Part I
Project 1: Wine Quality

```
>>> import pandas as pd
>>> import os

>>> # Referencing folders and data names
>>> path = os.getcwd()
>>> # We have to re-structure the path since we are in LaTeX
>>> path = os.path.abspath(os.path.join(path, os.pardir))
>>> file_name = 'winequality-red.csv'
>>> path_file = f'{path}/code_python/project1_wine/data/{file_name}'
>>> # We load the data and present an overview
>>> wine_data = pd.read_csv(path_file)
>>> 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
```

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.

Hello, World

There are a few things we could appreciate to check whether the data 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* is strongly correlated (negatively) to the *pH*, although their correlation is not *1*. 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

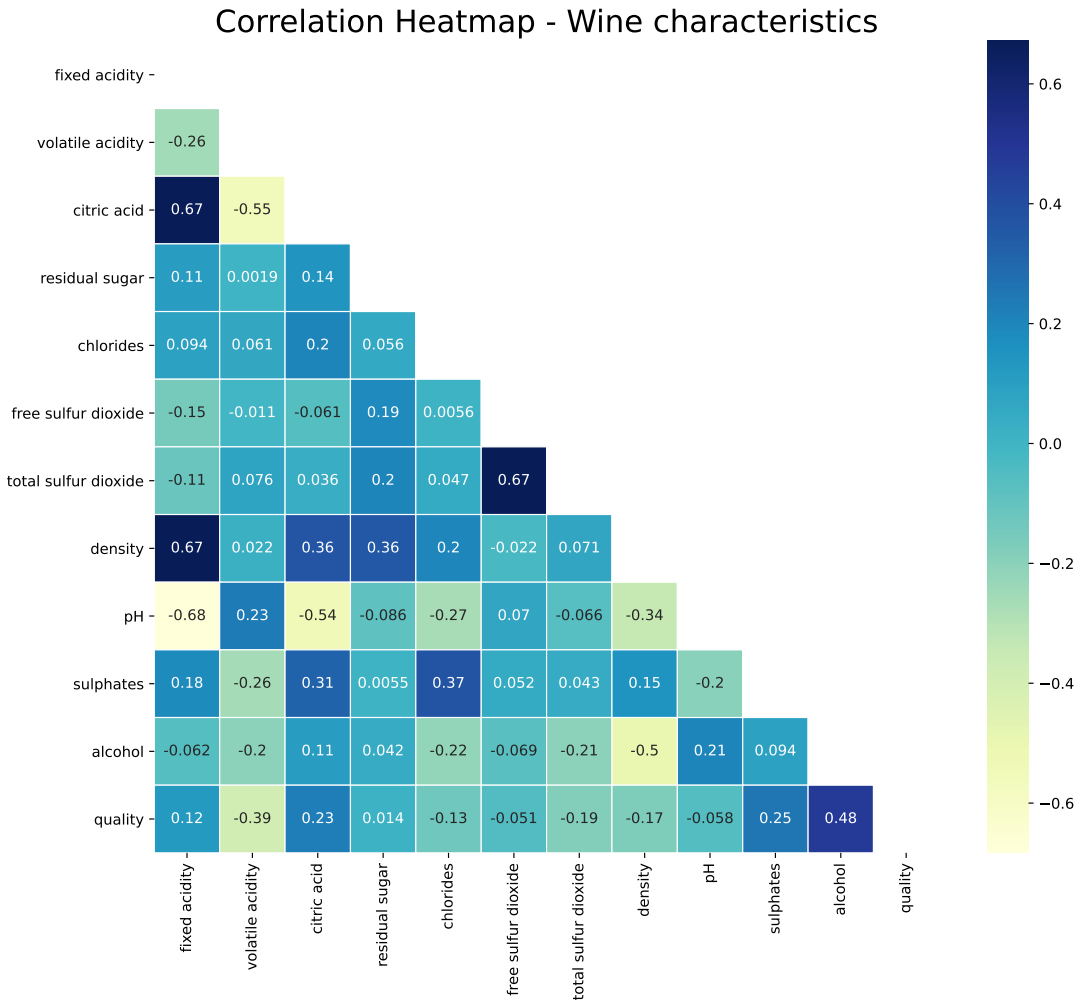

Figure 1: Correlation structure of wine characteristics.

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 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).

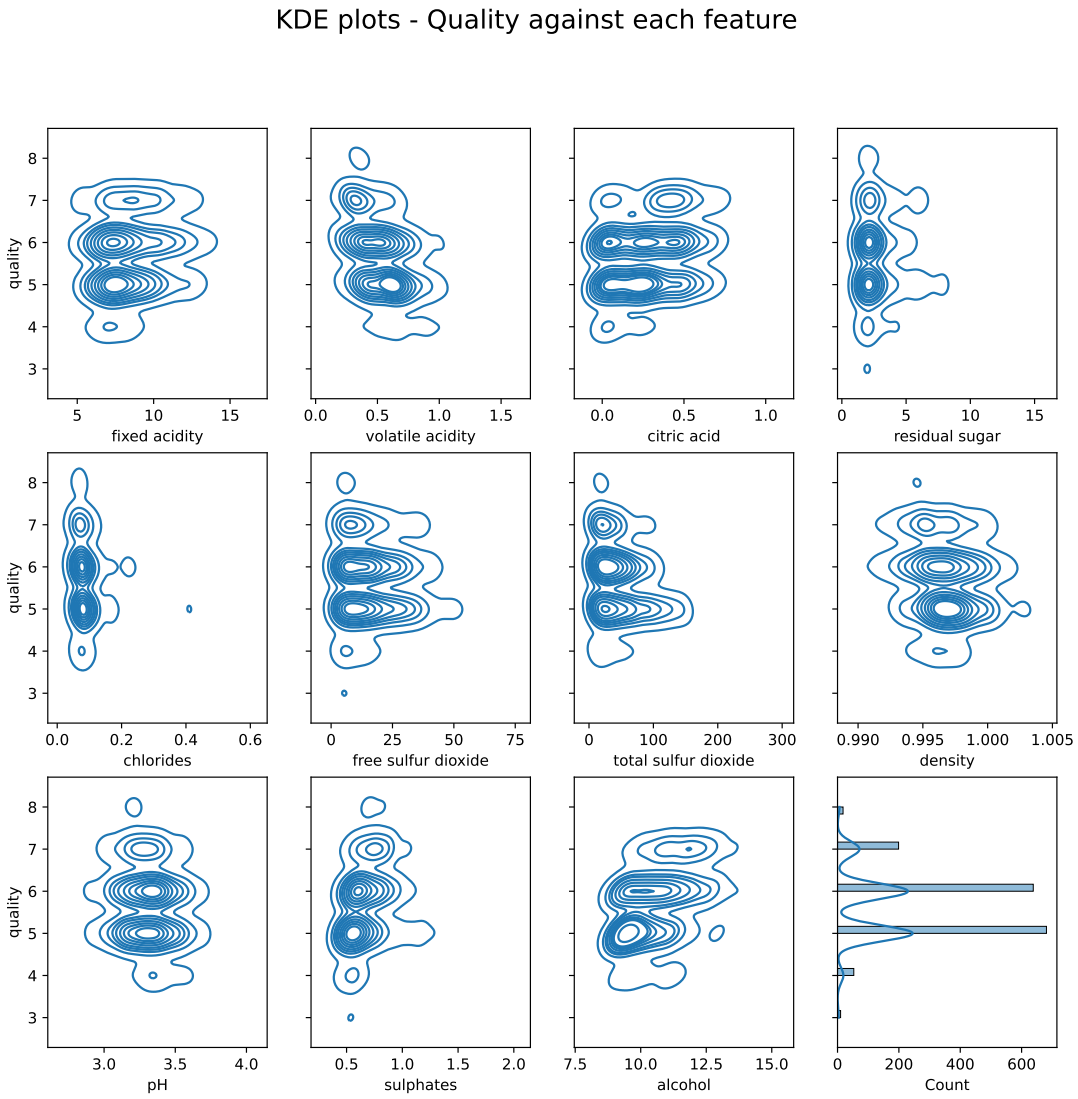

Figure 2: KDEs for wine features.

Part II

Project 2: Food Preferences

Part III

Project 3: Store Sales