Will Hughes, Brandon Ibarra, Breanna Wallace, Shaunghnessy Robertson

Data Science Bootcamp

Group 1 Project 2

Wines ETL Writeup

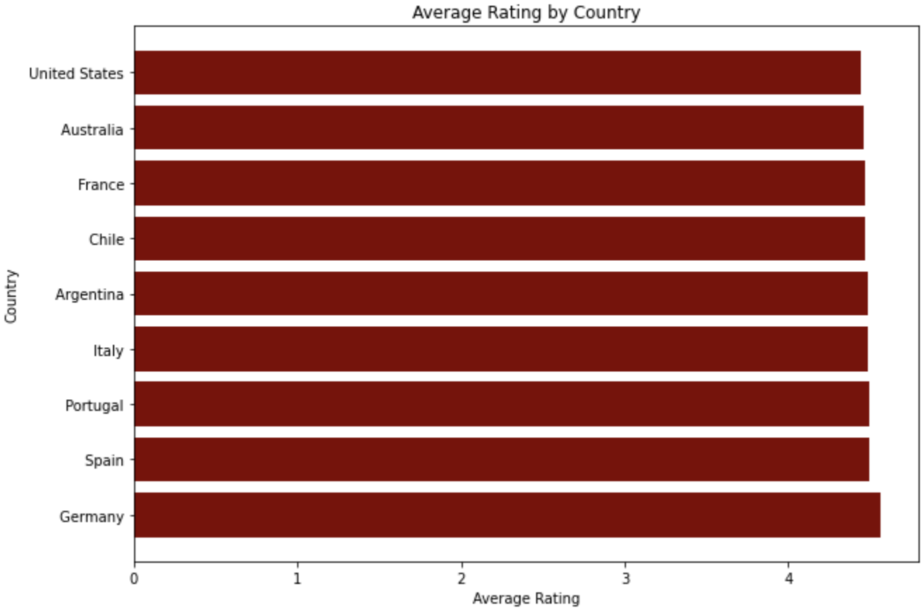
Wines come in many different types from red, white, rose, and sparkling. There are also many different names for wines like merlot which is a red wine from Bordeaux, France, a cabernet sauvignon which is also a red wine from the United States. Knowing the best price and the best rating from a bottle can really help narrow down what you might enjoy or what might be a great gift. Either a wine connoisseur or someone who is curious to try something new you are in the right place.

The first step we took to determine the best choice for someone who is either looking to try something different or return to enjoy their favorite wine was to web-scrape the website we used, Vivino. We extracted the information we needed to perform our analysis. To scrape the site, we used splinter and chrome driver. We ran into a few problems beforehand but were able to gain access to the website and successfully extract the information we needed. We then used beautifulsoup to pull data from the HTML code and turn it into readable text and strip characters. We then did “Title=result.find, Price=result.find” to help us organize our data to be able to put it into a data frame. Once we found the information for one of the wines, we were able to loop it to other wines to get data for those items as well. Once we received this data, we were able to create our data frame.

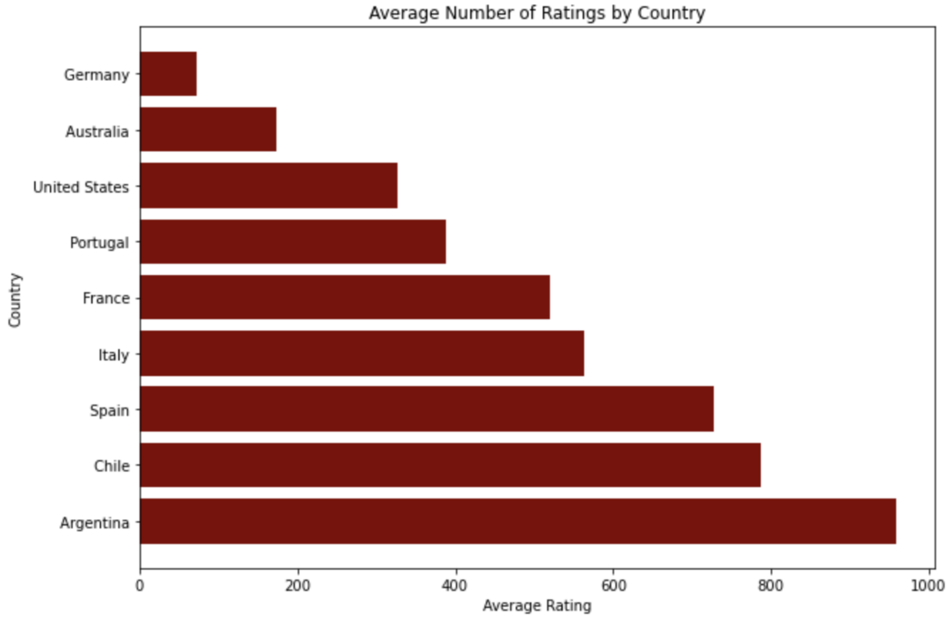
Once we were able to clean and extract the data for it to be put into data frames, we made different columns named brand, vintage, price, rating, region, and country. We then were able to find the location of each wine to put with its country to help us analyze it better. We then separated the year from the vintage, as well as cleaned up our data frames to end up with a rating, price number ratings, year, brand_id, region_id, country_id, and vintage_id.

Now that we’ve cleaned and normalized our data, we are ready to being the load progress in our database. We start by thinking about our database design and what sort of tables are needed for our database. We first created the table to put in our database then read it into pandas using python on jupyter notebook and looped our data to make a new data frame and column.

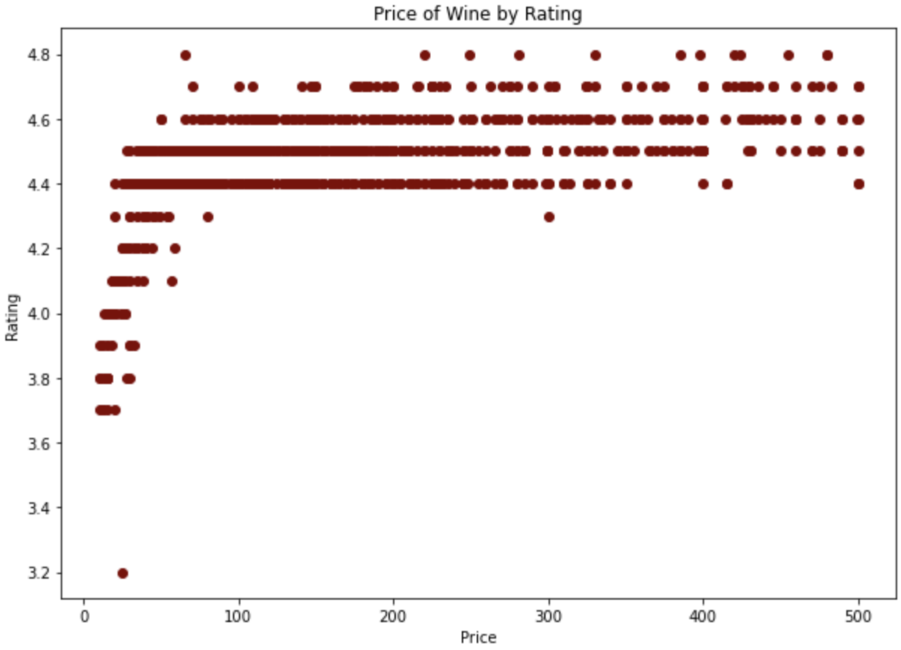
Finally, to visualize our findings, we produced graphs from our data set.



We can see from the bar graph above that, in comparison to the other countries, Germany has the highest average rating. However, as shown in the bar graph below, Germany has the lowest average number of reviews. That might be because Germany produces the fewest wines.



In the below scatterplot, we can observe that you may purchase wine that has received good reviews for $75 to $100. So, there are several options available for wine connoisseurs who seek high-quality wine at an affordable price.



Bonus II

In order to automate this process we would need to make a few key tweaks. First thing would be implementing Selenium's built in Scroll function which can be set to scroll to the bottom of the page, triggering the website to load more entries, sleep for a second or two to allow the page to finish loading, and then continue the process until there is nothing left to load.

The next thing we would add would be task-schedueling. Window's Task Scheduler allows the user to set programs and scripts to run at a selected time. We can also set this to run on a cloud device which would allow the script to run even if our computers are off.

Ideally, we would set the script to run some time in the middle of the night, once a week or so, and then do a left merge to add any new wines to the database.