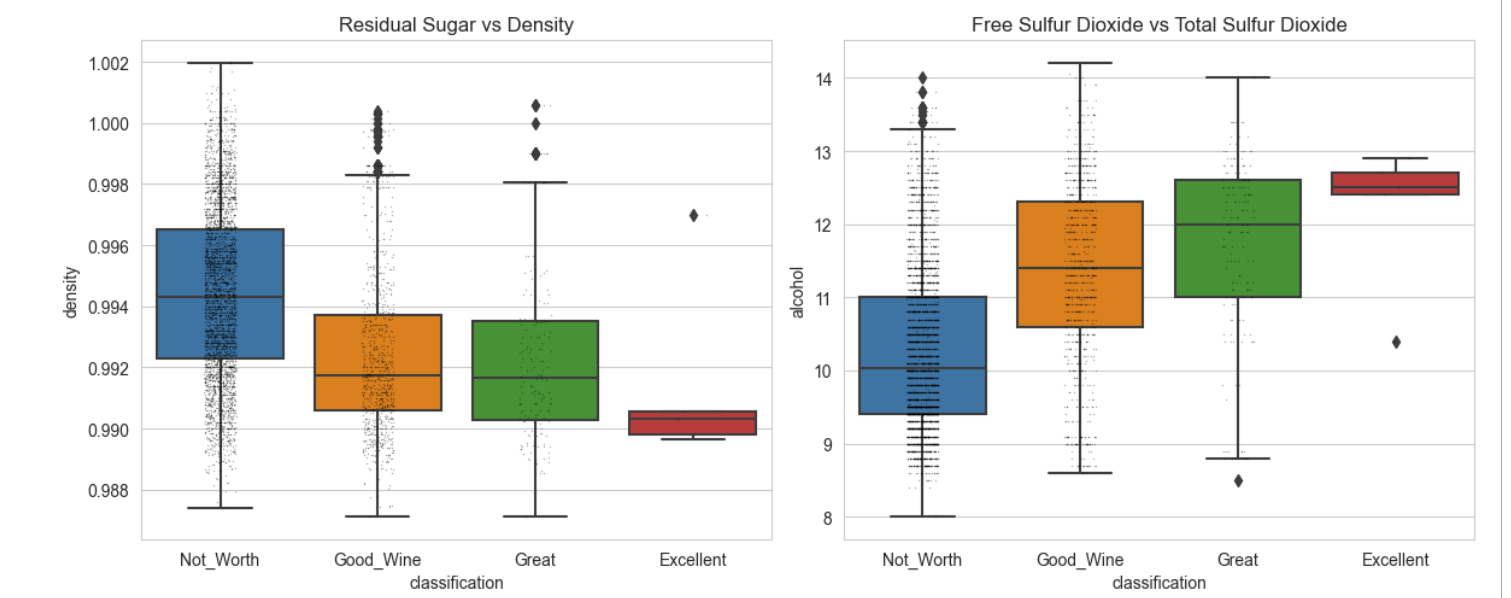
Another good correlation is between alcohol and density. However, it is a negative correlation with (r = -0.78) which is still good.

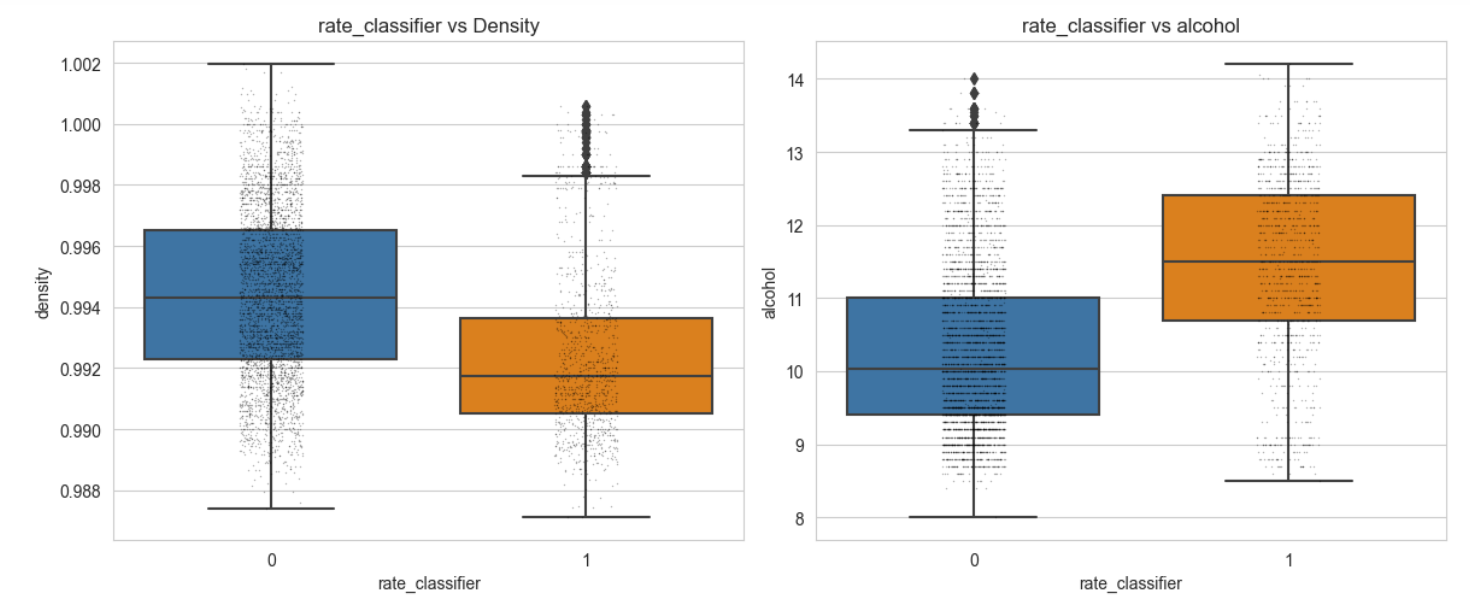
I made a plot with these variables who had a strong correlation and from visualization it can be seen that the quality of wine is better where alcohol is high and the residual sugar as well as density are low.

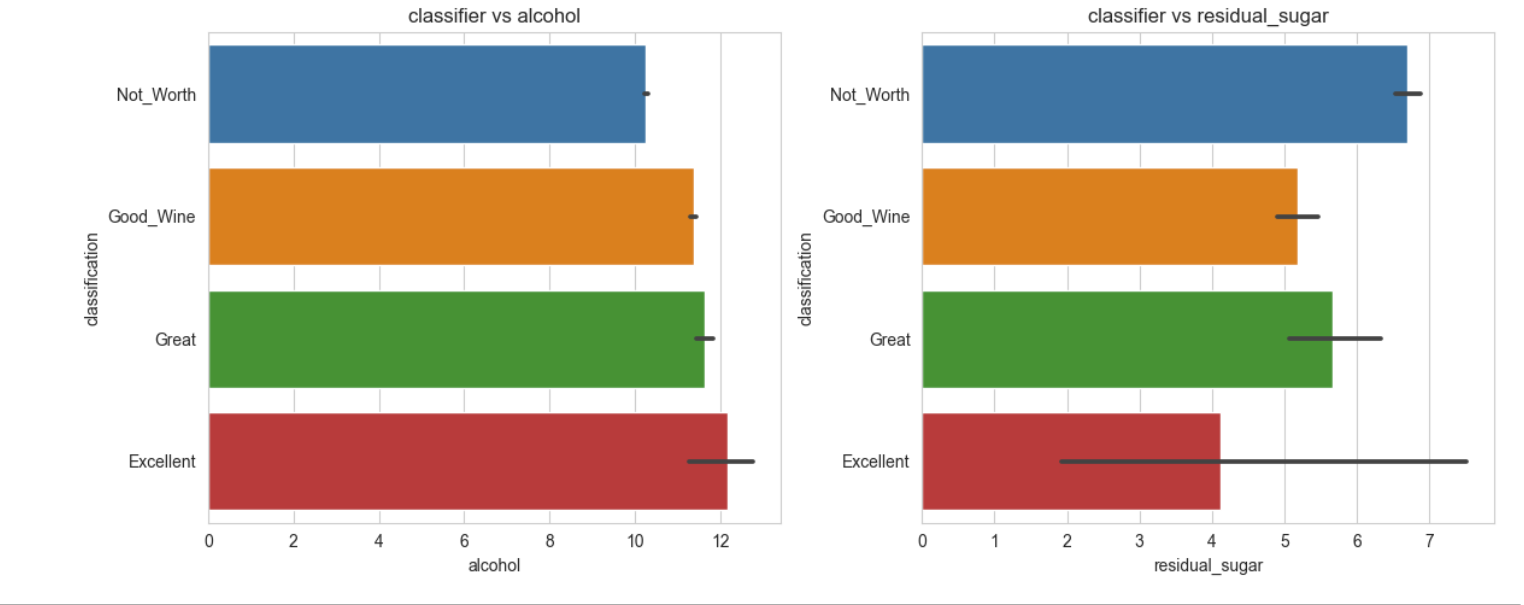
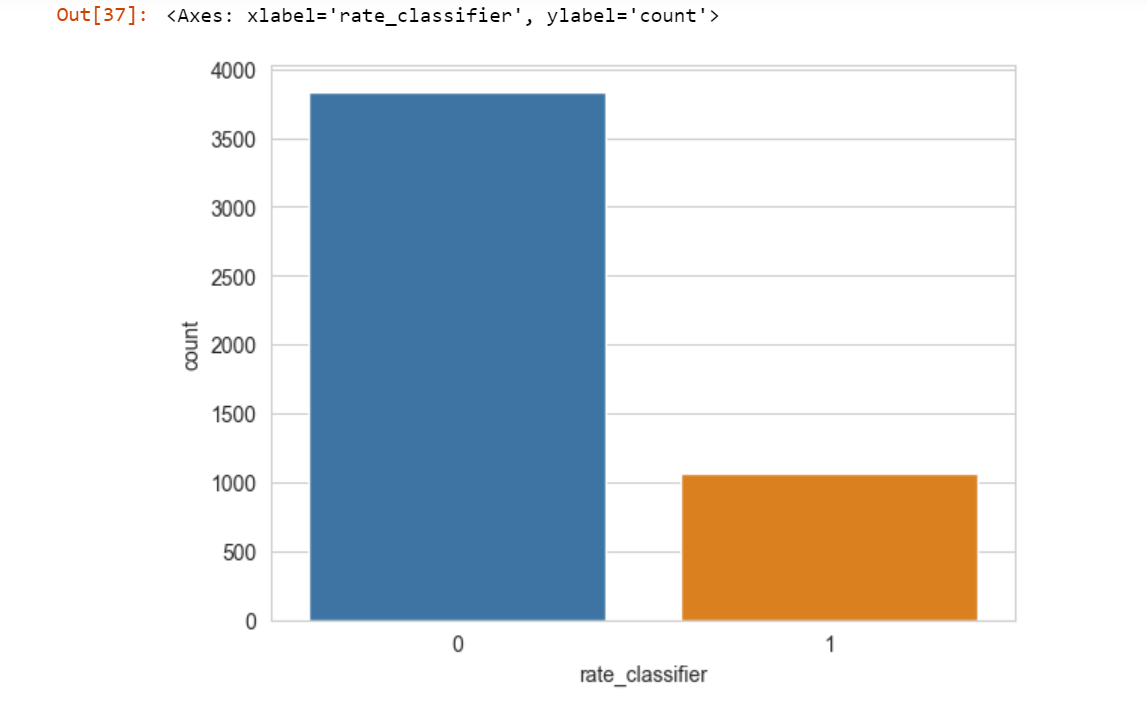
After, I divided into 2 categories for quality of wine starting from 0-6 as one value which will be 0, and from 7-10 as another value which will be 1. In this case we can do binomial logistic regression. I use 0-6 as any rating below 7 is considered a wine which is not worth to make. So, in this column it is said which wine is worth to drink and which is not. I have created another similar column in dataset to identify wine quality by rating, 10 being the highest and anything below 7 is worth to drink. We can understand that most of wines are not worth, the ratio of good wine would be 1050 / (3780+1050). From calculation the percentage of wine which is worth drinking is 21.74%. The 78.26% is wine not worth drinking (3780 / (3780+1050).



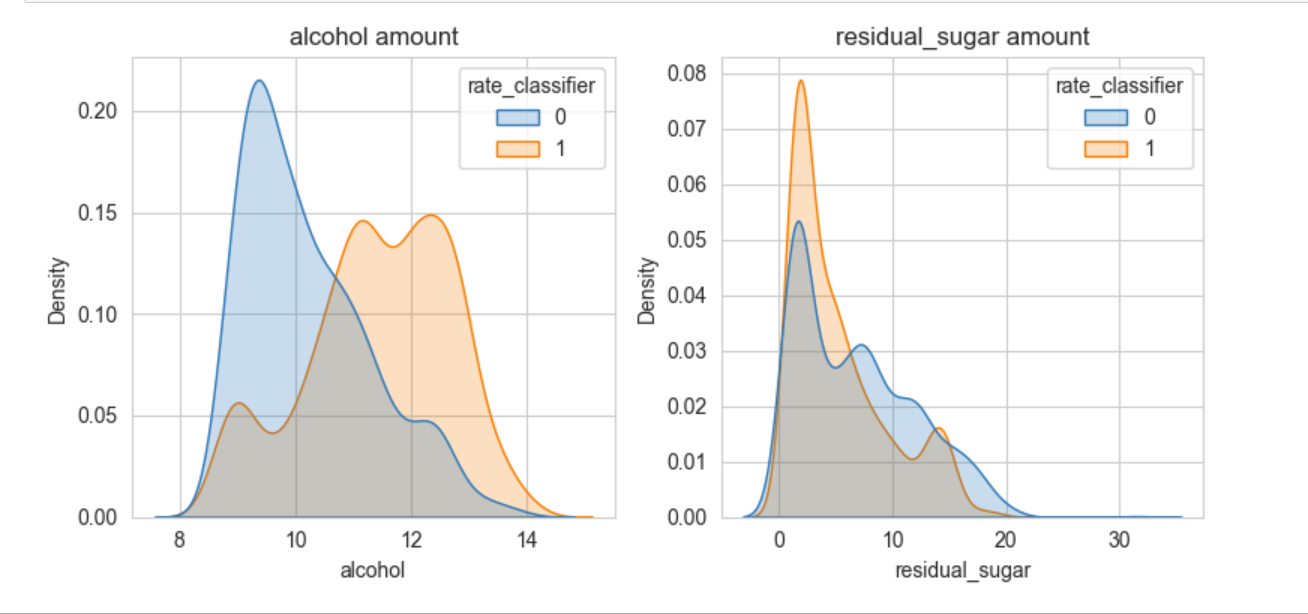
We can try a plot with boxplot to see the same in a different way.





As we have a smaller number of good wines, I picked 1060 random worthless wines so the total number of good wine and worthless wine can be equal, to calculate the sum of alcohol and residual sugar used to make those wines. It is obvious that residual sugar was used much more in wine which is worth and hence it has an important role on the quality of wine, whereas the alcohol amount was more used in wine which is considered good wine. There is approximately 10 percent difference in the alcohol between the classifiers and 22 percent in residual sugar.



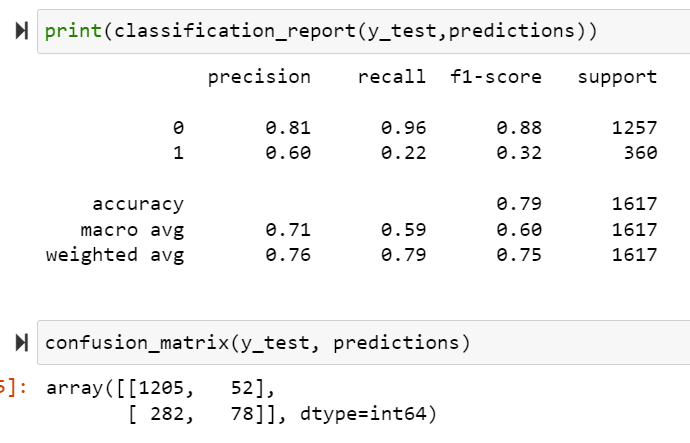
1. As in the data given, we can say that out dependent variable is ‘quality’ which is the rating of wine and we must understand the quality of wine given other X variables. In this case quality is the best dependent variable. As we need 2 binary values with shall use rate classifier variable as dependent variable as we divided quality into 2 groups.

Data needs to be split into two parts – Training and Testing sets. Our test data will be 1/3 of the data. Training set is used to understand the estimation of parameters and testing set is used to check the accuracy of the model.

Recall and Precision shall be high in order to have a good accuracy of the model.

If we need to increase recall, I should set the threshold below 0.5 and for the precision the value can be 0.7 for the model to be accurate.

I used statsmodels and sklearn libraries to go through regression models.

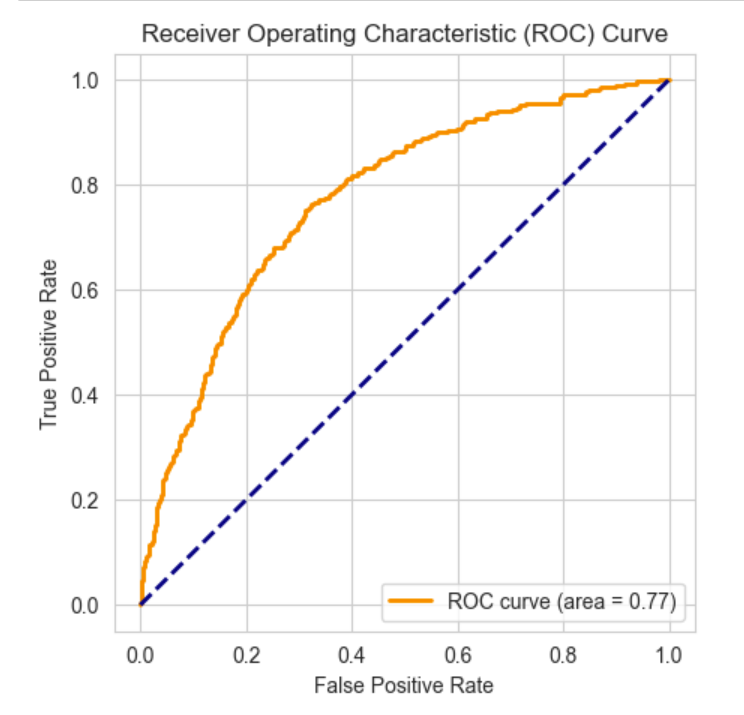


We must check the matrix to see how many observations lay in TP, TN, FN, FP.

As we can see the accuracy of the model is 0.79 which is the True-Positive + True Negative divided by the total number of observations.

For rate\_classifier = 1 we can say that there are high numbers of False Negatives as recall is equal to 0.22 which is lower than 0.5, however for the precision we have a value of 0.60 which is higher than 0.5 which means that we don’t have much numbers as False Positives.

I also built a visualization of these values with a ROC curve. With ROC you can understand how many observations were predicted correctly. Better the area of ROC curves between the line the better the model.



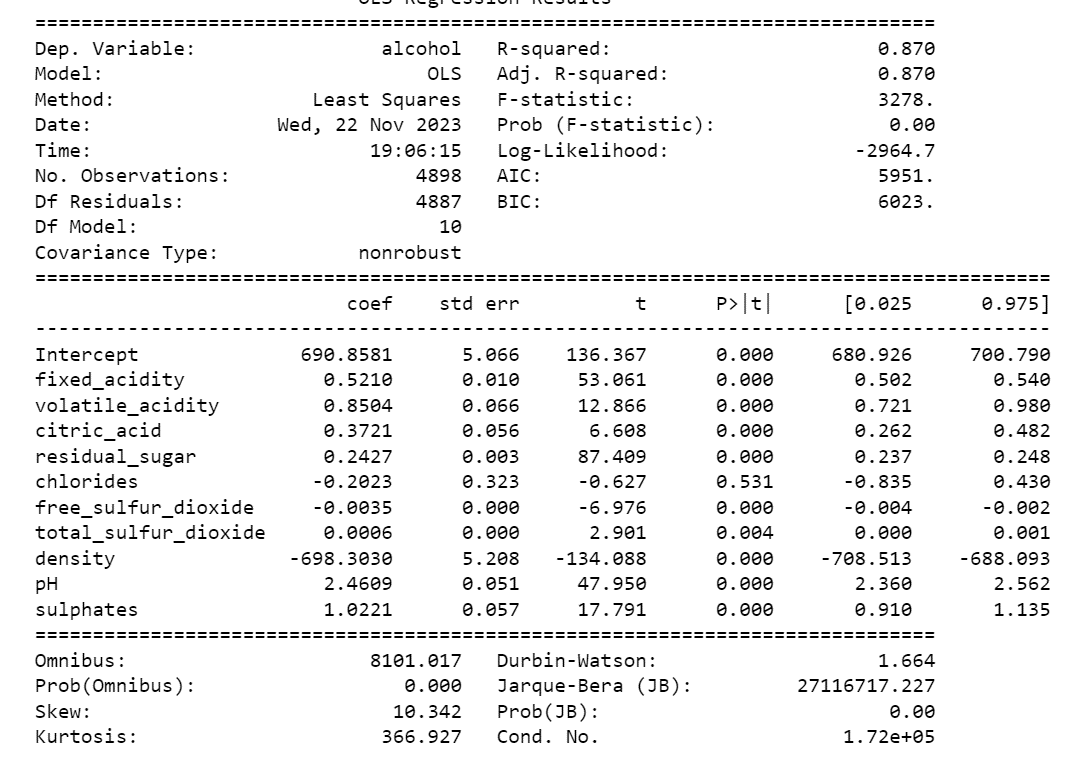
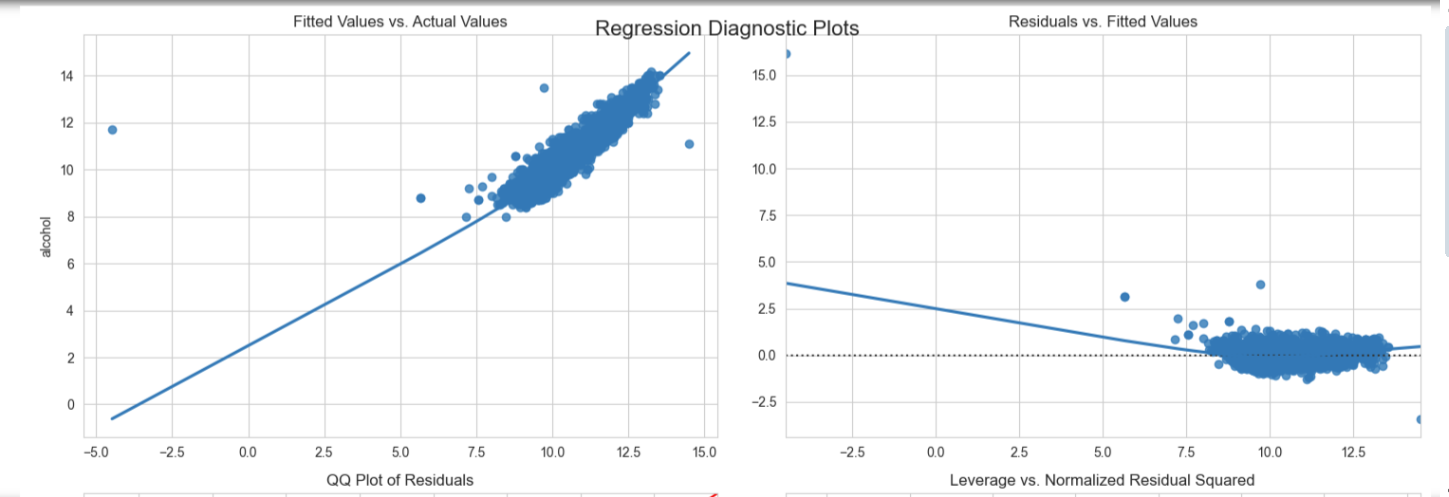
We will also use linear regression to predict the alcohol based on all variables. We can use multiple regression with logarithmic if the residuals are not normally distributed (distribution is skewed) and polynomial terms.

I prefer using statsmodels library as it has amazing plots for regression assumptions.

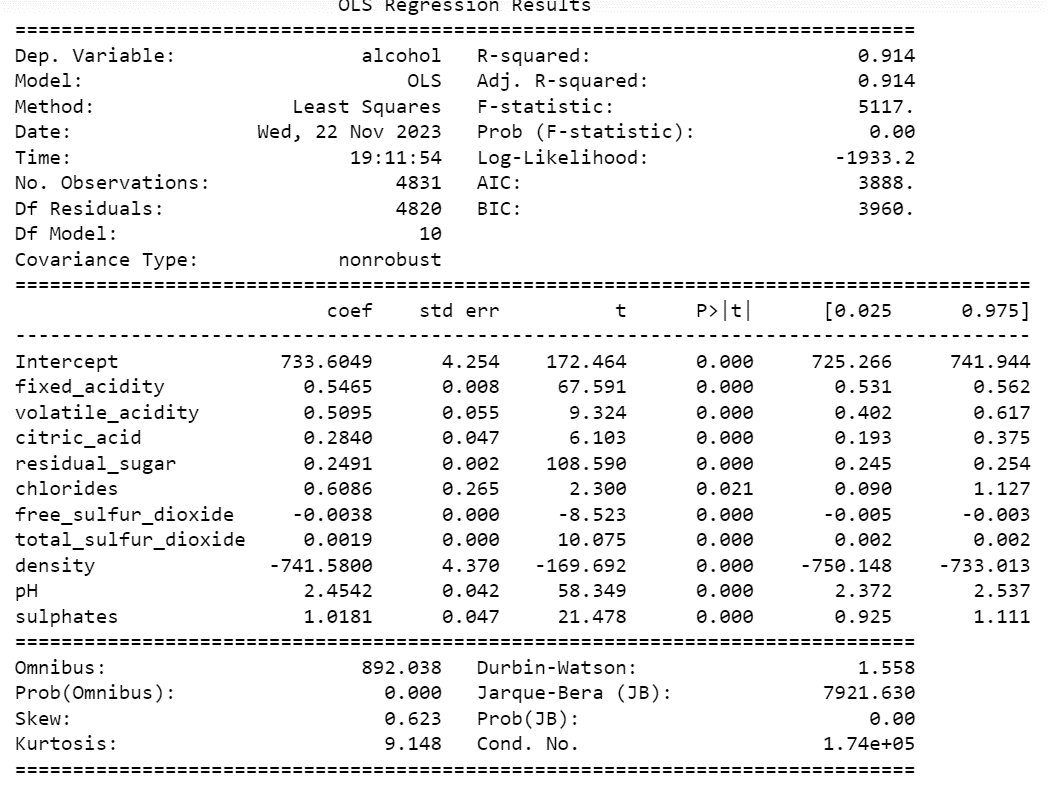
If we also want to predict alcohol. In this case multiple linear regression must be used and alcohol will be our Y variable. In this case outliers will be removed so R squared will be high. Without removing outliers R squared is 0.870. With removed data it is 0.914.

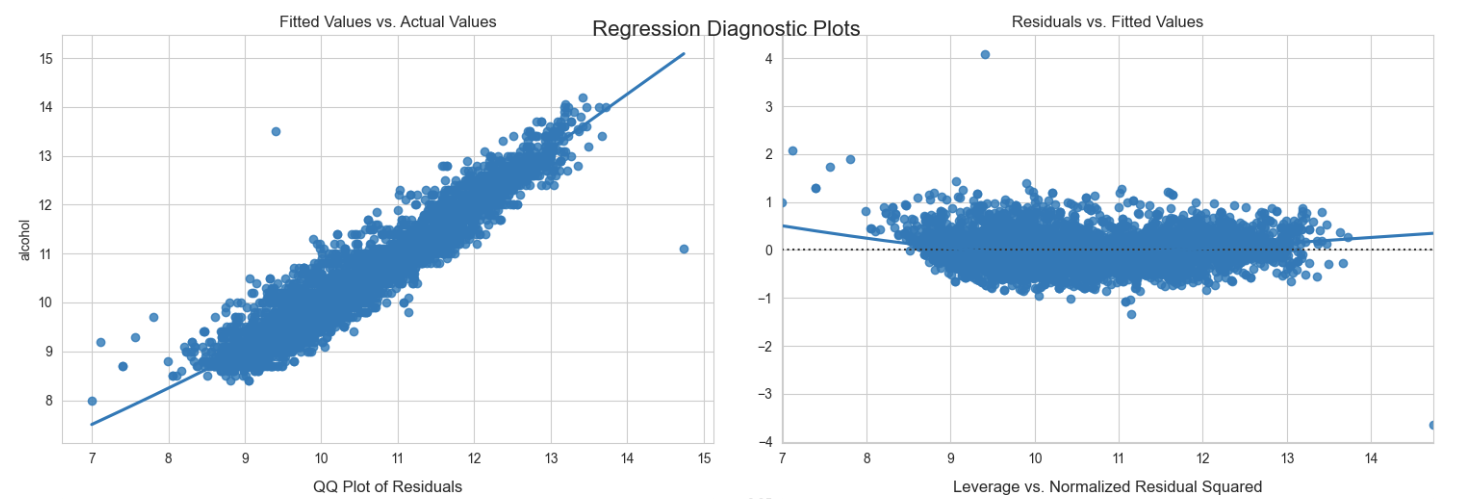
And you can see the assumptions’ change, without outliers it is much clearer.

After I built the regression model, we can see that the coefficients stand for this: For each 1 unit increase in Y (Dependent) variable our independent variable increases by X.

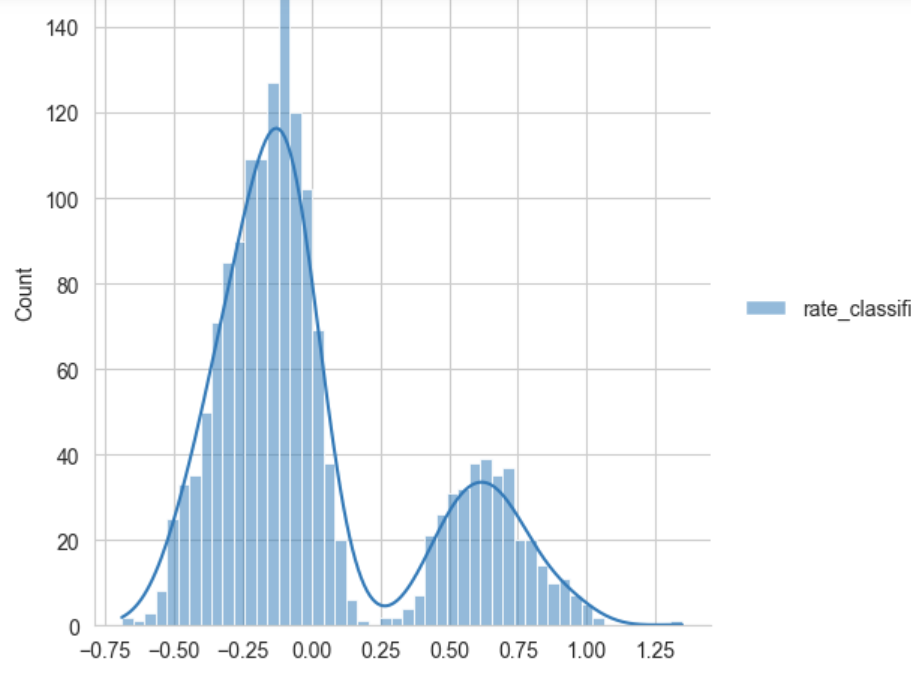
 

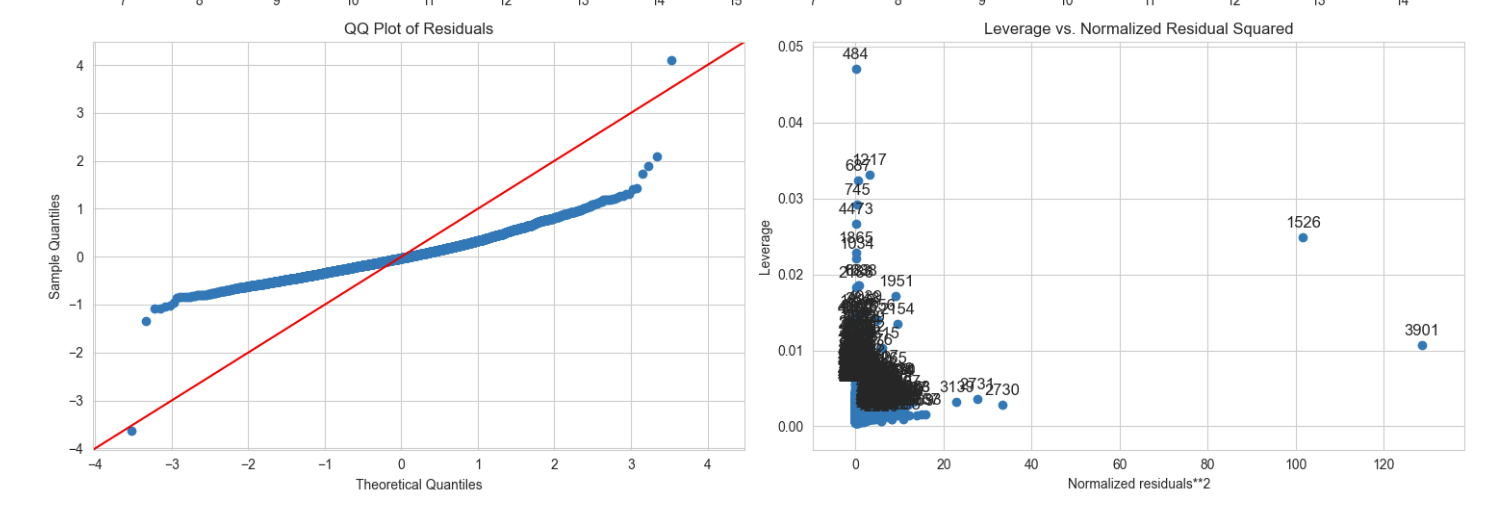
After I removed the outliers, I had these plots and with better model.





We can see that the fitted values align very good with the actual values in the line as well as residuals and fitted values. In this case we have a good linearity of the data and it is homoscedastic. But as for the QQ plot we can say that the data is not normally distributed for dependent variable being the alcohol.



We can see that our histogram shows a not normally distrubuted data and QQ plot verifies it.

Useful links:

http://www.sthda.com/english/articles/36-classification-methods-essentials/151-logistic-regression-essentials-in-r/

https://www.lexjansen.com/wuss/2018/130\_Final\_Paper\_PDF.pdf

https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009

https://www.simplilearn.com/tutorials/machine-learning-tutorial/linear-regression-vs-logistic-regression

https://www.javatpoint.com/linear-regression-vs-logistic-regression-in-machine-learning#:~:text=Linear%20Regression%20is%20used%20for%20solving%20Regression%20problem.,the%20values%20of%20categorical%20variables.

https://www.youtube.com/watch?v=yIYKR4sgzI8&t=385s

https://www.youtube.com/watch?v=nk2CQITm\_eo&t=14s

https://www.lexjansen.com/wuss/2018/130\_Final\_Paper\_PDF.pdf

https://www.geeksforgeeks.org/linear-regression-python-implementation/