Project Update #2

# Kanban Board:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **To-Do**  **WIP Limit** | **Preparation**  **3\*** | **Analyze**  **2\*** | **Deploy**  **2\*** | **Finished** |
| Understanding the business context | Define  business case  Research wine scoring standards | Compare data sets to Industry Standards |  |  |
| Understanding Data | Chose data set | Evaluate |  |  |
| Data Capture and Extraction | Installed Rstudio  Import Data sets to R |  |  | Data Imported |
| Data cleanup | Check for missing values  Validate data types  Check for additional spaces (trim data) |  |  | Data set ready for analysis |
| Modeling Data | Define Data Models to be used | Evaluate additional models aside from the Random Forest | Apply data models to data sets |  |
| Visualizing data | Add packages to R for data visualization | Created initial visualization | Select final set of visualization |  |
| Data and Analytics Validation/Testing | Define methods of validation  Define what data needs to be validated | Upload data to Excel | Validate data against results in Excel  Discuss with SME for validation/ feedback |  |
| Results sharing/write ups | Set meeting schedule  Created Github repository | Review code and plan |  |  |

# 

# Project Summary #2

|  |  |
| --- | --- |
| **Accomplishments**   * Cleaned up datasets * Created Heatmaps * Created Histograms * Created Correlation Matrices | **Working Well**   * Identified convenient meeting time * Creating initial visualizations |
| **Plans for next update**   * Machine learning techniques to see if we can train the system to pick a good wine   + Random Forest * Start writing report | **Issues**   * The datasets only have one column of data   + This issue was fixed on 11/19/2019 |

# Project Summary #1

|  |  |
| --- | --- |
| **Accomplishments**   * Defined goals * Chose data set * Generated Kanban Board * Set up recurring meetings | **Working Well**   * Identified convenient meeting time |
| **Plans for next update**   * Create a table with the variables and their description * Finding out if there are any missing values/ “Clean up” dataset * Look if there is a relationship between Wine Quality and Alcohol content in the wine * Create a correlation matrix | **Issues**   * N/A |

# Goals

## Main Goal:

To create a model that predicts the rating of a wine.

## Secondary Goals:

Understand which variables are responsible for the quality of the wine.

* What is the minimum set of properties and their values that define a high-quality wine?
* What are properties which make wine second-rate?

# Business Case

Evaluate a data set and creating a model that evaluates wine quality, will benefit the wineries by allowing them to evaluate small batches before beginning full production. Additional analytics provided will support different quality setups to drive higher ratings for red wine vs white

# Dataset

Wine Quality dataset is publicly available for research at the University of California, Irvine Machine Learning repository created by Paulo Cortez (Univ. Minho), Antonio Cerdeira, Fernando Almeida, Telmo Matos and José Reis (CVRVV) in 2009.

**Citation:** P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

Available at https://archive.ics.uci.edu/ml/datasets/Wine+Quality