**CEDB1160 Project**

| **Name** | **Date** |
| --- | --- |
| Samer Kareshe | 30 Nov 2019 |

**Resources**

* Python script for your analysis: WhiteWine\_Analysis.py
* Results figure/saved file:
  + Images\ correlation-winequality-white
  + Images\ whiteWineQuality
  + Distribution for Quality Differences(real-predicted)
  + V2 Distribution for Quality Differences(real-predicted)
* Dockerfile for your experiment: I will provide it by next week

**Research Question**

My approach for data analysis is focusing on enhancing the white wine quality based on the existing testing results for the below features

|  |  |
| --- | --- |
| Fixedacidity | volatileacidity |
| citricacid | residualsugar |
| chlorides | totalsulfurdioxide |
| freesulfurdioxide | Density |
| pH | sulphates |
| alcohol | quality |

**Abstract**

* **opportunity**: we have a 4,898 physical tests on white wine for all mentioned features
* **action**: increase Alcohol feature by 1% from the mean and apply machine learning to predict the quality, and finally compare the difference disruption for both schemes
* **resolution**: quality enhancement by increasing the wine alcohol feature

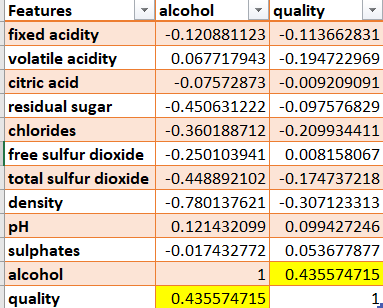
**Introduction**

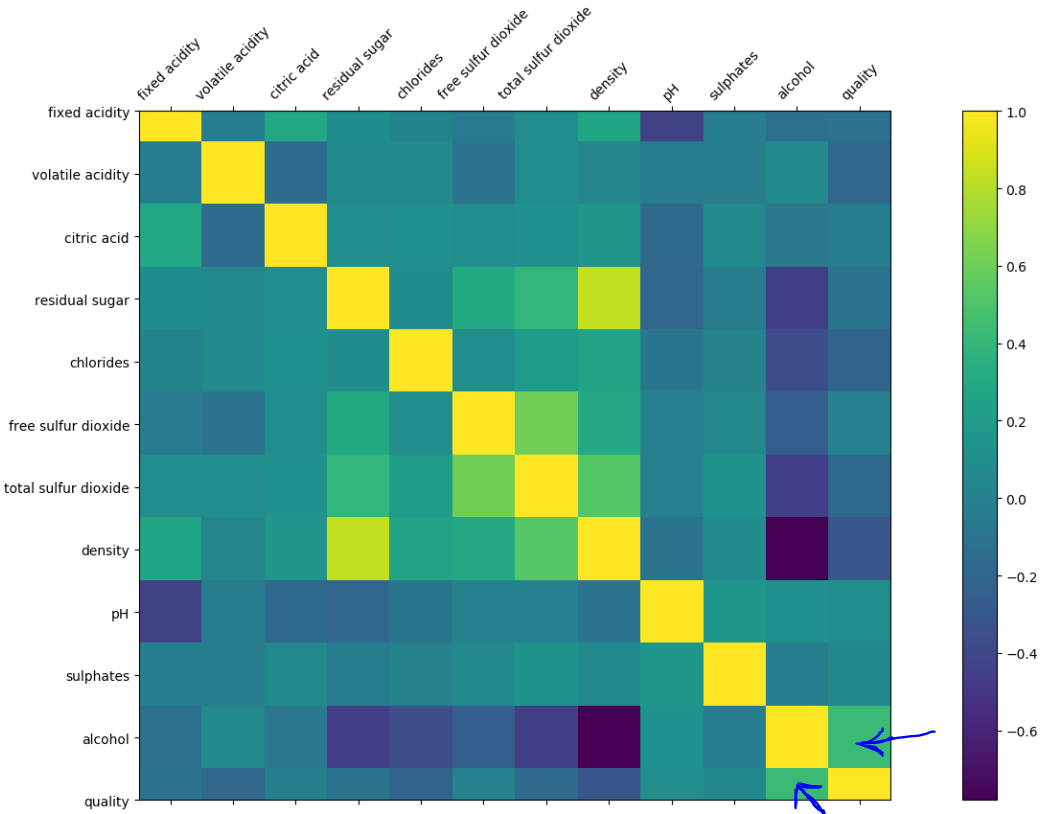
The white wine dataset is related to white vinho verde wine samples, from the north of Portugal. The goal is to model wine quality based on physicochemical tests

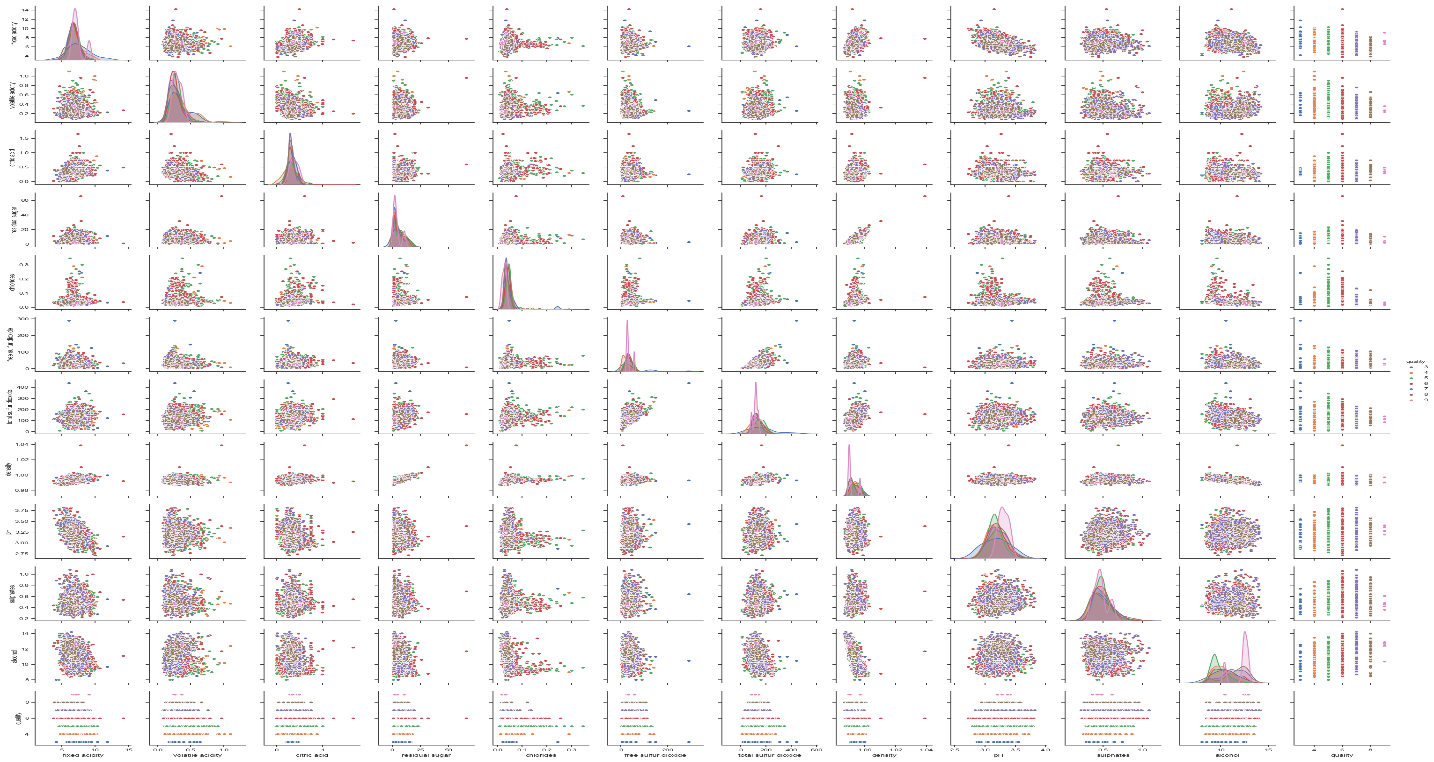
<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>

**Methods**

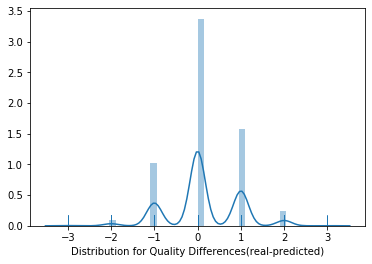
* conduct a correlation exercise and highlight the high correlated features with the quality one
* we have high correlation between Quality & Alcohol as below



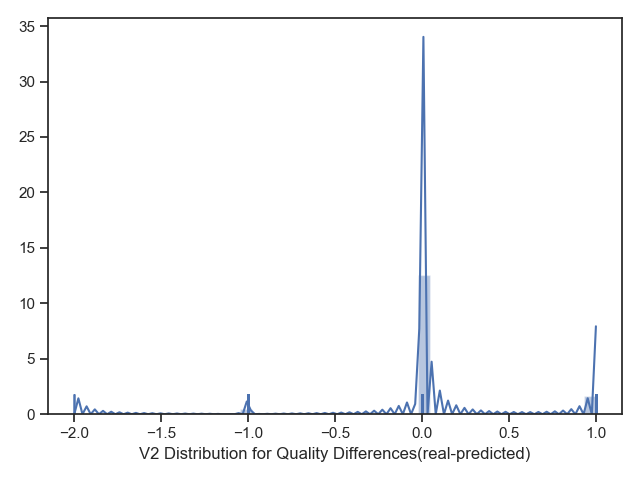


****

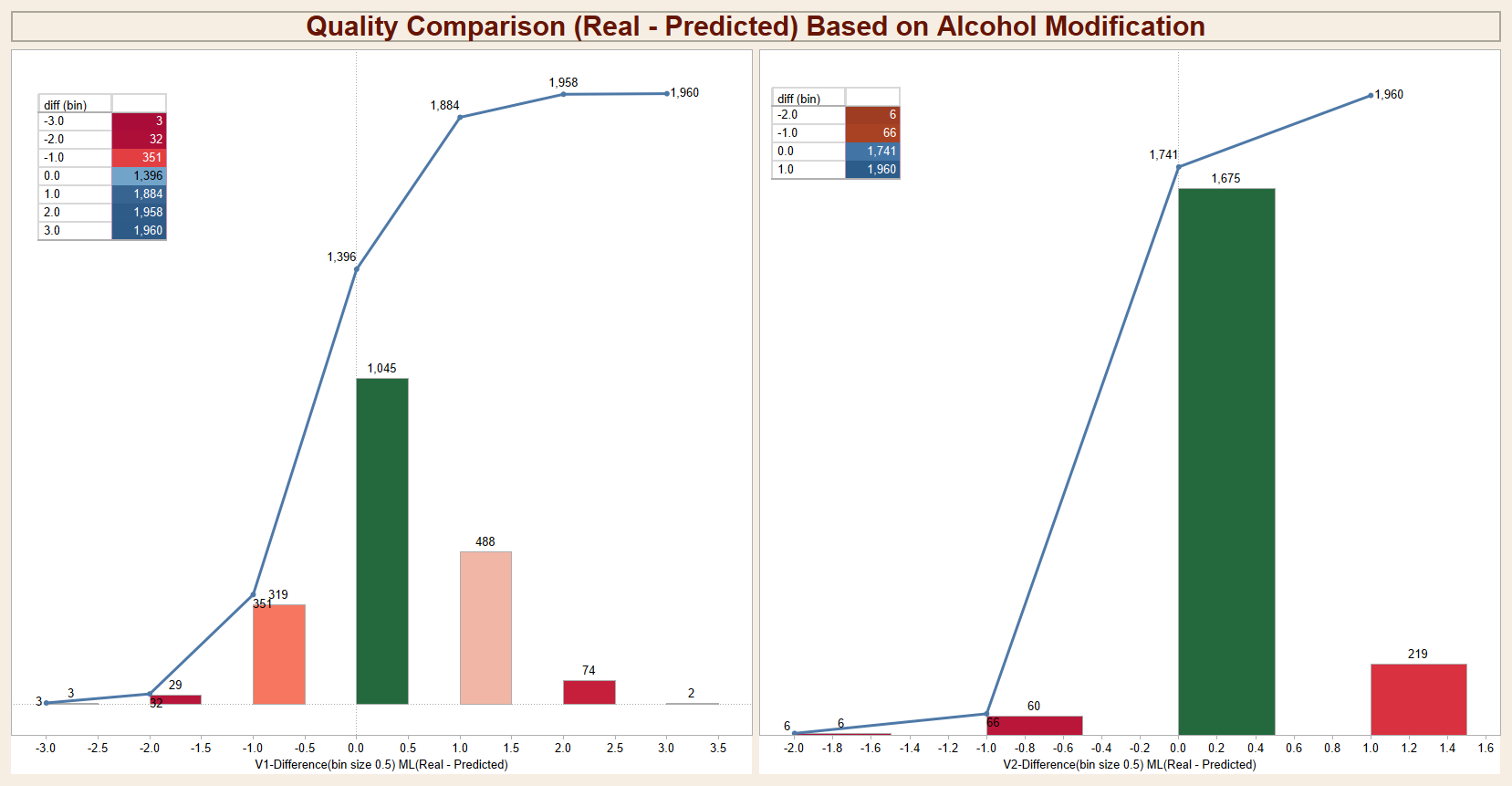
* apply machine learning and predicted the quality as a target and plot the distribution value for differences between real and predicated one

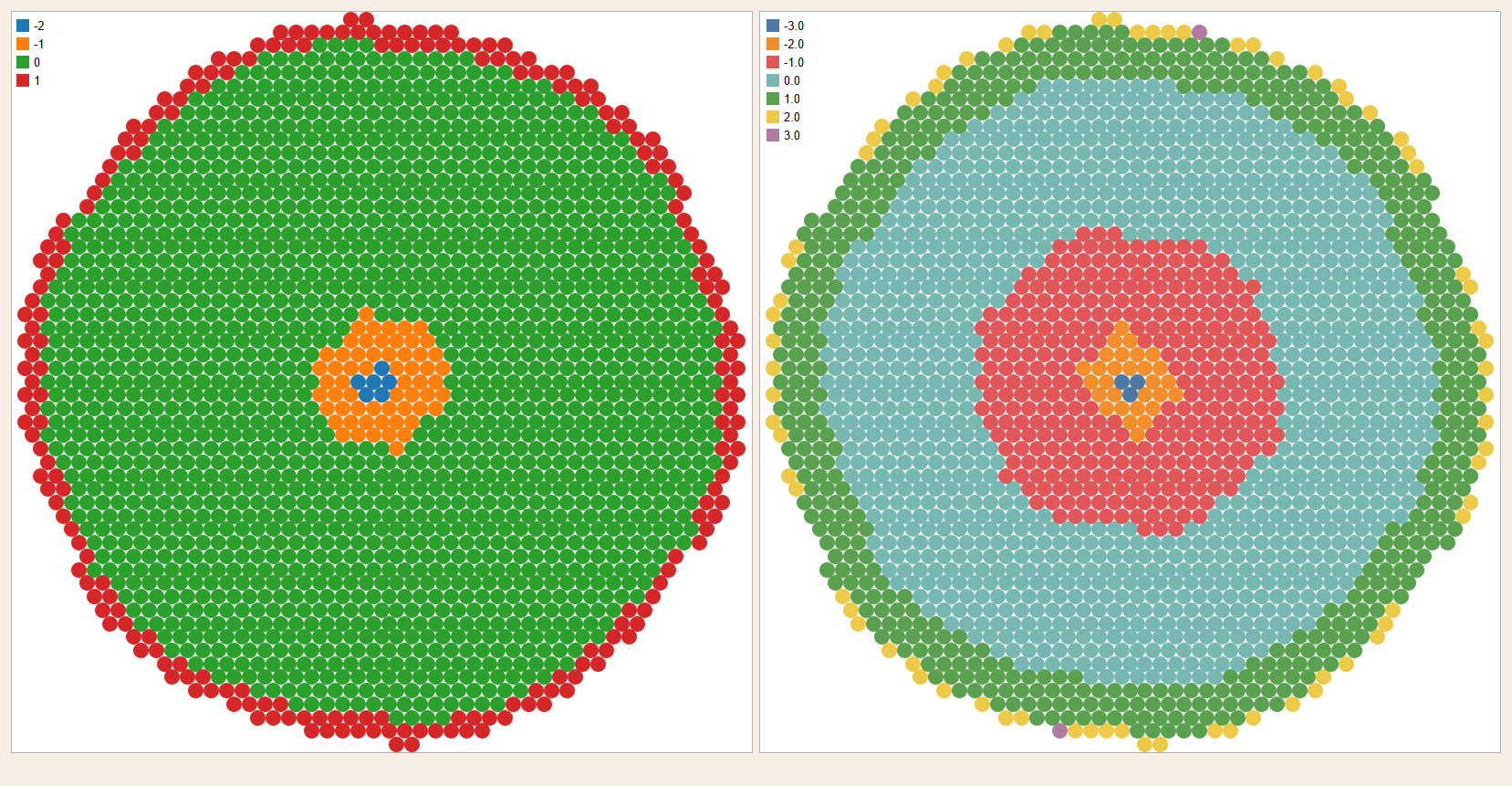


* replace alcohol feature by alcoholplus =alcohol + 1% (alcohol mean) and repeat the previous step
* as a result, the distribution plot became more accurate and less variation

****

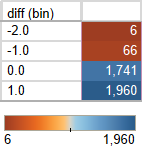
* I have saved the plots-based data and used Tableau BI on the top of them in order to provide more clear insights as below



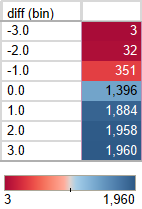


**Results**

* The differences variation rate has been enhanced by the assumption of alcohol feature modification
* First scheme distribution rages for differences



* Second scheme distribution rages for differences



* Conclusion: we could have better white wine quality by increasing alcohol %