MSD692 Data Science Practicum

Project Proposal

**Project Title:** Wine reviews analysis and visualization

**Description:** Analysis, modeling and visualization of wine review data.

**Data Science task types:**

Data acquisition, cleaning, preparation, normalization, modeling and visualization. Text analytics including sentiment analysis and predictive and topic modeling research.

**Data description:**

Wine review data from <https://www.kaggle.com/zynicide/wine-reviews>. The data consists of 130 thousand wine reviews from 11/24/2017. Data consists of wine rating, title, variety, description, country, province, region 1, region 2, winery, designation, price, taster name, taster twitter handle.

**Data Analysis methods used:**

Hoping to have time in the timeline for some visualizations in the Tableau reporting tool. Given current research with R having GGPLOT and Python having a Seaborne package for visualizations the Tableau work may not fit into the timeline. Also will discuss in the timeline the choosing of types of models to run but currently planning on doing exploratory data analysis on the data first and then choosing the model appropriately.

I want to look at different modeling ideas and see if there is a correlation between score and price which is mentioned on Kaggle already, but still want to repeat those results. I also wanted to look at breaking the words up in the description to features and see if any specific feature or features relate to pricing or wine score. Then using kernel density most likely using Python or Tableau I would like to product a map visualization on some of the key findings.

Research will be done on LDA, LPAR, Naïve Bayes, Random forest, feature extraction, kernel density and other techniques. Narrowing down what to use will be done in week 3 and other techniques and tools can be identified for use in the next Practicum class.

**Expected challenges and mitigation plan:**

While researching the data set I see issues with special characters in the data and some reports from Kaggle of duplication in wine review column. The data quality and preparation tasks will be given extra time in the timeline. Several text analytics ideas can be used for this data so in week 2 the actual methods will need to be finalized so that they can be given appropriate time in project timeline. It will be essential that by end of week three the models and tools are chosen so that correct data cleansing and time for modeling is reserved.

**Project timeline:**

Week 1 and 2 will be spent researching data sets and modeling techniques. End of week 1 will be spent on initial proposal. Week 2 will be spent doing more research and revising the project plan based on research and proposal feedback.

Weeks 3-6 will be spent on the data analysis including data preparation and cleanup. This is where final model will be chosen and run most likely will be using Python.

Week 7 I would like to spend completely focusing on visualizations in appropriate tool and perhaps Tableau if time permits.

Week 8 will be spent preparing presentation and posting and publishing final version.

**References:**

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. *J. VanderPlas.–O'Reilly Media*.