Kristen Logue, Trevor Kahl, David Primrose, Richard Bump

9/5/20

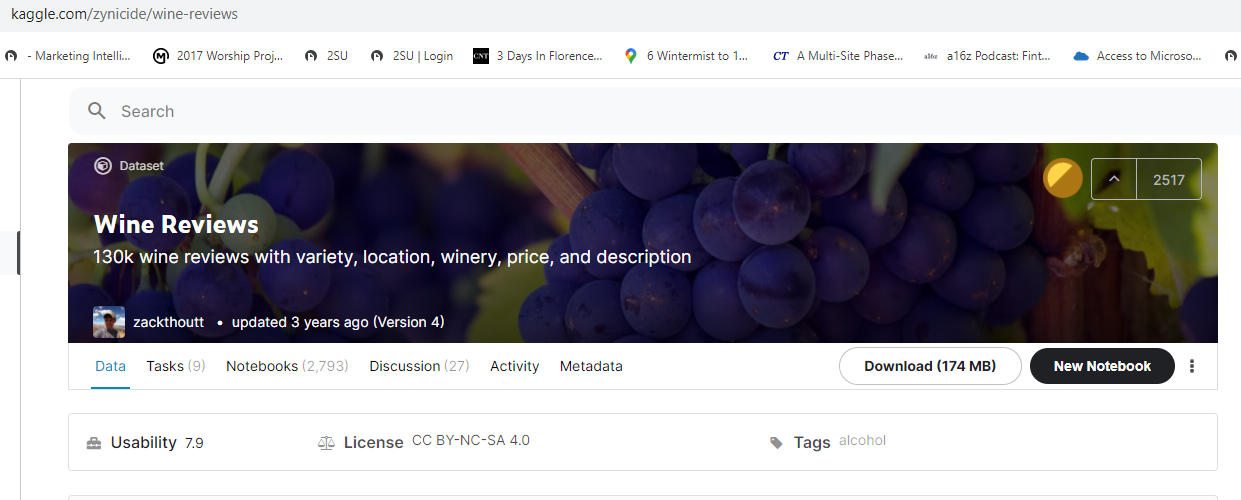
IST 652: Final Project

*# This assignment was completed as a group. David Primrose handled the initial data exploration and analysis of geocode. Richard Bump analyzed the effect of country of origin and GDP on wine. Kristen Logue analyzed the effect of weather on wine. Trevor Kahl analyzed Twitter sentiment on wine.*

# Wine Analysis

## Initial Data Exploration

Our team chose the Wine Reviews Dataset <https://www.kaggle.com/zynicide/wine-reviews>) as source data for our Scripting project. Not only was this interesting given our team affinity for wine but the timing appears to also fortunate--if not simply making the most of a bad situation. On one hand it is Covid-19. Covid 19, anxiety, and lockdown is driving a higher consumption of wine as we are confined at home. According to Lulie Halstead, CEO of Wine Intelligence, wine consumptions has increased from 9.3 times a month in October to 9.7 times a month by lockdown in March 2020. Yet…oversupply of grapes due to wine overproduction (<https://fox8.com/news/experts-say-wine-prices-could-drop-due-to-oversupply-of-grapes>) has led to lower wine prices. The combination has led to some affordable wine drinking opportunities as well as some interesting questions and decisions on wine.



The Kaggle wine review data set we chose includes 164,000 observations and 14 attributes. These include number, country, description (review) of the wine, designation, points, price, province, region1, region2, taster name, taster twitter handle, title, variety, and winery. The original source of this dataset is WineEnthusiast (<https://www.winemag.com/>). The data was scraped during the week of June 15th, 2017.



The data consists of 10 fields:

* *Numbers*: the location of the wine in the dataset
* *Points*: the number of points WineEnthusiast rated the wine on a scale of 1-100
* *Title*: the title of the wine review, which may also contain the vintage
* *Variety*: the type of grapes used to make the wine
* *Description*: a few sentences from a sommelier describing the wine
* *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
* *Region 2*: sometimes there are more specific regions specified within an area
* *Winery*: the winery that made the wine
* *Designation*: the vineyard within the winery where the grapes 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 above

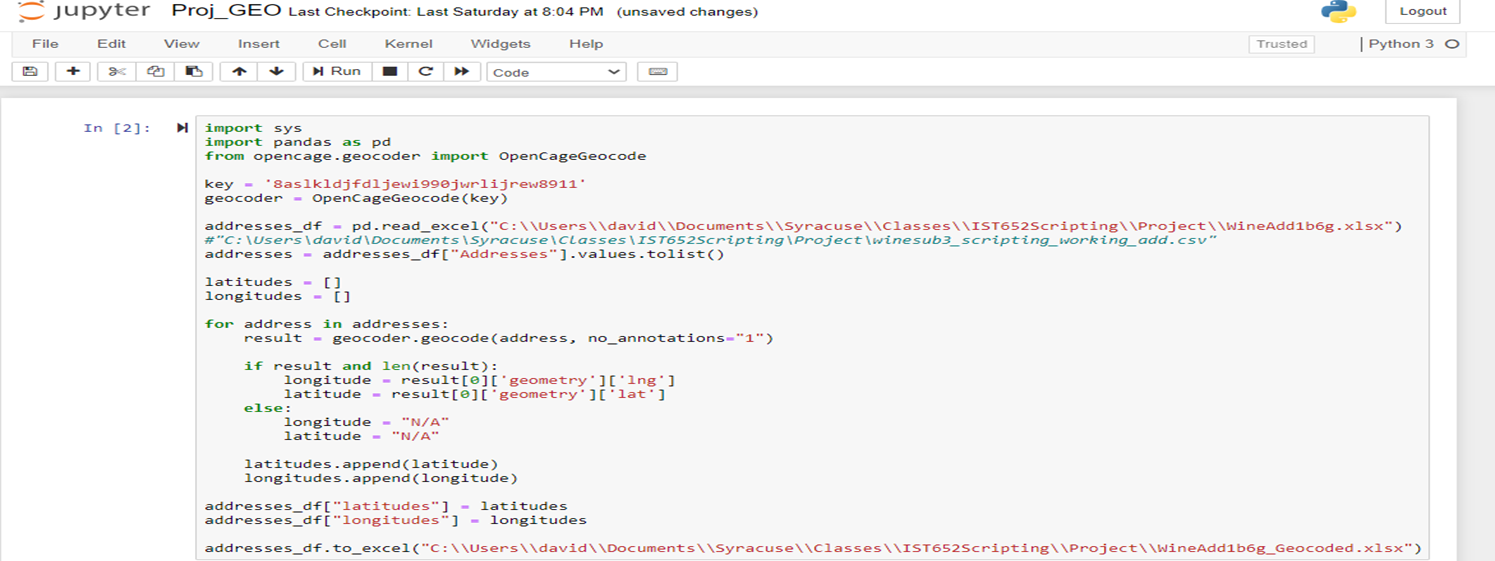
### Preparation

The Kaggle wine review data set we chose includes 164,000 observations and 14 attributes. As part of our cleaning and preparation we reduced the data size to 25,241 observations and 11 attributes. This was due to the volume of data, the limited amount of time, and the limits (code and cost) of geocoding at scale.

We removed “taster\_name”, “taster\_twitter\_handle”, and “title”. Additionally, we removed some observations where data was significantly incomplete, replaced NaN in regions with None, and filled in some gaps in pricing with averages of wines of similar variety, price point, and region. We also removed wines priced over $300. Hey, it’s Covid and we’re college kids! Who has that kind of money!?

### Adjustments

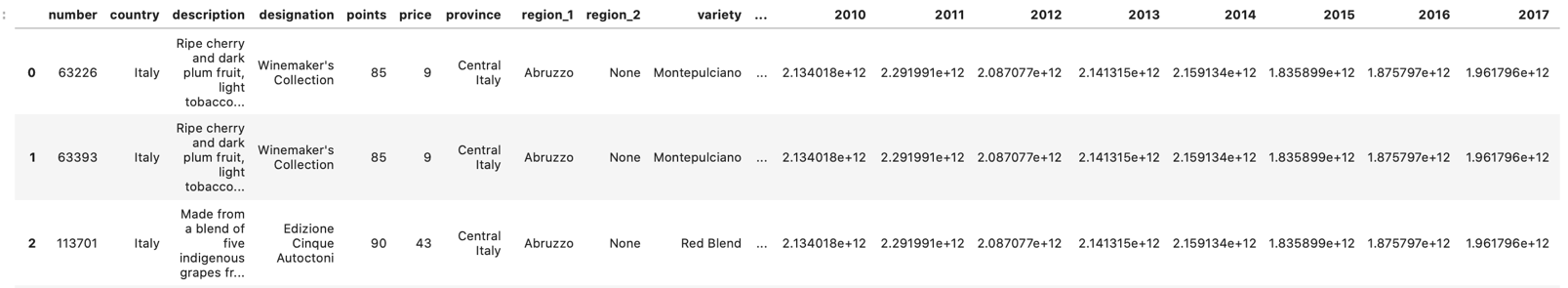
As part of our dataset adjustments we added geocode location, GDP, weather, and Twitter sentiment analysis. Our geocode source is Opencage (<https://opencagedata.com/dashboard>). After a search we modified a geocode example from Medium for our use (<https://medium.com/@lovespreadsheets/how-to-geocode-addresses-in-spreadsheets-using-python-780510615061>). This allowed us to add the geocode directly into the downloaded dataset. From here we were than able to reference it and load it into Python directly: Code for the geolocation is as follows:



## The Effect of the Economy of the Country of Origin on Wine

Many factors can be attributed to the price and quality of wine; some obvious and some not. We decided to look to see if the local economy might play a role as a defining factor. We found some data from the World Bank (https://databank.worldbank.org/reports.aspx?source=2&series=[API\_NY.GDP.MKTP.CD\_DS2\_en\_csv\_v2\_1217511.csv](https://databank.worldbank.org/API_NY.GDP.MKTP.CD_DS2_en_csv_v2_1217511.csv)) that contained the gross domestic product for 263 countries starting in 1960 and running annually through 2019 which we will join to the source wine review data and look for correlations.

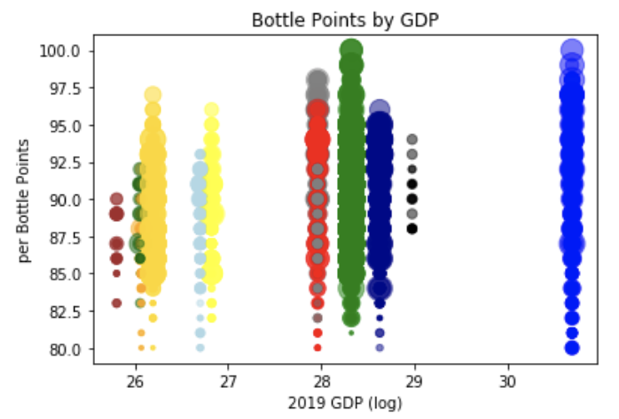
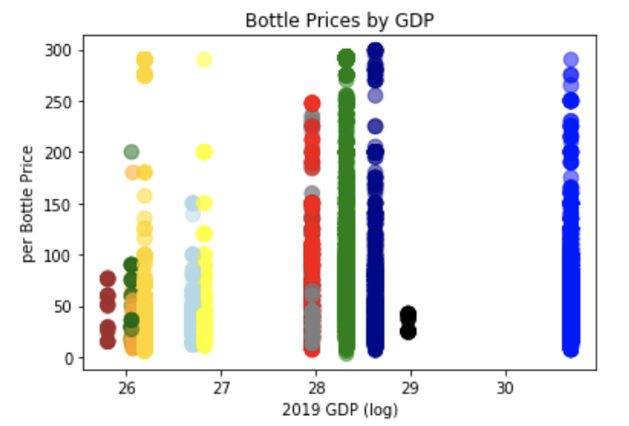
The dataset was already quite clean so not a lot of cleansing was necessary on the outset other than stripping off some rows from the top of the file and a couple of unnecessary rows with superfluous information. After the basics, there was some data transformation necessary to allow the two data frames to be joined by the country name. The join was a full left join with the base wine review data frame as to preserve all of the data in that data frame. The resulting data frame had all of the wines that we were reviewing from the twelve countries along with the GDP data dating back to 1960.



The next transformation was to calculate the compounded annual growth rate since 1991 which was added a column to the data frame. The final data transformation was to add in a color for the country of origin which was used for the purpose of graphing and charting which was also added as column to the data frame.

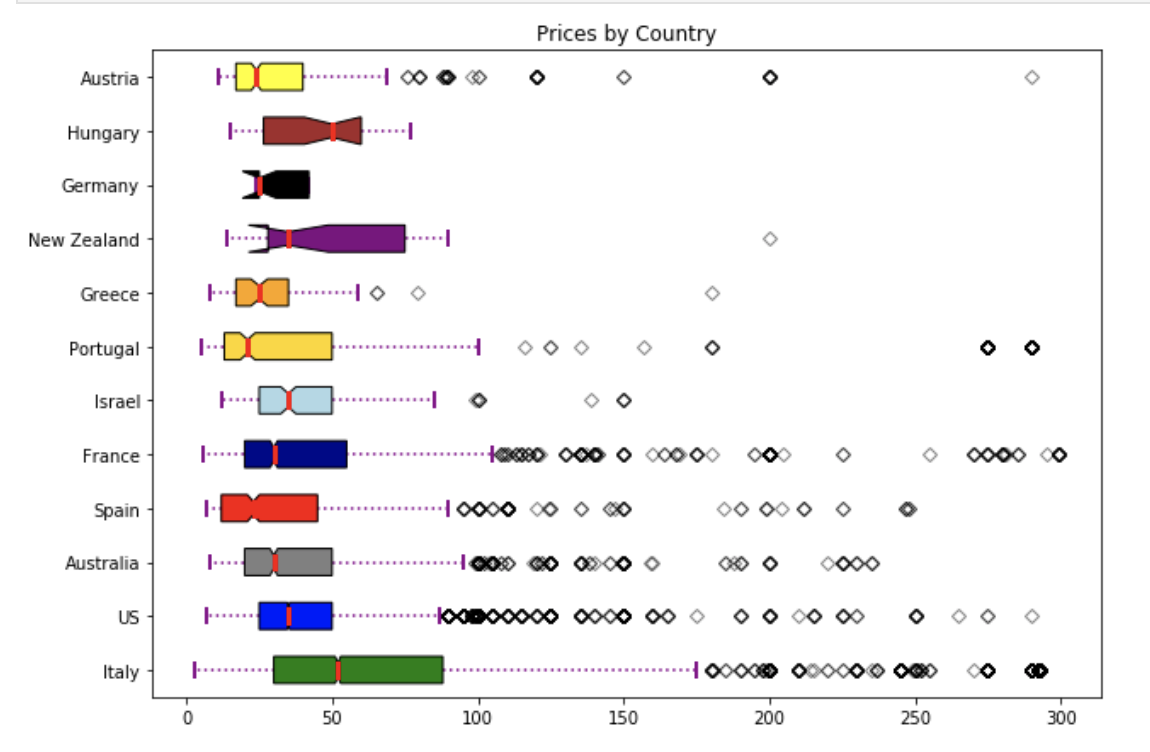
In looking at both wine prices and wine ratings relative to the GDP there does not really seem to be any perceptible correlations. There are some observations that are worth looking into though:

* Hungary, New Zealand and Portugal have the lowest GDP of the data set and it also on the lower scale for both prices and ratings.
* Outside of these three, the prices and ratings vary considerably
* This may be explainable by other factors such as sun, temperature or rain levels.

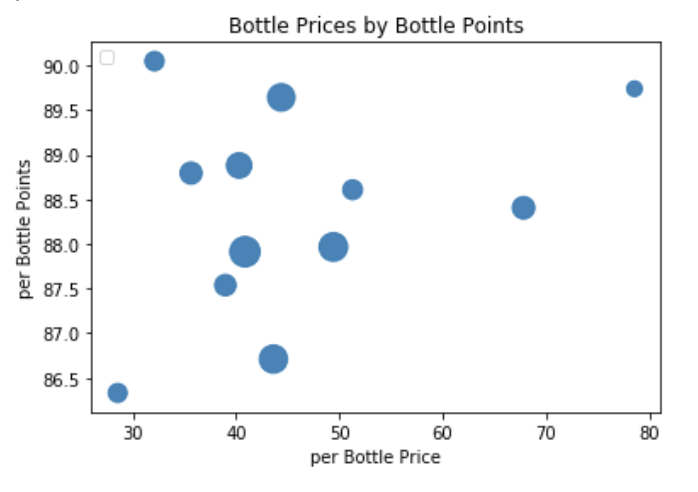


Next, we took the mean of prices, ratings, CAGR and GDP by country to view the data at a higher level. Again, the observations did not show anything too interesting, but we did notice a couple of things:

* The 5 countries (Italy, US, Australia, Spain and France), regardless of GDP, that have the most varieties of wine also have a high number of price outliers indicating that there are a lot of wineries charging very high prices for some of their bottles.
* This is not surprising as they have significantly more wineries and varieties than the other countries.



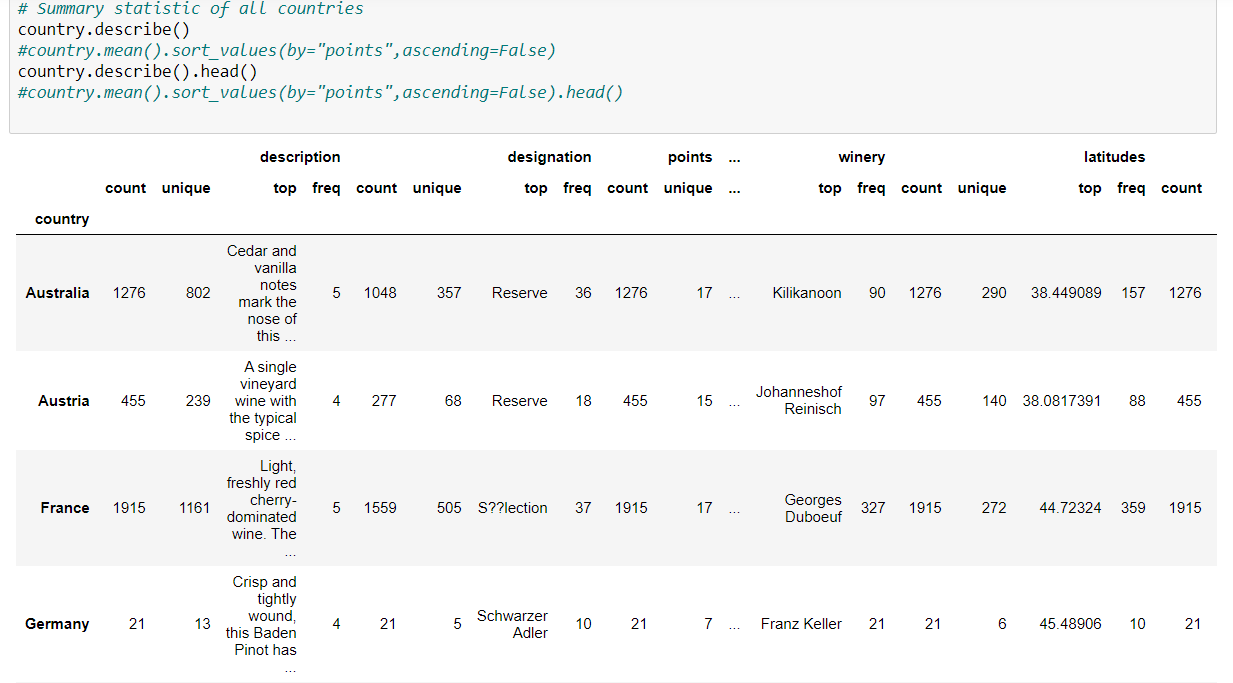
The final view we took at the data was to juxtapose the average prices to the average ratings and the compounded annual growth rate (CAGR) for that country. Italy and Portugal have on average the highest priced wine but not the highest rated wine nor the highest CAGR. The countries with the highest CAGR (Israel, New Zealand, Hungary and Australia) as illustrated by the size of the bubble are in the middle of the group for price and range from lower to the higher end of the ratings scale.



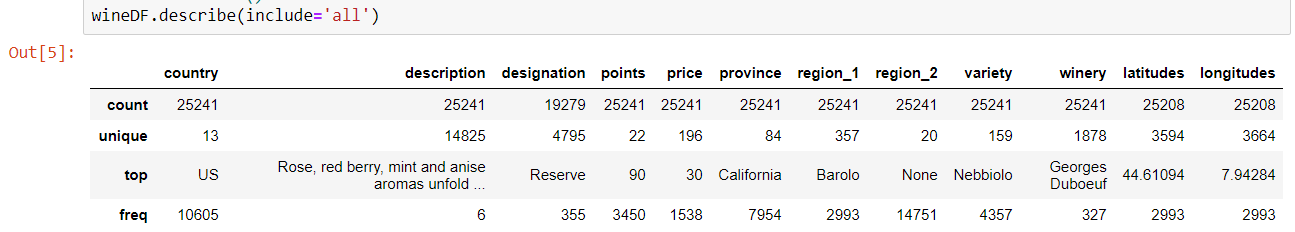
The assessment of the group is that there is no perceptible relationship between economic factors and the price or ratings variables.

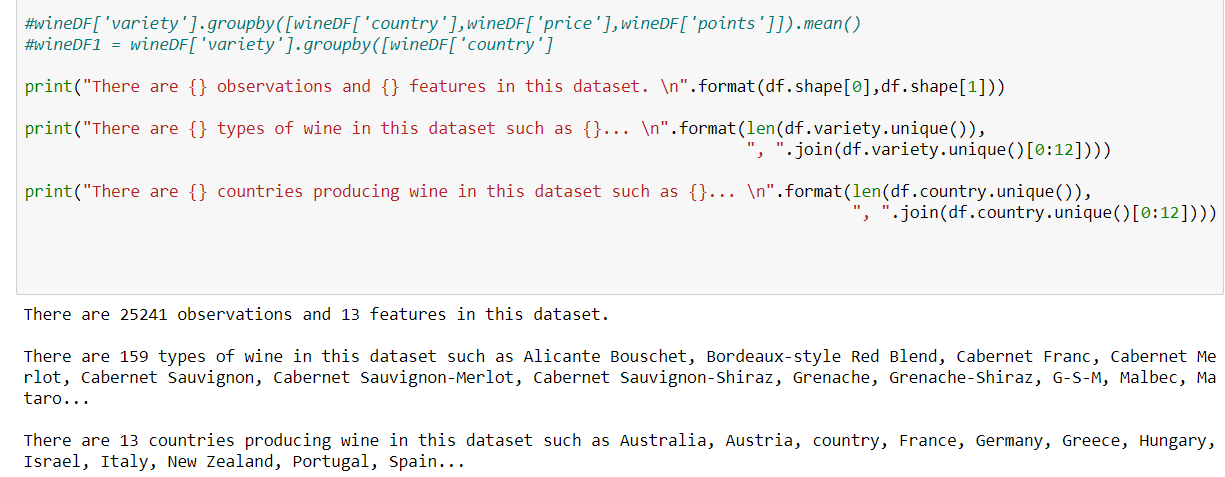
## Analysis of Geocode

Completing the geocoding allowed us to answer the question “where exactly is this wine coming from?”. It also allowed us to visualize interesting details from the dataset such as summary statistics of which an example is shown below:

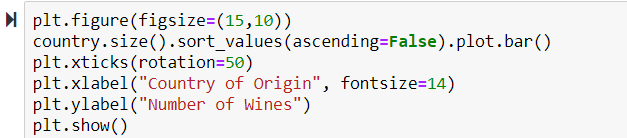


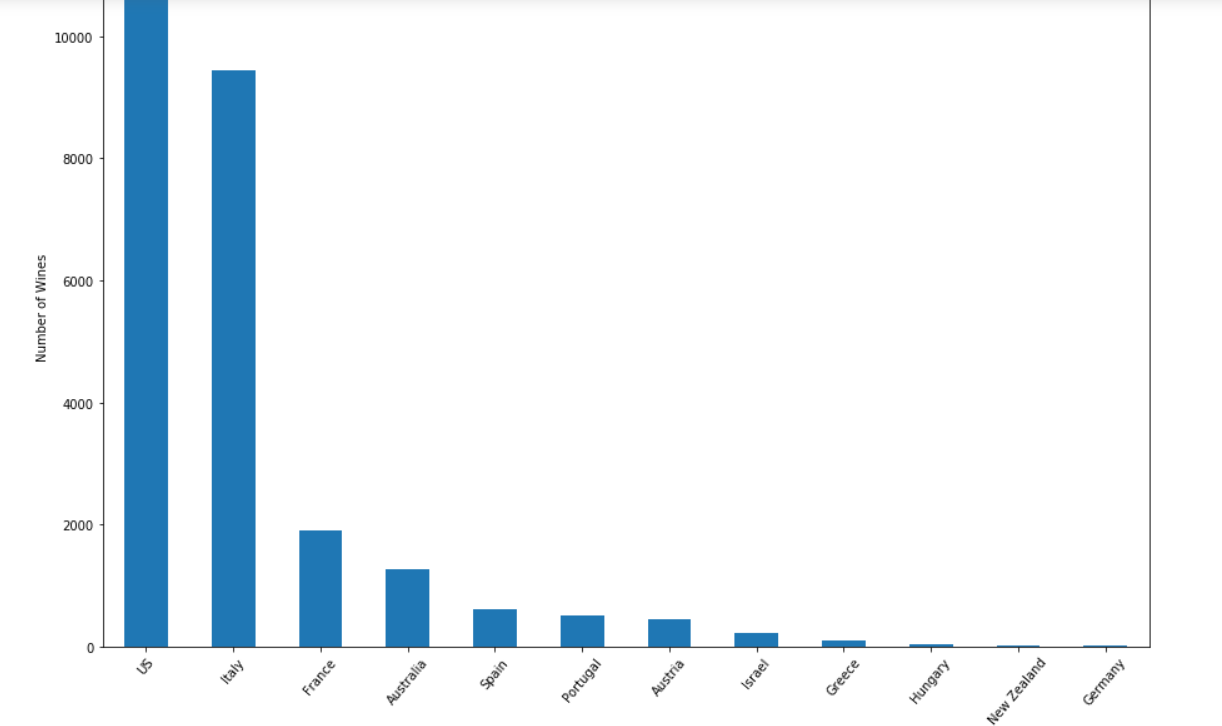
wineDF.describe(include='all')





How much wine are we talking about? Not hard to see who the largest number wine producers are:





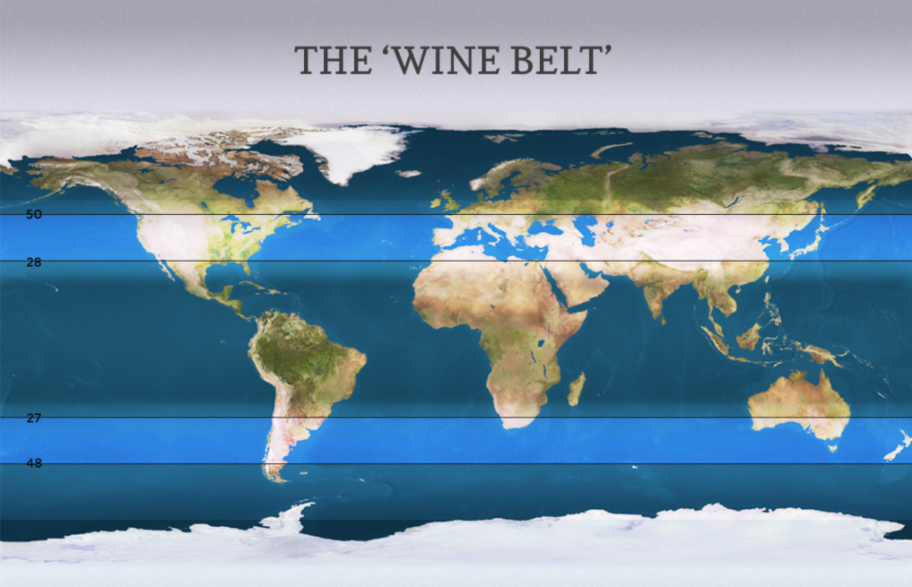
Where are we talking about…specifically?





## Does Weather Affect Wine Quality?​

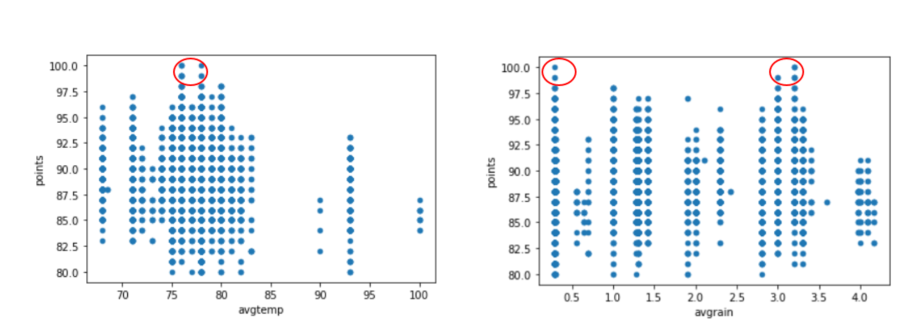
In order to get a better idea of the importance of weather on grape growth and, eventually, a high-quality wine, weather data was analyzed for each of the regions that are represented in the Wine Reviews data set. First, it is important to know what time of year has the largest impact on grape growth in a vineyard. According to [Wine Folly](https://winefolly.com/tips/start-planning-now-wine-harvest-season/), the wine harvest season happens over a two-month period, varying slightly based on the grape type, but more so on the location of the vineyard. Most wine vineyards are located on either the Northern or Southern ‘Wine Belt’. Vineyards in the North grow grapes that ripen and are ready for harvesting between August-October. Vineyards in the South experience this ripening between February-April.



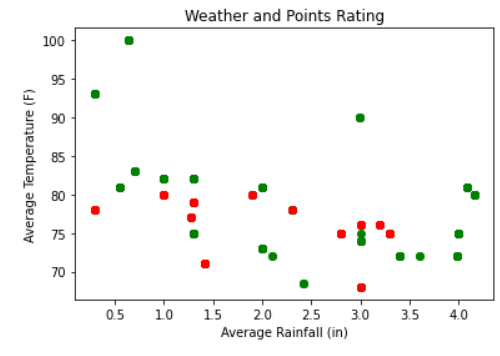
With this information in mind, the Wine Reviews data was split into two groups, those with vineyards in the Northern Hemisphere and those in the Southern Hemisphere. Weather averages for the Northern Hemisphere were pulled from September and Southern Hemisphere was pulled from March. Data on average temperature and rainfall for the ripening season of that area were pulled from [Current Results](https://www.currentresults.com/index.php) and joined to the Wine Reviews data set.

Initial data exploration found that the average temperature across all the vineyards in the data set during ripening season was 76.74 degrees Fahrenheit, and average rainfall was 1.82 inches. The lowest average temperature was 68 degrees Fahrenheit in Germany and parts of Austria. The highest average temperature was 100 degrees Fahrenheit in Arizona, USA. Average rainfall was more varied overall, with the lowest average coming from Adelaide District in California, USA at .29 inches and the highest average coming from Augusta, Missouri, USA at 4.16 inches.

In order to see if average rainfall or temperature at ripening season effects wine quality each wine’s awarded points can be used as a measure of quality. The Wine Reviews data set only encompasses highly rated wines (80 points or above), and there are many different varieties of wine represented. When taking the average temperature of each vineyard region in its ripening season the top rated (i.e. 99+ points) wines are grown in 76-78 degrees Fahrenheit. The two top rated wines represented here are a Cabernet Sauvignon from Napa Valley, California, USA and a Prugnolo Gentile from Vin Sano, Italy. Rainfall, on the other hand, has much higher variation. Those same top-rated wines, grown in temperatures within 2 degrees of each other, are grown in much different precipitation. The Cabernet Sauvignon grows with an average rainfall of .29 inches, where the Prugnolo Gentile has an average rainfall of 3-3.2 inches. Since vineyards grow their grapes outside, temperature could be a harder variable to control. Low rainfall, on the other hand, can be supplemented with watering, though it would raise costs.



Overall, the average temperature of a region during grape ripening season does seem to have an effect on wine quality, but the average rainfall does not have as clear a pattern. The best temperature to grow grapes that produce high quality wine is 76-78 degrees Fahrenheit.



Red: Points rating <= 94

Green: Points rating >= 95

## Does the size of the local economy affect the price or quality of wines?

After looking at the data, there does not seem to be correlation between the size of the local economy and either the price of the wine nor the rating of the wine. We did see some indications that price is related to the country or origin but related to factors other than the GDP.

## Does the growth of the local economy affect the price or quality of wines?

After looking at the data, there does not seem to be correlation between the growth of the local economy and either the price of the wine nor the rating of the wine. We did see some indications that price is related to the country or origin but related to factors other than the CAGR.

## What is the most popular wine on Twitter?

For the data completed for the sentiment analysis each category was taken directly off Twitter via API calls. Each tweet was found based on the hashtag of wine that analysis was to be done on. Each call was made to call and search for 2500 tweets. Once the tweets were called using the API they were stored in an empty dictionary. Once in the dictionary they were then parsed to remove special characters and other items.

Each tweet was then parsed for characters to then go over the sentiment of the tweet. In order to determine the sentiment of the tweet they had to determine a library to use for sentiment analysis. They used the Textblob library that was built on top of the NLTK package. Textblob uses movie review that were determined either positive or negative. Then based on this each tweet is passed through a polarity function. This then determines a polarity of the tweet between –1 and 1.

Then based on the polarity the tweet is determined either positive, negative, or neutral which means the tweet is at 0 for its polarity. Then given the polarity they are then able to determine the percentage of each positive, negative, and neutral tweets based on the number of tweets that are being analyzed in each case.

They looked at the varietals of the initial dataset that ranked high in average scores and searched for hashtags based on those varietals. They began with looking at #wine on Twitter in order to get a feel for how people feel about the drink with no varietal.

#Wine  
 Positive tweets percentage: 30.337078651685392 %

Negative tweets percentage: 2.247191011235955 %

Neutral tweets percentage: 67.41573033707866 % \

With this they could see that wine was overall seen as a neutral drink to Twitter. They then moved to looking at the varietals and began with #syrah. In the dataset they saw that the average rating of this wine was 93.33, which was one of the higher rated.

#Syrah  
 Positive tweets percentage: 30.337078651685392 %

Negative tweets percentage: 2.247191011235955 %

Neutral tweets percentage: 67.41573033707866 % \

This was similar to what they saw with wine which they found to interesting. Another interesting item to note was they saw that some of the negative tweets may not be considered negative by themselves, but the sentiment analysis saw them as such. They then looked at #malbec which had an average of 93.

#Malbec  
 Positive tweets percentage: 18.88888888888889 %

Negative tweets percentage: 1.1111111111111112 %

Neutral tweets percentage: 80.0 % \

#malbec seemed to show that less people to enjoy the wine according to the sentiment. It seemed that many users felt just alright given the choice to drink. They then choose to look at #pinotnoir which had a middle rating of 89.1. They chose to look this up given the relative popularity of the varietal in society.

#PinotNoir  
   
 Positive tweets percentage: 36.7816091954023 %

Negative tweets percentage:2.2988505747126435 %

Neutral tweets percentage: 60.91954022988506 % \

#pinotnoir showed the most positive reaction from Twitter and the least neutral reaction. Following this they then went to look into the least popular varietal of #lambrusco which had an average rating of 83.

#lAMBRUSCO  
 Positive tweets percentage: 28.571428571428573 %

Negative tweets percentage: 2.857142857142857 %

Neutral tweets percentage: 68.57142857142857 % \

This varietal was not a surprise as it followed its rating and did not impress the users of Twitter. They had planned to look further into the vineyards of that had high ratings but gathering tweets on these locations was not feasible via Twitter and they could not move forward with sentiment analysis of those items.

## Conclusion (All)

Overall, the average temperature of a region during grape ripening season does seem to have an effect on wine quality, but the average rainfall does not have as clear a pattern. When taking the average temperature of each vineyard region in its ripening season the top rated (i.e. 99+ points) wines are grown in 76-78 degrees Fahrenheit. High quality wines can be grown in areas with an average rainfall from .29 inches to 4.16 inches, but low rainfall can be supplemented with watering.

According to the data there appears to be no correlation for either the size or growth of the local economy and a wines price or points ratings. There were indications that the price of wine was related to the country of origin but this was not seen to been overwhelming.

Based on the sentiment of Twitters users they found that the most popular wine was #pinotnoir. This was interesting given that this was the most popular rated wine in the initial dataset. For future analysis the group would investigate a different sentiment analysis than Textblob given the large number of neutral tweets that did show in their analysis.