**ABSTRACT:**

Wine is a huge global market. With so many bottles changing hands each year, it is important to stay educated if you are looking to purchase. This paper begins by giving a brief statistical overview of wine ratings, and then performs comparison tests between ratings given by professionals and amateurs. The paper then looks at the most commonly rated wine regions to determine which regions produce on average better wines.

**INTRODUCTION:**

Each year, billions of bottles of wine are produced and consumed1. In Canada alone we consume over 600,000,000 bottles of wine each year, that’s right, six-hundred million bottles1. With so many bottles changing hands, it’s important for the consumer to be able to make an educated decision about what wines to buy. Fortunately, there are many easily available sources for wine ratings. In fact, there are professional wine reviewers, called sommeliers, that rate wines\*. Amateurs also enjoy reviewing bottles of wine; there is a website called Vivino2 where anyone can give their own opinions and ratings for a bottle of wine. This paper focusses on bringing to light some statistics about the world of wine and then looks at a comparison between the ratings of sommeliers and amateurs. We finish up by looking at the most popular regions for wine and whether there are some regions that are better than others for purchasing highly rated wines.

\* If you’re interested in seeing what it takes to become a sommelier, I strongly recommend watching the SOMM documentaries.

**METHODS:**

Data from Vivino was scraped [see work done by Kevin/Raman], to produce data for nearly eight thousand wines (n = 7731). The data was composed of red and white wines from 26 countries, with ratings between 3.0 to 5.0 (ratings out of 5.0). When the csv file was read in, all numeric data types were defined as float due to a lack of NA support by numpy3. The data had the following schema:

[title : string, winery : string, region : string, ratings : float, number of ratings : float].

The title contained data for the vintage in addition to the name of the wine, and the region contained the country and region. These parameters were extract such that the schema became

[wine : string, vintage : float, winery : string, country : string, region : string, ratings : float, number of ratings : float].

Data rated by sommeliers was also obtained as a dataset from Kaggle. This dataset was composed of red and white wines from the United States with ratings above 80/100. The original schema of the data was:

[country : string, description : string, designation : string, points : float, price : float, province : string, region\_1 : string, region\_2 : string, taster\_name : string, taster\_twitter\_handle : string, title : string, variety : string, winery : string]

After some data processing, the schema was:

[country : string, ratings : float, price : float, region : string, wine : string, variety : string, winery : string, vintage : float]

The points were converted to ratings out of 5 to match with the Vivino Data. For both datasets, any rows with missing or null values were dropped.

Figures were created as either scatter, histogram, or boxplot using matplotlib.

Summary statistics containging the mean, standard deviation, min, and max for the ratings of each dataset was obtained using the scipy stats module. The covariance between ratings and the number of ratings of the vivino set was evaluated using the linregress function of the scipy stats module. The covariance between price and ratings for the professional data set was obtained the same way.

To compare the two datasets, their ratings were scaled to a range of 0-1. A normality test was performed on the ratings of the datasets using the normaltest function for the scipy stats module and the variance was calculated from the levene function.

A Student’s T test was performed on the scaled data using the ttest\_ind function from the scipy stats module, followed by a Mann Whitney U test using the mannwhitneyu function.

For the regional comparisons, regions from the Vivino data set with at least 40 ratings total were selected, and the means of each region were calculated. The data was then split by region so that an ANOVA test could be performed. A pairwise Tukey HSD posthoc analysis was performed.

**RESULTS:**

To get a good idea of what the data looked like, some preliminary graphs were generated for both the Vivino and sommelier data set. Supplementary Figures 1-3 provide examples of some of these graphs (more graphs can be found in the figures folder provided).

Box plots and histograms (Figures 1 & 2) show the distribution of the ratings overall from both datasets.

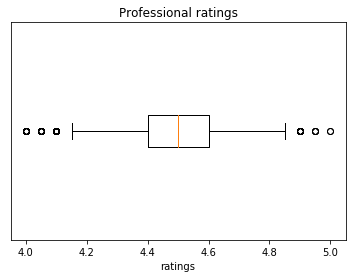
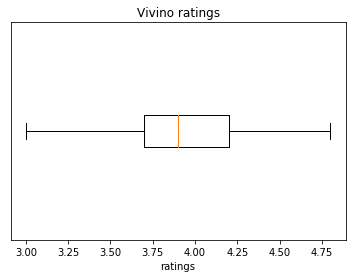


Figure A: Boxplot of Vivino Ratings Figure 1B: Boxplot of Professional Ratings

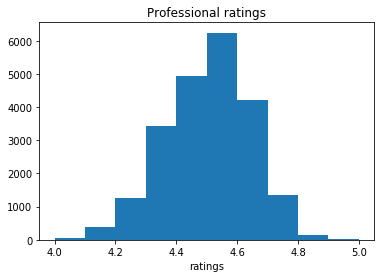
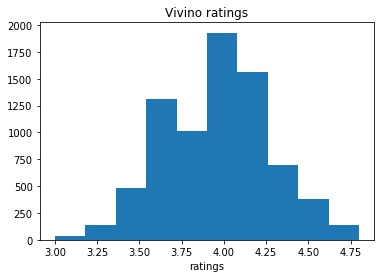


Figure A: Histogram of Vivino Ratings Figure 2B: Histogram of Professional Ratings

The preliminary stats for the Vivino and sommelier data are summarized in Table 1.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Set | Count | Mean | Standard Deviation | Minimum Value | Maximum Value |
| Vivino | 7693 | 3.952138 | 0.309134 | 3.0 | 4.8 |
| Professional | 22037 | 4.478393 | 0.139937 | 4.0 | 5.0 |

*Table 1: Summary Statistics of Vivino and Professional Ratings*

From the Vivino dataset, I thought it may be that people have a tendency to rate wines higher if it is a more popular bottle of wine. To determine if this was the case, a covariance test between number of ratings and ratings was performed. The result obtained was a correlation coefficient equal to 0.084. Since the value is close to zero, this indicates that there is no linear relationship between the number of ratings and the ratings.

I also suspected that the ratings the professionals give may increase linearly with price. The correlation coefficient for the covariance test between price and ratings was 0.39. There was therefore somewhat of a weak linear relationship between price and rating but less so than expected. Whether this relationship is due to sommeliers rating wines higher based on their price tag or whether higher quality wines are more expensive is an area for future research.

The ratings from both data sets were scaled to the range 0-1 so that their means could be compared. A normality test was performed on each data set, and both failed (Table 2). However, since we have a large number of samples and the histograms of the data looks roughly normal (Figure 2), it was determined that the data was normal enough.

|  |  |  |
| --- | --- | --- |
| Data Set | Mean | Normal Test p-value |
| Vivino | 0.53 | 9.9\*10-9 |
| Professional | 0.47 | 3.4\*10-50 |

*Table 2: Means and Normal Test results for Vivino and Professional Data Sets*

A levene test was then performed on the two data sets to determine if they had equal variance. The result was a p-value of 4.3\*1076, indicating that the two samples do not have equal variance.

Since the two samples did not have equal variance, we must use the version of a T-test that does not assume equal variance. The result of the T-test was a p-value of 4.3\*10-117, meaning that we conclude that the means of the scaled ratings from the Vivino and professional data sets are different.

A Mann WhitneyU test was also performed since it makes no assumptions of the probability distributions. The result was a p-value of 7.4\*10-97, also meaning that we conclude that the samples have different means.

I then looked at the regions of the Vivino data set that had more than 40 ratings total. This returned 40 different regions. These regions and their means are summarized in Table 3.

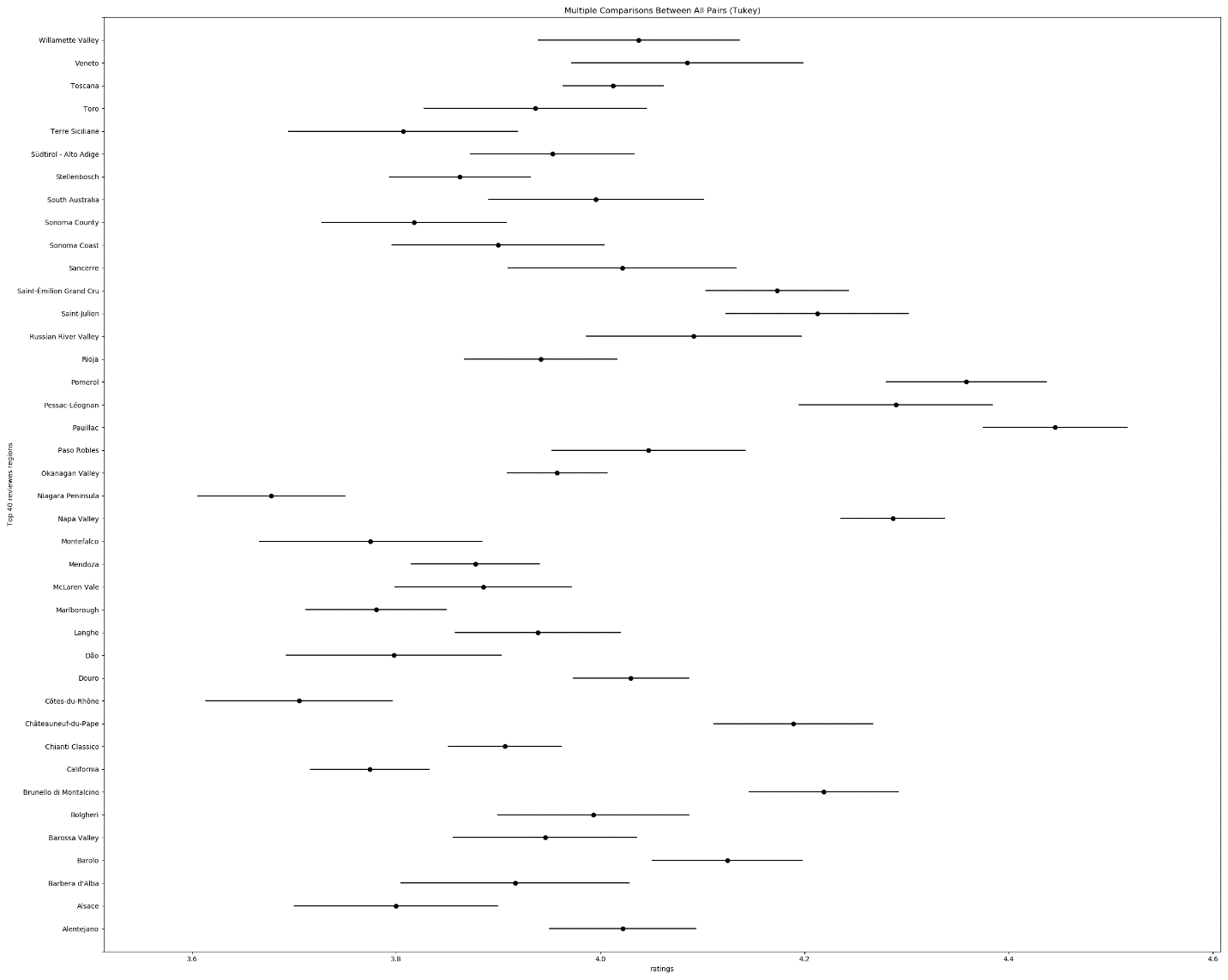
An ANOVA test was then performed on the ratings of each of the 40 most popularly rated regions. The resulting p-value was 4.6\*10-278, indicating that there is a difference between the means of the groups. Since a significant result was obtained, a post hoc analysis was performed using Tukey’s HSD test. The result of the post hoc analysis is summarized in Figure 3.

**DISCUSSION:**

This paper brought to light some insight surrounding the distributions of wines rated by sommeliers and those rated by amateurs on Vivino. It was found that the average rating given by the amateurs was higher than that of the professionals.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Region** | **Rating** |  | **Region** | **Rating** |
| Alentejano | 4.0 |  | Okanagan Valley | 4.0 |
| Alsace | 3.8 |  | Paso Robles | 4.0 |
| Barbera d'Alba | 3.9 |  | Pauillac | 4.4 |
| Barolo | 4.1 |  | Pessac-Léognan | 4.3 |
| Barossa Valley | 3.9 |  | Pomerol | 4.4 |
| Bolgheri | 4.0 |  | Rioja | 3.9 |
| Brunello di Montalcino | 4.2 |  | Russian River Valley | 4.1 |
| California | 3.8 |  | Saint-Julien | 4.2 |
| Chianti Classico | 3.9 |  | Saint-Émilion Grand Cru | 4.2 |
| Châteauneuf-du-Pape | 4.2 |  | Sancerre | 4.0 |
| Côtes-du-Rhône | 3.7 |  | Sonoma Coast | 3.9 |
| Douro | 4.0 |  | Sonoma County | 3.8 |
| Dão | 3.8 |  | South Australia | 4.0 |
| Langhe | 3.9 |  | Stellenbosch | 3.9 |
| Marlborough | 3.8 |  | Südtirol - Alto Adige | 4.0 |
| McLaren Vale | 3.9 |  | Terre Siciliane | 3.8 |
| Mendoza | 3.9 |  | Toro | 3.9 |
| Montefalco | 3.8 |  | Toscana | 4.0 |
| Napa Valley | 4.3 |  | Veneto | 4.1 |
| Niagara Peninsula | 3.7 |  | Willamette Valley | 4.0 |

*Table 3: Means of 40 Most Popular Regions*



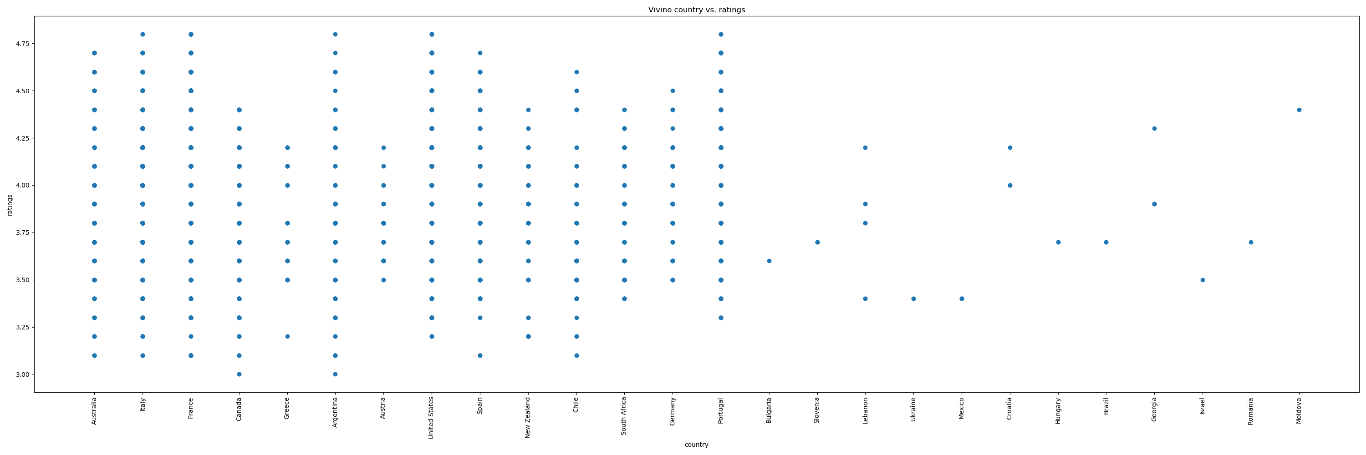
*Figure 3: Tukey’s HSD Test Post Hoc Analysis of Ratings for each Region*

Additionally, it was found that if you are looking for a bottle of wine, the best region to buy is from Pauillac. This region had a statistically higher average rating amongst Vivino users than nearly all other popularly rated regions. This makes sense since the Pauillac region is in Bordeaux, which is commonly believed to contain some of the best wines in the world. The region that was found to be worse than the other regions is the Niagra peninsula. This personally does not surprise me since Canada is only getting started with wine production and that region gets very cold in the winter, making it not ideal for many grape varietals. Some of the ratings that surprised me was California and Cotes-du-Rhones. Both California and Cotes-du-Rhones are big wine producers, and are generally thought to be quite good. Also of note, the Okanagan Valley performed fairly well with an average of 4.0. This is in the same league as regions such as Sonoma and Rioja, both prestigious regions.

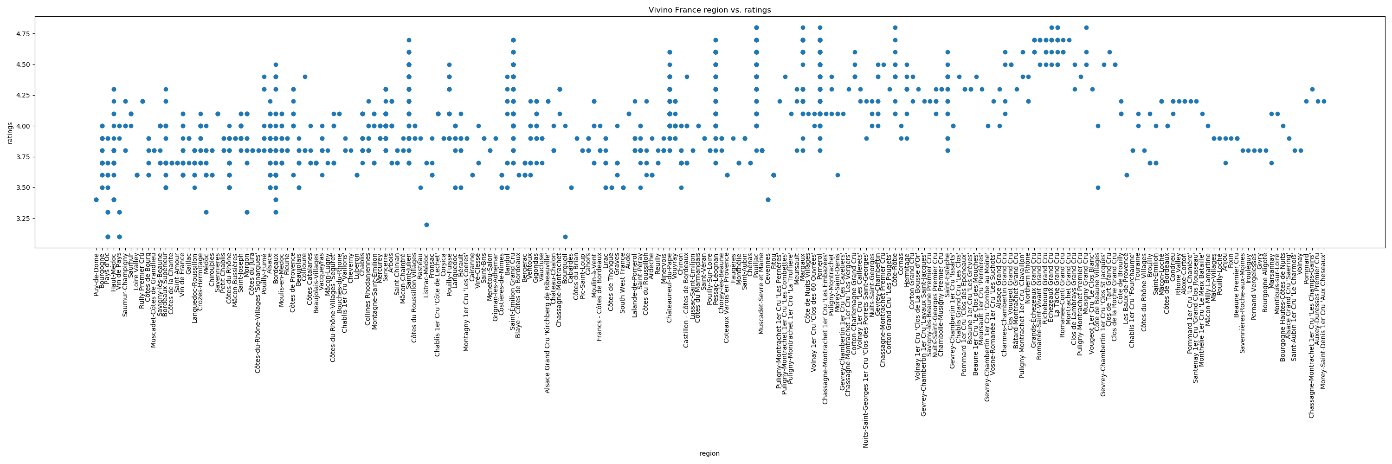
**REFERENCES:**

1. <https://www.alcohol.org/guides/beer-wine-production-consumption-worldwide/>
2. <https://www.vivino.com/>
3. <https://pandas.pydata.org/pandas-docs/stable/user_guide/gotchas.html#support-for-integer-na>

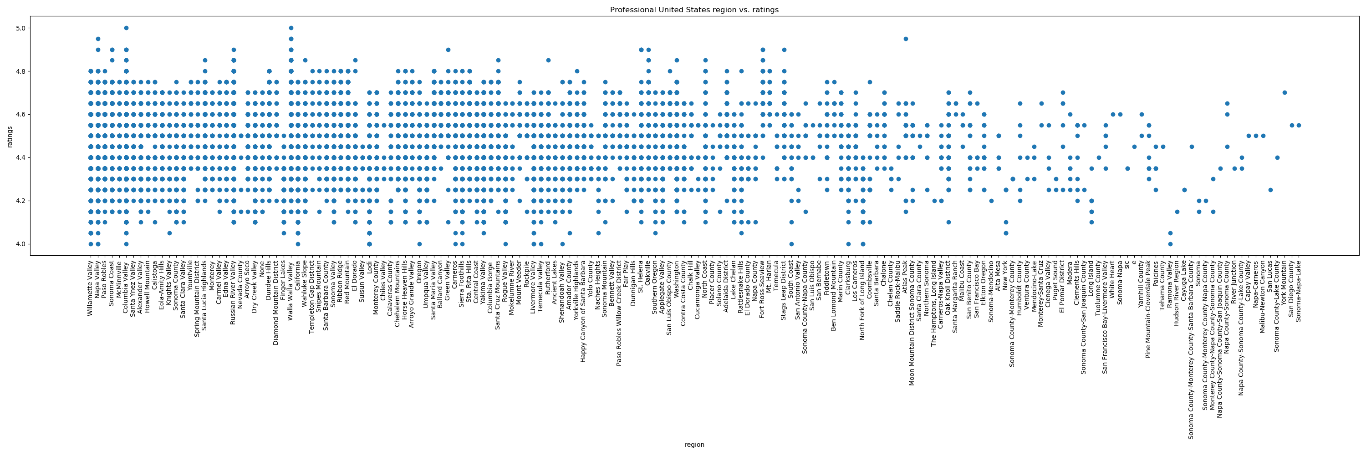
**SUPPLEMENTAL DATA:**



Supplementary Figure : Vivino data showing the distribution of ratings for each country



Supplementary Figure 2: Vivino data showing the distribution of ratings for each region in France



Supplementary Figure 3: Sommelier data showing the distribution of ratings for each region in the United States