Intro to Data Science Wine Analysis Project
Ivan Novikov

Intro

I chose to do my final project on the wine quality data set from the UCI machine learning repository. I chose this project because I like wine and I recently learned about something in the wine world known as *natural wine.* Natural wine is a nebulous term without a concrete meaning but what a natural wine is essentially is one made with wine making methods that were used before the industrialization of wine making. One new tool of wine making is adding sulphates to the wine because they are a preservative. Conversely, an absence of sulphates is one of the signs of a natural wine. The question I aim to answer in this project is if the level of sulphates has anything to do with the quality of wine. More specifically, does the fact that few sulphates are added to natural wine increase anything about its quality?

Data set description

There are two data sets, one for red wine, the other for white wine. I did regressions on both but I am far more interested in the questions about red wine.

Both of the data sets share the same features:

1 - fixed acidity
2 - volatile acidity
3 - citric acid
4 - residual sugar
5 - chlorides
6 - free sulfur dioxide
7 - total sulfur dioxide
8 - density
9 - pH
10 - sulphates
11 - alcohol
Output variable (based on sensory data):
12 - quality (score between 0 and 10)

In my data preprocessing I added a row Id in KNIME.

The original data set was a single column where the individual cells in the rows were separated by ‘;’, so I created a new csv with individual columns for the features.

There ate 1597 records in the red wine data set, 4896 records in the white wine data set.

The data sets are complete with no missing values, likely because they are from an academic study.

All of the features are numeric. All of the features are continuous except for quality, which is discrete.

Modeling Methodology

The models that I created for this data set are all linear regressions. This decision appropriate because regressions allow us to see which variables are most important in predicting the target variable with relative ease. It is also appropriate because all our features are numeric.

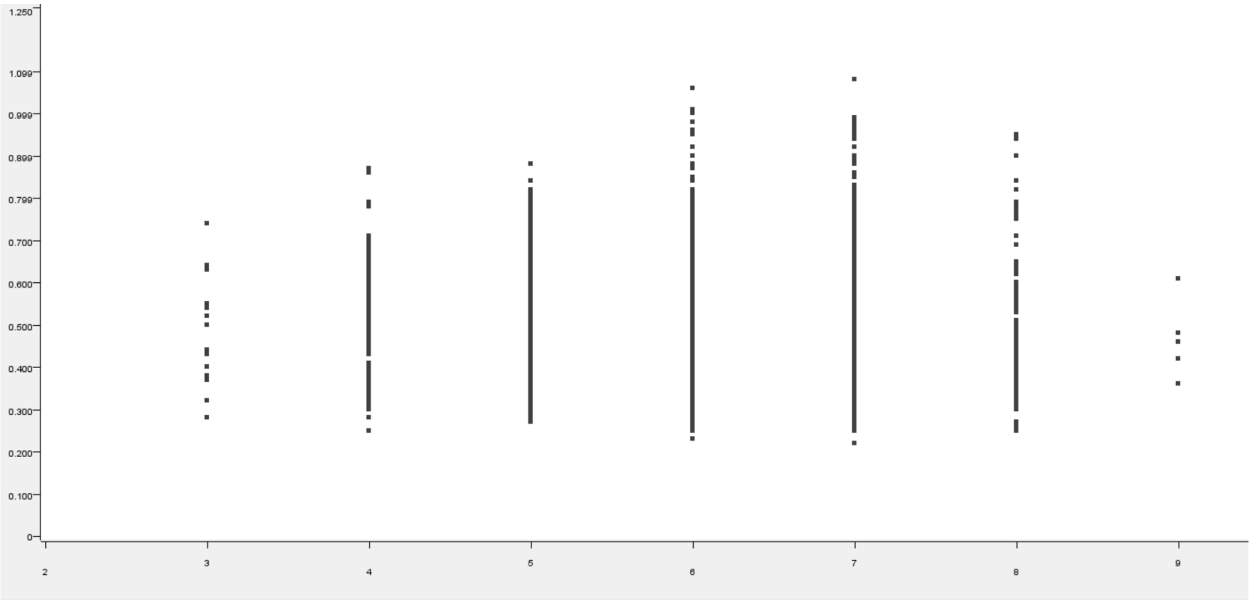
Because the quality variable is discrete, a decision tree could also be made for this data set. This however would not answer our question as directly. It could answer the possible question of how all of the features interact to create a good wine but this is not our goal.

I did not use any unsupervised learning methods because my analysis is not open ended. I have a specific question about how a target variable interacts with a predictor variable. I am not looking for patterns in the data to gain insight about the dataset as a whole.

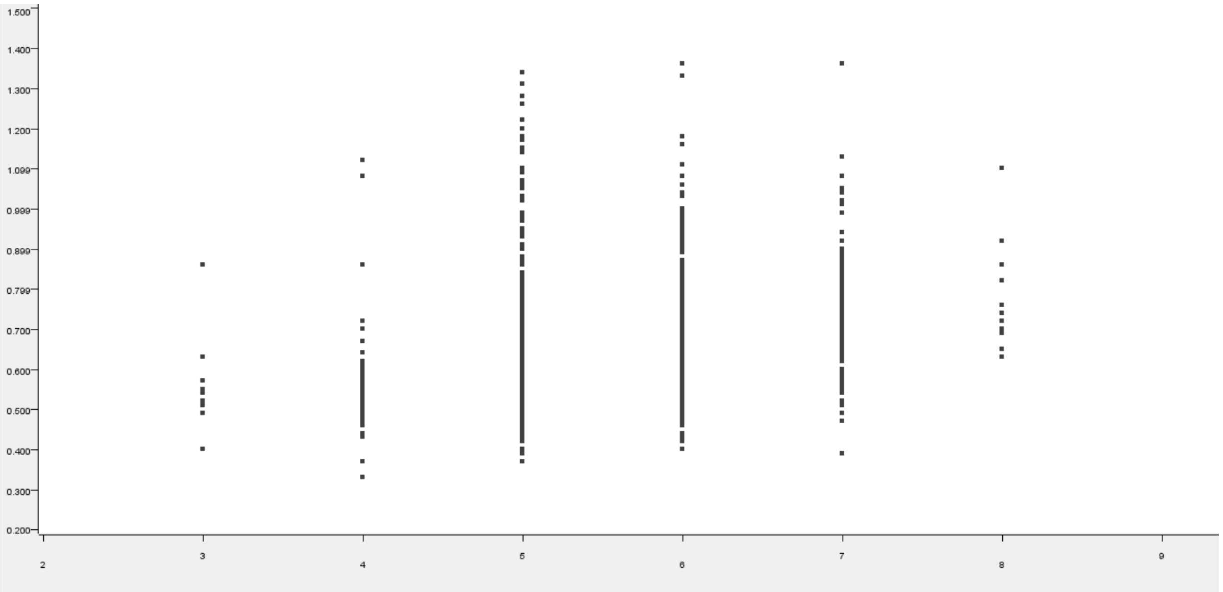
Linear Regression

Linear regression is a supervised machine learning prediction method which assigns variables with weights based on how correlated they are with the target feature. In a simple case with a single variable, linear regression creates a line of best fit on a plane where y axis represents values for the target variable, and x axis represents the values of the independent variable. The line is created so that the total distance between every data point and the line is minimized. The distance between the line and each individual data point is squared and then added together, this is called the sum of squared residuals. Reducing this number as low as it can go is how the slope and intercept of the line are decided. Dividing this number by the number of data points also gives us an important value, the mean squared error or MSE. The MSE is how the accuracy of the model is judged. The higher the MSE, the less the ability of the model to accurately predict. The Square root of the MSE lets us know how far the average point is away from our data points. The Square MSE is essentially how good the model is in terms of the data points that we are dealing with. This is how the model from this project works but instead of being on a 2-dimensional plane, our model and line is on a many dimensional plane.

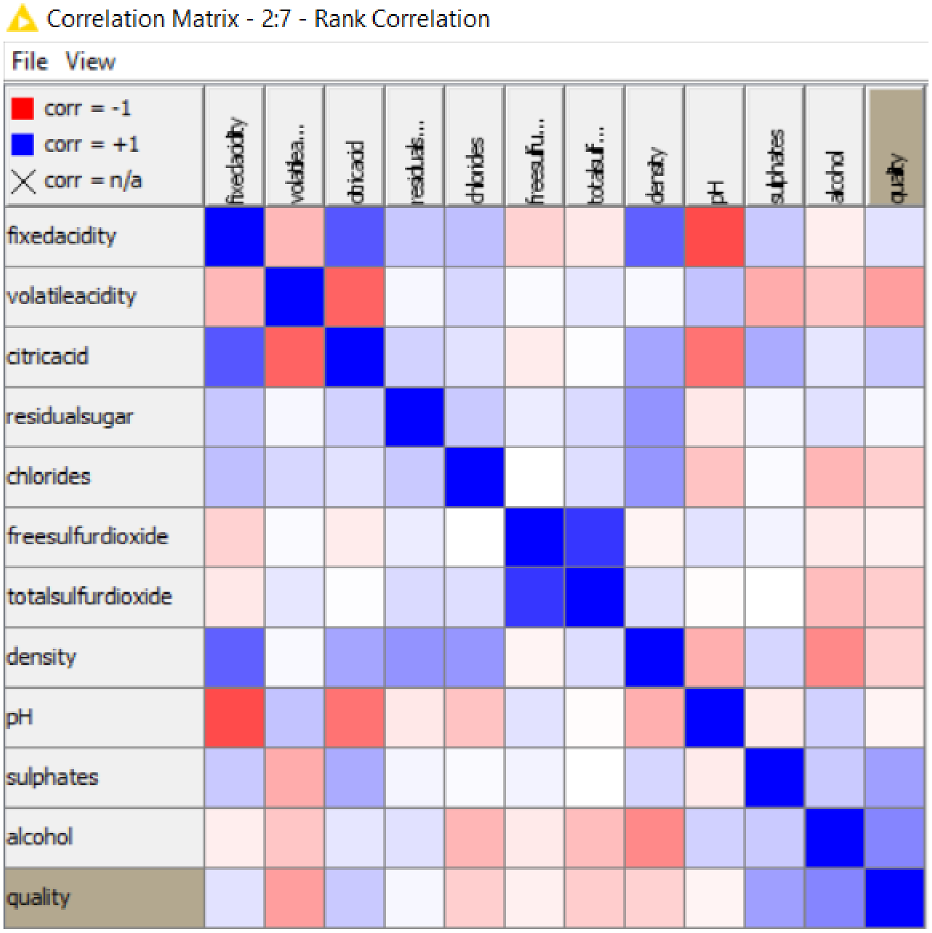
Interesting Finds in Initial Data exploration.



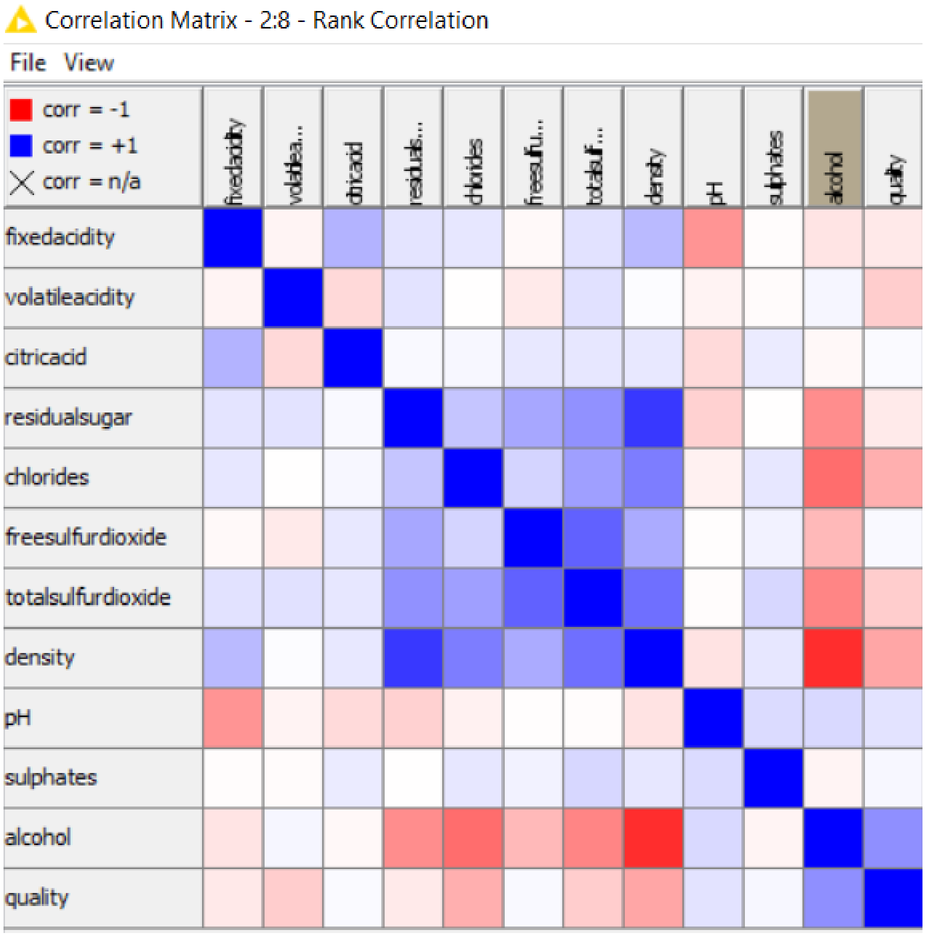
This is the a scatter plot for the white wine data set. The X axis is quality, the y axis is sulphates. It is interesting how wines in the middle of pack represent a large range of sulphates levels. Also this scatter plot alone answers the central question of our analysis. Notice how the wines with the highest quality have a small, low range of sulphate values.



This is a similar scatter plot for the red wine data set. The middling red wines also represent a large range of sulphate values. The quality red ones however have a larger and higher range of sulphate levels.



This is a feature in KNIME called correlation rank that allows us to look at the correlations between coefficients. The deep blue color represents a perfect positive correlation, the deep red color represents a perfect negative correlation. The deeper the color the stronger the correlation, the lighter the color the weaker the correlation. This first table is for the red wine data set. With this particular statistical method, the correlation is .377 and a middling blue color. From this visualization we can see that there is a decent correlation between sulphates and wine.



For white wine we see that there is almost no correlation between quality and the level of sulphates. The white square represents a correlation value of .0333.

Regression Models

For the regressions, please refer to the .HTML file attached along with this project. The .HTML file is jupyter workbook with the regressions, the sequential logic behind them, and possible interpretations of their results.

Conclusion

The conclusions are different for red and white wine. Our original question is if sulphates have anything to do with wine quality. More specifically, is there a negative relationship between the level of sulphates and the quality of the wine. For both red and white wine, we have parts of our analysis that point different ways.

With red wine, our final regression does suggest that there is a significant negative relationship between quality and sulphates. However, the correlation that we found in KNIME seems to say that there is a significant positive relationship between quality and sulphates.

With white wine, according to our regressions there seems to be little relationship with sulphates and quality. Even though there is not a relationship over the range of qualities, we see from our scatter plot that high quality white wines have low levels of sulphates.