‘Exploratory Analysis of Wine Quality Evaluation by Physicochemical factors’

Authors:

Prajwal Chinchmalatpure, Bindu Latha Banisetti, Venkata Sai Ujwala Bayana, Harshitha Somala, Abhishek Reddy Andluru

Goal:

The goal is to determine which set of factors/attributes plays a major role in determining the quality of Red and White wine individually. Then compare potential quality-affecting attributes between the red and white wines and see if there are any common attributes among them. We start with finding the correlation between the eleven attributes of the dataset which represents eleven factors that can potentially determine the quality of the wine. Then we can try to correlate each attribute with the actual quality of the wine. The two datasets are associated with Portugal's red and white "Vinho Verde" wine. We got 12 attributes in each dataset. The initial 11 attributes represent the potential quality determining factors for a wine. And the 12th attribute represents the quality of the wine on a scale of 1 to 10. 1 being the lowest and 10 being the highest quality wine. The data was acquired from May, 2004 to February, 2007. The sample wine data is recorded from laboratory and sensory analysis. The attributes of the dataset are as follows:

Fixed Acidity: Tartaric Acid content in wine (g/dm3).

Volatile Acidity: Acetic Acid content in wine (g/dm3).

Citric Acid: Citric acid content in wine (g/dm3). Contributes to the wine's crispness.

Residual Sugar: Sugar content left in the wine after fermentation (g/dm3).

Chlorides: Sodium Chloride (salt) content in wine (g/dm3)

Free Sulfur Dioxide: SO2 content in free form (g/dm3)

Total Sulfur Dioxide: Total Amount of SO2(g/dm3). SO2 acts as an antioxidant & antimicrobial agent. Too much SO2 leads to deriving a pungent smell.

Density: Density of Wine (g/dm3)

pH: pH of Wine on a scale of 0-14. 0 means highly Acidic, while 14 means highly basic.

Sulphates: Potassium Sulphate content in wine (measured in g/dm3). Contributes to the formation of SO2.

Alcohol: Alcohol content in wine (in terms of % volume)

Quality: Wine Quality graded on a scale of 1 - 10 (Higher is better)

Premise:

The wine business is investing in new technology for both the production and sale of wine to promote its growth. In this setting, wine certification and quality evaluation are critical. Quality assessment is a common aspect of the certification process, and it may be used to enhance winemaking (by identifying the most important factors) and stratify wines.

We will be performing some initial exploratory analysis on the dataset starting with importing two datasets into the environment. Performing different analytics such as univariate, bivariate, and multivariate analysis to see the relation between the attributes. We will use boxplots, pair plots, and histograms to better understand how the attributes relate to each other.

We do not know if the data set requires any tidying currently.

The project’s goal is to identify the determining power of different factors for the quality of the wine. This can help manufacturers target marketing the wine by modeling consumer tastes from niche markets. We visualize the relationship between attributes and target attributes in Python using various libraries like NumPy, pandas, seaborn, and matplotlib. We try to determine the impact of these attributes on quality from the Exploratory Data Analysis that we have done. We also plan on finding and implementing an appropriate machine learning model as our extended goal if time permits us to do so.

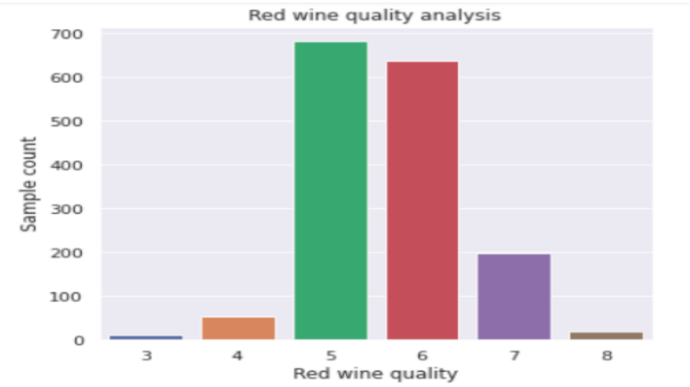
Challenges:

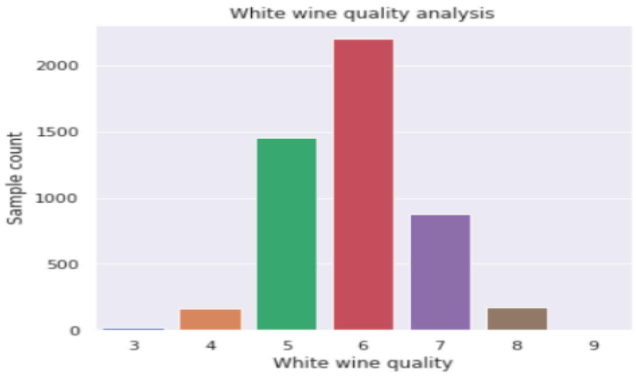
1. Determining whether the outliers have an impact on the overall data modeling.
2. As there are many factors that can have high correlation with each other. This leads to redundancy of data which we want to avoid.
3. Determining the best fitting data model.

Plans to overcome:

1. We would like to get started with plotting various attributes, computing the impact of the outlier on the plots and then making an informed decision whether to keep or let go of outliers for each attribute.
2. We can drop one of the two factors having a high degree of correlation and thus avoid redundancy.
3. We are contemplating multiple linear regression currently. But we will research more for other potential models.

Preliminary results:





The two plots above represent the distribution of wine quality samples for Red wine and White wine respectively. We loaded the CSV files for red & white wines and generated the above plots. Most red wine samples fall under the quality of 5 and 6. While for the white wine we have most samples with quality 6 and the second most samples with quality 5. For both the wine types; the quality of 3 is found in the least number of samples.

References:

1. https://archive.ics.uci.edu/ml/datasets/wine+quality
2. https://www.sciencedirect.com/science/article/pii/S0167923609001377?via%3Dihub