Exploratory Data Analysis for Machine Learning – Final Course Project

Wine Quality

Author: Vasile Lucian Popa

Overview:

- Data description and data dictionary;
- Initial plan for data exploration;
- Data cleaning and feature engineering;
- Key findings and insights;
- Hypothesis formulated about data;
- Formal significance test on hypothesis;
- Conclusion and suggestions;
- Summary about dataset and future steps.

Data description

- This datasets is related to red variants of the Portuguese "Vinho Verde" wine. The dataset describes the amount of various chemicals present in wine and their effect on it's quality. The datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).Your task is to predict the quality of wine using the given data. A simple yet challenging project, to anticipate the quality of wine.
- The complexity arises due to the fact that the dataset has fewer samples, & is highly imbalanced.

- Data source: https://www.kaggle.com/yasserh/wine-quality-dataset

Data dictionary:

1. fixed acidity
   most acids involved with wine or fixed or nonvolatile (do not evaporate readily);

2. volatile acidity
   the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste;

3. citric acid
   found in small quantities, citric acid can add ‘freshness’ and flavor to wines;

4. residual sugar
   the amount of sugar remaining after fermentation stops, it’s rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet;

5. chlorides the amount of salt in the wine;

6. free sulfur dioxide
   the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine;

7. total sulfur dioxide
   amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine;

8. density
   the density of water is close to that of water depending on the percent alcohol and sugar content;

9. pH
   describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3–4 on the pH scale;

10. sulphates
    a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant;

11. alcohol
    the percent alcohol content of the wine;

12. quality
    output variable (based on sensory data, score between 0 and 10);

13. id

Initial plan for data exploration:

1. Understanding the wine data column;
2. Perform basic data check (info, describe, shape etc.);
3. Perform relation analysis by graphical approach;
4. Cleaning the data;
5. Formulate hypothesis for the data;

Data cleaning and feature engineering:

- The wine quality dataset doesn't have any missing values/rows/cells for any of the variables/feature. It seems that data has been collected neatly or prior cleaning has been performed before publishing the dataset.
- Dropped ID column since is not relevant to the analysis.
- Got rid of the extreme outliers(dropped rows below 1% and above 99% quantile)
- Check for missing values: The wine quality dataset doesn't have any missing values/rows/cells for any of the variables/feature.
- It seems that data has been collected neatly or prior cleaning has been performed before publishing the dataset.
- Create a new categorical response variable/feature ('overall') from existing 'quality' variable:

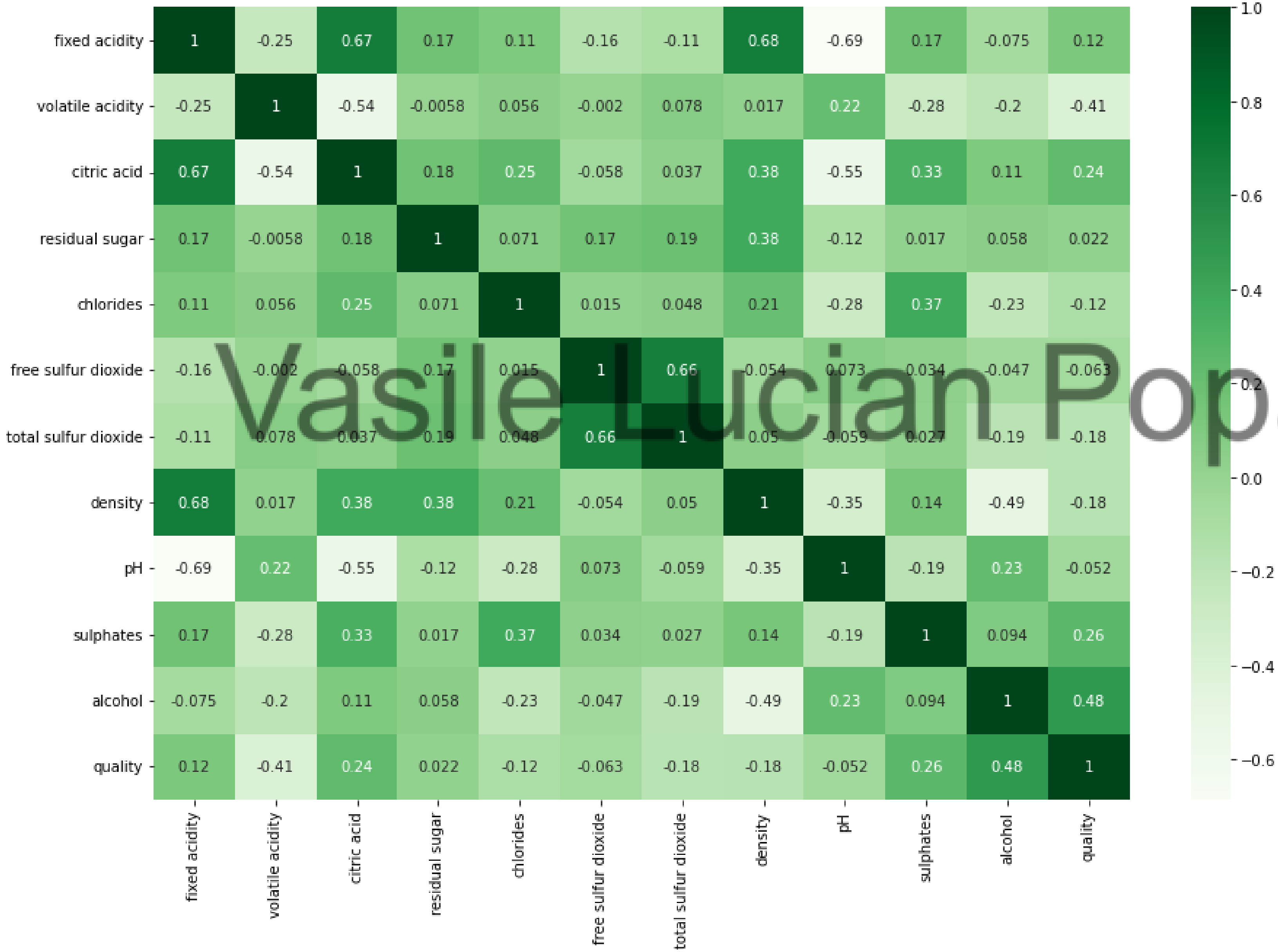
Good = [8,9]

Medium = [5,6,7]

Poor = [3,4]

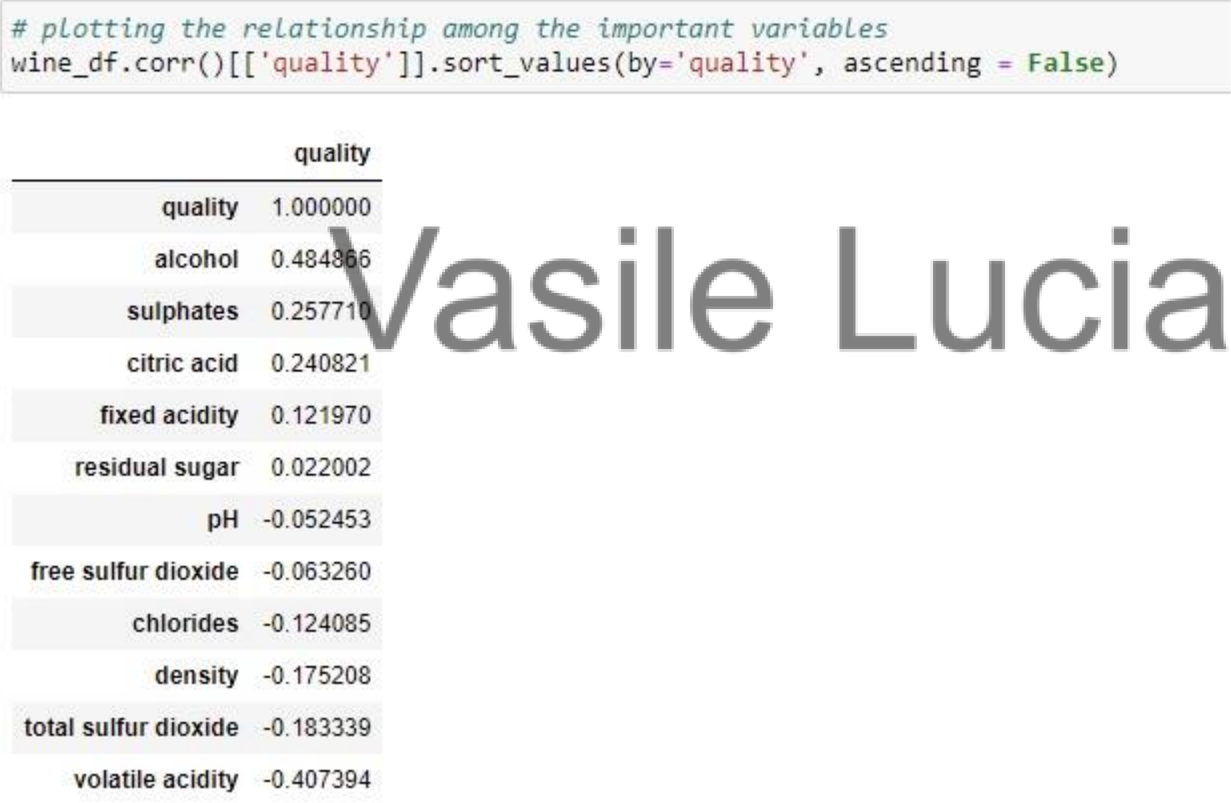
Key findings and insights

- Check the strength of the correlation among the variables



	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
fixed acidity	1	-0.25	0.67	0.17	0.11	-0.16	-0.11	0.68	-0.69	0.17	-0.075	0.12
volatile acidity	-0.25	1	-0.54	-0.0058	0.056	-0.002	0.078	0.017	0.22	-0.28	-0.2	-0.41
citric acid	0.67	-0.54	1	0.18	0.25	-0.058	0.037	0.38	-0.55	0.33	0.11	0.24
residual sugar	0.17	-0.0058	0.18	1	0.071	0.17	0.19	0.38	-0.12	0.017	0.058	0.022
chlorides	0.11	0.056	0.25	0.071	1	0.015	0.048	0.21	-0.28	0.37	-0.23	-0.12
free sulfur dioxide	-0.16	-0.002	-0.058	0.17	0.015	1	0.66	-0.054	0.073	0.034	-0.047	-0.063
total sulfur dioxide	-0.11	0.078	0.037	0.19	0.048	0.66	1	0.05	-0.059	0.027	-0.19	-0.18
density	0.68	0.017	0.38	0.38	0.21	-0.054	0.05	1	-0.35	0.14	-0.49	-0.18
pH	-0.69	0.22	-0.55	-0.12	-0.28	0.073	-0.059	-0.35	1	-0.19	0.23	-0.052
sulphates	0.17	-0.28	0.33	0.017	0.37	0.034	0.027	0.14	-0.19	1	0.094	0.26
alcohol	-0.075	-0.2	0.11	0.058	-0.23	-0.047	-0.19	-0.49	0.23	0.094	1	0.48
quality	0.12	-0.41	0.24	0.022	-0.12	-0.063	-0.18	-0.18	-0.052	0.26	0.48	1

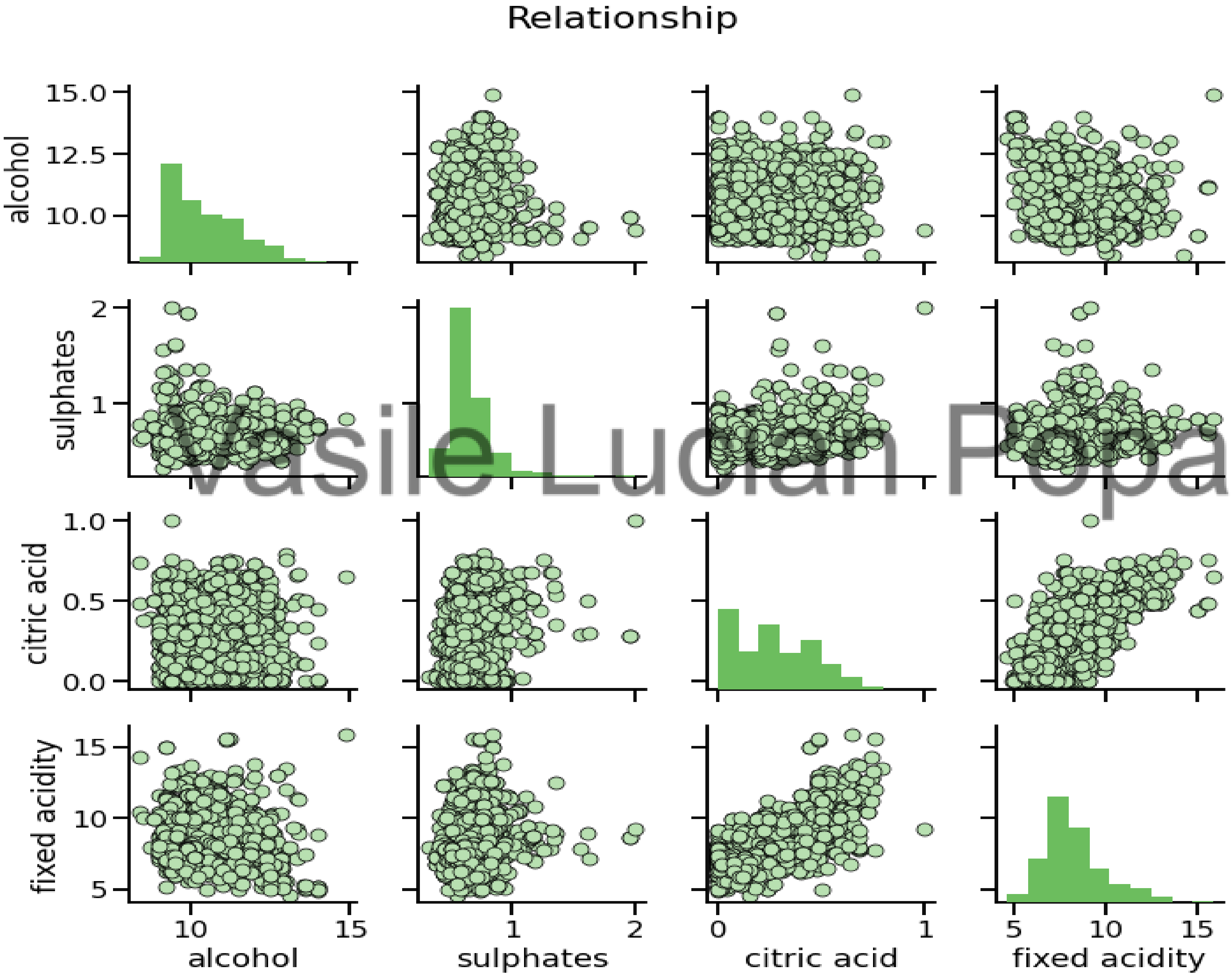
- We can observe that, the 'alcohol, sulphates, citric acid & fixed acidity' have maximum correlation with response variable 'quality'.
- This means that, they need to be further analyzed for detailed pattern and correlation exploration. Hence, we will use only these 4 variables in our future analysis.



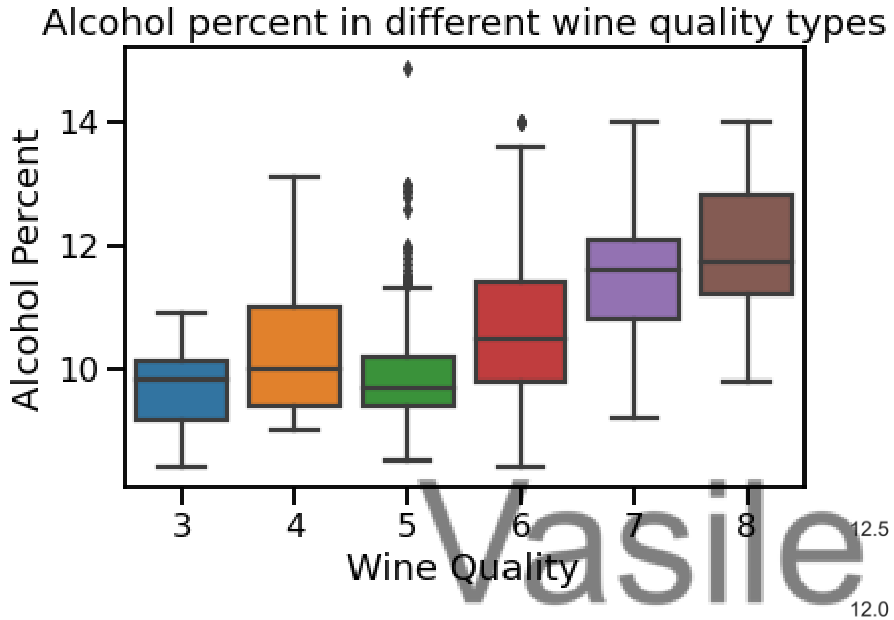
```python
# plotting the relationship among the important variables
wine_df.corr()[['quality']].sort_values(by='quality', ascending = False)
```

	quality
quality	1.000000
alcohol	0.484866
sulphates	0.257710
citric acid	0.240821
fixed acidity	0.121970
residual sugar	0.022002
pH	-0.052453
free sulfur dioxide	-0.063260
chlorides	-0.124085
density	-0.175208
total sulfur dioxide	-0.183339
volatile acidity	-0.407394

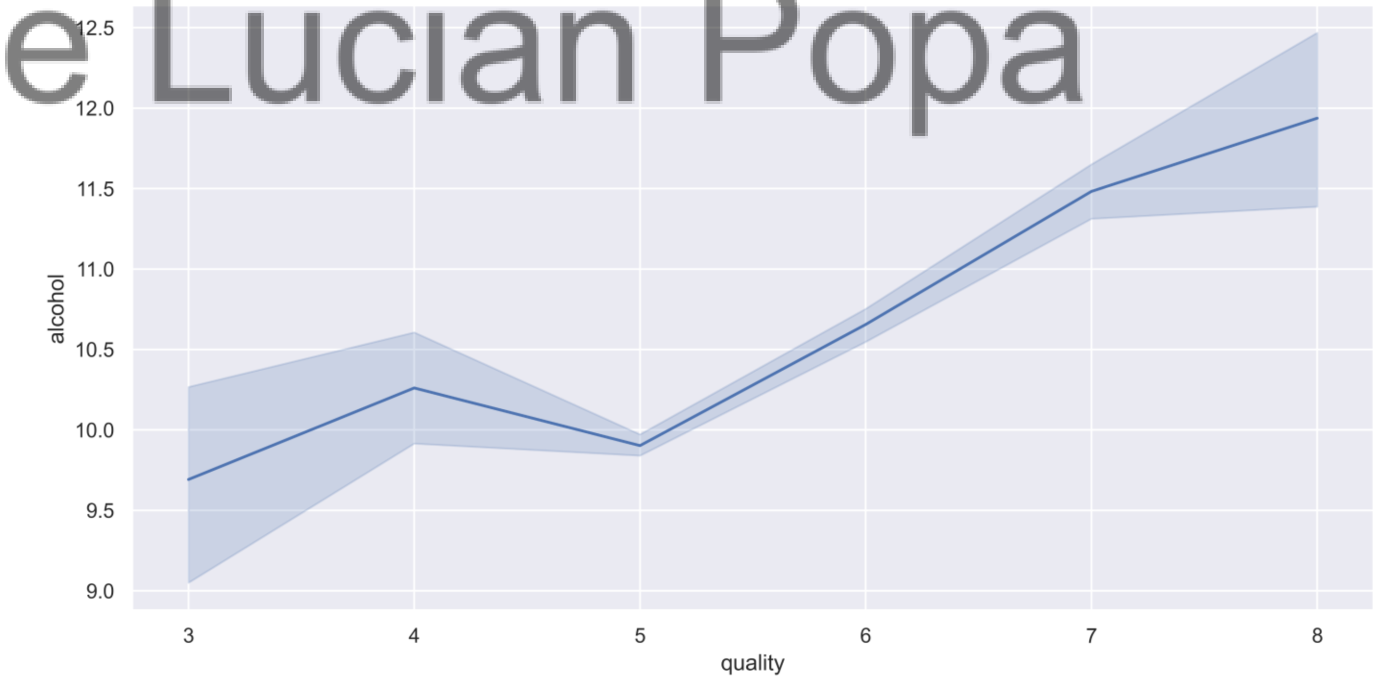
- Plot relationship between the 'alcohol, sulphates, citric acid & fixed acidity':



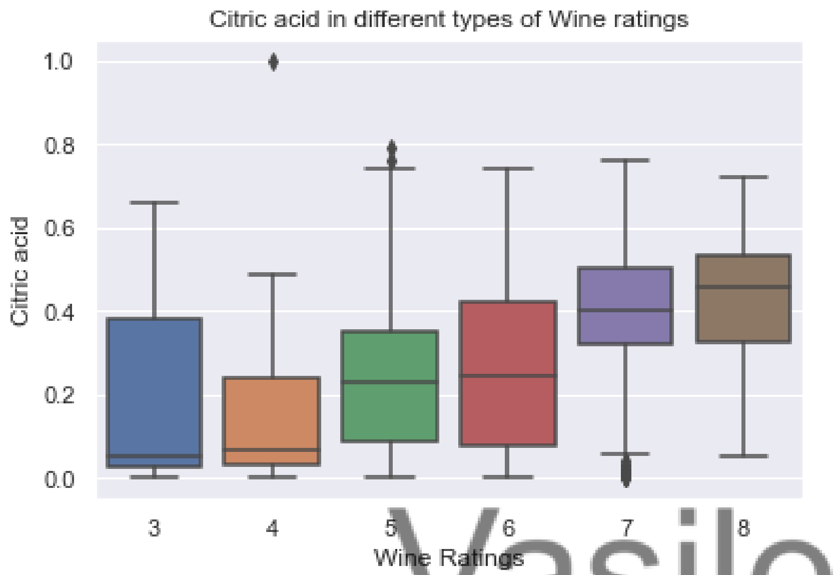
Analysis of alcohol percentage with wine quality:



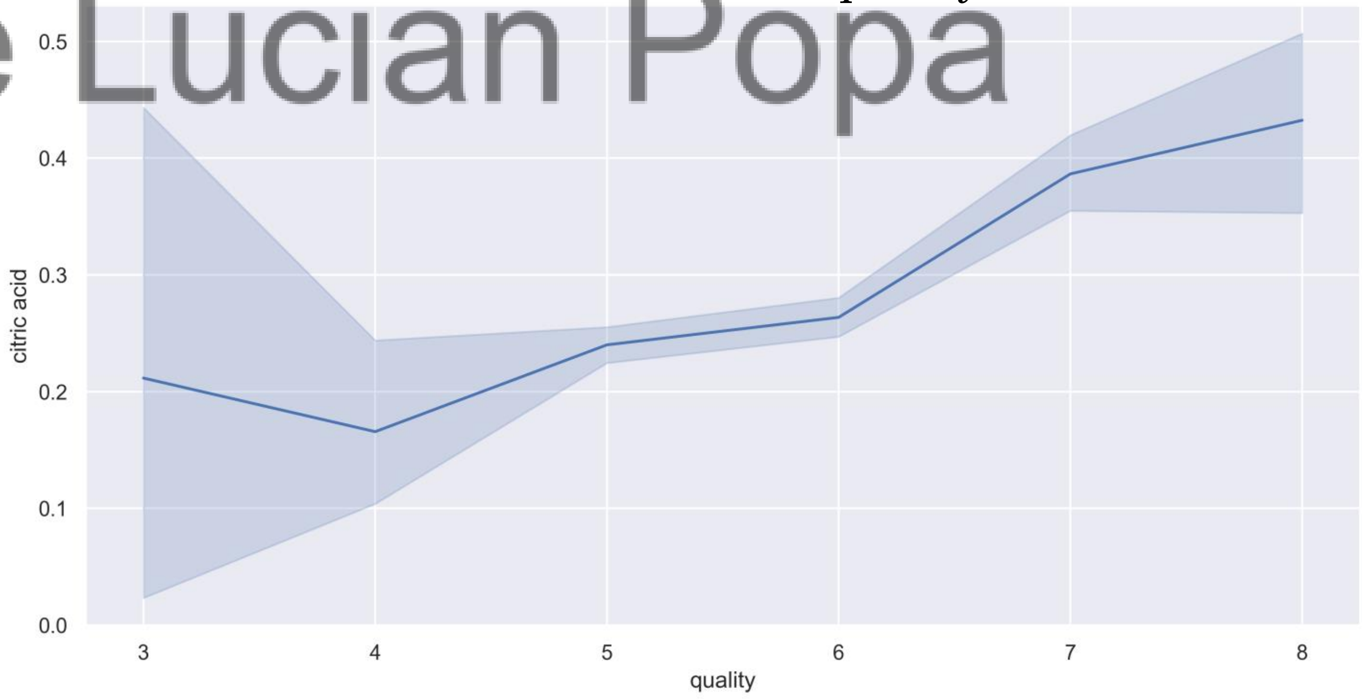
Alcohol % plot on the quality scale:



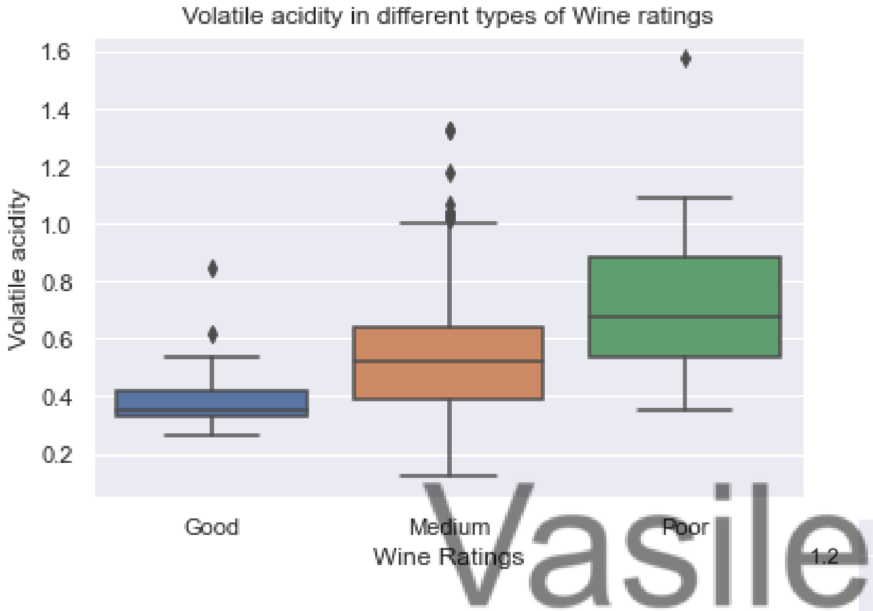
Analysis of citric acid with wine quality:



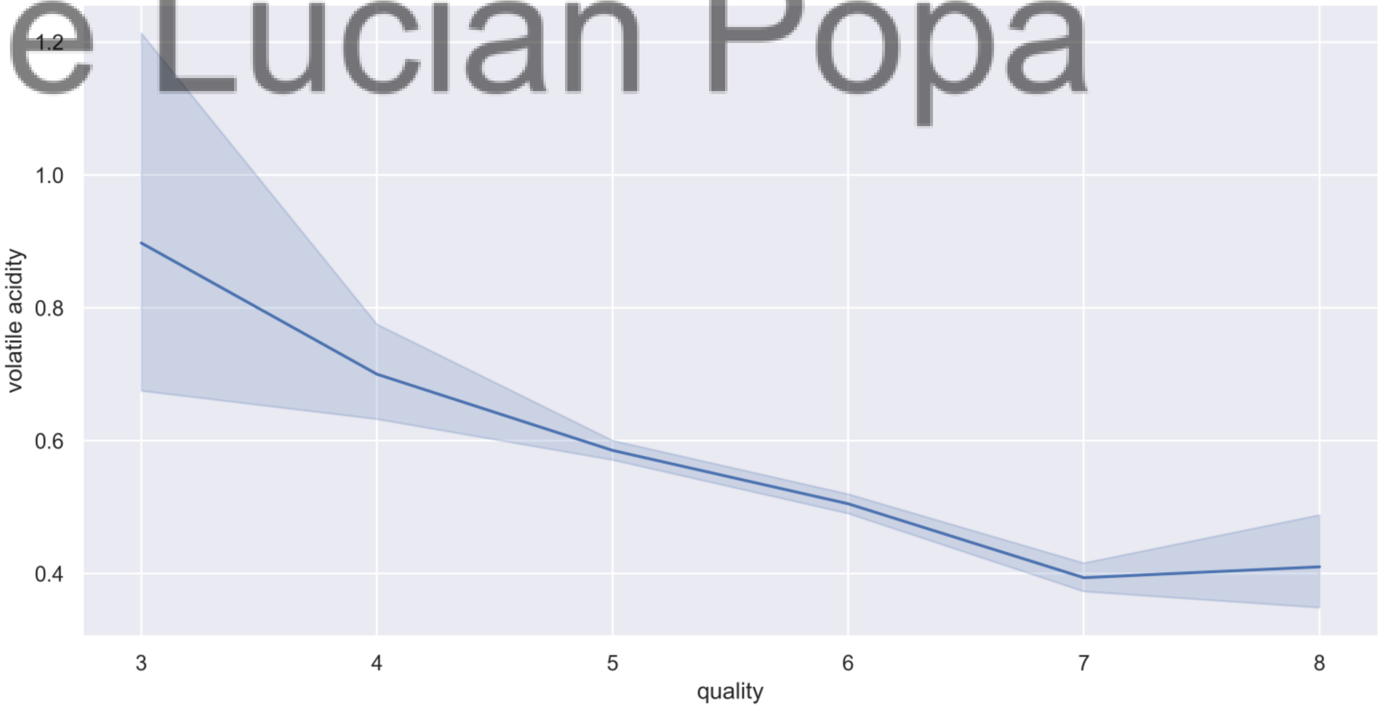
Citric acid on different quality levels:



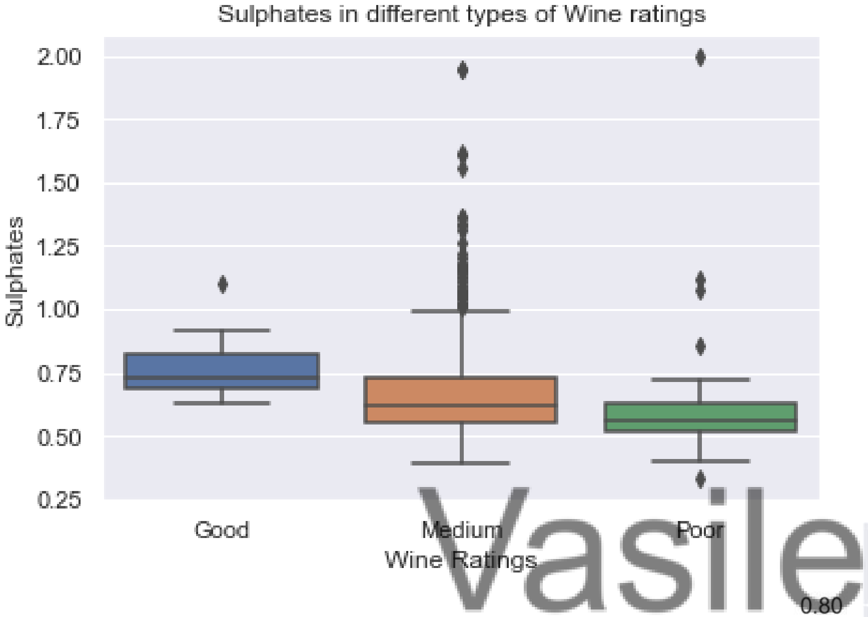
Analysis of volatile acidity with wine quality:



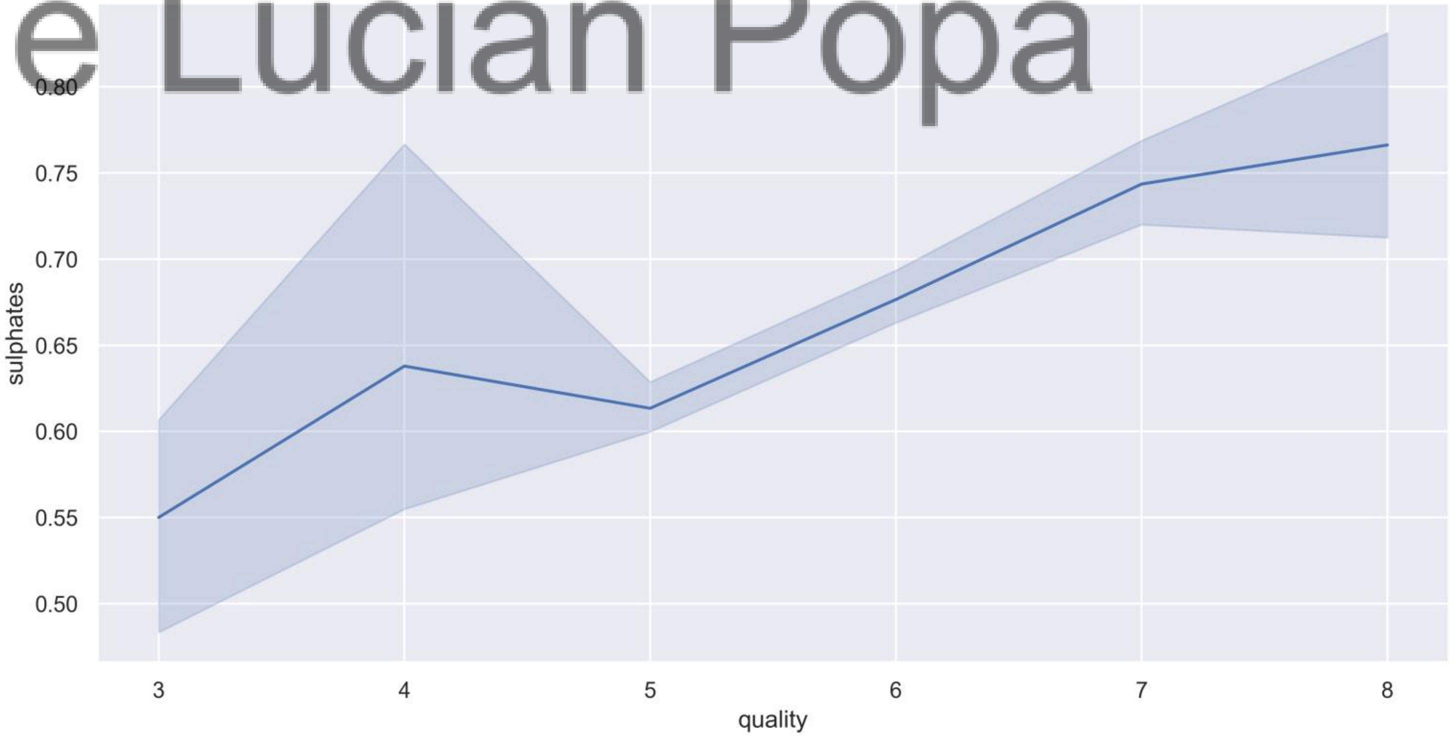
Volatile acidity vs quality:



Analysis of sulphates with wine quality:



Sulphates vs. qulity:

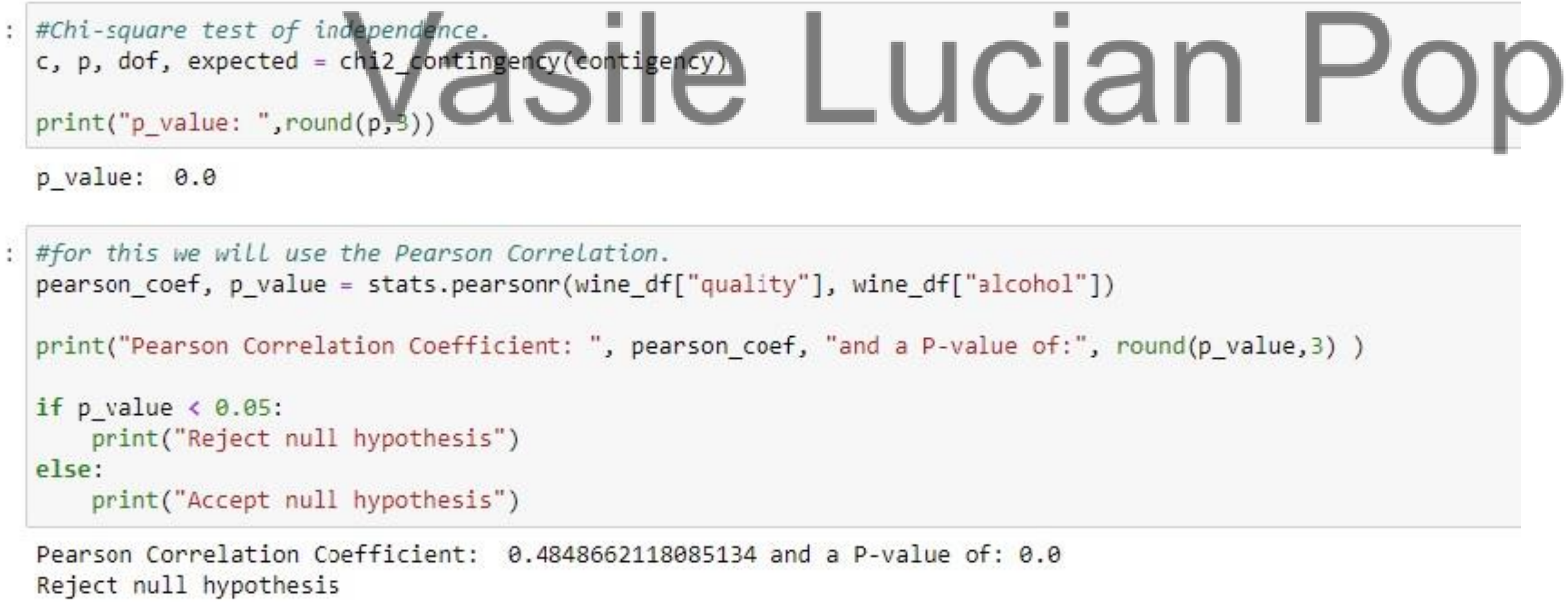


Alcohol vs. Quality Hypothesis

- Hypothesis formulated about data and Formal significance test on hypothesis;

Hypothesis:

- * Null: Increase in the alcohol % doesn't affect the quality of the wine.
- * Alternative: Increase in alcohol % does affect quality.



```
#Chi-square test of independence.
c, p, dof, expected = chi2_contingency(contigency)

print("p_value: ",round(p,3))
```

```
p_value:  0.0
```

```
#for this we will use the Pearson Correlation.
pearson_coef, p_value = stats.pearsonr(wine_df["quality"], wine_df["alcohol"])

print("Pearson Correlation Coefficient: ", pearson_coef, "and a P-value of:", round(p_value,3) )

if p_value < 0.05:
    print("Reject null hypothesis")
else:
    print("Accept null hypothesis")
```

```
Pearson Correlation Coefficient:  0.4848662118085134 and a P-value of: 0.0
Reject null hypothesis
```

Conclusion

- 1. As alcohol level increase ==> Quality increases
- 2. As chlorides level decreases ==> Quality increases
- 3. As citric acid level increases ==> Quality increases
- 4. As density decreases ==> Quality increases
- 5. As the volatile acidity decreases ==> Quality increases
- 6. As sulphates increases ==> Quality increases

But since, only this four contributes towards wine quality :

- `alcohol, citric acid, volatile acidity, sulphates `
- * Increase in the alcohol qty, increases the quality of the wine.
- * Increase in the citric of the wine, increases the quality of the wine.
- * Decrease in the volatile acidity of the wine, increases the quality of the wine.
- * Increase in sulphates, increases the quality of the wine.

Suggestions

- The usage of this analysis will help to understand whether by modifying the variables, it is possible to increase the quality of the wine on the market. If you can control your variables, then you can predict the quality of your wine and obtain more profits.

- For further analysis of this dataset, data should be standardized, and find more insights and valuable information should be extracted out of it. This dataset is well cleaned and prepared, but it's missing volume. More data should be added in order to obtain better insights out of it.