Project Proposal

Team Members: Xinyang Liu(xl9qw), Boh young Suh(bs6ea), Ni Zhuang(nz4gg)

Data

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

1. Identify the Problem

Not everyone can identify wines through blind tasting like a master sommelier. However, with a predictive model, we will be able to identify wine more easily based on the description and some important features. Some wine lovers who are not professional in wine tasting and want to make better purchase decisions can show high interest in our research and deep learning to predict wine ratings. Successful development from a training model to predict wine quality will enable wine tasters to better predict variety of wines and this will also enrich their wine tasting experience.

2. Define Objectives and Metrics

- What seems to be associated with high review scores and high prices?
- Can the descriptions on wine bottles affect their scores or prices?
- Is there a positive correlation between review scores and prices?

The main objective of our project is to optimize buyers' wine selection process, so that non-professionals can also find the ideal wine. Our obtained model should demonstrate the relations between variables and be able to predict the score and price of wine based on given information. We will measure how well our model predicts the wine ratings by performing cross validations, etc.

3. Understand the State-of-the-Art

Kaggle shows 150K with variety, location, winery, price and description and our team will be using this dataset to build our model. We plan to use both parametric and non-parametric, linear and non-linear statistical method to analyze our data. In addition, with a descriptive column in our data, we will also make use of text mining to uncover the underlying relationship between each variables.

4. Define Hypotheses and Approach

Hypothesis

- The descriptions on wine bottles affect the wine's score and price (Whether or not certain words are included)

- There is a positive correlation between review scores and prices
- The place of origin will affect the wine’s prices and ratings

Our group plans to apply various data mining methods learned in class such as basic linear model, text mining, KNN, Random Forest to test our hypothesis mentioned above. With text mining, if descriptions with more positive words lead to better scores or prices, then it supports our assumption that the descriptions on wine bottles affect the wine’s score and price. And in terms of relation between review scores and prices, if they are significant and have positive coefficients in the model obtained for each other, we can conclude that a positive correlation exists between the two variables.