

Wine Quality

-Sonya @AUG2021



Wine is an alcoholic drink that is made up of fermented grapes. How can you differentiate the wine according to their quality?

According to experts, the wine is differentiated according to its smell, flavor, and color, but how the ML to do this job?

About the project

- 1. Import & process dataset
- 2. Data Exploration
- 3. Model Test & Predict
 - Analysis using Linear Regression
 - Analysis using Classification

--- Endnote

Import & Process Data

What I do

- Get info
- Checking that missing values
- Get describe
- Check shape
- Drop duplicate
- Check unique values

```
1 df_ice.head()
executed in 33ms. finished 15:36:55 2021-08-12
```

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pН	sulphates	alcohol	quality
0	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5
1	7.8	0.88	0.00	2.6	0.098	25.0	67.0	0.9968	3.20	0.68	9.8	5
2	7.8	0.76	0.04	2.3	0.092	15.0	54.0	0.9970	3.26	0.65	9.8	5
3	11.2	0.28	0.56	1.9	0.075	17.0	60.0	0.9980	3.16	0.58	9.8	6
4	7.4	0.70	0.00	1.9	0.076	11.0	34.0	0.9978	3.51	0.56	9.4	5

```
1 df_fruit.head()
executed in 26ms, finished 15:37:00 2021-08-12
```

	fixed_acidity	volatile_acidity	citric_acid	residual_sugar	chlorides	free_sulfur_dioxide	total_sulfur_dioxide	density	pН	sulphates	alcohol	quality
0	7.0	0.27	0.36	20.7	0.045	45.0	170.0	1.0010	3.00	0.45	8.8	6
1	6.3	0.30	0.34	1.6	0.049	14.0	132.0	0.9940	3.30	0.49	9.5	6
2	8.1	0.28	0.40	6.9	0.050	30.0	97.0	0.9951	3.26	0.44	10.1	6
3	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6
4	7.2	0.23	0.32	8.5	0.058	47.0	186.0	0.9956	3.19	0.40	9.9	6

Desired dataframe

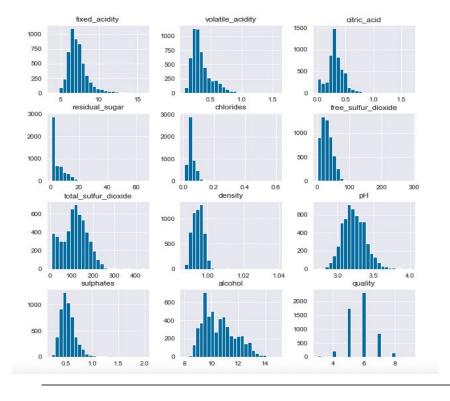
- Append the category it needed
- Append datasets
- Output for worklog

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 5320 entries, 0 to 4897
Data columns (total 13 columns):
                          Non-Null Count Dtype
    Column
    fixed acidity
                                        float64
                          5320 non-null
    volatile acidity
                                        float64
                          5320 non-null
    citric acid
                          5320 non-null
                                        float64
    residual sugar
                          5320 non-null float64
    chlorides
                          5320 non-null
                                        float64
    free sulfur dioxide
                          5320 non-null float64
    total sulfur dioxide
                                        float64
                          5320 non-null
    density
                          5320 non-null
                                        float64
                                        float64
    pН
                          5320 non-null
    sulphates
                          5320 non-null
                                        float64
    alcohol
                                        float64
                          5320 non-null
                          5320 non-null
    quality
                                        int64
    category
                          5320 non-null
                                         object
dtypes: float64(11), int64(1), object(1)
memory usage: 741.9+ KB
```

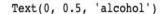
```
# output the edited dataset and save for worklog

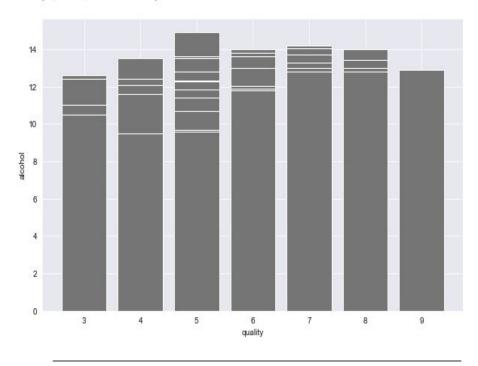
outputpath='/Users/sonyaWU/Desktop/wine_quality_1.xlsx'
wine_df.to_excel(outputpath,index=False,header=True)

executed in 839ms, finished 15:08:30 2021-08-11
```



Distribute on features





Value of alcohol make change

Exploration

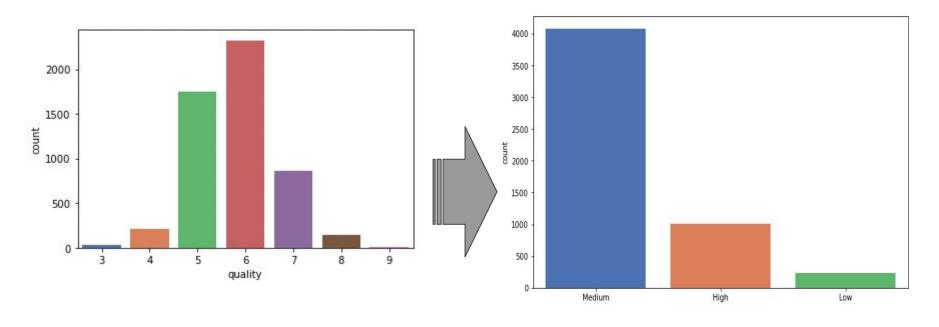
Correlation

Compare these three charts, we can see **Sulfur dioxide** is highly correlated with the Quality of a wine. And followed by **Density & Alcohol**.

					Correlatio	n Matrix o	f both win	es Quality	,				
fixed_acidity -	1	0.21	0.33	-0.1	0.29	-0.28	-0.33	0.48	-0.27	0.3	-0.1	-0.08	1
volatile_acidity		1	-0.38	-0.16	0.37	-0.35	-0.4	0.31			-0.065	-0.27	- 0.
citric_acid *		-0.38	1						-0.34		-0.0055	0.098	- 0.
residual_sugar -	-0.1	-0.16			-0.12	0.4			-0.23	-0.17	-0.31	-0.057	
chlorides -				-0.12	1	-0.19	-0.27		0.026		-0.27	-0.2	- 0.
ree_sulfur_dioxide	-0.28	-0.35	0.13		-0.19	1	0.72	0.0062	-0.14	-0.2	-0.17	0.054	- 0.
tal_sulfur_dioxide -	-0.33	-0.4	0.19		-0.27	0.72		0.0067	-0.22	-0.28	-0.25	-0.05	
density -						0.0062	0.0067	1	0.034		-0.67	-0.33	- 0.
pH -	-0.27	0.25	-0.34	-0.23	0.026	-0.14	-0.22	0.034	1	0.17	0.097	0.04	
sulphates -	0.3		0.059	-0.17	0.41	-0.2	-0.28	0.28	0.17	1	-0.017	0.042	
alcohol -	-0.1	-0.065	-0.0055	-0.31	-0.27	-0.17	-0.25	-0.67		-0.017	1	0.47	
quality -	-0.08	-0.27	0.098	-0.057	-0.2		-0.05	-0.33	0.04	0.042	0.47	1	
	fixed acidity -	volatile_acidity -	dtric_acid -	residual sugar -	chlorides -	e_sulfur_dioxide -	ital sulfur dioxide -	density -	£	sulphates -	alcohol -	quality	

					Correlatio	n Matrix o	f Fruit_win	e Quality)			
fixed_acidity -		-0.019		0.084	0.024	-0.058	0.082		-0.43	-0.017	-0.11	-0.12
volatile_acidity	-0.019	1	-0.16	0.098	0.086	-0.1	0.1	0.061	-0.047	-0.021	0.047	-0.19
citric_acid *		-0.16	1	0.11	0.13	0.092	0.12	0.16	-0.18	0.049	-0.077	0.0071
residual_sugar -	0.084	0.098	0.11		0.076			0.82	-0.17	-0.021	-0.4	-0.12
chlorides -	0.024	0.086	0.13	0.076		0.1			-0.091	0.018	-0.36	-0.22
free_sulfur_dioxide	-0.058	-0.1	0.092		0.1	1			-0.0077	0.038	-0.25	0.011
otal_sulfur_dioxide -	0.082	0.1	0.12			0.62			0.0082	0.14	-0.45	-0.18
density -		0.061	0.16						-0.064	0.082	-0.76	-0.34
pH -	-0.43	-0.047	-0.18	-0.17	-0.091	-0.0077	0.0082	-0.064	1	0.14	0.093	0.12
sulphates -	-0.017	-0.021	0.049	-0.021	0.018	0.038	0.14	0.082	0.14		-0.023	0.053
alcohol -	-0.11	0.047	-0.077	-0.4	-0.36	-0.25	-0.45	-0.76	0.093	-0.023	1	0.46
quality -	-0.12	-0.19	0.0071	-0.12	-0.22	0.011	-0.18	-0.34	0.12	0.053	0.46	
	fixed acidity -	volatile_acidity -	dtric_acid -	residual sugar -	chlorides -	ee_sulfur_dioxide -	tal_sulfur_dioxide -	density -	¥	sulphates -	alcohol -	quality

					Correlation	on Matrix o	of Ice_win	e Quality)				
fixed_acidity -	1	-0.26	0.67	0.11	0.086	-0.14	-0.1	0.67	-0.69	0.19	-0.062	0.12
volatile_acidity -	-0.26	1	-0.55	-0.0024	0.055	-0.021	0.072	0.024	0.25	-0.26	-0.2	-0.4
citric_acid -	0.67	-0.55		0.14		-0.048	0.047		-0.55		0.11	
residual_sugar -	0.11	-0.0024		1	0.027				-0.083	-0.012	0.063	0.014
chlorides -	0.086	0.055		0.027	1	0.00075	0.046		-0.27		-0.22	-0.13
free_sulfur_dioxide	-0.14	-0.021	-0.048		0.00075			-0.018	0.057	0.054	-0.08	-0.05
total_sulfur_dioxide -	-0.1	0.072	0.047		0.046	0.67	1	0.078	-0.079	0.035	-0.22	-0.18
density -		0.024				-0.018	0.078	1	-0.36		-0.5	-0.18
pH -	-0.69	0.25	-0.55	-0.083	-0.27	0.057	-0.079	-0.36		-0.21		-0.055
sulphates -		-0.26		-0.012	0.39	0.054	0.035		-0.21		0.092	
alcohol -	-0.062	-0.2	0.11	0.063	-0.22	-0.08	-0.22	-0.5		0.092		
quality -	0.12	-0.4		0.014	-0.13	-0.05	-0.18	-0.18	-0.055			
	fixed_acidity -	volatile_acidity -	atric_acid -	residual_sugar	chlorides -	e_sulfur_dioxide	sulfur dioxide -	density -	¥	sulphates -	alcohol -	quality -



Set quality standards

3,4 = Low

5,6 = Medium

7,8,9 = High

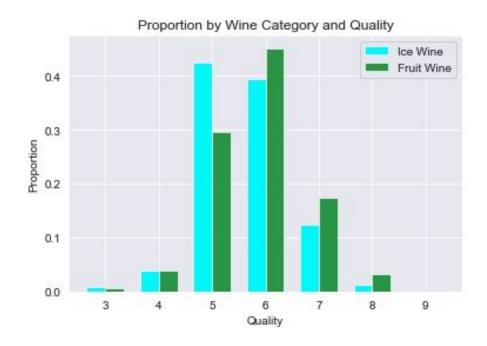
High: 1009

Medium: 4075

Low:236

Category & Quality Proportion

- Counts for each rating and category
- Total counts for each category
- Proportions by dividing rating counts by total of samples



Model Test & Predict

Linear Regression

Try in datasets for df_ice, or df_fruit, result comes difference.

The coefficients denote the impact of each on the quality of wine.

- Every Alcohol measure increase will lead to increase of 0.33 in quality.
- The Chlorides increasing will be decrease the quality.

```
In [190]:
                print('Mean Absolute Error:', metrics.mean absolute error(test pred, y test))
              2 print('Mean Squared Error:', metrics.mean squared error(test pred, y test))
              3 print('Root Mean Squared Error:', np.sqrt(metrics.mean squared error(test pred, y test)))
           executed in 4ms, finished 07:55:12 2021-08-13
           Mean Absolute Error: 0.5327418793324455
           Mean Squared Error: 0.44458329185943474
           Root Mean Squared Error: 0.6667707940960182
In [191]:
                coefficients = pd.DataFrame(lr.coef , features)
              2 coefficients.columns = ['Coefficient']
                print(coefficients)
           executed in 9ms, finished 07:55:14 2021-08-13
                                  Coefficient
           fixed acidity
                                     0.053091
           volatile acidity
                                    -1.147530
           citric acid
                                    -0.211109
           chlorides
                                    -1.292045
           total sulfur dioxide
                                    -0.001856
           density
                                   -21.321908
           sulphates
                                     0.876361
           alcohol
                                     0.273638
```

Random Forest Classifier

It shows a massive increase in accuracy when we reduced the number of classes while still managing to retain meaning of the output.

```
from sklearn.preprocessing import MinMaxScaler

# creating normalization object
norm = MinMaxScaler()

norm_fit = norm.fit(X_train)
new_xtrain = norm_fit.transform(X_train)
new_xtest = norm_fit.transform(X_test)

print(new_xtrain)
executed in 25ms, finished 17:25:44 2021-08-12
```

4.2.2 Apply model

```
v 1 # importing modules
2 from sklearn.ensemble import RandomForestClassifier
3 from sklearn.metrics import classification_report
executed in 5ms, finished 17:34:12 2021-08-12
```

```
f = RandomForestClassifier()
f.fit(X_train, y_train)
g = f.predict(X_test)
print("Accuracy for RandomForestClassifier:",
metrics.accuracy_score(y_test, g))
executed in 213ms, finished 07:46:43 2021-08-13
```

Accuracy for RandomForestClassifier: 0.5588235294117647

Coming up later:

- Follow up using multiple modelling for classifications
- Random Forest Classifier gives the 88% accuracy

Thank you Mr. Terresiu for sharing data information in Lulu Island Winery.