Project 2 Proposal: Sommelier Statistics

Chris Pham, Ben Arnoldy, Matt Thielen

Summary:
Wine is a drink that is consumed worldwide by many different cultures. Many people have different criteria when they decide to purchase a bottle of wine. Thanks to Kaggle's data set, we have data to look into the global wine industry through 'WineEnthusiast'. With this data we plan to pursue the answers to the following questions while keeping an eye out for other interesting findings along the way:

- Does the price of wine correlate with better ratings?
- What region has the most expensive wine?
- What region has the best rated wine?
- What type of wine is rated the best?
- Do ratings correlate to U.S. wine import quantities from country of origin for the vintage year, changes in import quantities, or other derivable import measures?
- Do ratings correlate to major climate deviations in the wine's country of origin for its vintage year?
- Do individual sommeliers display a preference for a certain type or country of origin of wine? If a sommelier seems particularly biased, can we attempt to determine why?
- How well do clusters of wine titles and descriptions correlate with geographical origin?
- Does the vineyard's Köppen climate classification correlate to wine rating?

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

Critical Parameters

- *Points*: the number of points WineEnthusiast rated the wine on a scale of 1-100 (though they say they only post reviews for wines that score >=80)
- *Title*: the title of the wine review, which often contains the vintage if you're interested in extracting that feature
- *Variety*: the type of grapes used to make the wine (ie Pinot Noir)
- *Description*: a few sentences from a sommelier describing the wine's taste, smell, look, feel, etc.
- *Country*: the country that the wine is from

- *Province*: the province or state that the wine is from
- *Region 1*: the wine growing area in a province or state (ie Napa)
- *Region 2*: sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
- *Winery*: the winery that made the wine
- *Designation*: the vineyard within the winery where the grapes that made the wine are from
- *Price*: the cost for a bottle of the wine
- *Taster Name*: name of the person who tasted and reviewed the wine
- *Taster Twitter Handle*: Twitter handle for the person who tasted and reviewed the wine

Preliminary data cleaning will include assigning appropriate data types, categorizing categorical variables, counting word frequency in titles and descriptions, and parsing wine vintage from the title.

Supplemental Datasets:
https://dataweb.usitc.gov/ (Free Account Needed)
U.S. International Trade Commission (U.S. wine imports by year and country of origin)

https://en.wikipedia.org/wiki/K%C3%B6ppen_climate_classification
Köppen climate classification (data series will need to be programmatically assigned based on wine origin data)

https://www.ncdc.noaa.gov/cdo-web/datasets (Global Summary of the Month/Year)
https://www1.ncdc.noaa.gov/pub/data/cdo/documentation/gsom-gsoy_documentation.pdf (Documentation)
Global Historical Climate Data

Approach:
Each series within the data set will be standardized (assigned a common data type, blank values will be appropriately filled in, etc.). Then we'll approach each question independently, cleaning and joining supplemental datasets and programmatically creating new data series as required. Some questions will require single or multi-variate regression, while others will require descriptive cluster analysis or word frequency analysis. Ultimately, our intent is to explore the data, discovering and presenting new descriptive insights and may therefore exclude insignificant or uninteresting findings from the final report.

Graph 1: Scatterplot of points vs price across 129,971 reviews. This quick graph generated in R gives us a rough sense of the data shape on two key variables and guesses as to the correlation. There are also some extreme price outliers like a bottle of wine that costs more than $3,000. Most of the wines are under $500. It looks like higher ratings do seem to generally command higher prices.

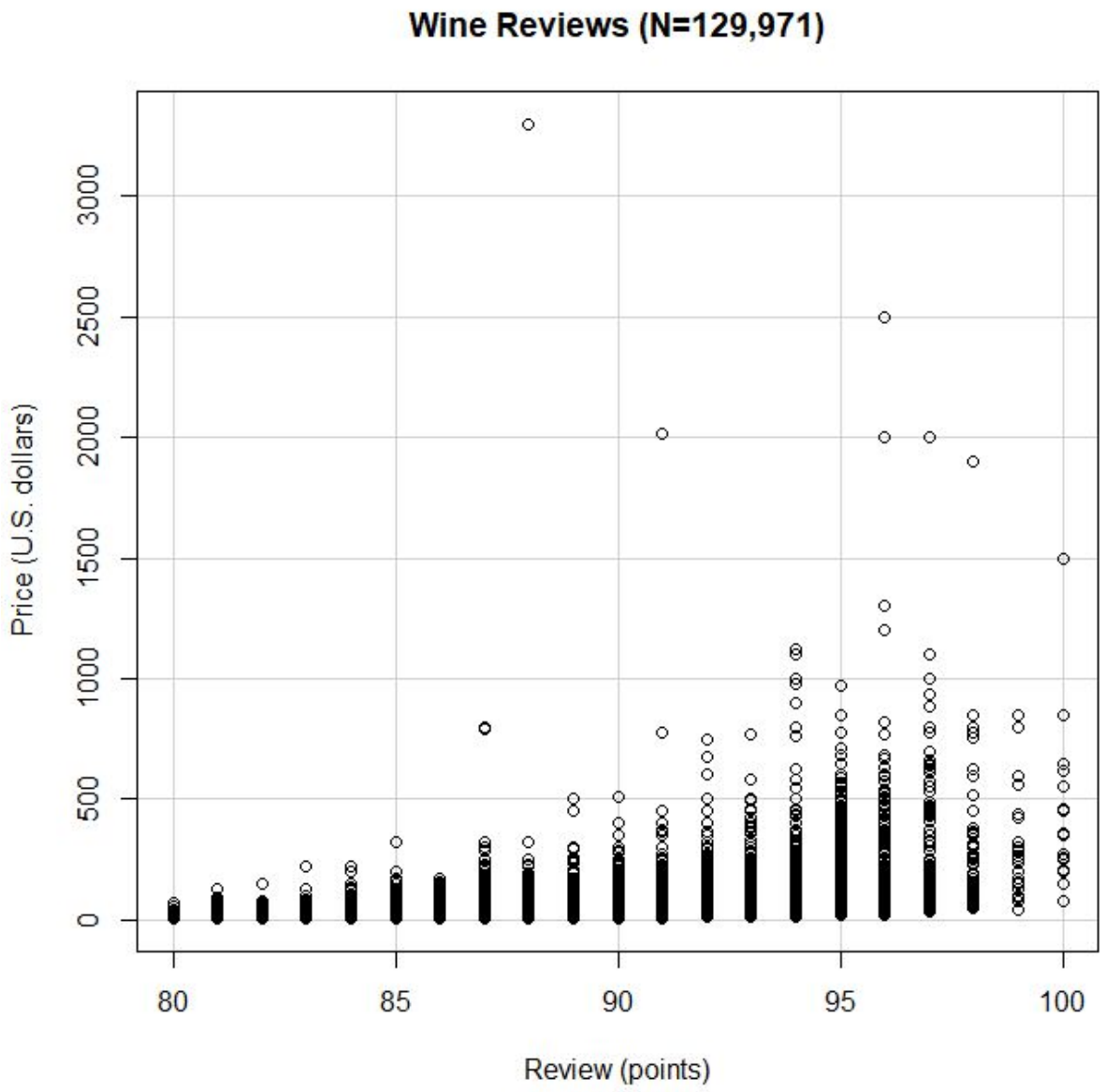

Table 1: Missing Datapoints: This gives us a sense of where data gaps are significant in the primary dataset.

Points	0
Title	0
Variety	1
Description	0
Country	63
Province	63
Region 1	21,247
Region 2	79,460
Winery	0
Designation	37,467
Price	8,996
Taster Name	26,244
Taster Twitter Handle	31,213