Wino Dino Project

Executive Summary

Through the use of a 2017 JSON database hosted on Kaggle, we sought to analyze a large subset of wines along the dimensions of price, country, rating, and variety. These analyses would then serve as visualizations on our dynamic webpage that allows users to filter the dataset along several pertinent criteria, including: country, province, winery, variety, rating, and price.

Our analyses sought to satisfy the conjectures of our hypotheses along three dimensions, namely:

1. Is there a positive trend between average rating and price group?
2. What is the rank order of countries based on average price per bottle of wine?
3. What is the rank order of countries based on the number of discrete varieties of wine they produce?

With the use of Jupyter Notebook, Pandas, and Matplotlib, we found that there is indeed a positive trend between average rating and price group where higher priced wines generally confer a higher average rating. At the same time, our hypotheses that France and Italy would lead the world in average cost per bottle of wine and the number of unique varieties was refuted. In fact, the United States produced, by far, the most variety at 231 types and Italy, Australia, and the United States commanded the highest average prices per bottle of wine from their respective countries (calculated in constant currency, USD).

We utilized JavaScript, HTML, and CSS to produce a webpage that allows for the dynamic filterability of our wine dataset so that the stakeholder (be they customers, collectors, or retailers/distributers) can navigate the large assortment along the criteria listed in the first paragraph.

In the future, we plan on retiring our stagnant dataset and connecting to the Global Wine Score API in order to remain up-to-date for our stakeholders. We also plan to add an additional column, the “Price/Rating Ratio.” We believe this column will help our stakeholders more efficiently sort and discover wines based on their relative value and believe the column will be exceptionally well-suited when employed after filtering along several criteria.

Project Purpose

The purpose of our project is to provide a simple and intuitive interface to navigate through a database of wines with the use of several filters. In so doing, this allows the user (customer, collector, or retailer/distributer) to tailor their search to only the most relevant fermented grapes.

Data Munging

The database was acquired from Kaggle as a JSON file with approximately 130,000 rows of discrete wine data. Thereafter, the file was uploaded to Firebase and pulled via Pandas into Jupyter Notebook where data manipulation began to take place.

Once the data was uploaded into a Pandas data frame, all rows with empty or “NaN” cells in any of the 10 columns were removed so our visualizations would be composed of only wines where we had the greatest amount of information valuable to our intended users. In all, this process removed approximately half of the rows present in the raw data, leaving our working database with 70,175 unique wines with full information.

Hypothesis and Expectations

Due to our group’s relatively sheltered exposure to the universe of wines, we initially hypothesized that the total number of varieties of wines would not exceed 30. This was a hypothesis that emerged organically and was not pursued with a visualization. Areas of research we explore via visualizations include the average rating (in points) based on price ranging from less than $10/bottle to greater than $40/bottle in constant currency USD, the average price per bottle of wine for seven countries, and the number of varieties of wines for each of those countries. Our hypotheses for each are below:

1. Average Rating to Price Group

Our expectation was that the trend would persist relatively linearly with a positive correlation where increases in price yielded wines with a higher average rating.

1. Average Price per Country

We presumed that bottles of wine from France would carry the greatest price tag per average bottle, followed by Italian and Argentinian wines. On the low end, we hypothesized that wines hailing from the United States would command the lowest price.

1. Number of Varieties of Wine per Country

Somewhat paralleling our expectations with average price per country, we initially suspected that France and Italy would possess the greatest number of unique wine varieties while the United States varieties would comprise the smallest amount of the 641 types. We assumed this as our experience in wine aisles at grocery stores suggested that there are probably around a dozen.

Technical Information

As stated prior, our data was manipulated via Pandas within Jupyter Notebook after retrieving the raw data from our Firebase repository. The complete list of libraries used in our analysis of the database is: Pandas, Matplotlib, NumPy, SQLAlchemy, Flask, and Flask\_SQLAlchemy. Technical information on the following webpage and each of the visualizations is organized below:

1. Visualization of Hypothesis 1 – Average Rating to Price Group

The price column of the cleaned wine data frame was subject to binning where a secondary “price\_groups” column was created with the categories “<10,” “10-14,” “15-19,” “20-24,” “25-29,” “30-34,” “35-39,” “40+.” A new variable was then created that grouped the initial cleaned wine data frame by “price\_groups” with an adjacent column that calculated the average points per group. These data points were then plotted graphically as a line with Matplotlib with the binned price groups along the x-axis and the average rating in perpendicularity.



1. Visualization of Hypothesis 2 – Average Price per Country

The cleaned wine dataset was again subject to binning, this time along the “country” column. With seven countries remaining of the original 44 after data munging, a concise horizontal bar chart developed from grouping the unique countries by the average price of bottles of wine in each of their respective areas using Matplotlib.



1. Visualization of Hypothesis 3 – Number of Varieties of Wine per Country

Similar to the average price per country, the number of varieties of wine per country was ascertained by binning the seven unique countries in our cleaned wine data frame and then pairing them with their corresponding magnitude of unique wine varieties. This distilled dataset was then plotted with a horizontal bar graph using Matplotlib.



1. Project Webpage

The wine data is hosted on Heroku and is built with JavaScript, CSS, and HTML. D3 is utilized within the JavaScript file to present our cleaned wine dataset in a dynamically filterable nature where users are provided the ability to orient the aggregate data along country, province, winery, price, variety, and rating. The resulting visuals are a pie chart and a filtered table.

Javascript “ForEach” formulae are activated in series to generate new data frames from our filters once the clickable event listener is triggered. These data frames are then visually represented within the HTML as the “tbody.”

Conclusions

The accuracy of our hypotheses was somewhat mixed. Although the trend of average rating to price range was positive and largely linear, our presumptions about the rank order of average price by country and variety count by country were erroneous.

In the case of average price of a bottle of wine by country, Italy was the most expensive country which somewhat coincided with our initial belief, but French wines ranked fourth in average cost and Argentinian wines the least pricey at seventh. In addition, our assumption that the United States would have the most affordable bottle of wine was refuted by the data which indicates that the United States has the third most expensive bottles of wine in our dataset, beating out the average French bottle.

Along a similar vein, we once again underestimated the United States’ prolific wine industry. Our initial assumption supposed that the United States would generate the least variety of unique wines out of all the countries in question, but it was in fact the opposite. The United States effectively doubled the number of varieties that France produces and generates 56 more varieties than Italy. However, Italy and France are ranked number two and number three, respectively.

Future Work

We intend to build upon the analysis we have conducted on our cleaned wine dataset. In order of importance, we plan on connecting to the Global Wine Score API so we are not working with stale data and can remain up-to-date on the newest issues from wineries around the globe. Thereafter, we will add an additional column to the filterable table within our webpage that calculates the “Price/Rating Ratio,” an indication of relative value among groups of wine. The table can then be filtered along any of the criteria that are currently present on our webpage, such that a user can compare wines of a certain price range and variety based on the relative value they can expect to receive. We believe this will inform all stakeholders of our application, from customers, to collectors, to retailers and distributors.