Using Wine Reviews to Predict Quality and Type

Nate Hoffelmeyer, Sam Harvey, Erin Cali

IST 718

Introduction

"Accept what life offers you and try to drink from every cup. All wines should be tasted; some should only be sipped, but with others, drink the whole bottle." –Paulo Coelho

What makes a good wine? Thinking of a good wine you've had can conjure images of many different things. From the physical traits - aroma, body, hue. To the taste – dry, sweet, tannic, fruity. In that as well you have to take into account the wines origin, what varietal(s) it May contain, and what vintage it may be. Even sommeliers can have vastly different perspectives on what a good wine is.

This poses a deeper mystery to the inquisitive mind. Are good wines really that different? Are there similarities that are consistent through the highest rated wines and the way they are described? Do the best wines come from the same regions of the wine world? What about the human factor? Do the same sommeliers laud the same kinds of wines? When presented with so many variables, there has to be a common thread amongst them.

Data Preparation

The data used for this analysis was taken from Kaggle at https://www.kaggle.com/zynicide/wine-reviews. Wine descriptions from 150,930 expert sommeliers were scraped from the Wine Enthusiast website and included the attributes describing the wine and the taster. The attributes describing the wine were: review title, winery, country, the vineyard within the winery from whence the grapes came referred to as designation, points Wine Enthusiast has rated the wine, bottle price, province or state of the wine referred to as province, region of the wine, and the grape varietal used in the wine. Attributes for the taster were their name and Twitter handle. During data preparation, null and duplicate values were removed which reduced the number of descriptions to 92,393.

This analysis presented an opportunity to explore a dataset that provided an enormous potential in order to predict the possible quality of a group of wines easily. When attempting to analyze a dataset this vast; some difficult decisions had to be made in reducing the amount of information being processed for the purpose of clarity and being able to attain intended visual results. Wines were rated on a scale of 1-100, but the lowest score in the dataset was 80. The distribution of scores can be seen in Figure A with the majority of the wines centering around the mean of 87.75.

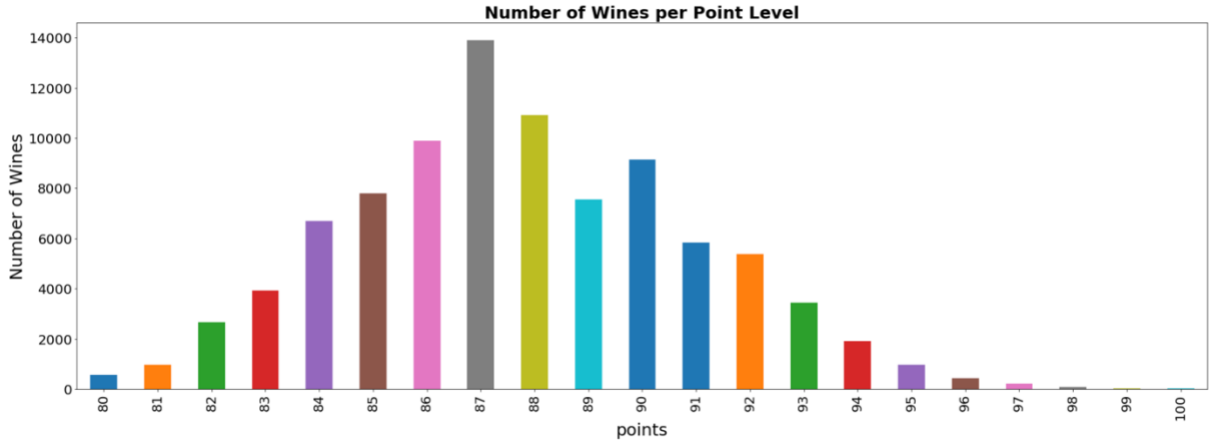

Figure A: Number of wines categorized by points

This range of 20 scores was discretized into 5 buckets: under average wines as 1 for 80-84 points, average wines as 2 for 84-88 points, good wines as 3 for 88-92 points, very good wines as 4 for 92-96 points, and excellent wines as 5 for 96-100 points. The distribution among the simplified points can be seen in Figure B.

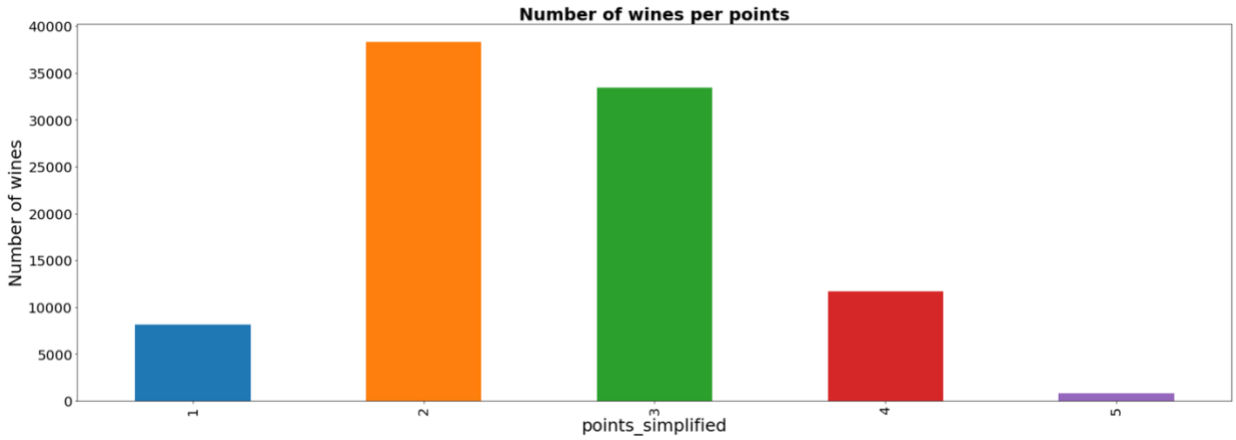

Figure B: Number of wines categorized by simplified points.

Next, the length of the descriptions was calculated. Boxplots were created showing the distribution of description lengths by original scores and simplified scores.

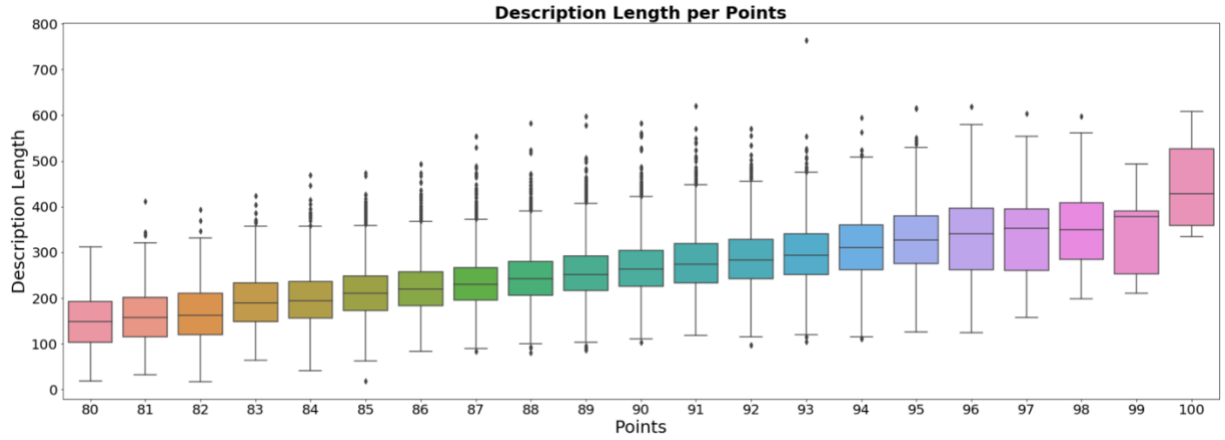

Figure C: Boxplot of description lengths by points

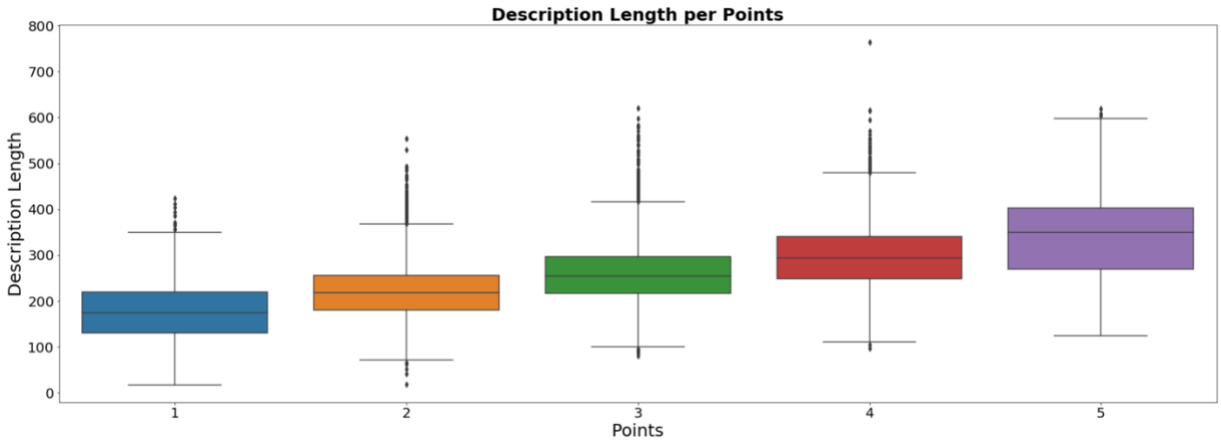

Figure D: Boxplot of description lengths by simplified points

Figures C and D show that the higher the rating, the longer the description. With distributions shown for each point value, there is more variability in Figure C. In Figure D, there is a clear positive correlation between points and length of description.

Problem Specification

Wine reviews use a lot of words, but how much information do they contain? A wide range of language is used in describing wine and the same wine could be described in very different ways depending on the taster. Through this analysis, wine descriptions were used to predict quality of the wine based on points and whether a wine was red or white using random forest classifier, multinomial naïve Bayes, and logistic regression.

Analysis

Analysis began by examining the points given to each wine and whether the description of the wines can predict these points. For simplification, a separate data frame, dp, was created isolating the descriptions and points. The descriptions were turned into a vector from which a sparse matrix was created. Another vector was created using term-frequency inverse document

frequency (TF-IDF) to account for the importance of different words. Random forest classifier was applied first, once with the count vector and then with the TF-IDF vector. These models, as well as subsequent models were trained on 90% of the data and tested on the remaining 10%. Each yielded 96% precision and recall. Next, multinomial naïve Bayes was performed with both vectors. For the count vector, precision and recall were 70% while the TG-IDF vector had a 70% precision and 67% recall. Finally, logistic regressions were run on both vectors. Count vector returned a precision and recall were 81%. The TG-IDF vector had a 74% precision and 73% recall. Clearly, random forest classifier is the best model despite the computational requirements.

From Agiorgitiko to Zweigelt, over four hundred types of wine were described in the dataset making it difficult to visualize the varieties. However, only 41 countries are represented in the data, making it possible, as seen in Figure D, to group varieties by country and visualize the number of wine varieties offered by country.

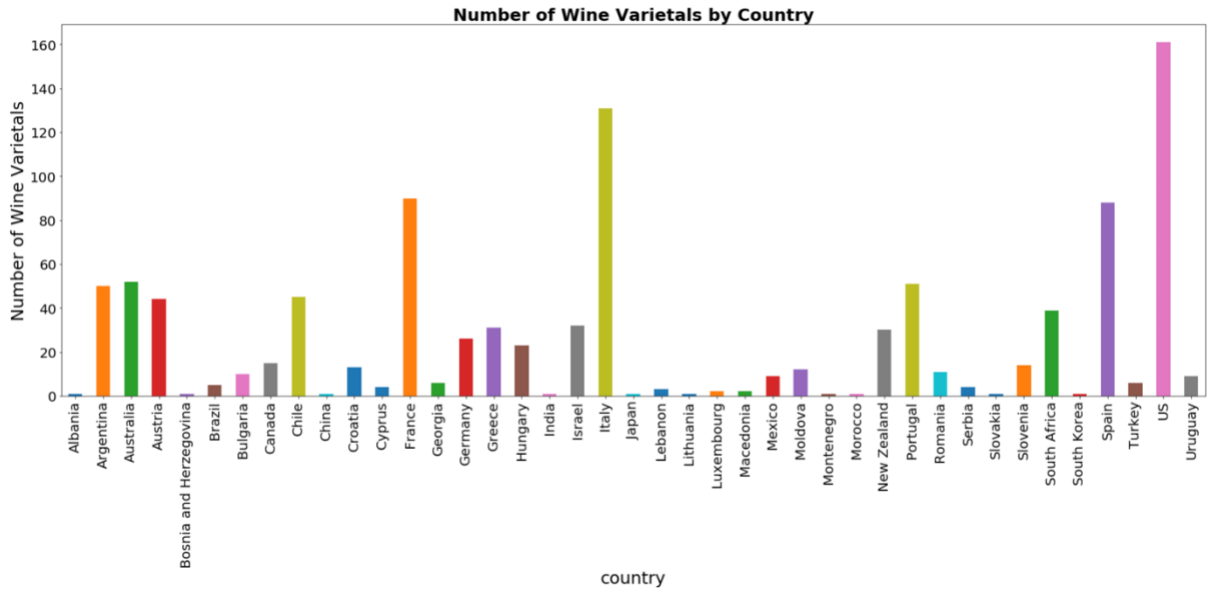

Figure E: Number of wines available from each country

Analysis on the ability to predict wine type, whether a wine is red or white, is conducted using random forest classifier. Using the count vector, 96% precision and recall were achieved. The same results were attained using the TF-IDF vector.

Observations

Observations made after completion of the models presented information in the potential refinement of the model processing. Models were found to have a distribution of wine by points slightly skewed right. Rarer wines tended to have a documentation of a higher point value, while wines with a higher point value generally tended to have longer descriptions as well, thereby offering more data to be utilized by the models. It can then be expected that most wines featured would be centered around the average quality, and can be assumed that the better quality a wine, the more enjoyable it will be to document and identify. These observations help set the basis for further tuning and refining the

models available, and what refinement should be made in order to help present a more accurate output of the models.

Recommendations

Future models can be further refined, thereby improving precision and recall. The Random forest classifier (RFC) had the highest percentage of all three models, but still contained room for further improvement. While the random forest classifer had the highest percentage, it is more computationally intense than Logistic or Multinomial models. Raising the precision of these is possible with still being less resource heavy than random forest.

Further refinement of the models presents a possibility of improving precision and recall for the RFC to 96%. Improvement of precision and recall of MNB or LE models may be less computationally intense than RFC, but with a potential loss of accuracy over it. A Support Vector Machine or Kernel Support Vector Machine may be run to potentially improve output at a cost to computational intensity that may prove not worthwhile for those model types.

References

(n.d.). Tf-idf :: A Single-Page Tutorial - Information Retrieval and Text Mining. Retrieved from http://www.tfidf.com/

Thoutt, Z. (2017, December). Kaggle: Your Home for Data Science. *Wine Reviews.* Retrieved from http://www.kaggle.com/zynicide/wine-reviews