Part I
Project 1: Wine quality

Introduction

In the paragraphs to come we will discuss different approaches and models to be used in the dataset used in [1].

The following dataset consist of different a sample of wines with different characteristics and their relevant quality. The box below has an overview description of the dataset we will analyse. A further description of each variable can be found in Table 1.

```
>>> wine_data.info()
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1599 entries, 0 to 1598
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype
---  ------                --------------  -----
 0   fixed acidity         1599 non-null   float64
 1   volatile acidity      1599 non-null   float64
 2   citric acid           1599 non-null   float64
 3   residual sugar        1599 non-null   float64
 4   chlorides             1599 non-null   float64
 5   free sulfur dioxide   1599 non-null   float64
 6   total sulfur dioxide  1599 non-null   float64
 7   density               1599 non-null   float64
 8   pH                    1599 non-null   float64
 9   sulphates             1599 non-null   float64
 10  alcohol               1599 non-null   float64
 11  quality               1599 non-null   int64
dtypes: float64(11), int64(1)
memory usage: 150.0 KB
>>> wine_data.describe().round(decimals=2).transpose()
                       count   mean    std   min    25%    50%    75%     max
fixed acidity         1599.0   8.32   1.74  4.60   7.10   7.90   9.20   15.90
volatile acidity      1599.0   0.53   0.18  0.12   0.39   0.52   0.64    1.58
citric acid           1599.0   0.27   0.19  0.00   0.09   0.26   0.42    1.00
residual sugar        1599.0   2.54   1.41  0.90   1.90   2.20   2.60   15.50
chlorides             1599.0   0.09   0.05  0.01   0.07   0.08   0.09    0.61
free sulfur dioxide   1599.0  15.87  10.46  1.00   7.00  14.00  21.00   72.00
total sulfur dioxide  1599.0  46.47  32.90  6.00  22.00  38.00  62.00  289.00
density               1599.0   1.00   0.00  0.99   1.00   1.00   1.00    1.00
pH                    1599.0   3.31   0.15  2.74   3.21   3.31   3.40    4.01
sulphates             1599.0   0.66   0.17  0.33   0.55   0.62   0.73    2.00
alcohol               1599.0  10.42   1.07  8.40   9.50  10.20  11.10   14.90
quality               1599.0   5.64   0.81  3.00   5.00   6.00   6.00    8.00
```

Exploratory analysis

To avoid unnecessary analysis, first we will perform a few checks to get a deeper understanding of the data we are using. To that end, we first check that the correlation structure amongst the variables (see fig. 1), including the quality of the wine. This should give us a general idea of how and if the variables are related to one another.

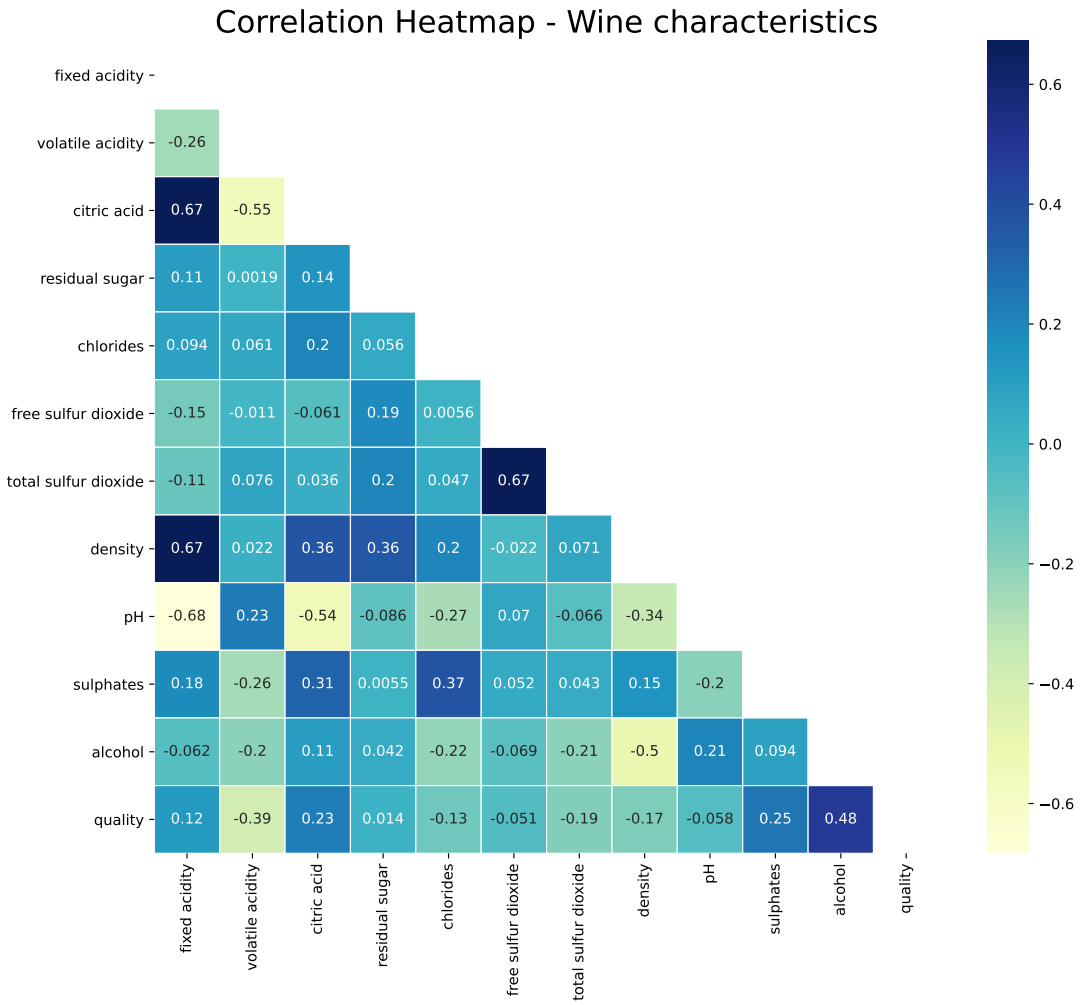

Figure 1: Correlation structure of wine characteristics.

There are a few things we could check whether the data in the data to assess whether what we are looking relates somehow to our prior knowledge about the topic. This prior knowledge might help us identify certain links, that might not be obvious if the data were not labeled.

At first glance, one can note that *fixed acidity* and *citric acidity* are strongly correlated (negatively) to the *pH*, although their correlation is not *-1*. This should relate to our prior knowledge, given that *pH* is directly related to the acidity of a solution. Additionally, we can see that *alcohol* correlates negatively with the *density* of the wine. This makes sense, given that the wine is a solution of different solutes, those in all likelihood are "heavier" than the alcohol solute. Hence, the more the alcohol content increases in the wine, the less dense it becomes.

Another interesting characteristic of the wine that is highly correlated to its quality is the alcohol content.

Another visual analysis we could perform is to look at the KDEs (Kernel Density Estimates). These we could imagine as a cross-sectional cut in a bi-variate probability distribution. Figure 2 shows us that even though the wines in the sample range from quality 3 to 8 (see subplot in row 3, column 4), it would seem that there are mostly 4 important groups. The earlier stated fact, ad priori, gives us an interesting thing we should have in mind when trying to fit any kind of model. I.e., the tails of the quality (grades 3 and 8) will be underrepresented, then most model we could think of fitting will have trouble predicting a grade close to 3 and 8 and beyond (to each direction).

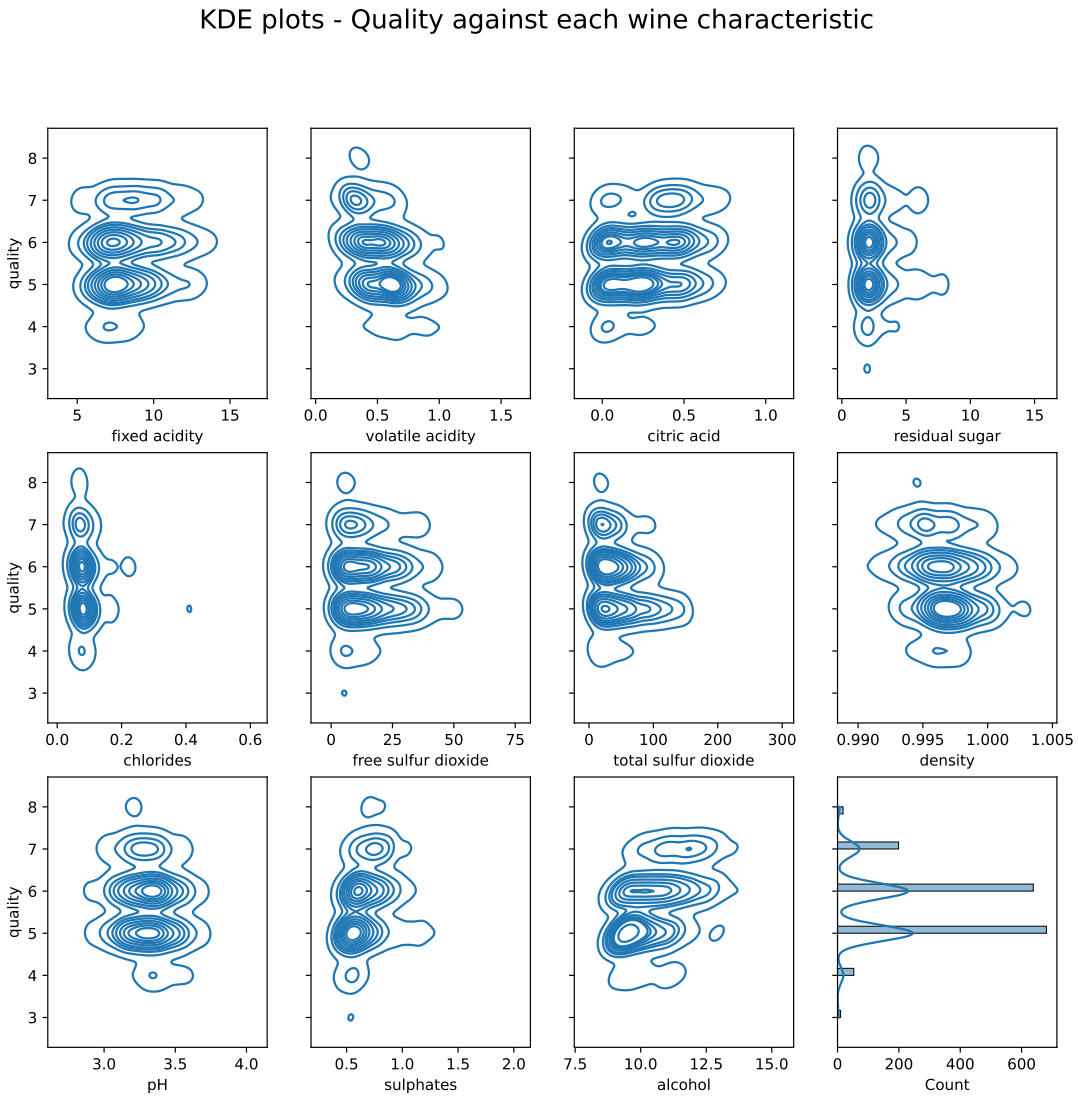

Figure 2: KDEs for wine features.

Analyses

Among the many things that we could analyse in these data[1], we will stick to predicting the quality of a wine given its characteristics.

Conclusion

[1] All these characteristics are important in a wine. We used all these statistical models to try to understand to which degree they are important to determine a wine‘s quality. These analyses might be really important for a winemaker to know. Given that by affecting the inherent characteristics of the wine will most likely have an impact on the quality of that wine.

[1]Some other analyses, like clustering the wine sample do carry the same importance. I.e., we could try clustering the wine sample, to find that certain wines belong or were produced by the same vineyard due to the similarities in the grapes which will most likely relate to their characteristics when turned into wine. Even though, this might be interesting it is not really useful to a researcher/data scientist trying to add value to the winemaking processes.

Part II

Project 2: Food Preferences

Part III

Project 3: Store Sales

References

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. *Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553*, 2009.

Table 1: Description of wine characteristics.

Characteristic	Description
fixed acidity	most acids involved with wine or fixed or nonvolatile (do not evaporate readily).
volatile acidity	the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.
citric acidity	found in small quantities, citric acid can add "freshness" and flavor to wines.
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.
chlorides	the amount of salt in the wine.
free sulfur dioxide	the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bi-sulfate ion; it prevents microbial growth and the oxidation of wine.
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.
density	the density of water is close to that of water depending on the percent alcohol and sugar content.
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.
sulphates	a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.
alcohol	the percent alcohol content of the wine.
quality	output variable (based on sensory data, score between 0 and 10).