BEST VALUE WINE

Discovering high quality wine

ABSTRACT

Thousands of wineries exist across the world. Each region and each winery have a distinct style. Some wine-producing regions can charge a premium price. Other, less renowned regions, must sell at a more modest price. The variance in price for a bottle of wine is extremely large (e.g. £4 - £3,000). The aim of this study is, a) to determine if Price is a good predictor of Quality and, b) to discover wineries which make high quality wine and sell it at a reasonable price.

Steve Williamson

Capstone Applied Data Science Project

Introduction and Business Problem

I am conducting a feasibility study into starting a wine merchant business. The wine trade is a competitive business and the supply chain is extremely complicated as there are thousands of wines to choose from and thousands of wine producers across the world. Customers have many options for purchasing wine, such as: Supermarket, specialist wine merchant shop, online wine clubs and subscription services.

In order to succeed in this competitive market, I need to have a Unique Selling Point (USP). I hope to make this my ability to source high quality wines at a price that is affordable to the every-day wine drinker (e.g. £10 - £50). These wines would be of the same quality as equivalent wines that sell for much higher prices. There are thousands of wines produced (operating from wineries) across the world. I need a model which will help me identify those high-quality wine producers, who are unable to charge the premium prices of well-known wine brands and regions. It is not feasible to taste all those wines, so I must make my selection based on geographic regions, expert ratings and other relevant characteristics.

The primary output will be:

- a ranking list of wine producing regions that offer the best value wine (from Data Set)
- The identification of wineries within these regions (from Four Square)

The selection of wineries will be based on the following:

- Regions / Provinces that have been rated highly for the quality of their wine
- Wineries, within those regions, which are rated highly by wine drinkers

Data Sources

My primary data source will be a data set of Wine Ratings, I have downloaded this as a csv file from Kaggle. It contains the ratings of 129,970 different wines, and has the following columns:

- Country
- Description
- Designation
- Points
- Price
- Province
- Region 1
- Region 2
- Taster Name
- Taster Twitter Handle
- Title
- Variety
- Winery

There are some interesting features in this data set worth exploring, for example::

- Points (the quality rating of the wine, scored on a scale 1 - 100)
- Price (the average selling price of a bottle)

Points are scored by an expert wine critic (Taster). **Price** is determined by various factors, such as region, name of wine producer, grape variety and marketing. Some key insights I am looking to discover are:

- How strong is the correlation between Price and Points (Quality)?
- Is there a correlation between Province / Region and Quality?
- Is there a correlation between Province / Region and Price?

Data from FourSquare and Google

I will use the **Google geo-coordinate API** to obtain the Latitude and Longitude of the regions and specific wineries.

I will use FourSquare to:

- Identify Wineries within each region
- obtain the ratings of individual wineries

I will also use Foursquare to elaborate my list of recommended wineries with other information, relevant to the purchase of wine.

Methodology

The steps in the process were as follows:

- Data Pre-processing
- Exploratory Data Analysis
- Statistical Analysis
- Identification of high-quality wine producing regions
- Identification of new (high-quality) wineries

Data Pre-Processing

Remove Noise from Dataset

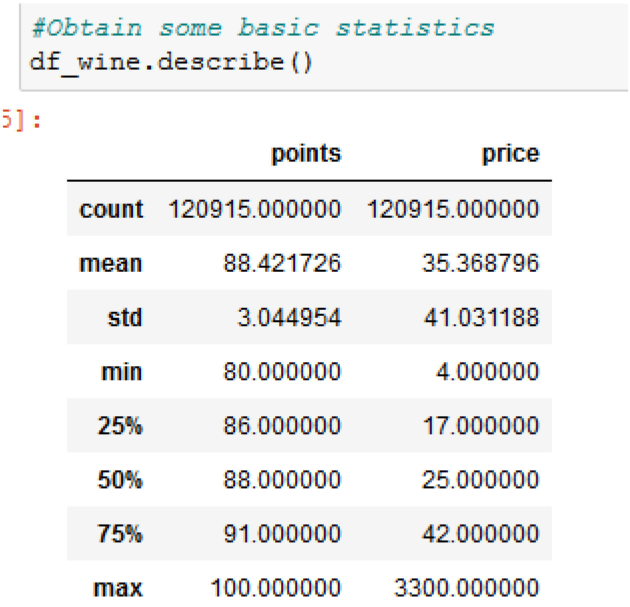
The key features I need for data analysis and subsequently completing the project are:

- Country
- Province
- Points
- Price

I removed those samples (rows) that had null values in any of those columns. This reduced the data set from 129,971 to 120,915 rows. This was only a small reduction of samples and still leaves me with a large data set, suitable for analysis.

Exploratory Data Analysis

I first obtained some basic statistical information on the data-set. Points and Price are the only two numerical data points.



```
#Obtain some basic statistics
df_wine.describe()
```

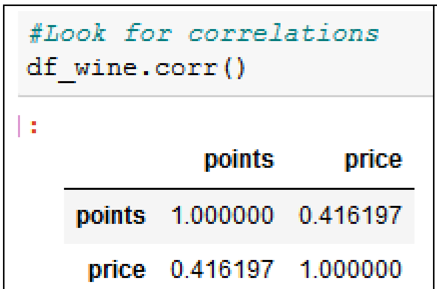
5]:

	points	price
count	120915.000000	120915.000000
mean	88.421726	35.368796
std	3.044954	41.031188
min	80.000000	4.000000
25%	86.000000	17.000000
50%	88.000000	25.000000
75%	91.000000	42.000000
max	100.000000	3300.000000

Price / Points correlation

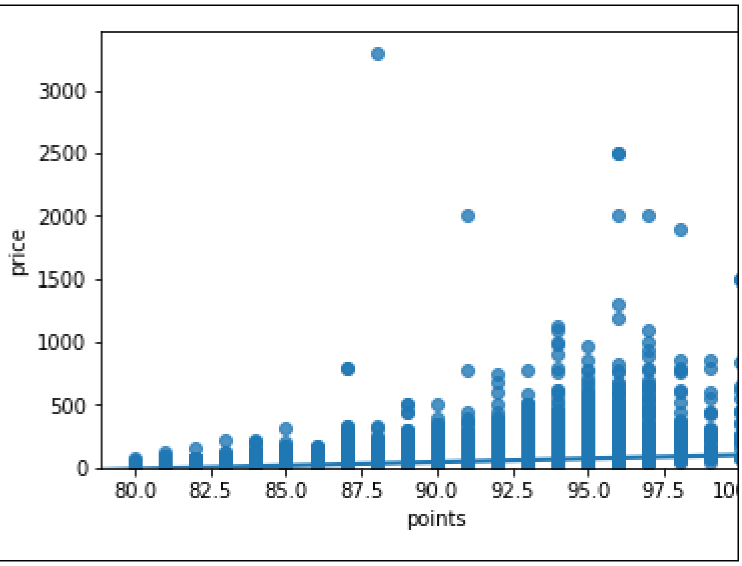
The first two features I explored was Points and Price. Points represents quality and these are assigned by an expert wine taster. I wanted to identify the strength of the correlation between Price and Quality (Points). Applying the **.corr()** to the data frame provided the **co-relation co-efficient**, which was:

0.41



```
#Look for correlations
df_wine.corr()
```

```
]:
             points     price
points     1.000000  0.416197
 price     0.416197  1.000000
```



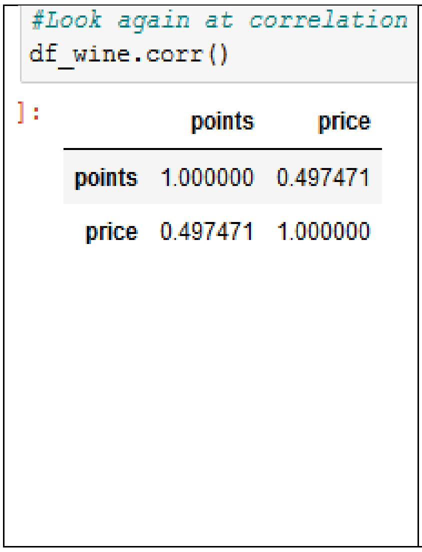
Remove Outliers

The Standard Deviation (from above descriptive table) showed the following:

Stat	**Points**	**Price**
StdDev	4	40

This indicates a narrow range of values for points and a wide range of values for price. A Scatter Plot with a regression line illustrated a small number of extreme outliers on price (e.g. bottles of wine in excess of £500). I decided to remove those outliers. As my primary focus was to be quality (Points), I decided to remove all rows where the price was in excess of **£50**.

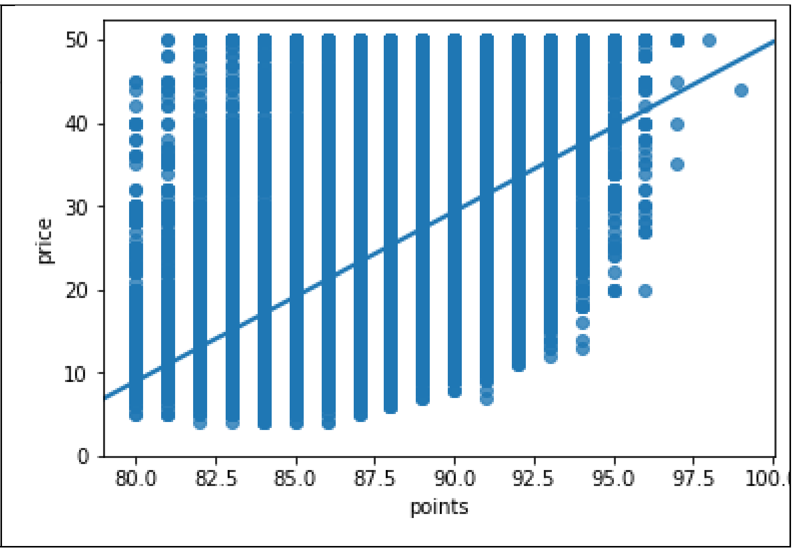
This resulted in a reduction of 19,774 samples. Re-running the correlation function and replotting the regression plot, yielded the following:



```
#Look again at correlation
df_wine.corr()
```

]:

	points	price
points	1.000000	0.497471
price	0.497471	1.000000



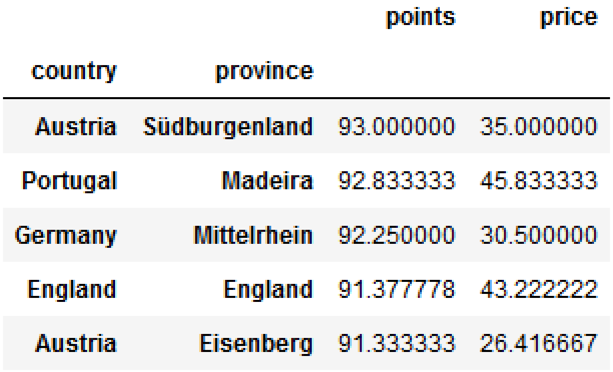
Removing outliers makes the statistical model more reliable. As can be seen, the co-relation is still weak, confirming the null hypothesis, that Price is **not** a good indicator of quality.

Results

The Best value wine producing Regions

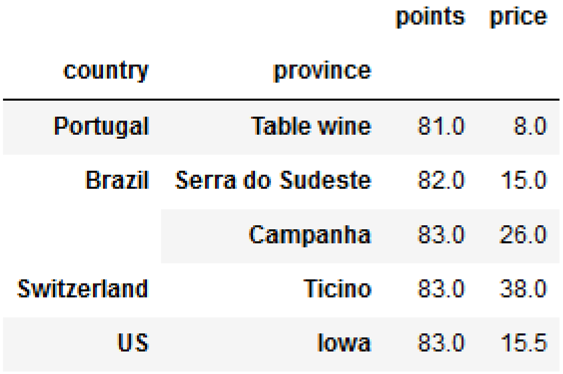
From the remaining data set (containing 101,141 samples), I produced some summary statistics by Country and Province, using the .groupby() and mean() functions. I took the mean points score and price for each region and looked at the Top 5 and the Bottom 5. This produced the following results:

Top 5 Countries / Provinces



		points	price
country	province		
Austria	Südburgenland	93.000000	35.000000
Portugal	Madeira	92.833333	45.833333
Germany	Mittelrhein	92.250000	30.500000
England	England	91.377778	43.222222
Austria	Eisenberg	91.333333	26.416667

Bottom 5 Countries / Provinces



		points	price
country	province		
Portugal	Table wine	81.0	8.0
Brazil	Serra do Sudeste	82.0	15.0
	Campanha	83.0	26.0
Switzerland	Ticino	83.0	38.0
US	Iowa	83.0	15.5

Portugal appeared in **both** the Top 5 and the Bottom 5. This highlights the importance of using Province. Country alone will not be a good predictor of quality.

I now have two data sets:

- Main data set consisting of one row per wine
- Provinces data set, consisting of the following features:
  - Country
  - Province
  - Mean Points score
  - Mean Price
  - Latitude
  - Longitude

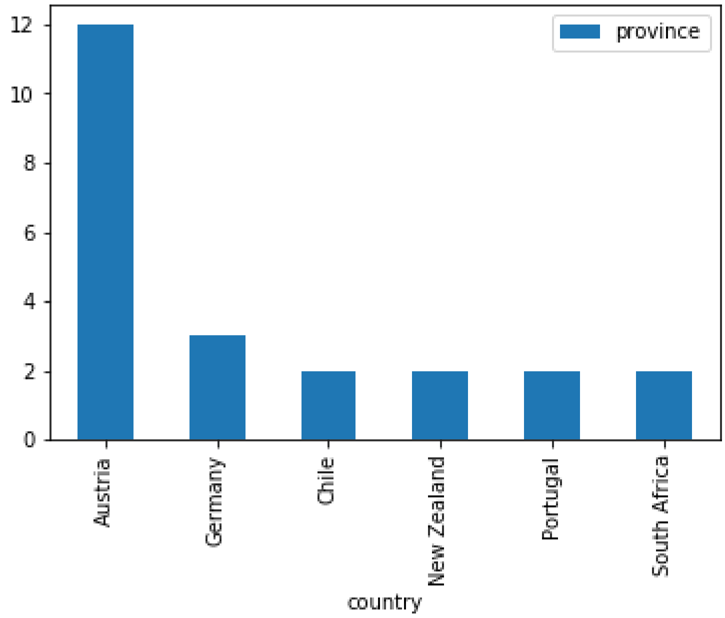
The Latitude and Longitude are values were obtained by calling the Google Geo-code API and searching on Province and Country.

Further refinement of Data Set

The Provinces (Regions) data set was refined further by:

- removing those Provinces that had a mean point score of less than 90
- Removing the rows where a country only had one province with a mean score of 90 or more

The following Bar Chart shows the countries and the number of provinces in each country that rank as high-quality wine producing regions:

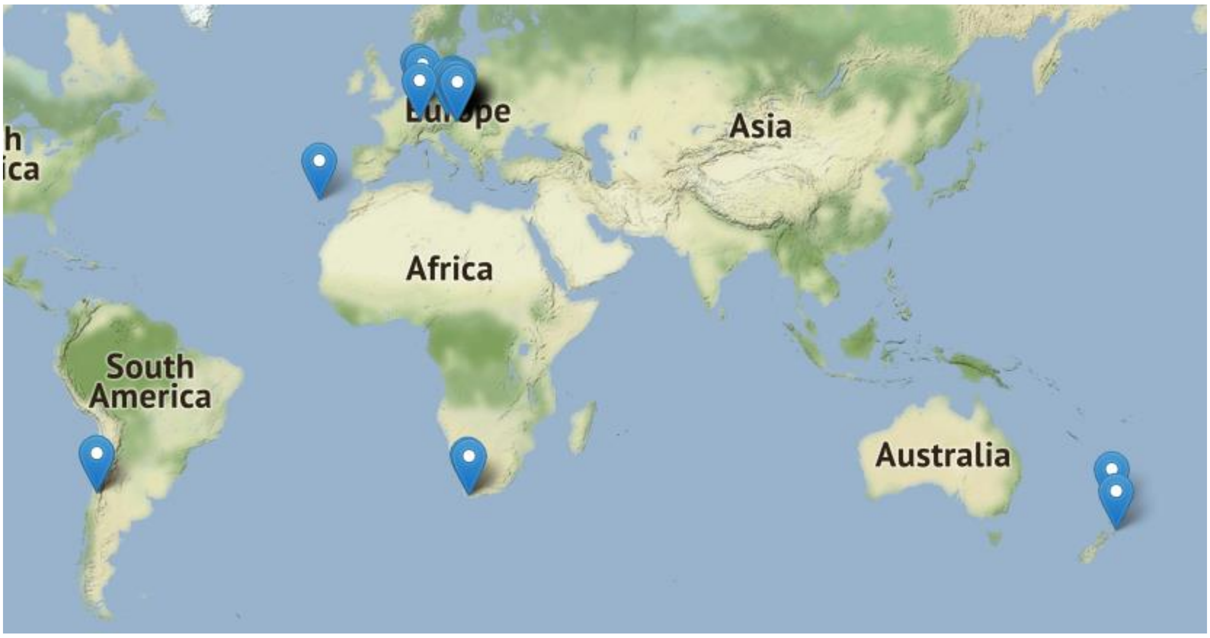


As can be seen, Austria is the Country with the greatest number of high-quality provinces.

It should be remembered that the focus of this project is on **Best Value, Quality wine**. If it was purely on quality, then this result may have been different. At an earlier stage, I removed all wines that sold in excess of £50 per bottle. Even though the co-relation between Points and Price is weak, it would likely have altered this result. A table, ranking the Countries / Provinces before removing the high price wines is in the Notebook.

Plotting Provinces on the world Map

Having obtained the geo-coordinates, I plotted these provinces on a world map (created using folium)



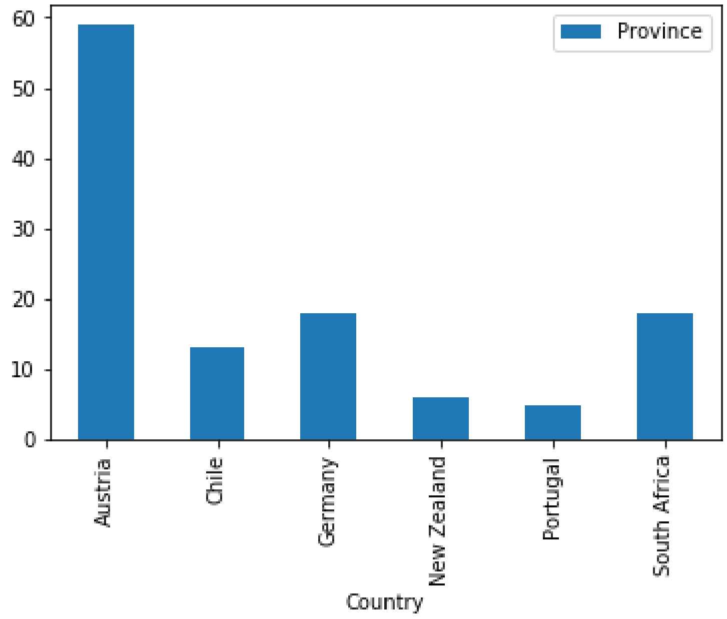
Discover new (high quality) Wineries

The next stage of the project involves identifying wineries in these provinces. Many of these wineries **are not** on the original data set. This step was achieved as follows:

- Update the Provinces data frame with the Get geo-co-ordinates for each province (using a Google maps API)
- Explore each region (using the FourSquare **explore** API), and search for "Winery"
- Build a new data set, made up of those newly discovered wineries
- For each of these newly discovered wineries, use another FourSquare API (**venue details**) to get further information on each winery, specifically:
  - Rating
  - Url

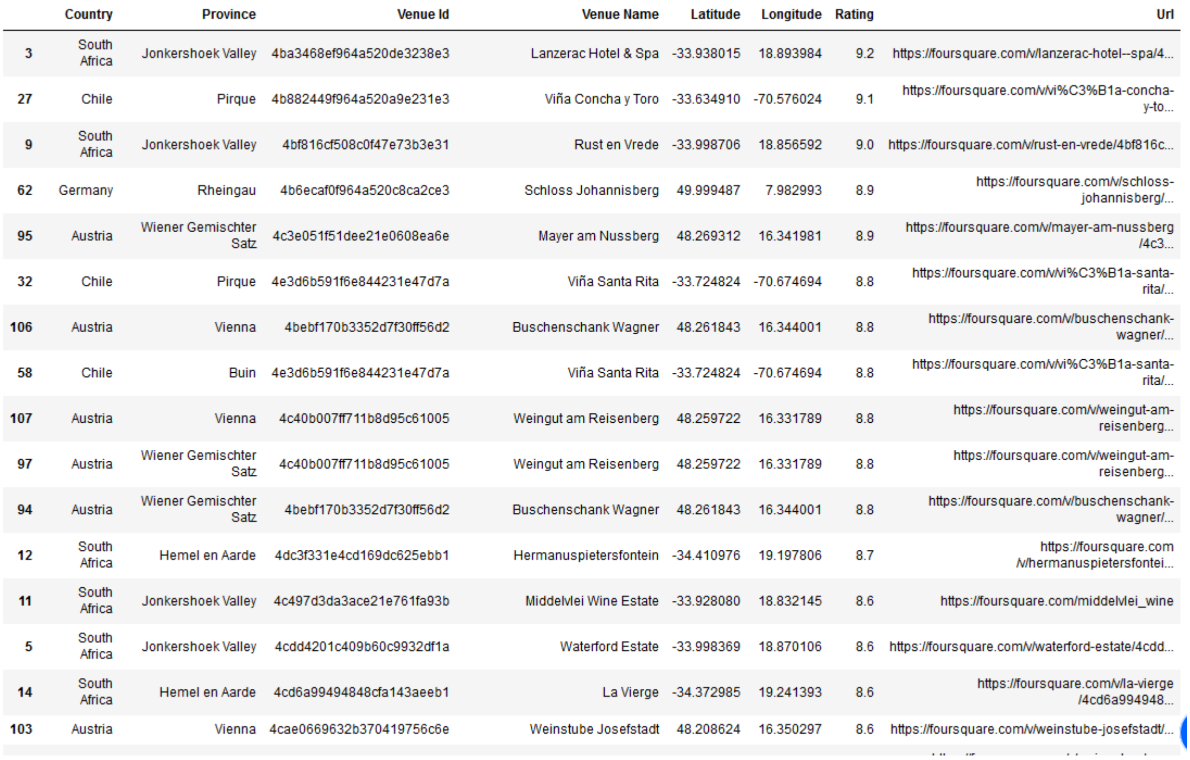
The rating may be useful to perform another level of ranking on the data and the url will be useful when plotting these wineries on a map.

Newly discovered wineries

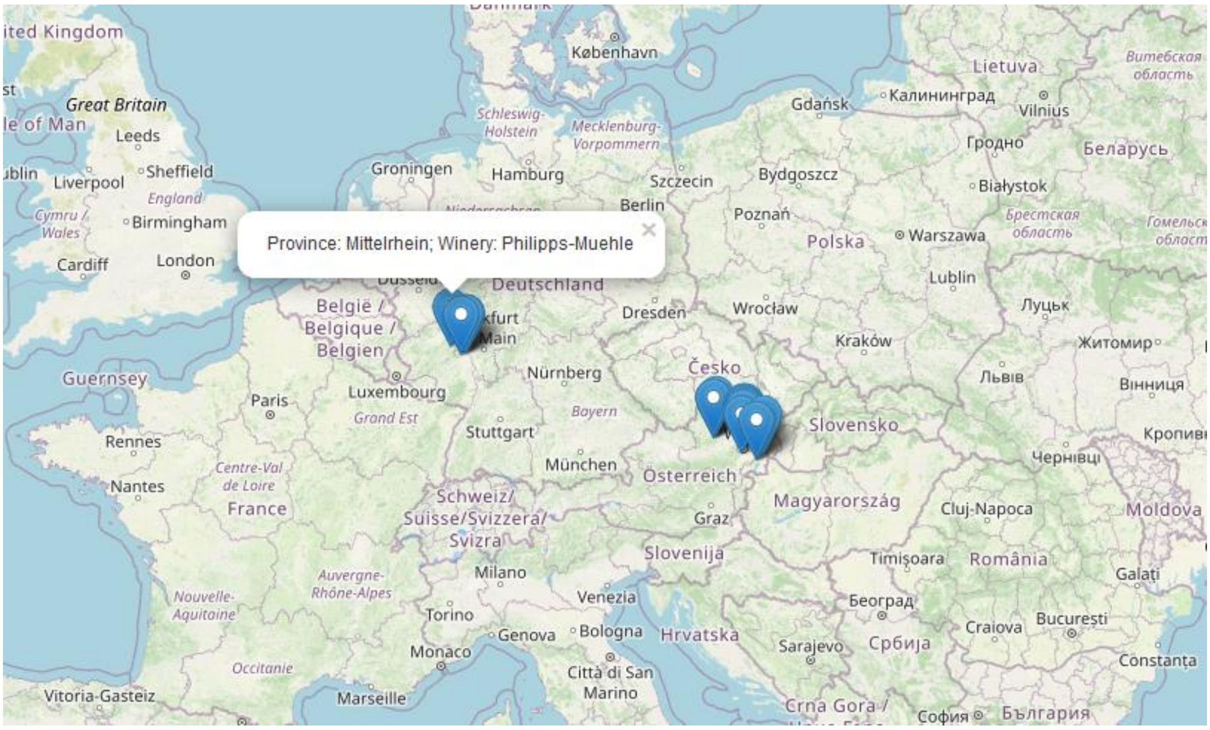


Ranking of New Wineries

The results of the discovery exercise are below. 119 new wineries were identified, and the ratings were in the range 6.3 to 9.2. The top 50 from this list were plotted on a world map and a snap-shot of the Europe region is below (only wineries with a rating score greater than 8 made it onto this final list.



	Country	Province	Venue Id	Venue Name	Latitude	Longitude	Rating	Url
3	South Africa	Jonkershoek Valley	4ba3468ef964a520de3238e3	Lanzerac Hotel & Spa	-33.938015	18.893984	9.2	https://foursquare.com/v/lanzerac-hotel--spa/4...
27	Chile	Pirque	4b882449f964a520a9e231e3	Viña Concha y Toro	-33.634910	-70.576024	9.1	https://foursquare.com/v/vi%C3%B1a-concha-y-to...
9	South Africa	Jonkershoek Valley	4bf816cf508c0f47e73b3e31	Rust en Vrede	-33.998706	18.856592	9.0	https://foursquare.com/v/rust-en-vrede/4bf816c...
62	Germany	Rheingau	4b6ecaf0f964a520c8ca2ce3	Schloss Johannisberg	49.999487	7.982993	8.9	https://foursquare.com/v/schloss-johannisberg/...
95	Austria	Wiener Gemischter Satz	4c3e051f51dee21e0608ea6e	Mayer am Nussberg	48.269312	16.341981	8.9	https://foursquare.com/v/mayer-am-nussberg/4c3...
32	Chile	Pirque	4e3d6b591f6e844231e47d7a	Viña Santa Rita	-33.724824	-70.674694	8.8	https://foursquare.com/v/vi%C3%B1a-santa-rita/...
106	Austria	Vienna	4bebf170b3352d7f30ff56d2	Buschenschank Wagner	48.261843	16.344001	8.8	https://foursquare.com/v/buschenschank-wagner/...
58	Chile	Buin	4e3d6b591f6e844231e47d7a	Viña Santa Rita	-33.724824	-70.674694	8.8	https://foursquare.com/v/vi%C3%B1a-santa-rita/...
107	Austria	Vienna	4c40b007ff711b8d95c61005	Weingut am Reisenberg	48.259722	16.331789	8.8	https://foursquare.com/v/weingut-am-reisenberg...
97	Austria	Wiener Gemischter Satz	4c40b007ff711b8d95c61005	Weingut am Reisenberg	48.259722	16.331789	8.8	https://foursquare.com/v/weingut-am-reisenberg...
94	Austria	Wiener Gemischter Satz	4bebf170b3352d7f30ff56d2	Buschenschank Wagner	48.261843	16.344001	8.8	https://foursquare.com/v/buschenschank-wagner/...
12	South Africa	Hemel en Aarde	4dc3f331e4cd169dc625ebb1	Hermanuspietersfontein	-34.410976	19.197806	8.7	https://foursquare.com/v/hermanuspietersfontei...
11	South Africa	Jonkershoek Valley	4c497d3da3ace21e761fa93b	MiddelVlei Wine Estate	-33.928080	18.832145	8.6	https://foursquare.com/middelvlei_wine
5	South Africa	Jonkershoek Valley	4cdd4201c409b60c9932df1a	Waterford Estate	-33.998369	18.870106	8.6	https://foursquare.com/v/waterford-estate/4cdd...
14	South Africa	Hemel en Aarde	4cd6a99494848cfa143aeeb1	La Vierge	-34.372985	19.241393	8.6	https://foursquare.com/v/la-vierge/4cd6a994948...
103	Austria	Vienna	4cae0669632b370419756c6e	Weinstube Josefstadt	48.208624	16.350297	8.6	https://foursquare.com/v/weinstube-josefstadt/...



Province: Mittelrhein; Winery: Philipps-Muehle

Map, with the top wineries marked on it. By clicking on the map, you see the name of the winery.

Discussion

Machine Learning

I did not include a Machine Learning Model in this project. If Price were a good indicator of quality, I would have had a Use Case for a Linear Regression model. Alternatively, I could have done some exploratory analysis on other features. However, the main purpose of this project is to identify regions that produce high quality wine.

The remainder of the analysis focussed on refining the data set and elaborating it with external data in order to identify those wineries that exist in high quality wine producing regions.

Statistical Analysis

Identifying Correlations

There were two numeric features in the data set, Price and Points (quality). The analysis showed that the correlation was weak (i.e. price is not a predictor of quality).

Countries and Regions

Examining the mean points score of each Country / Province provided a ranking of the best wine producing regions. This analysis showed that country alone could not be used to predict good quality wines. For example, Portugal was in the Top 5 and the Bottom 5. Province is a narrow geographic region.

Other possible predictors

Another possible predictor would have been elevation of the winery (i.e. number of meters above sea level). Many wine producers claim that vineyards on higher altitude regions produce better quality wine. It would have been useful to confirm this statement. That would have been possible as follows:

- Use FourSquare or Google to get the elevation of each winery on the original list (120,000 samples) and add that column to the data set
- Do a correlation analysis, i.e. a scatter chart with Points and Elevation

That was not done as part of this project but may be useful to investigate further.

Next Steps

The aim of this project is to identify several high-quality wine producers. The output of this exercise (i.e. Top 50 wine producers) would be input for more a detailed business plan, e.g.

- What would be the cost of sourcing wine from those producers?
- How would we market these (not well known) wines?

Conclusion

I generated the following Data Frames during this exercise.

Data Set	**Description**
Wines	One row (sample) per wine, with ratings and prices
Provinces	One row per Country / Province (including the mean Points score and the geo-coordinates for each province)
New wineries	One row for each of the newly discovered wineries (including the FourSquare rating and the web site url)

It was discovered at an early stage that the correlation between Price and Points (wine quality) was weak. This was a positive finding because the purpose of this exercise is to discover **high quality, good value** wines. It there was a strong co-correlation between Price and Quality, this would have become a very different exercise.

There are a set of Countries and Provinces that have a high proportion of quality wines, which sell at an affordable. This project identified all wineries in these high-quality provinces. Ratings were obtained from Four Square and these were used to rank those wineries.

119 wineries were discovered. There ratings were in the range: **6.3** to **9.2**. To further improve the quality of the results, only the Top 50 were plotted on the world Map (see above). This meant that only those with a rating of 8 or higher were selected.