Assignment 1: Jimmy Wrangler, Data Explorer

Nishil Parmar (2917887)

September 13, 2018

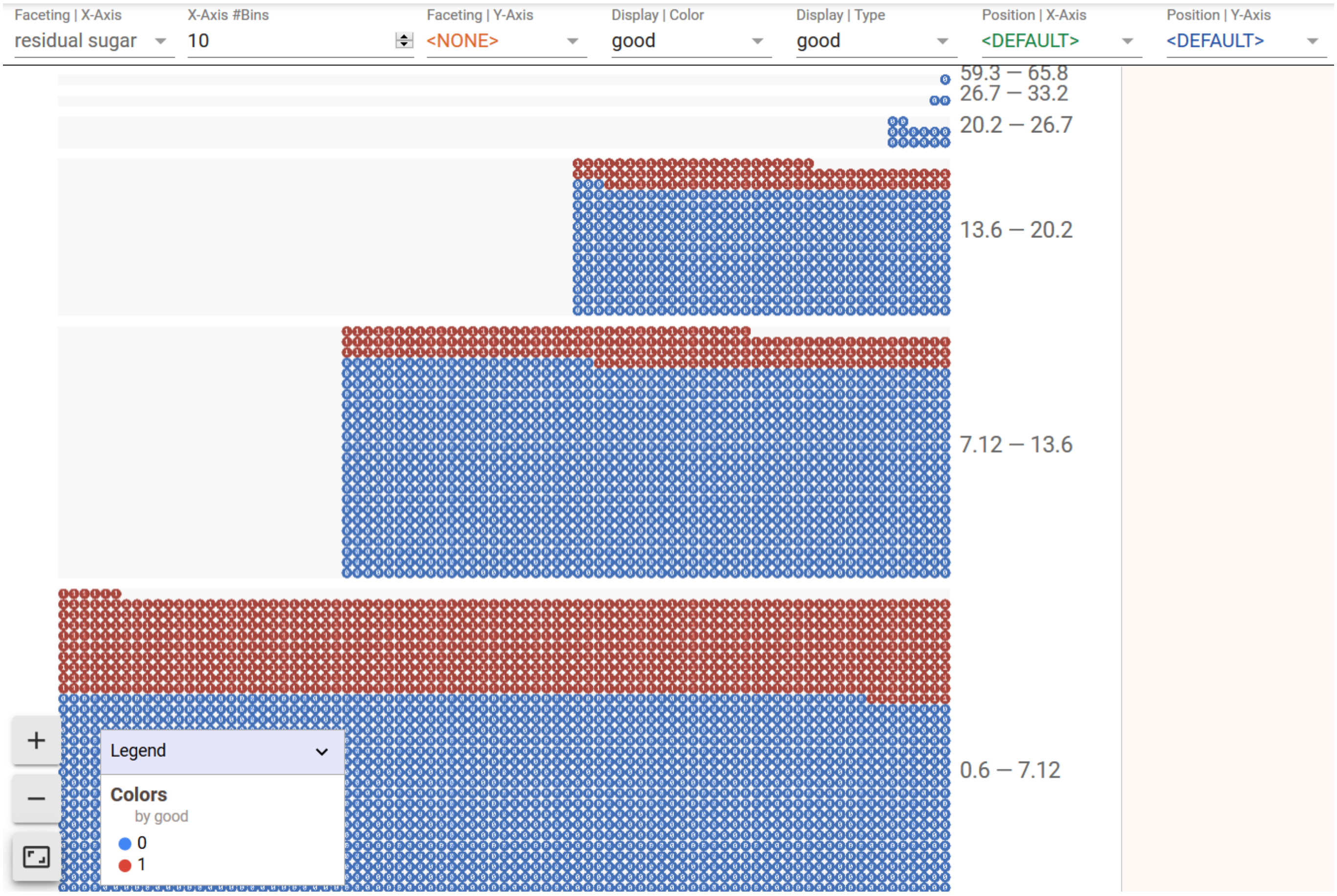
Introduction

This report describes the approach i used to search for apt public data sets that could be used for this assignment, its preparation and combination, visualizations and results produced using Jupyter Notebooks.

Data Sets

I used Kaggle.com to search for public data sets. I ended my research with ”Red Wine Quality” dataset from UCI Machine Learning repository and a dataset for ”White Wine Quality”. These datasets can be viewed as classification and regression tasks. The inspiration behind using this datasets was to use Data Science to determine a wine of good quality based on its chemical properties like alcohol percent, pH level, residual sugar, etc. Both the datasets had same columns so combining them was fairly easy just by using pandas.concat() function.

I used Google Facets to perform data exploratory analysis on my datasets. Initial exploration revealed some interesting results.



Faceting | X-Axis residual sugar
X-Axis #Bins 10
Faceting | Y-Axis <NONE>
Display | Color good
Display | Type good
Position | X-Axis <DEFAULT>
Position | Y-Axis <DEFAULT>

Figure 1: Exploratory analysis showing association between good quality wine and residual sugar

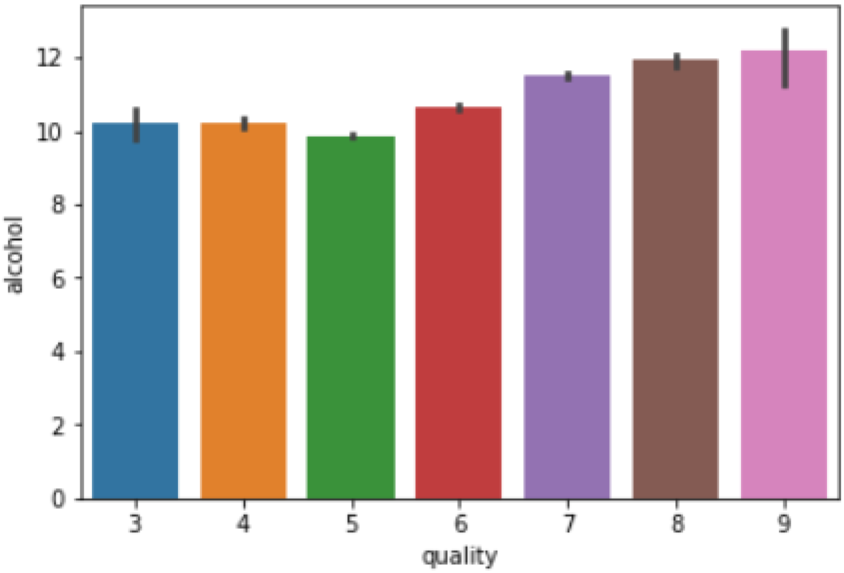
Data Preparation

After combining the datasets using pandas.concat() function into a pandas dataframe, i checked for duplicates and dropped them. I checked rows for missing attribute values and found none. If there were missing attribute values i planned to drop those records or replace the missing values with mean value if the attribute is numeric.

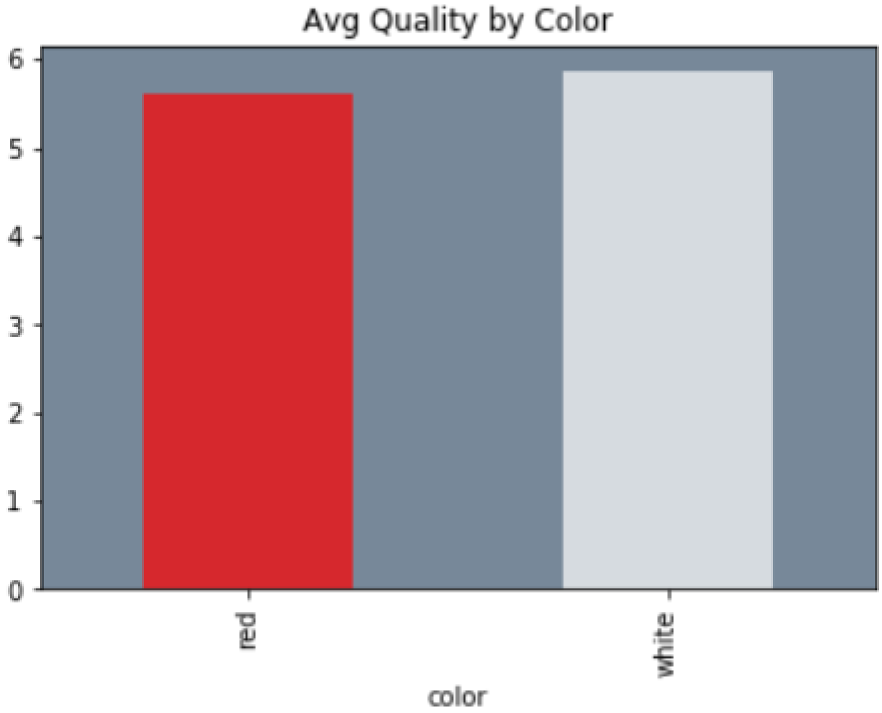
Results

While creating visualizations, i tried to find out which are the most and least significant attributes in terms of wine quality.

- Higher alcohol percent in wine is associated with good quality wines



- White Wine is associated with slightly higher quality



References

[1] https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
[2] https://www.kaggle.com/aleixdorca/wine-quality