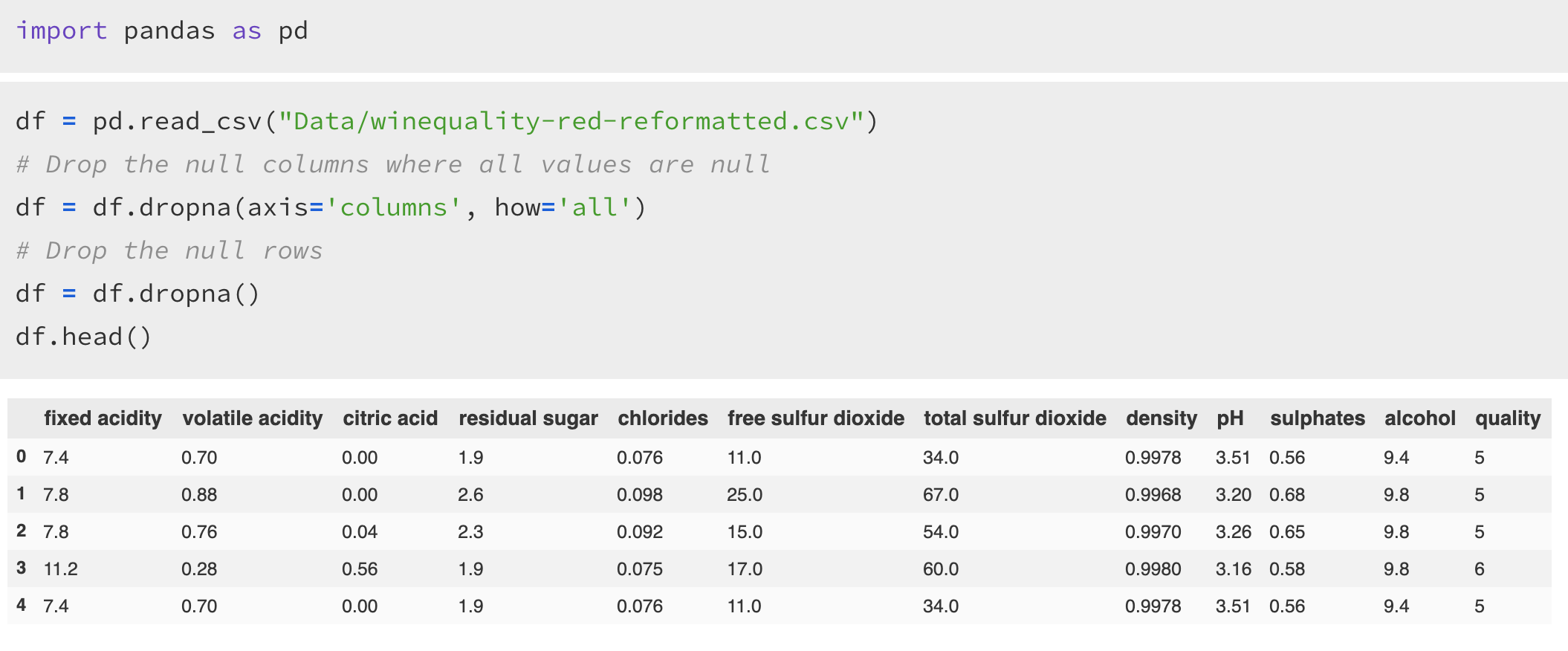
RED WINE

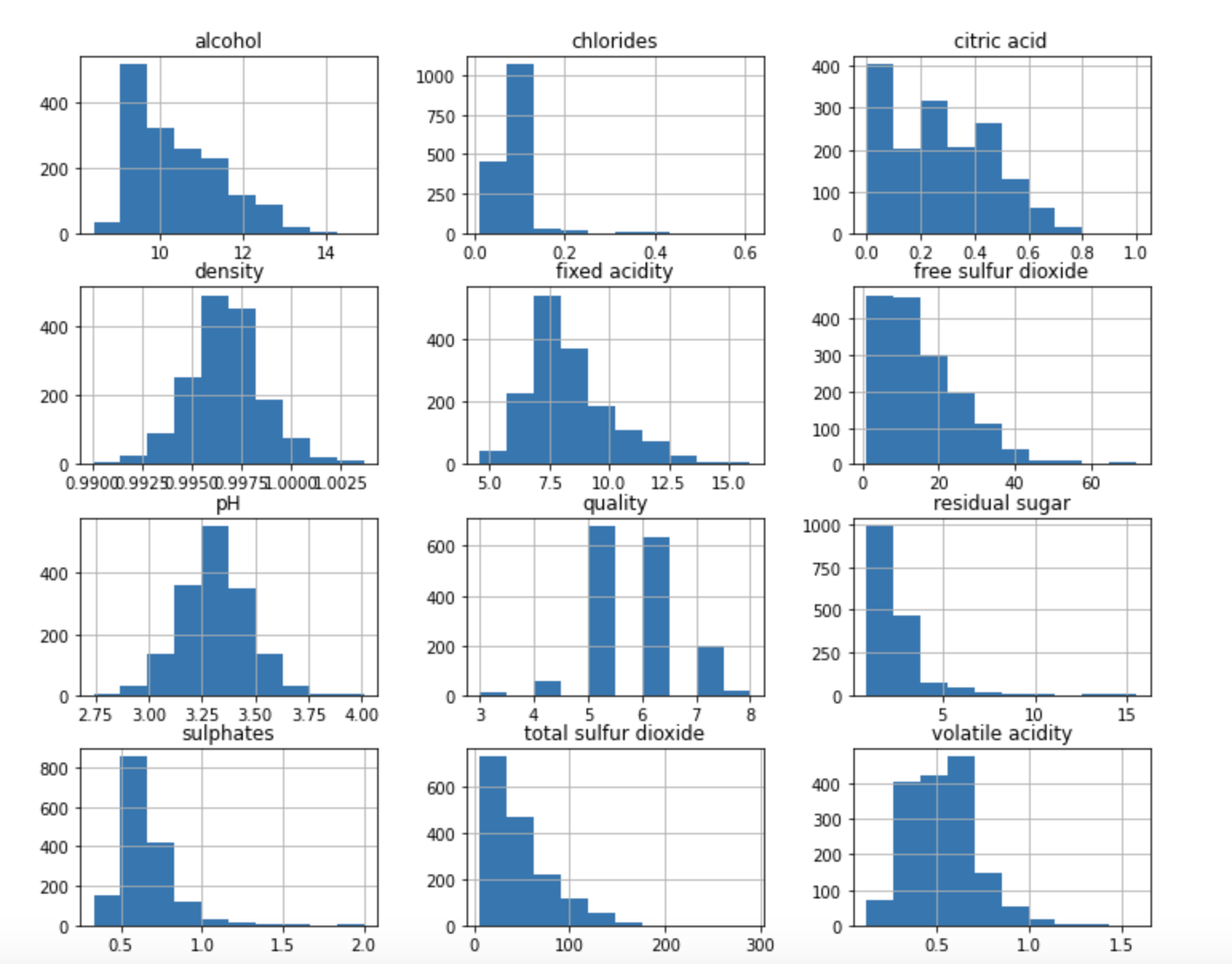
**Objective**: Building a regression model that would predict quality of red wine based on physicochemical properties: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol.

**Exploring the red wine dataset**:

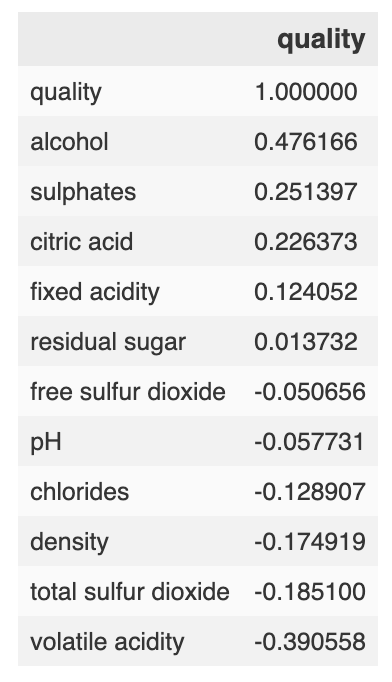


* There are 1599 records in this dataset. There are in total 12 variables, of which “quality” is our target that has a scale of 0-10.

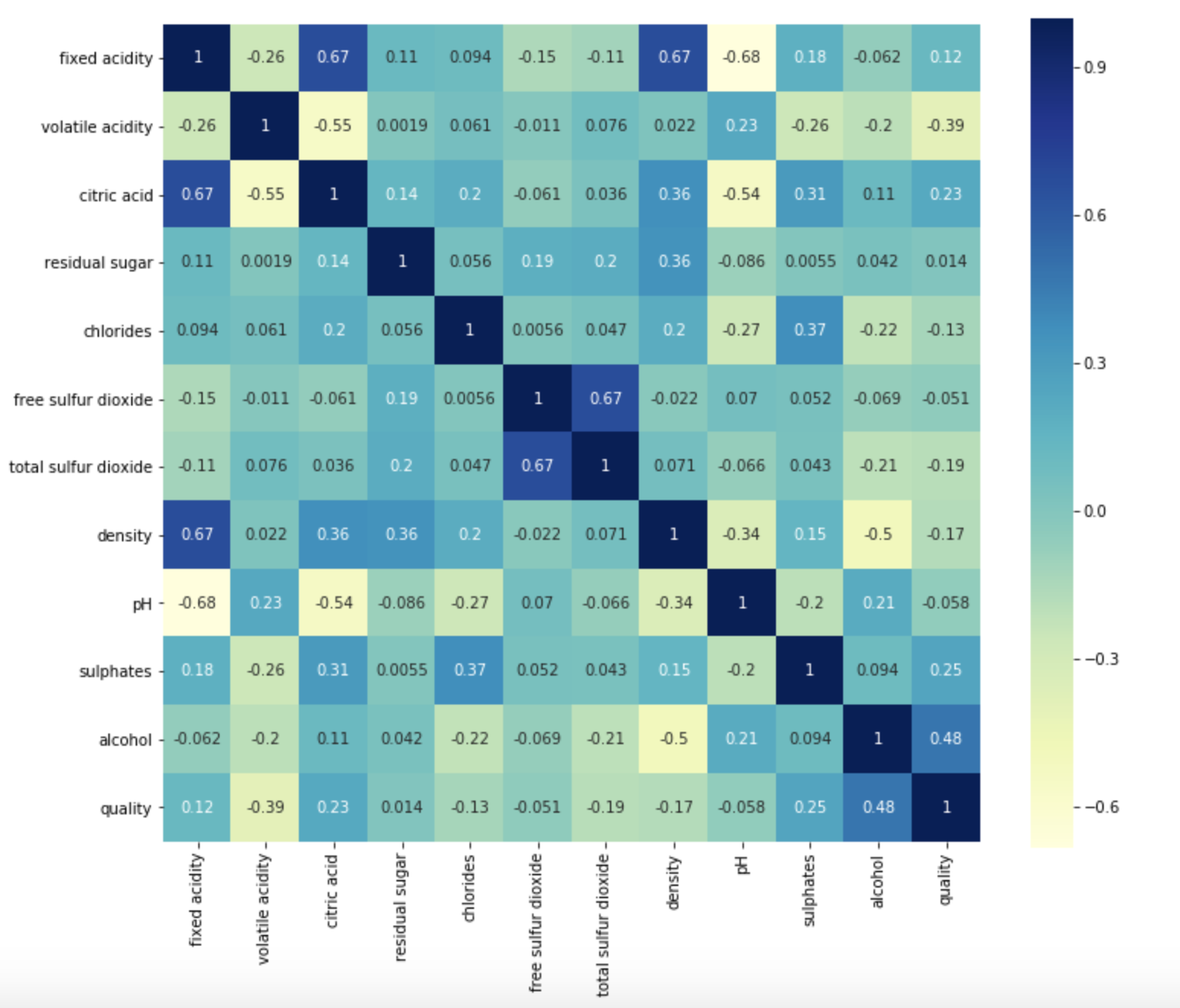
**Distribution of each variable**:



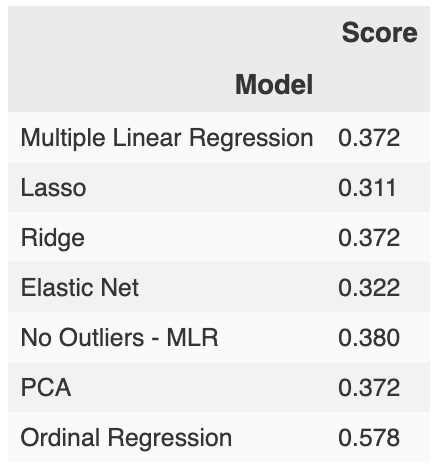
**Correlation between each independent variable and quality**:



**Visualizing the correlation between each variable using seaborn heatmap**:



**Test Scores by Different Regression Models**:



* For each model, data was split into training and testing data and were scaled and normalized using MinMaxScaler, which transforms each feature so that it lies within the range 0 and 1.
* **Multiple Linear Regression**: With Scikit-Learn, we created a model using Linear Regression, fitted the model to the training data, and quantified the model using the scaled data. Score came out to ~0.372, which is very low and weak.
* Proceeded to try other linear regression models, specifically **Lasso, Ridge, and Elastic** **Net** to see if we could improve our performance.

*[“Lasso will eliminate many features, and reduce overfitting in your linear model. Ridge will reduce the impact of features that are not important in predicting your y values. Elastic Net combines feature elimination from Lasso and feature coefficient reduction from the Ridge model to improve your model's predictions.”](https://medium.com/@vijay.swamy1/lasso-versus-ridge-versus-elastic-net-1d57cfc64b58" \l ":~:text=Lasso%20will%20eliminate%20many%20features,to%20improve%20your%20model's%20predictions)*

However, results were not any better than what we got initially with our multiple linear regression model.

* Next, we thought outliers could be an issue so we removed outliers and conducted the multiple linear regression model again to see if our model performed better. It only improved by 0.01, making no significant difference.
* Looking back to the seaborn heatmap we had created earlier, we noticed that there was some correlation between variables, and to address any multicollinearity in our dataset, we tested Principal Component Regression. Again, no difference in score.
* Finally, after taking another closer look at our dataset, it seemed that none of the preceding models were fitting given that our output variable (quality) was ordinal. Hence, we thought an ordinal regression would be more appropriate and deliver a better model as “quality” is measured on a 10-point scale, with 0 being bad and 10 being excellent. Although not robust, we see a significant improvement in our score.