Introduction and Data Provenance

YZ Analytics

Introduction

The first traces of wine were found in Georgia in 6000BCE (Watson, 2010). In **Figure 1** we show a map of more than 10 thousand wineries operational today.

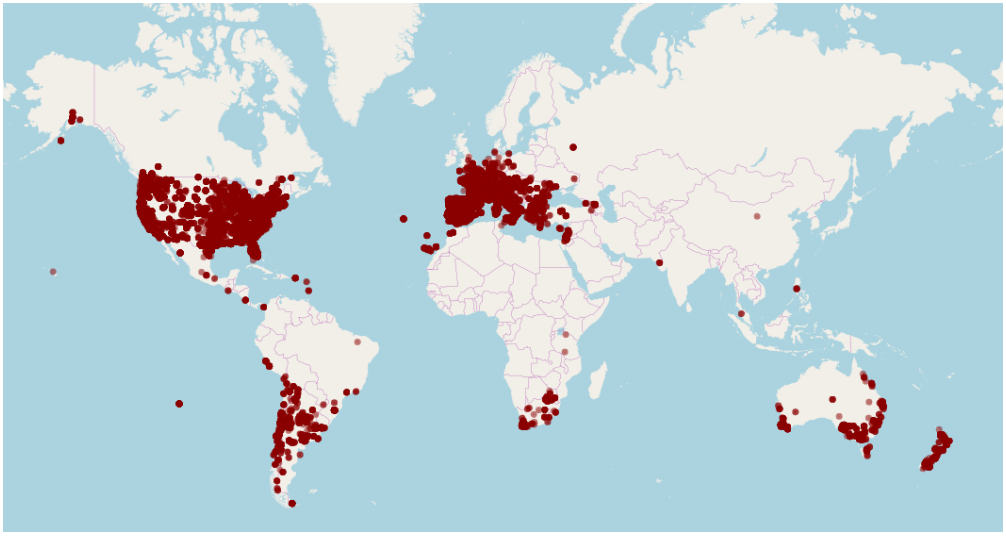

Figure 1: Map of more than 10000 wineries present in our dataset.

For more than 8000 years, wine has been a relevant part of civilization, and yet, many people still do not think twice when they say "I'm not a wine person". Even then, there are many people that are also able to comfortably say: "This juicy, fruit-forward wine delectates the palate."

This report and the accompanying web application aim for two things: 1) to showcase an advanced usage of visualization techniques in the context of a product for wine exploration, and 2) to expand and develop statistical techniques for a seamless incorporation into our product. Therefore, our main goal is to provide an immersive experience which leverages technology and builds upon the statistical/technical knowledge of a bachelor-level student, all within the context of wine.

Data Provenance

In this project we will be working with Kaggle's wine review data, found here: https://www.kaggle.com/zynicide/wine-reviews (Thoutt, 2017). There are close to 130,000 wine reviews. A small glimpse of what the data looks like is available below for convenience:

```
## Observations: 129,971
## Variables: 15
## $ X1                    <dbl> 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12...
## $ country               <chr> "Italy", "Portugal", "US", "US", "US", "...
## $ description           <chr> "Aromas include tropical fruit, broom, b...
## $ designation           <chr> "Vulkà Bianco", "Avidagos", NA, "Reserve...
## $ points                <dbl> 87, 87, 87, 87, 87, 87, 87, 87, 87, 87, ...
## $ price                 <dbl> NA, 15, 14, 13, 65, 15, 16, 24, 12, 27, ...
```

```
## $ province              <chr> "Sicily & Sardinia", "Douro", "Oregon", ...
## $ region_1              <chr> "Etna", NA, "Willamette Valley", "Lake M...
## $ region_2              <chr> NA, NA, "Willamette Valley", NA, "Willam...
## $ taster_name           <chr> "Kerin O'Keefe", "Roger Voss", "Paul Gre...
## $ taster_twitter_handle <chr> "@kerinokeefe", "@vossroger", "@paulgwin...
## $ title                 <chr> "Nicosia 2013 Vulkà Bianco  (Etna)", "Qu...
## $ variety               <chr> "White Blend", "Portuguese Red", "Pinot ...
## $ winery                <chr> "Nicosia", "Quinta dos Avidagos", "Rains...
## $ year                  <dbl> 2013, 2011, 2013, NA, 2012, 2011, 2013, ...
```

Next we proceed with a discussion of the variables that are available in our dataset.

Variables:

Our dataset contains almost 130K observations, one for each wine review, and 18 variables.

Below, a description of the seventeen variables is included:

- id - The unique obervation identifier for reviews in our dataset
- country - The country that the wine is from
- description - A few sentences from a sommelier describing the wine's taste, smell, look, feel, etc
- designation - The vineyard within the winery where the grapes that made the wine are from
- 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)
- price - The cost for a bottle of the wine
- 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 (i.e. Rutherford inside the Napa Valley), but this value can sometimes be blank
- variety - The type of grapes used to make the wine (ie Pinot Noir)
- winery - The winery that made the wine
- title - The title of the wine review, which often contains the vintage if you're interested in extracting that feature
- year - pulled from the vintage in the title
- taster_name - name of the taster
- taster_twitter_handle - Twitter handle of the taster
- address - A combination of the winery and the country
- latitude - geocoded latitude
- longitude - geocoded longitude

An Assessment of Data Quality

Frequently undergraduate analyses are done with carefully curated data or fail to consider the quality of the datasets used until the conclusion. We decided to assess the quality of our data through research of wine production as this was decisive in creating our product.

In **Figure 2** we show a map of the most relevant wine producing regions in the world. It is clearly evident that our dataset does not represent the global production of wine because the presence of observations originating from either Russia and China is lacking.

Apart from the representativeness of our dataset, we consider

References

Nabi, Javaid. (2018). "Machine Learning - Text Processing." Towards Data Science. https://towardsdatascience.com/machine-learning-text-processing-1d5a2d638958 Thoutt, Zack. (2017). "Wine Reviews." Kaggle. https://www.kaggle.com/zynicide/wine-reviews 2018. "Predicting Wine Ratings Using Light-

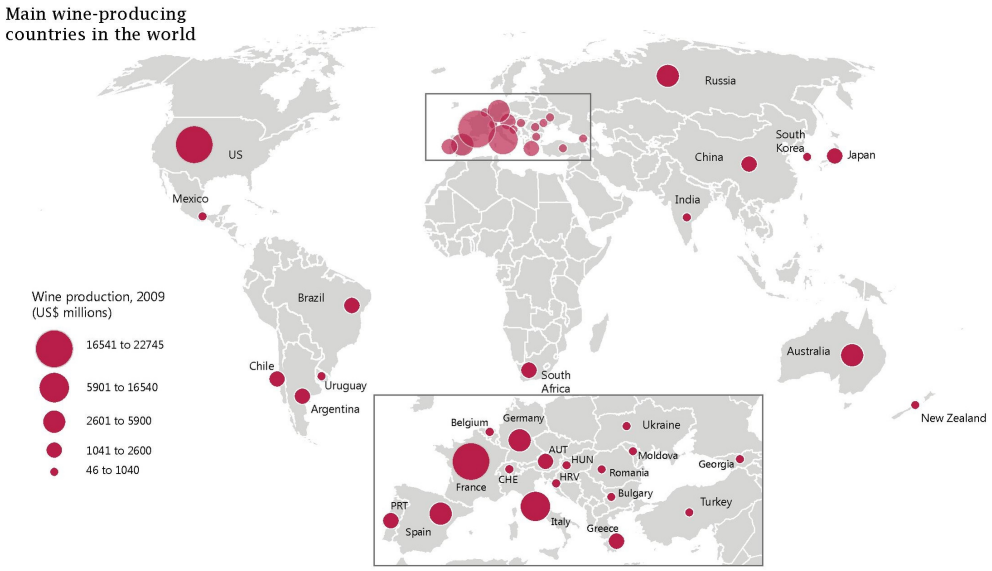

Figure 2: Map from wikipedia with most relevant wine regions. Note the difference between the figure generated from our dataset (Figure 1) and this map.

GBM + Text2Vec." Kaggle. https://www.kaggle.com/nnnnick/predicting-wine-ratings-using-lightgbm-text2vec 2019. WineEnthusiast. https://www.winemag.com/?s=&drink_type=wine&page=0 Watson, Ivan. (2010) "Unearthing Georgia's Wine Heritage." CNN. http://edition.cnn.com/2010/WORLD/europe/04/20/georgia.wine.heritage/