Humberto Leon, Joao Lucena, Armando Zubillaga
DS3001 Foundations of Data Science
Professor Kyumin Lee
April 21, 2020

# Project Proposal

- **The name of your team and the team members.**

Name: Wine Tasters

Members: Humberto Leon, Joao Lucena, Armando Zubillaga

- **What is the need? Who wants or benefits?**

The need of our project is to predict the wine points based on features such as region, country, description, and grape variety on the dataset. Essentially, wineries and wine enthusiasts would benefit from this analysis so sellers can value their produced wine more accurately and enthusiasts would know which wine to buy according to their preferences.

- **What data (or datasets)?**

We are going to be using a dataset of Wine Reviews extracted from Kaggle website. The creator built the data sets from the WineEnthusiast magazine during the week of June 15th,  2017. It also has a data set from the week of November 22nd, 2017 with additional information. The idea would be to combine both datasets to get more historical data.

https://www.kaggle.com/zynicide/wine-reviews

- **What is your "data science" toolkit? You should list specific tools / packages you will use.**

We will use Jupyter Notebook as our “data science” toolkit where we will be writing our code in Python. Pandas and Numpy will be used to import the csv file and for data cleaning. Matplotlib and Seaborn will be used for data visualization. Scikit-learn will be used to apply machine learning algorithms to the data in our project.

- **Preliminary sketch of what you hope to build**

- Import data from csv files
- Clean data by extracting only columns needed for project
- Clean up duplicate entries of same wine

- Separate wine ratings in clusters
- Use multiple linear regression to predict wine points
- Analyze regressions to decide what factors influence wine ratings the most
- Predict wine ratings based on the factors chosen