Statistics Report

Analysis of Wine data

Mamidi Ratna Abhishek

Problem Definition

A wine dataset is provided. The task is to analyze data and build a regression model to predict the quality of the wine.

Abstract

The main goal of this report is to extract maximum knowledge from the Wine data in different ways. The data is analyzed and the plots are shown. A regression model is built to predict the quality of wine using the features provided. The assumptions of regression are also checked.

Methodology

1. Description of data
2. Preprocess data
3. Visualize data
4. Build a Regression model
5. Check Regression Assumptions
6. Goodness of fit

Description of data

1. **Name of the data**: Wine data from UCI Machine learning repository
2. **Number of data points**: 4898
3. **Number of features**: 11
4. **Target attribute**: Quality of wine
5. **Range of target attribute:** 3 to 9

Data

fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide
0.308	0.186	0.217	0.308	0.107	0.15
0.24	0.216	0.205	0.015	0.119	0.042
0.413	0.196	0.241	0.097	0.122	0.098
0.327	0.147	0.193	0.121	0.145	0.157
0.327	0.147	0.193	0.121	0.145	0.157 Export to plot.ly »

total sulfur dioxide	density	pH	sulphates	alcohol	quality
0.374	0.268	0.255	0.267	0.129	6.0
0.285	0.133	0.527	0.314	0.242	6.0
0.204	0.154	0.491	0.256	0.339	6.0
0.411	0.164	0.427	0.209	0.306	6.0
0.411	0.164	0.427	0.209	0.306	6.0 Export to plot.ly »

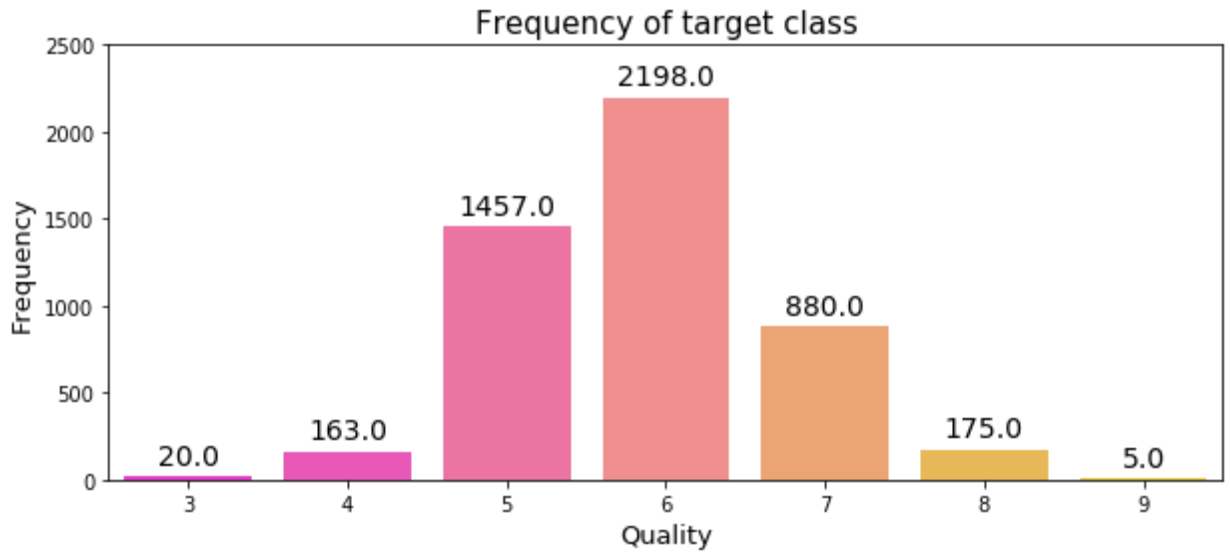
Features

1. Fixed acidity
2. Volatile acidity
3. Citric acid
4. Residual sugar
5. Chlorides
6. Free sulfur dioxide
7. Total sulfur dioxide
8. Density
9. pH
10. Sulphates
11. Alcohol

Target Attribute

- Quality of wine

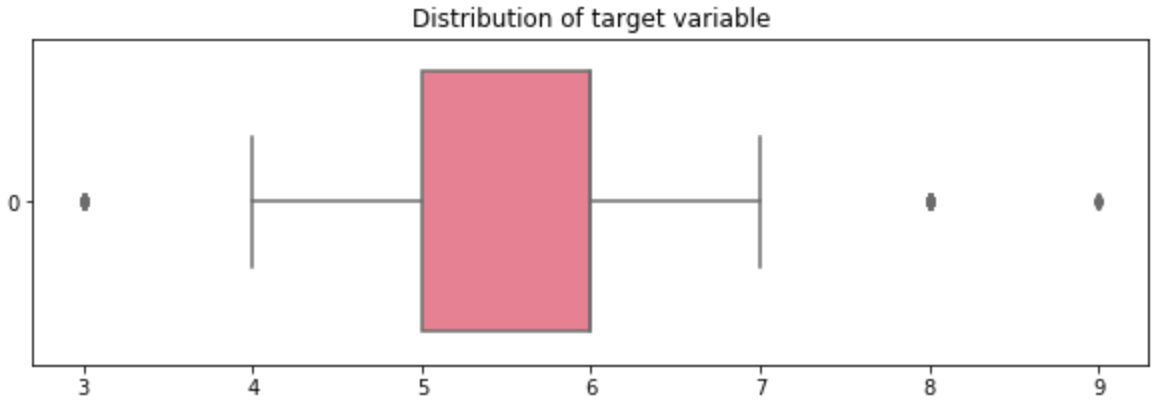
Distribution of Target Attribute



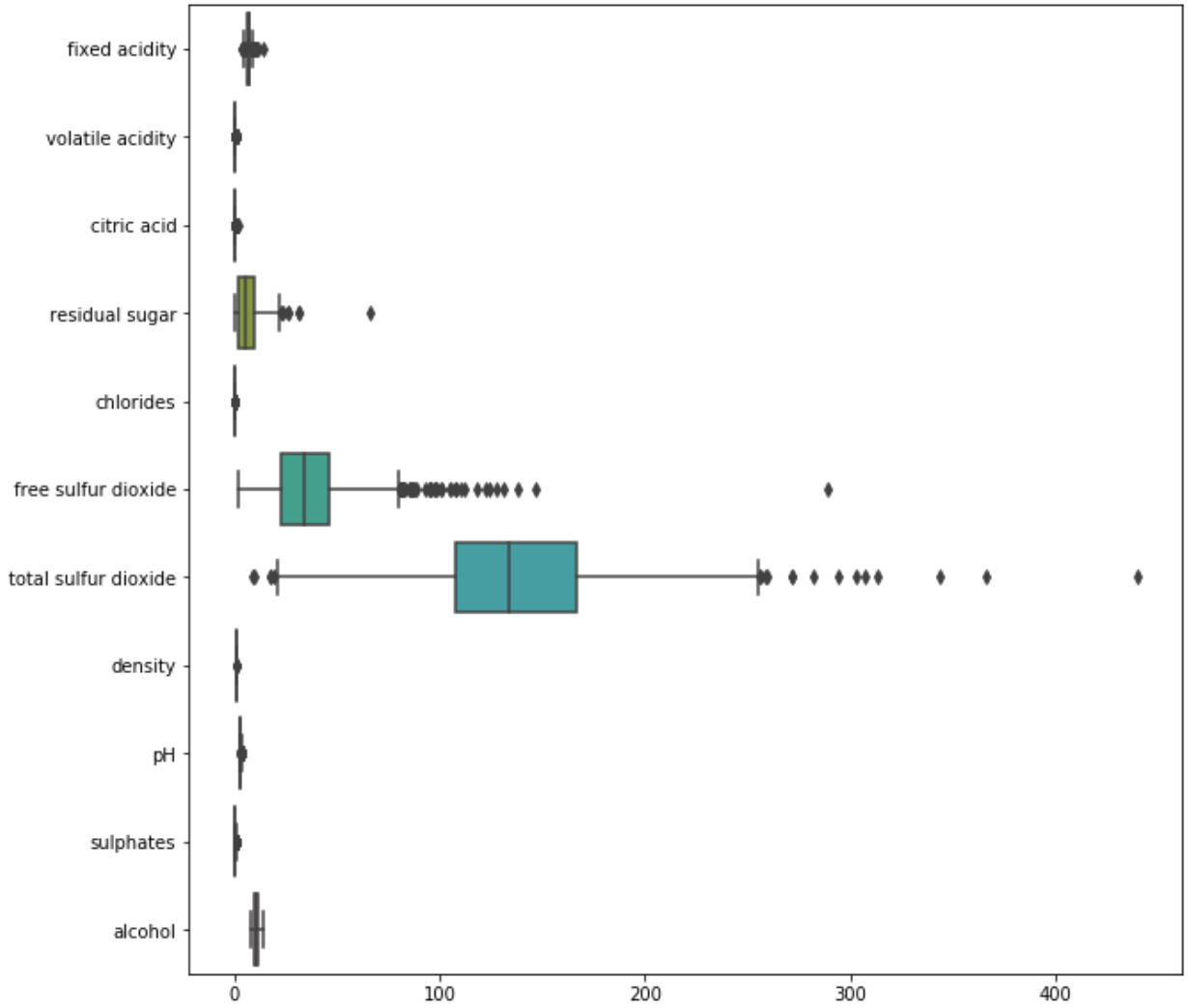
- The quality of wine ranges from 3 to 9
- The data is not balanced. The number of data points having quality 6 is very high and quality 3 and 9 are very low.
- This may affect the model.

Statistics of Target Attribute

Index	Description
count	4898.0
mean	5.87790935075541
std	0.8856385749678312
min	3.0
25%	5.0
50%	6.0
75%	6.0
max	9.0



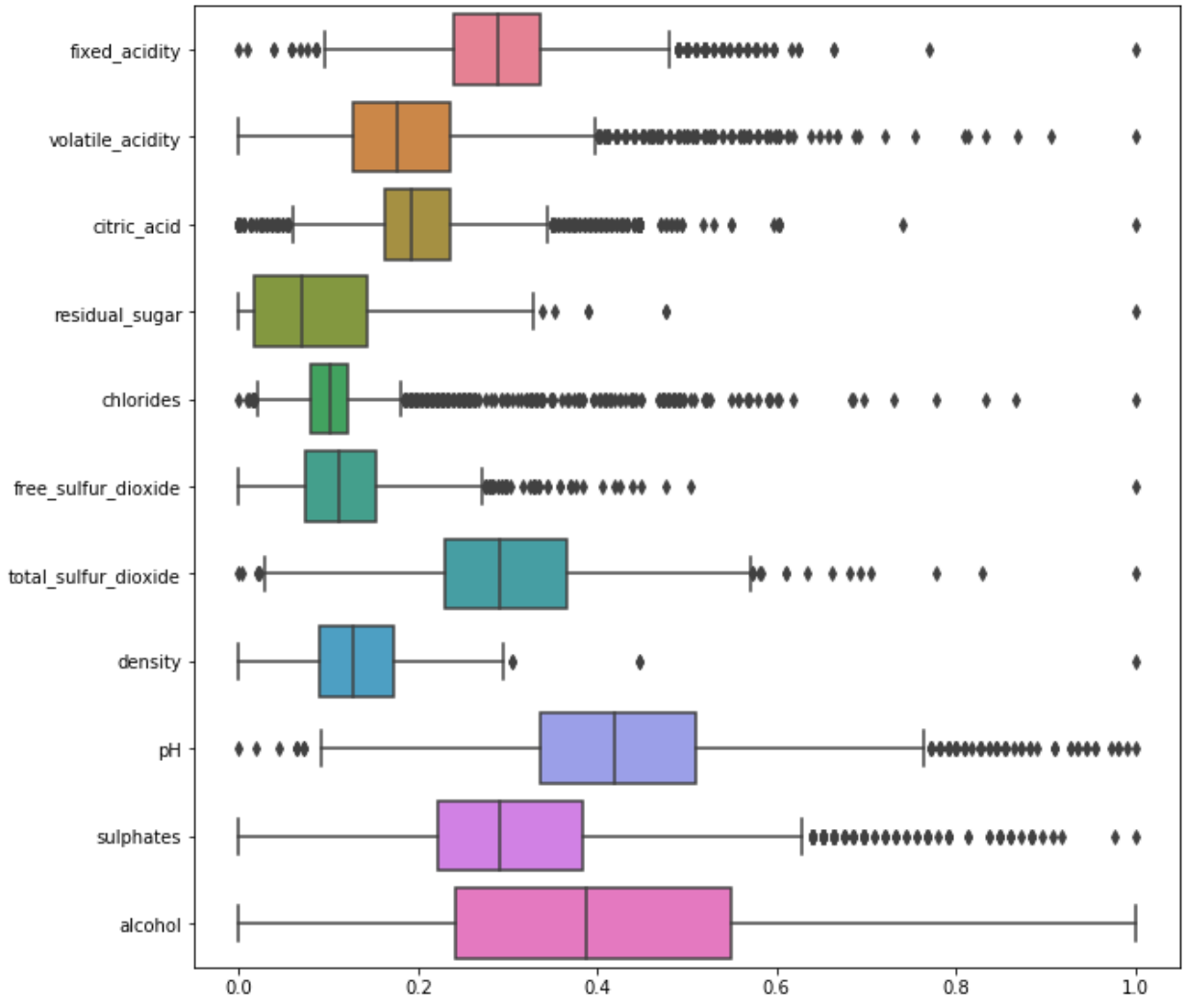
Preprocess data



- If we observe the above boxplot, the range of features is different from each other.
- We can normalize the data. All the variables range from 0 to 1 after normalization and don't lose any information.

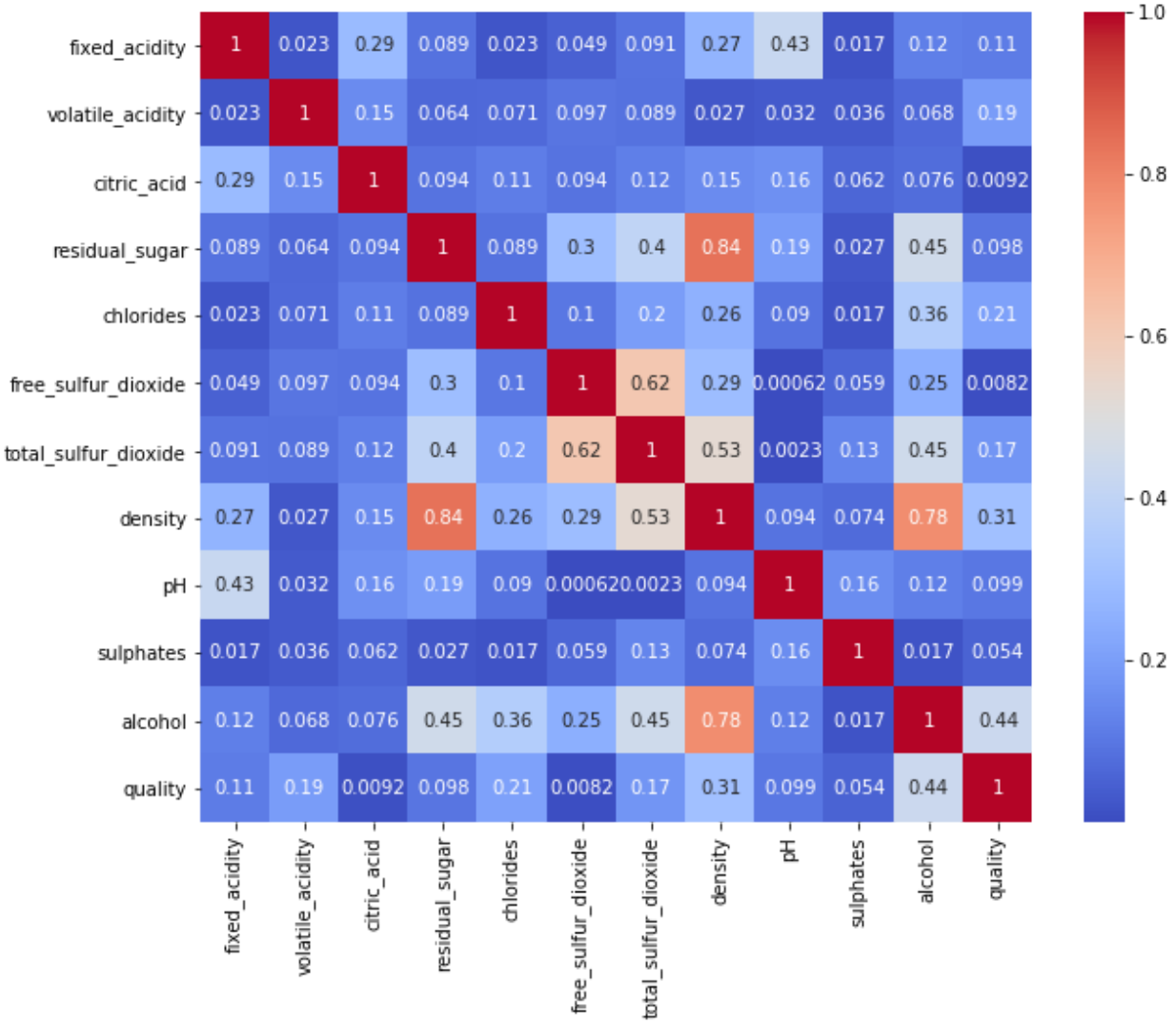
Distribution of features - After normalization

- This looks good than before and very easy to understand the distribution of data.



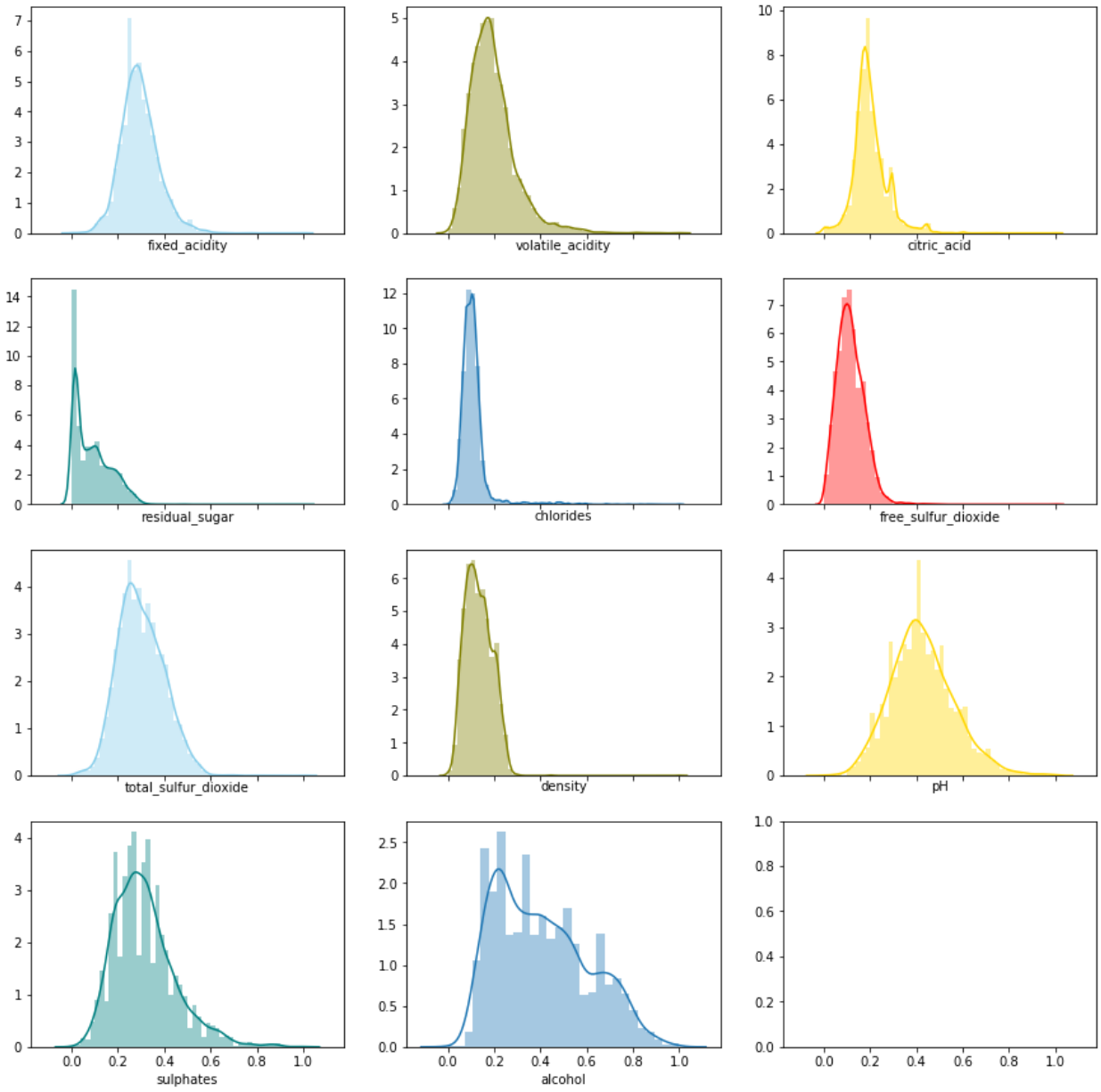
Visualize data

Correlation between features



- The correlation between “density” and “residual sugar” is 0.84.
- The correlation between “alcohol” and “density” is 0.78.
- The correlation between “total sulfur dioxide” and “free sulfur dioxide” is 0.62.
- These are the three pairs of features having a high correlation(>0.5).

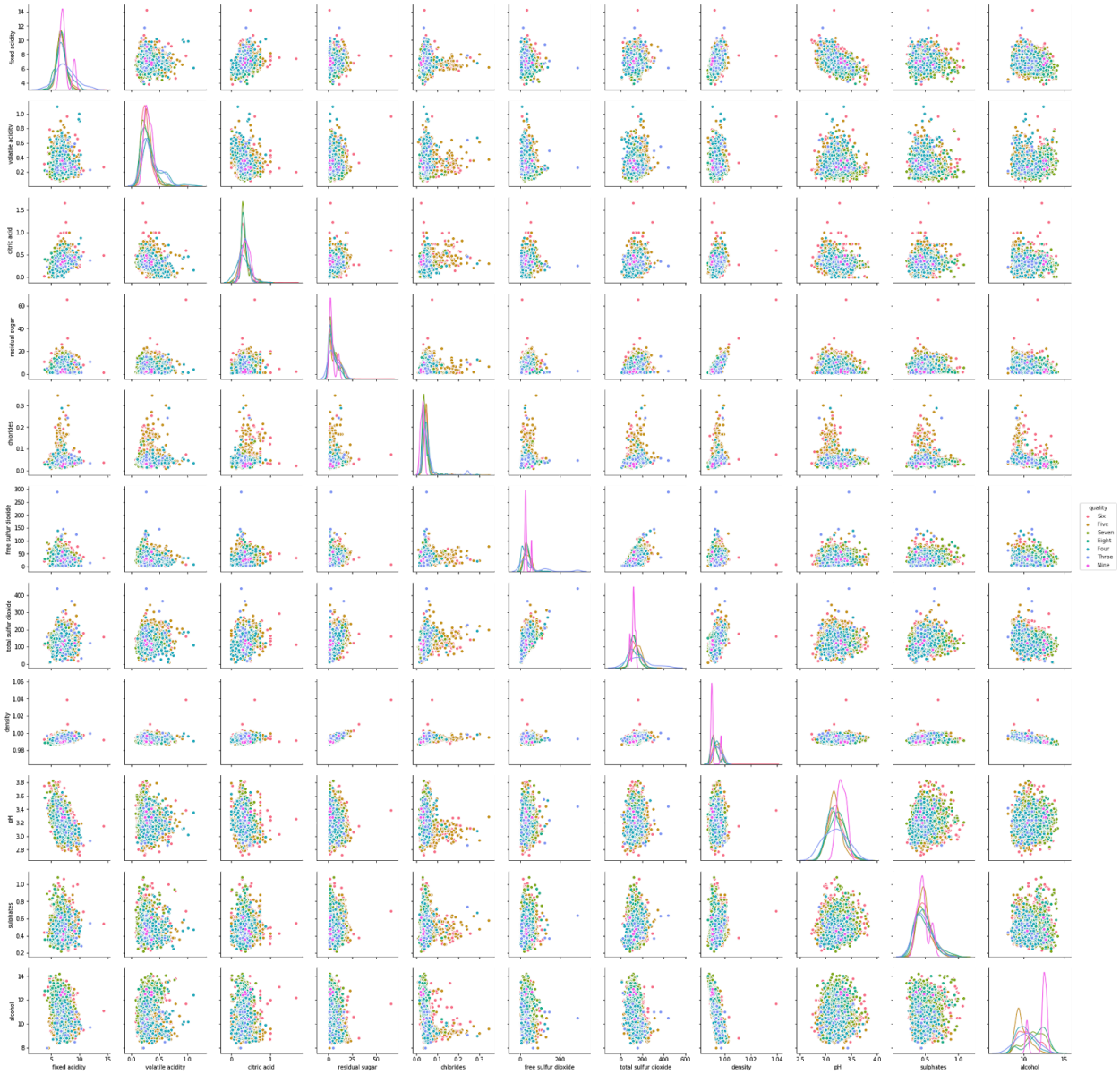
Distribution of each feature



- If we observe the distribution of all features, they follow a Normal distribution.
- There is some fluctuation in the features “sulphates” and “alcohol”.

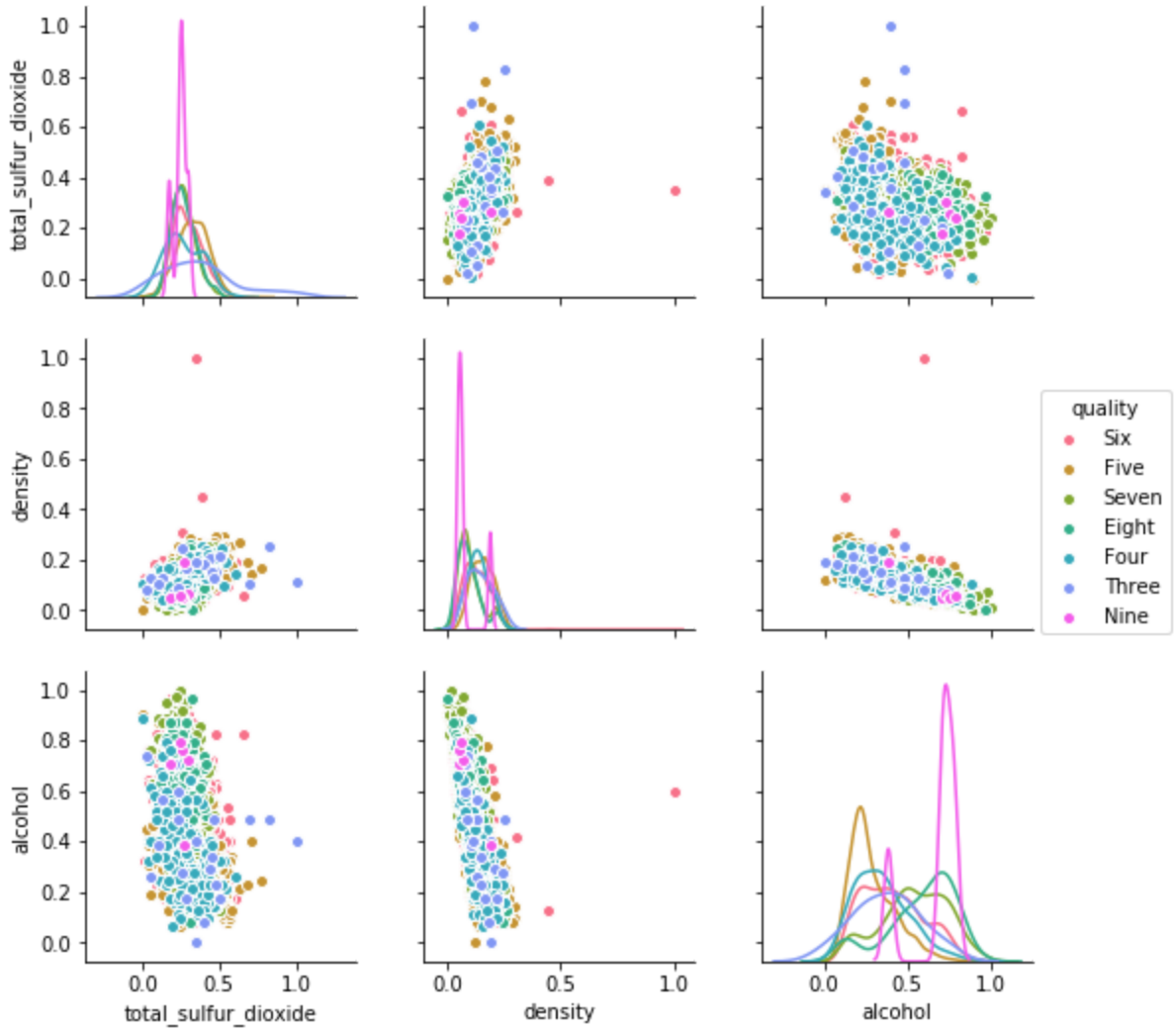
Pair plot between features

- This is to understand the relation between features.



- From this plot, we can see how different features are correlated with each other.
- In the above plot, the features that are plotted on the x-axis and y-axis are in the given order itself.

Pair plot between correlated features

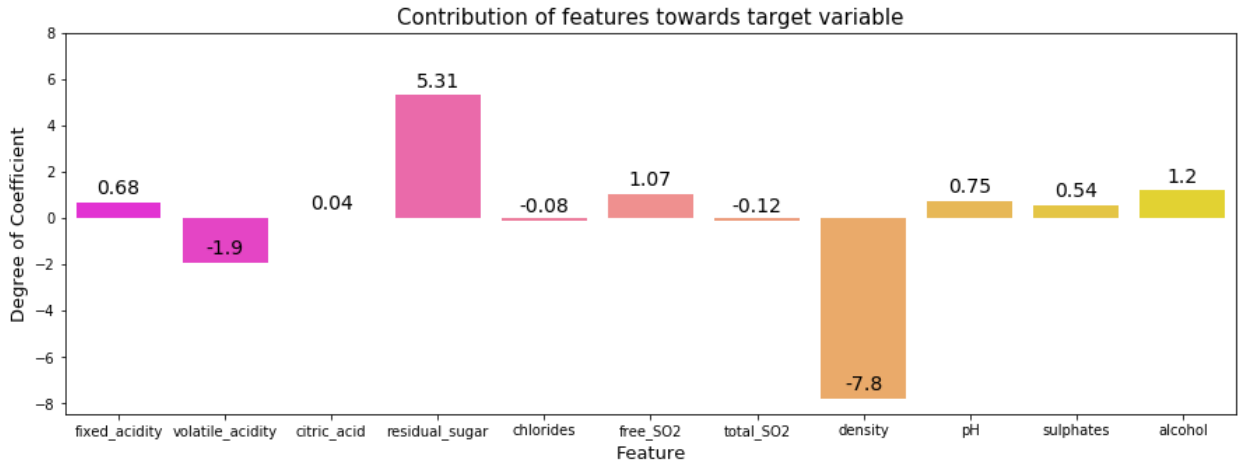


- As we have seen above in the correlation plot, there is a high correlation(>0.5) in between some of the features.
- Here, we can visualize how these features are correlated.
- If we observe carefully, we cannot separate the data points of different quality easily, because all the data points of various quality are overlapped.

Build a Regression model

Linear Regression using Gradient Descent:

- The method of Linear Regression that finds the coefficients of different features using Gradient Descent optimization, is fit to the data to see how independent variables are contributing to the dependent variable.
- The below plot shows the coefficients of features(contribution).



- If we observe, the coefficient of density is 7.8(absolute value), citric acid is 0.04, chlorides is 0.08(absolute value) and total sulfur dioxide is 0.12(absolute value).
- This is to understand the contribution of different features.

Ordinary Least Squares(OLS):

- In statistics, ordinary least squares (OLS) is a type of linear least squares method for estimating the unknown parameters in a linear regression model.
- The OLS method corresponds to minimizing the sum of squared differences between the observed and predicted values. This minimization leads to the estimators of the parameters of the model.
- The results of OLS Regression are shown below:

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.282
Model:	OLS	**Adj. R-squared:**	0.280
Method:	Least Squares	**F-statistic:**	174.3
Date:	Fri, 30 Nov 2018	**Prob (F-statistic):**	0.00
Time:	01:53:26	**Log-Likelihood:**	-5543.7
No. Observations:	4898	**AIC:**	1.111e+04
Df Residuals:	4886	**BIC:**	1.119e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.5509	0.107	51.650	0.000	5.340	5.762
fixed_acidity	0.6814	0.217	3.139	0.002	0.256	1.107
volatile_acidity	-1.9004	0.116	-16.373	0.000	-2.128	-1.673
citric_acid	0.0367	0.159	0.231	0.818	-0.275	0.348
residual_sugar	5.3127	0.491	10.825	0.000	4.351	6.275
chlorides	-0.0833	0.184	-0.452	0.651	-0.444	0.278
free_sulfur_dioxide	1.0713	0.242	4.422	0.000	0.596	1.546
total_sulfur_dioxide	-0.1232	0.163	-0.756	0.450	-0.443	0.196
density	-7.7952	0.989	-7.879	0.000	-9.735	-5.856
pH	0.7550	0.116	6.513	0.000	0.528	0.982
sulphates	0.5431	0.086	6.291	0.000	0.374	0.712
alcohol	1.1995	0.150	7.988	0.000	0.905	1.494

- The R-squared is 0.282 and Adjusted R-squared is 0.280.
- If p-value > 0.05, we fail to reject the null hypothesis, otherwise we reject the null hypothesis.

- The p-values of the features “citric acid” and “chlorides”, is greater than 0.05. Also, the contribution of these features is very less.
- So, we can remove the remove the features from the data.
- Let’s fit the model again and see if there would be any change. The results of OLS Regression are shown after the removal of these two features from the data.

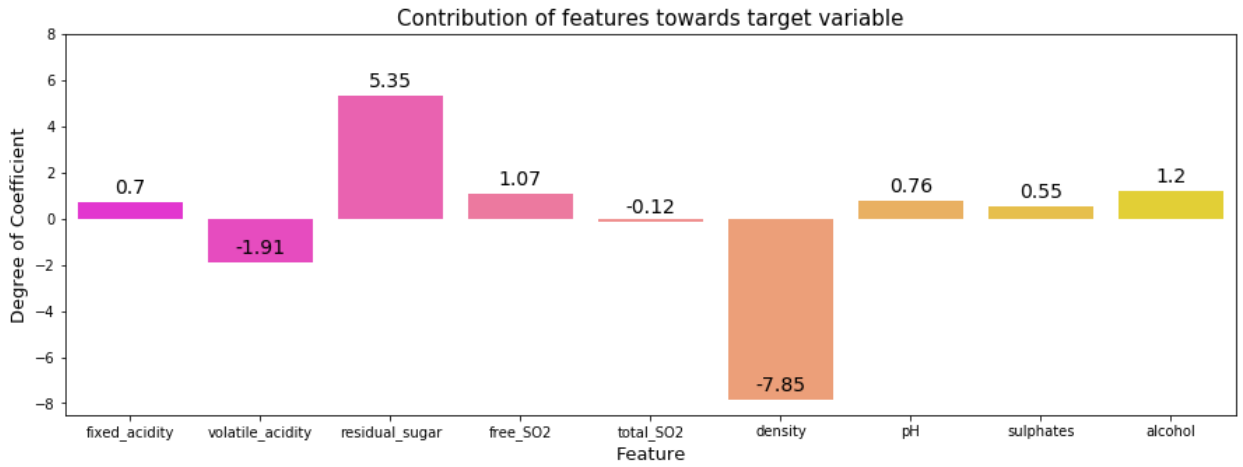
OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.282
Model:	OLS	**Adj. R-squared:**	0.281
Method:	Least Squares	**F-statistic:**	213.1
Date:	Fri, 30 Nov 2018	**Prob (F-statistic):**	0.00
Time:	01:53:43	**Log-Likelihood:**	-5543.9
No. Observations:	4898	**AIC:**	1.111e+04
Df Residuals:	4888	**BIC:**	1.117e+04
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.5444	0.104	53.160	0.000	5.340	5.749
fixed_acidity	0.7024	0.213	3.303	0.001	0.286	1.119
volatile_acidity	-1.9091	0.114	-16.761	0.000	-2.132	-1.686
residual_sugar	5.3505	0.480	11.148	0.000	4.410	6.291
free_sulfur_dioxide	1.0699	0.242	4.422	0.000	0.596	1.544
total_sulfur_dioxide	-0.1226	0.163	-0.753	0.451	-0.442	0.196
density	-7.8530	0.972	-8.078	0.000	-9.759	-5.947
pH	0.7614	0.114	6.695	0.000	0.538	0.984
sulphates	0.5451	0.086	6.324	0.000	0.376	0.714
alcohol	1.2027	0.149	8.046	0.000	0.910	1.496

- There is no change in the values of R-squared and there is an increase of 0.001 of Adjusted R-squared. So, there is no harm in removing those features from the data. Also, the contribution of these features to predict the quality of the wine is very less as shown before. Now, we are left with 9 features.

Contribution of features after removing two features



- The coefficients of features have changed a little bit.

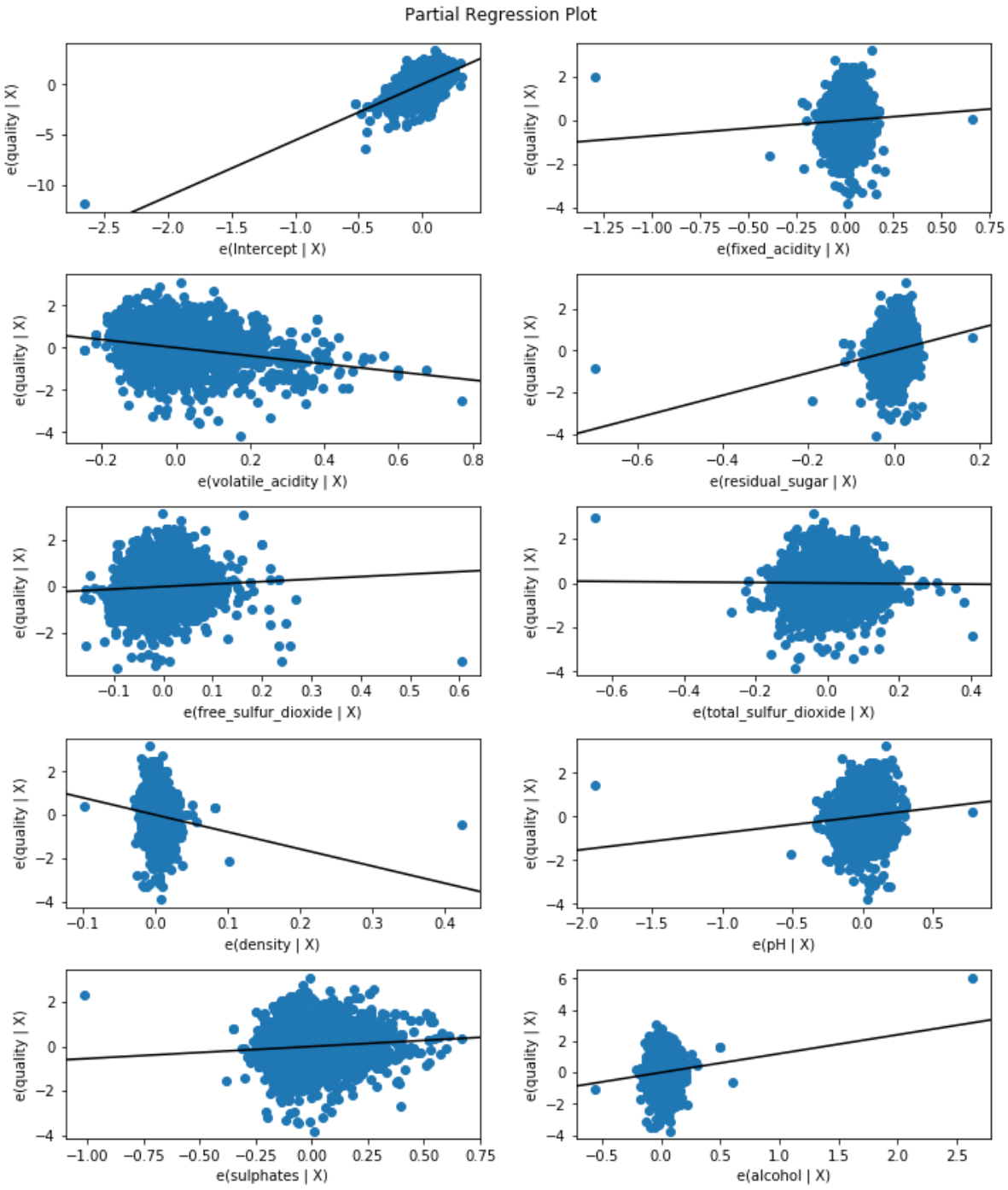
Check Regression Assumptions

1. Linearity
2. Homoscedasticity
3. Correlation of errors
4. Normality of errors.

- Let's check each condition using the predicted values and the errors/residuals.
- Residuals are the difference between "true value" and the "predicted value".

Note: **The two features “chlorides” and “citric acid” are removed from the data.**

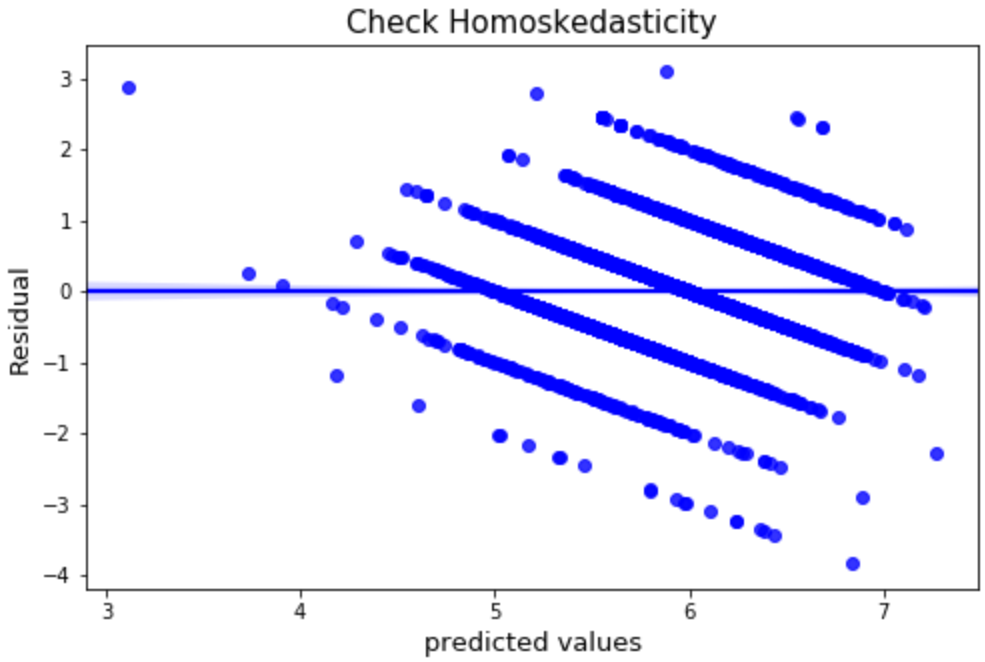
Linearity



- If we observe carefully, all the partial residual plots between the independent variable and dependent variable are linear.
- Linearity condition is satisfied.

Homoskedasticity

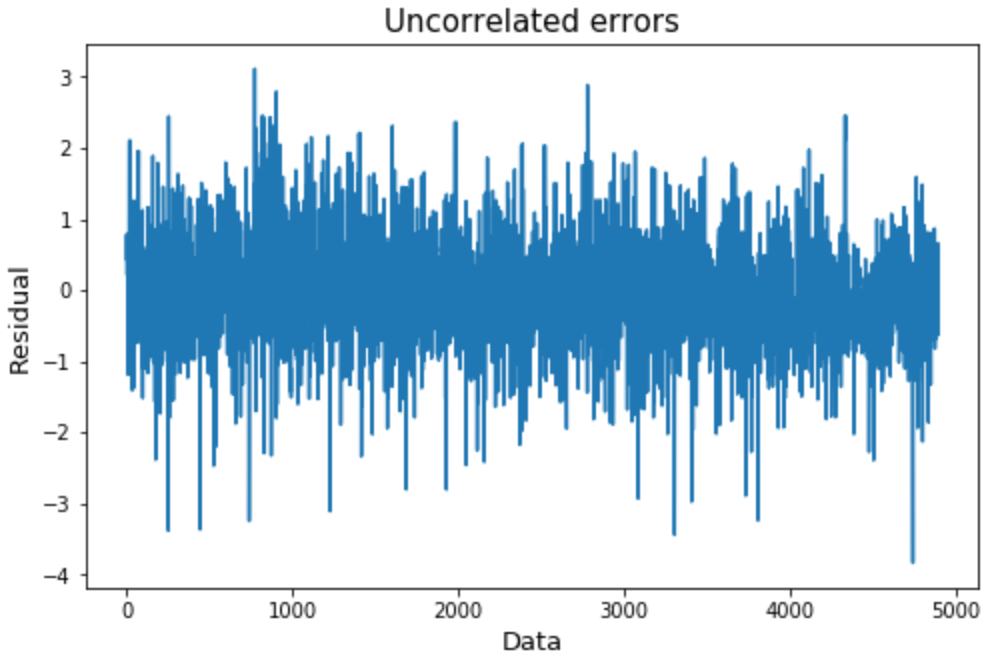
- To check homoskedasticity, we plot the residuals vs predicted values/fitted values.
- If we see any kind of funnel shape, we can say that there is heteroskedasticity.



- The points are not random. Also, we can see the shape of a funnel to the right, which confirms that there is heteroskedasticity.
- It means that the variance of Y across all X is not the same.
- We can conclude that, Homoskedasticity condition doesn't hold in this case.

Correlation of errors

- If there is no correlation between errors, then the model is good.

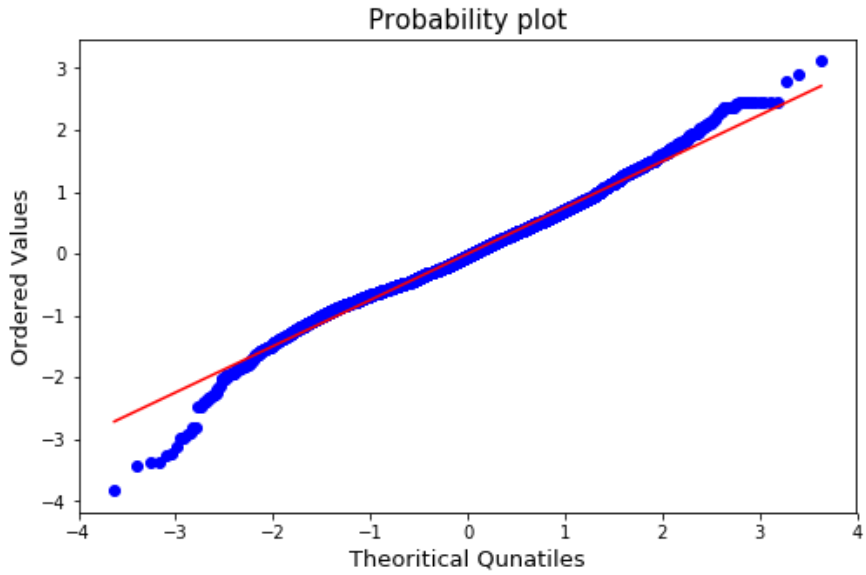


- If we observe, there is no correlation/pattern between errors. It is purely random.
- We can also check this condition using the Durbin-Watson test:
    - If DW = 2, then there is no correlation.
    - If DW < 2, then the errors are positively correlated.
    - If DW > 2, then the errors are negatively correlated.
- If we perform Durbin-Watson test, the value of DW is 1.621.
- According to the test, we can say that the errors are positively correlated.
- However, this is a point estimate for perfect uncorrelation of errors(DW=2). So, we won't get DW as 2 on real data. If it around 2, then we can conclude that the errors are uncorrelated.

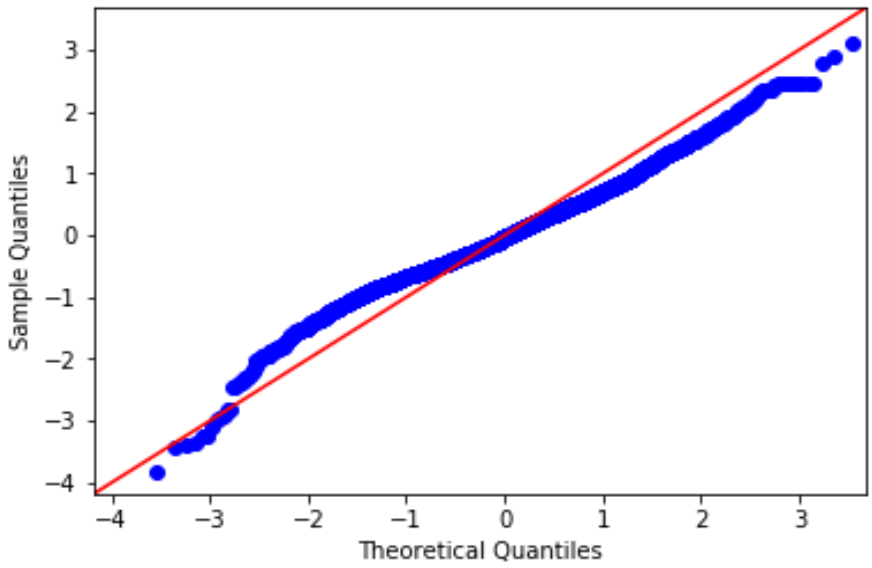
Normality of error terms

- This can be checked by plotting probability probability plot(p-p plot) or Quantile-Quantile plot(Q-Q plot).

Probability-Probability plot



Quantile-Quantile plot



- If we observe the above plots, we can conclude that the errors are following a Normal distribution, because the plot shows the fluctuation around the line and there is not much deviation.
- The graph is linear.

Linear Regression Assumption: Multicollinearity

- If the independent variables are independent of each other, then we say there is no multicollinearity.
- This can be tested in different ways:
    1. **Correlation plot**: If we observe the plot, there is multicollinearity between variables.
    2. **Variation Inflation Factor:** With VIF > 10 there is an indication that multicollinearity may be present. With VIF > 100 there is certainly multicollinearity among the variables.
- We can conclude that multicollinearity among variables exists.
- If multicollinearity is found in the data, centring the data, that is deducting the mean score might help to solve the problem. Other alternatives to tackle the problems is conducting a **factor analysis/Principal Component Analysis(PCA)** and rotating the factors to ensure the independence of the factors in the linear regression analysis.
- We can do the same analysis after applying PCA on the data. We can see some improvements in the model as there won't be any multicollinearity.
- The results of OLS Regression are shown below after transforming the feature variables using PCA.

OLS Regression Results

Dep. Variable:	quality	**R-squared:**	0.282
Model:	OLS	**Adj. R-squared:**	0.280
Method:	Least Squares	**F-statistic:**	174.3
Date:	Fri, 30 Nov 2018	**Prob (F-statistic):**	0.00
Time:	10:47:45	**Log-Likelihood:**	-5543.7
No. Observations:	4898	**AIC:**	1.111e+04
Df Residuals:	4886	**BIC:**	1.119e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	5.8779	0.011	547.502	0.000	5.857	5.899
feature_1	1.6886	0.049	34.258	0.000	1.592	1.785
feature_2	0.2988	0.072	4.145	0.000	0.157	0.440
feature_3	0.4878	0.085	5.753	0.000	0.322	0.654
feature_4	-1.0113	0.104	-9.735	0.000	-1.215	-0.808
feature_5	2.0724	0.116	17.803	0.000	1.844	2.301
feature_6	-1.3769	0.142	-9.693	0.000	-1.655	-1.098
feature_7	1.3978	0.154	9.093	0.000	1.096	1.699
feature_8	0.6638	0.167	3.973	0.000	0.336	0.991
feature_9	0.6649	0.195	3.406	0.001	0.282	1.048
feature_10	0.9257	0.266	3.477	0.001	0.404	1.448
feature_11	-9.0805	1.117	-8.127	0.000	-11.271	-6.890

- If we observe the p-values of transformed features, all the p-values are less than 0.05, which shows that multicollinearity problem is solved.

Discussion

- We can do alot of changes to improve the accuracy of the model. Some of the conditions above are violated. If we can transform the variables accordingly, we can achieve good results. If we observe the R-squared score, it is 0.282. It is able to explain only 28% of the variance, which is poor. So, there is a scope to apply different methods to get better models.
- I have applied different popular Regression methods on the data to compare the results we got. The below table shows the comparison of the R-squared of different methods.

Method name	R-squared
Linear Regression	0.28200
Ridge Regression	0.277162
KNN	**0.597895**
Decision Tree	0.418447
Bayesian Regression	0.281631
SVM	0.275995

- If we compare R-square, KNN outperformed on all the Regression methods. Also, all the methods performed better than LinearRegression. So, we can conclude there is alot of scope to improve the Linear Regression model.

Conclusion

- We have visualized wine dataset in all possible ways and they are shown in the form of plots.
- A Linear Regression model is built to predict the target variable. Some improvements have been done on the model by removing some features that are not contributing and the data is transformed using Principal Component Analysis(PCA).
- The test of assumptions for Linear Regression is also checked and they are analyzed properly.
- In the end, Linear Regression is compared with different other popular Regression methods, in which KNN performed well owhen compared to others.
- There is alot of scope to increase the performance of Linear Regression model.
- We can increase the samples to build a robust model. Also, we can add some more features that contribute to the wine quality.