CM3010 Midterm Report

1. Finding and critiquing the dataset

1.1 Choosing a source of open data

Dataset: Wine Magazine Reviews

- **Dataset description**: This is a dataset of wine reviews derived from data that was scraped from WineEnthusiast, an online magazine, "the premier destination for all things wine. From acclaimed wine ratings and reads to wine storage, glasses and more." - as stated on the website: https://www.wineenthusiast.com/ratings/

  - "Search the official Wine Enthusiast Ratings database to find the best reviews for all your favourite wine, beer and spirits. Our team of reviewers blind taste more than 24,000 wines from across the globe every year. Our free access to wine, beer and spirit ratings can be filtered by score, price, vintage, grape variety and region, making it a flexible tool to easily find your next favourite bottle."

- **Data fields**:
  ['country', 'description', 'designation', 'points', 'price', 'province', 'region_1', 'region_2', 'taster_name', 'taster_twitter_handle', 'title', 'variety', 'winery']

- **Description** as per metadata posted on Kaggle
  - **country**: The country that the wine is from
  - **description**: The review (text)
  - **designation**: The vineyard within the winery where the grapes that made the wine are from
  - **points**: The number of points WineEnthusiast rated the wine on a scale of 1-100 (though it is said only reviews for wines that score more than 80 are posted)
  - **price**: The cost for a bottle of the wine
  - **province**: The province or state that the wine is from
  - **region_1**: The wine growing area in a province or state (ie Napa)
  - **region_2**: Sometimes there are more specific regions specified within a wine growing area (ie Rutherford inside the Napa Valley), but this value can sometimes be blank
  - **taster_name**: Name of the reviewer/taster
  - **taster_twitter_handle**: Twitter handle of the reviewer/taster
  - **title**: The title of the wine review, which often contains the vintage if you're interested in extracting that feature
  - **variety**: The type of grapes used to make the wine (ie Pinot Noir)
  - **winery**: The winery that made the wine

- **Link to data** (kaggle): https://www.kaggle.com/datasets/zynicide/wine-reviews

1.2 Assessing the dataset

1.104 Identifying data sources for a given domain

- **Quality:**
  - The data has been scraped from WineEnthusiast by the Kaggle user during the week of June 15 2017, and updated with reviews again on November 22 2017 after feedback from users of the dataset. The update included the title of each review from which the year, taster's name and Twitter handle could be parsed out of, to rectify duplicate entry issues.
  - This information is gathered from the acknowledgements and update found on the data card on the kaggle link from where I have downloaded the dataset : https://www.kaggle.com/datasets/zynicide/wine-reviews
  - Based on these factors, the data is fairly reliable as there seems to have been effort into ensuring its accuracy and addressing identified issues:
    - *Reputation of the source*: Wine Enthusiast is a respected publication in the wine industry, suggesting their reviews are likely credible.
    - *Reviewer expertise*: The dataset includes information on tasters' names and Twitter handles, allowing for potential verification of their expertise.
    - *Data maintenance*: The dataset has been updated based on user feedback, demonstrating a commitment to quality control.
    - *Consistency*: The inclusion of titles for each review provides a way to check for inconsistencies or errors within the data.
    - *Examining sample reviews*: Manual review of a sample of the data reflects quality and consistency.

- **Detail:**
  - There is sufficient detail on the various aspects of the wines that are broken down to individual fields
  - This is helpful for a comprehensive understanding of the wines. The data includes granular details like grape variety, specific vineyard designations, and even sub-regions within wineries, providing a rich picture of each wine's origin and potential characteristics.
  - This level of detail is valuable for informing nuanced analyses and decision-making, allowing exploration of factors like regional trends, producer variations - making it useful for wine recommendations, investment strategies, research on wine quality for sommeliers, wine collectors and researchers.
  - The kaggle user in fact, who produced this dataset has indicated that "this dataset offers some great opportunities for sentiment analysis and other text related predictive models. [Their goal being] to create a model that can identify the variety, winery, and location of a wine based on a description."

- **Documentation:**
  - The documentation was clear and concise and provided comprehensive context of the data, even with some statistical analysis in the metadata.
  - The kaggle user even included the code for the scraper (here) in the case that users have any more specific questions about data collection that have not been addressed.

- **Interrelation:**
  - Connecting this dataset to others would significantly enhance its value and enable a broader range of analyses. Possibilities include:
    - *Retailer data*: Linking reviews to sales data from wine retailers could explore consumer preferences, price sensitivity, and regional trends. Potential sources include online wine stores or point-of-sale systems.
    - *Expert ratings*: Comparing Wine Enthusiast scores with ratings from other publications or experts could assess consistency and potential biases.
    - *Grape variety characteristics*: Incorporating information about grape varieties' typical characteristics could enhance understanding of taste descriptions and ratings.

  - Relative ease of connection based on the fact that data is well broken down into distinct fields, and compatible formats (csv, json) are provided.

- **Use:**
  - Potential uses of this dataset include:
    - *Recommendation systems*: Recommending wines based on user preferences, price range, or desired style.
    - *Predicting wine quality and price*: Using machine learning to model relationships between features and scores/prices.
    - *Exploring regional trends*: Analysing ratings to uncover regions producing exceptional wines.
    - *Understanding consumer preferences*: Identifying popular grape varieties, styles, and price points.
    - *Tracking wine trends over time*: Analysing changes in ratings and prices to spot emerging regions or styles.

- Missing information that could enhance the dataset:
  - *Vintage and date of publishing* (of the review):
    - Year of wine production is often important for quality and price.
    - This is included in the review titles, but would be helpful to be parsed out into a separate field especially as there are repetitions in the data of reviews of the same wine, but of different vintage and published at different times as well.
  - *Wine characteristics*: Information about alcohol content, acidity, sweetness, body, and tannins.
  - *Reviewer demographics*: Information about reviewer age, gender, and location to assess potential biases.

1.206 Exploring terms of use and licenses on publicly-available data

Explicit Licensing and Ownership:

- Kaggle Dataset Page: The Kaggle dataset page itself doesn't explicitly mention any licenses or ownership of the data. This is quite common on Kaggle, as users upload datasets they may not necessarily own, relying on the original source's licensing.

- Wine Enthusiast Terms of Use: Wine Enthusiast has a dedicated "Terms of Use" page on their website: https://www.wineenthusiast.com/terms-of-use/ . It outlines restrictions on how you can use their content, including data.

  - Key Points from Wine Enthusiast Terms of Use:
    - *Prohibited Uses*: You cannot "share, store, display or otherwise use the Content without Wine Enthusiast's express written permission" (including commercial use).
    - *Limited Commercial Use*: Commercial users can use up to 500 individual pieces of "Review Content" (including ratings, scores, tasting notes) but need written permission for exceeding this limit.
    - *Attribution*: You must cite Wine Enthusiast as the source when using their Review Content.

- Possible Restrictions and Their Significance:
  - *Non-commercial use*: This dataset seems suitable for personal exploration and non-commercial research. Sharing findings freely or using them for educational purposes should not pose any issues.

- Separation of Information:
  - *Kaggle Page*: Unfortunately, the Kaggle page doesn't explicitly link to Wine Enthusiast's Terms of Use or provide clear ownership information. This highlights the potential disconnect between uploaded data and its original licensing terms.

1.3 Explanation of interests

This dataset on wine reviews is intriguing due to its rich and diverse information encompassing various aspects of wine tasting and production. It provides a comprehensive view of wines from different countries, provinces, wineries, and varieties, as well as insights into the preferences of various tasters. This dataset can be valuable for wine enthusiasts, producers, and researchers alike.

As someone with a budding interest in wine and limited background knowledge, the dataset caught my attention. Given the immense diversity and complexity of the wine world, it seemed like an intriguing source to gain insights into the factors that make a wine preferable and contribute to the high ratings. I am particularly curious about the geographical statistics, exploring where wines, especially specific types, are predominantly produced. The inclusion of pricing information adds an extra layer of interest, prompting an analysis of the relationship between pricing and the perceived quality or preference of wines, as well as their production dynamics.

Questions to Ask of the Dataset:

1. Identification of Exceptional Wines:
   - What are the wines that received the highest ratings across different locations, varieties, and tasters?

2. Understanding Variety Popularity:
   - Which wine varieties are most commonly reviewed, and what are their average ratings?

3. Geographical Analysis:
   - How do wine ratings vary across different countries and provinces?
   - Which countries or provinces produce the highest-rated wines on average?

4. Top Wineries Insights:
   - What are the top-rated wineries based on both review count and average points?
   - Are there correlations between the popularity of wineries and their geographical locations?

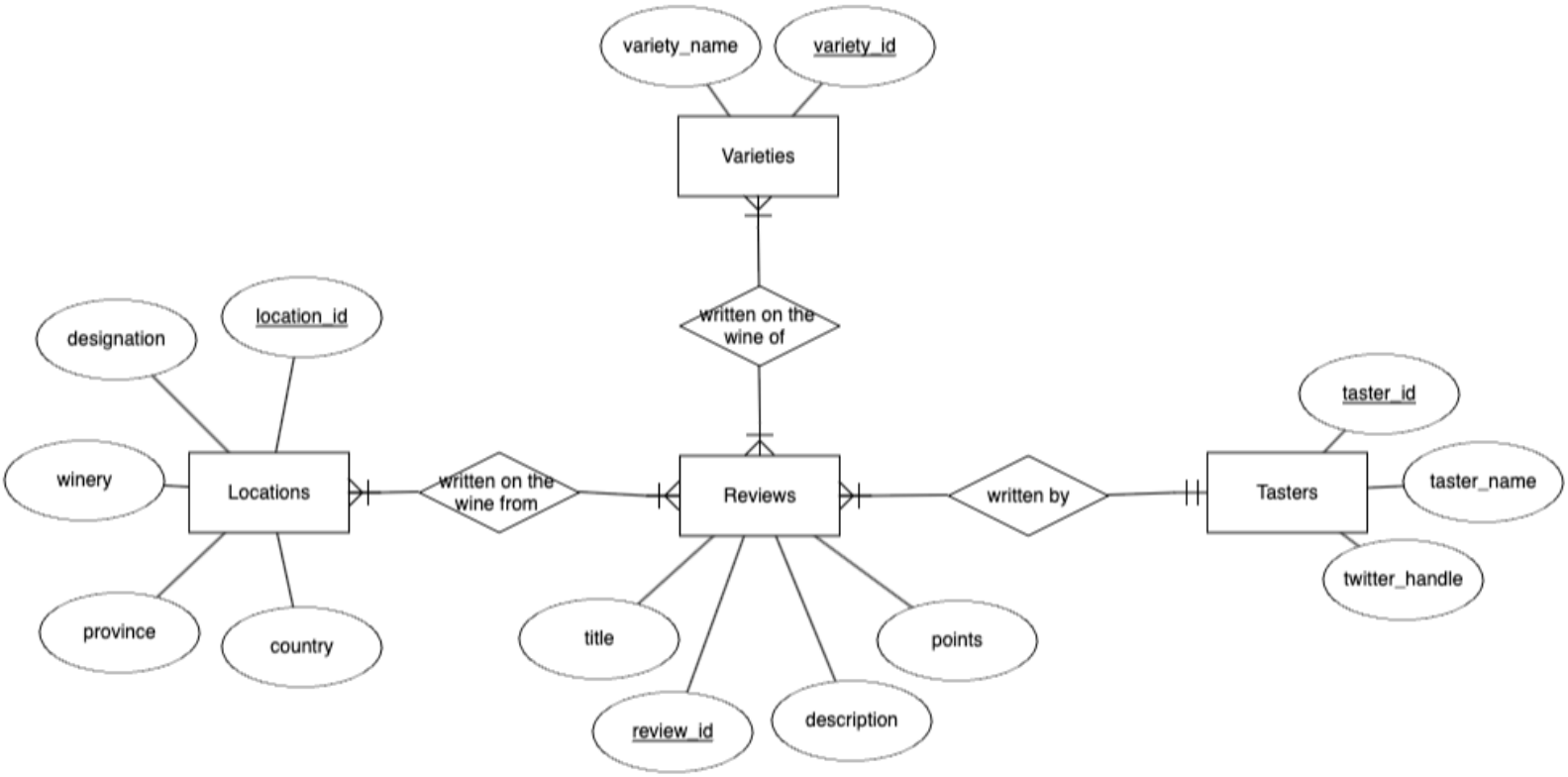
5. Price:
   - How does the pricing of wines correlate with their perceived quality or preference
   - Does it offer insights into production dynamics or market trends?

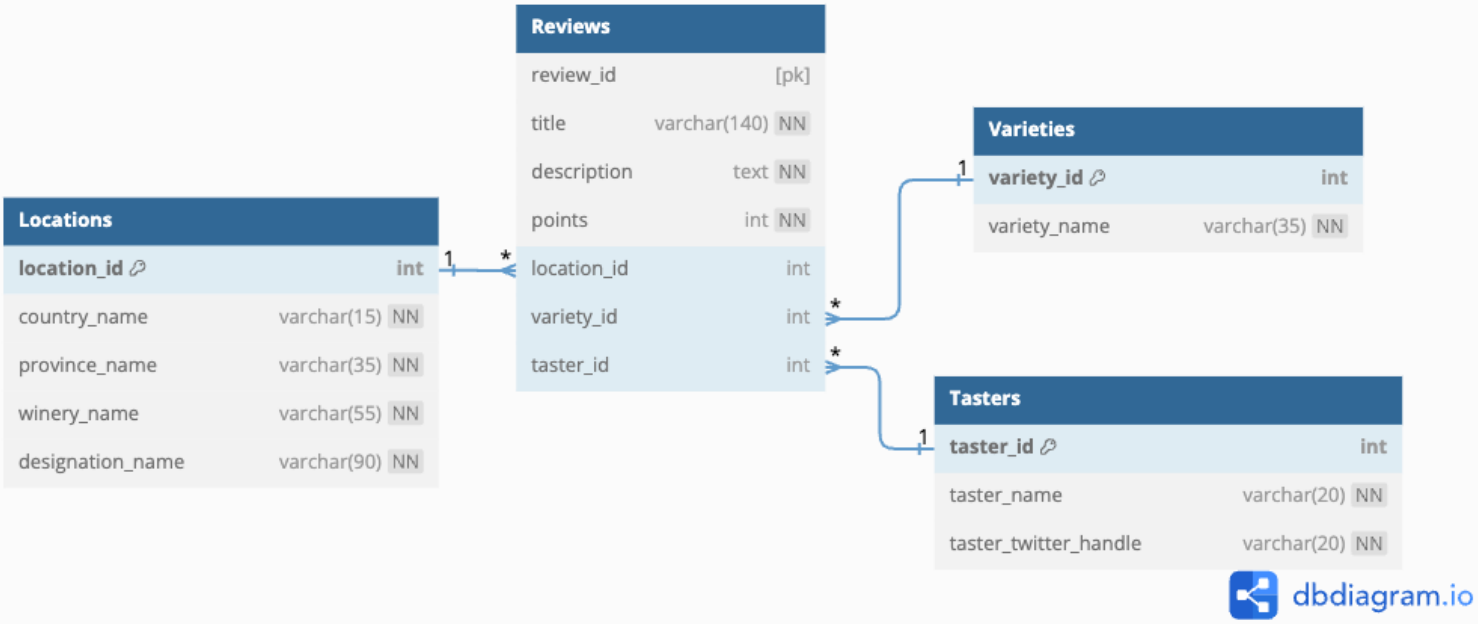
6. Statistical Analysis of Location Details:
   - How many provinces, wineries, and designations are associated with each country?
   - What is the distribution of wine-related entities across different geographical locations?

By leveraging a database application, these questions can be answered efficiently and provide valuable insights into the world of wines, facilitating informed decision-making for wine producers, aiding enthusiasts in discovering new favourites, and enabling researchers to identify trends within the wine industry.

2. Data Model

2.1 E/R model (with cardinality)





Implementing a subset of the data (Data processing)

Further inspection of the dataset showed that there were repeated rows of the same wine, of course with multiple reviews of the same wine, but also with repeated reviews from different years by the same taster and also varying prices. In order to streamline data for the purpose of this assignment I decided to:

- **Drop of price column**
  - Due to the difficulty of tagging a specific price to a wine considering they can be purchased at different prices
  - The wines in this table can only identified by review title, although they can be summarised to a unique combination of variety and designation, along with the vintage - which is unfortunately part of the title and not a separate column

- **Picking only 1 review per distinct wine (as per title) with maximum points**
  - As mentioned above, in order to streamline the data to only one review per distinct wine (as per title), I retained only the review for each respective title with maximum points.
  - After dropping price and selecting rows based on this condition, I was left with distinct rows, making it easier to normalise the data into tables as compared to when I tried to work with the unfiltered data.

- **Scale down of dataset**
  - The sample data used is of the upper quartile with regards to the points provided in the reviews, as I sliced the data for only reviews with points > 95. The range of points being from 80 - 100, this is the top quarter.
  - This is in part to accommodate to the capacity of the Coursera environment, considering even the cleaned data still had over 30 thousand rows.
  - This subset of the data has a manageable about 1500 rows and is still of interest as they detail the reviews with the maximum points, possibly representing details of the best/ highest rated wines.

Details of the data cleaning can be found in the following screenshots:

2. Data cleaning (general)

2.1 Drop region_1, region_2 and price columns

Due to large amounts of missing data

```
df1 = df_all.drop(columns=['region_1', 'region_2', 'price'])
df1.head()
```

	Unnamed: 0	country	description	designation	points	province	taster_name	taster_twitter_handle	title	variety	winery
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	Sicily & Sardinia	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	Douro	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	Oregon	Paul Gregutt	@paulgwine	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	Michigan	Alexander Peartree	NaN	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian
4	4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	Oregon	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks

2.2 Drop the rows with NA elements

```
df2 = df1.dropna()
df2.head()
```

	Unnamed: 0	country	description	designation	points	province	taster_name	taster_twitter_handle	title	variety	winery
0	0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	Sicily & Sardinia	Kerin O'Keefe	@kerinokeefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	Douro	Roger Voss	@vossroger	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
4	4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	Oregon	Paul Gregutt	@paulgwine	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks
5	5	Spain	Blackberry and raspberry aromas show a typical...	Ars In Vitro	87	Northern Spain	Michael Schachner	@wineschach	Tandem 2011 Ars In Vitro Tempranillo-Merlot (N...	Tempranillo-Merlot	Tandem
6	6	Italy	Here's a bright, informal red that opens with ...	Belsito	87	Sicily & Sardinia	Kerin O'Keefe	@kerinokeefe	Terre di Giurfo 2013 Belsito Frappato (Vittoria)	Frappato	Terre di Giurfo

2.5 Check for duplicates

(5849 duplicate entries found)

```
df_dup = df3.groupby(df3.columns.tolist(), as_index=False).size()
df_dup.loc[df_dup['size'] > 1]
```

	country	description	designation	points	province	taster_name	taster_twitter_handle	title	variety	winery	size
12	Argentina	A bit of dust and leather for openers, with sh...	Altosur	85	Mendoza Province	Michael Schachner	@wineschach	Finca Sophenia 2007 Altosur Malbec (Tupungato)	Malbec	Finca Sophenia	2
34	Argentina	A deep, developed wine with cola, licorice and...	Henry Gran Guarda No. 1	92	Mendoza Province	Michael Schachner	@wineschach	Lagarde 2007 Henry Gran Guarda No. 1 Red (Mend...	Red Blend	Lagarde	2
41	Argentina	A fairly oaky bouquet with overt wood grain, v...	Encuentro	88	Mendoza Province	Michael Schachner	@wineschach	Rutini 2011 Encuentro Cabernet Sauvignon (Mend...	Cabernet Sauvignon	Rutini	2
51	Argentina	A full, oily nose with aromas of orange blosso...	Las Compuertas	86	Mendoza Province	Michael Schachner	@wineschach	Luigi Bosca 2014 Las Compuertas Riesling (Lujá...	Riesling	Luigi Bosca	2
58	Argentina	A heady but attractive bouquet of marzipan, ba...	Single Vineyard Reserva	92	Mendoza Province	Michael Schachner	@wineschach	Lamadrid 2012 Single Vineyard Reserva Malbec (...	Malbec	Lamadrid	2
...	...	...	...	...	...	...	...	...	...	...	...
65174	Uruguay	Inky cool herbal black-fruit aromas set up a c...	100 Años Viñas Mendoza	87	Montevideo	Michael Schachner	@wineschach	Gimenez Mendez 2015 100 Años Viñas Mendoza Mal...	Malbec	Gimenez Mendez	2
65175	Uruguay	Inky purple in color, with herbal, minty, toas...	Las Brujas	87	Canelones	Michael Schachner	@wineschach	Gimenez Mendez 2015 Las Brujas Red (Canelones)	Red Blend	Gimenez Mendez	2
65179	Uruguay	No kidding that this was aged in oak (Criado e...	Criado en Roble	87	Canelones	Michael Schachner	@wineschach	Montes Toscanini 2015 Criado en Roble Tannat (...	Tannat	Montes Toscanini	2
65199	Uruguay	Rusty in color, this maturing Tannat-led blend...	Preludio Barrel Select Lot N°92	87	Juanico	Michael Schachner	@wineschach	Familia Deicas 2009 Preludio Barrel Select Lot...	Red Blend	Familia Deicas	2
65205	Uruguay	This blend of Pinot Noir and Chardonnay is ora...	Reserve Collection Blanc de Noir	84	Canelones	Michael Schachner	@wineschach	Marichal 2015 Reserve Collection Blanc de Noir...	Rosé	Marichal	2

5849 rows × 11 columns

2.6 Drop duplicates

```
df4 = df3.drop_duplicates()
print(df3.shape)
print(df4.shape)
print(df_dup.shape)
print(df3.shape[0] - df4.shape[0], "entries removed")
```

```
(71066, 10)
(65217, 10)
(65217, 11)
5849 entries removed
```

Checking for repeated entries

e.g. same title with different descriptions/ points

- As seen in below (df_repeat_title), there are repeat entries with the same title.
- When comparing entries with the same title and description (df_repeat_title_des) - no repetitions are found, indicating that similar titles have varying descriptions.
- When comparing entries with the same title and points (df_points), repetitions are found.
- When grouping the titles (df_diff_points), repetitions are also found .
  - Therefore, there are similar entries with the same title and points, same title and varying points and varying descriptions
  - i.e. we have something like {title1, description1, points1}, {title1, description2, points1}, {title1, description3, points3}, and so on.

Are there entries with repeated titles? - Yes

```
df_repeat_title = df_entries[['title']].value_counts().reset_index(name = 'repeated')
df_repeat_title.loc[df_repeat_title['repeated'] > 1]
```

	title	repeated
0	Segura Viudas NV Aria Estate Extra Dry Sparkli...	7
1	Segura Viudas NV Extra Dry Sparkling (Cava)	7
2	Ruinart NV Brut Rosé (Champagne)	6
3	Bailly-Lapierre NV Brut (Crémant de Bourgogne)	6
4	Jacquart NV Brut Mosaïque (Champagne)	5
...	...	...
536	Chanson Père et Fils 2013 Bastion Premier Cru ...	2
537	Ca' del Bosco NV Cuvée Prestige Sparkling (Fra...	2
538	Jean-Baptiste Adam NV Les Natures Sparkling (C...	2
539	Masottina NV Brut (Conegliano Valdobbiadene P...	2
540	Taittinger NV Folies de la Marquetterie Brut ...	2

541 rows × 2 columns

Are there entries with the same title and description? - No

```
df_repeat_title_des = df_entries[['title', 'description']].value_counts().reset_index(name = 'repeated')
df_repeat_title_des.loc[df_repeat_title_des['repeated'] > 1]
```

	title	description	repeated

Are there entries with the same title and points? - Yes

```
df_points = df4[['title', 'points']].value_counts().reset_index(name = 'same review points')
df_points
```

	title	points	same review points
0	Segura Viudas NV Extra Dry Sparkling (Cava)	86	4
1	Bailly-Lapierre NV Brut (Crémant de Bourgogne)	90	4
2	Boizel NV Ultime Extra Brut (Champagne)	90	3
3	Henriet-Bazin NV Blanc de Noirs Grand Cru Brut...	91	3
4	Bodegas Dios Baco S.L. NV Oxford 1.970 Pedro X...	91	3
...	...	...	...
65032	Domaine de la Sanglière 2015 Juliette Rosé (Me...	85	1
65033	Domaine de la Sanglière 2015 La Riviera Rosé (...	89	1
65034	Domaine de la Sanglière 2015 La Sanglière Rosé...	87	1
65035	Domaine de la Sanglière 2016 Breezette Rosé (C...	87	1
65036	Štoka 2011 Izbrani Teran (Kras)	90	1

65037 rows × 3 columns

Are there entries with the same title and different points? - Yes

```
df_diff_points = df_points[['title']].value_counts().reset_index(name = 'var_points')
df_diff_points
```

	title	var_points
0	Segura Viudas NV Aria Estate Extra Dry Sparkli...	5
1	Ruinart NV Brut Rosé (Champagne)	4
2	Segura Viudas NV Reserva Heredad Sparkling (Cava)	4
3	Henri Abele NV Brut (Champagne)	4
4	Taittinger NV Nocturne Sec (Champagne)	4
...	...	...
64527	Domaine des Croix 2011 La Vigne au Saint (Cor...	1
64528	Domaine des Croix 2014 Grèves (Corton)	1
64529	Domaine des Croix 2014 La Vigne au Saint (Cor...	1
64530	Domaine des Croix 2014 Les Bressandes Premier ...	1
64531	Štoka 2011 Izbrani Teran (Kras)	1

Simplify data

- Retain only the first instance of the entries with the maximum points

This removes repetition of descriptions and points, leaving only unique title values with their respective descriptions and points

```
df_max = df4.loc[df4.groupby('title')['points'].idxmax()]
print(df4.shape)
print(df_max.shape)
print(df4.shape[0] - df_max.shape[0], "entries removed")
df_max.head()
```

```
(65217, 10)
(64532, 10)
685 entries removed
```

	country	description	designation	points	province	taster_name	taster_twitter_handle	title	variety	winery
63807	Spain	The previous two years we did not find this wi...	Rosé	82	Catalonia	Michael Schachner	@wineschach	1+1=3 2008 Rosé Cabernet Sauvignon (Penedès)	Cabernet Sauvignon	1+1=3
55163	Spain	Spiced apple and toast aromas are clean and di...	Brut	87	Catalonia	Michael Schachner	@wineschach	1+1=3 NV Brut Sparkling (Cava)	Sparkling Blend	1+1=3
33657	Spain	Clean, fresh apple aromas and a minerally, cit...	Cygnus Brut Nature Reserva Made With Organic G...	89	Catalonia	Michael Schachner	@wineschach	1+1=3 NV Cygnus Brut Nature Reserva Made With ...	Sparkling Blend	1+1=3
20319	Spain	A dusty, yeasty nose is simplistic but friendl...	Rosé	86	Catalonia	Michael Schachner	@wineschach	1+1=3 NV Rosé Sparkling (Cava)	Sparkling Blend	1+1=3
122898	US	Juicy and fresh, this deeply colored wine offe...	All Profits to Charity	89	California	Jim Gordon	@gordone_cellars	100 Percent Wine 2012 All Profits to Charity R...	Red Blend	100 Percent Wine

No repeat reviews (with the same title)

```
df_reviews = df_max[['title']].value_counts().reset_index(name='reviews')
df_reviews
```

	title	reviews
0	1+1=3 2008 Rosé Cabernet Sauvignon (Penedès)	1
1	Naggiar 2012 Il Nonno Estate Reserve Red (Sier...	1
2	Naggiar 2011 Estate Grown Mourvèdre (Sierra Fo...	1
3	Naggiar 2011 Estate Malbec (Sierra Foothills)	1
4	Naggiar 2011 Estate Petite Sirah (Sierra Footh...	1
...	...	...
64527	Domaine de la Sanglière 2014 Juliette Rosé (Me...	1
64528	Domaine de la Sanglière 2015 Breezette Rosé (C...	1
64529	Domaine de la Sanglière 2015 Cuvée Spéciale Ro...	1

Condense data

To only reviews with points in the upper quarter (max - (max-min)/4)

```
[ ] import numpy as np
    max_points = max(df6['points'])
    min_points = min(df6['points'])
    q = max_points - (max_points-min_points)/4
    q
```

```
95.0
```

```
[ ] df_final = df6.loc[df6['points'] >= 95]
    df_final
```

	country	description	designation	points	province	taster_name	taster_twitter_handle	title	variety	winery
99316	Italy	Here's a delicious red that opens with appeali...	Vigna Piaggia	95	Tuscany	Kerin O'Keefe	@kerinokeefe	Abbadia Ardenga 2012 Vigna Piaggia (Brunello ...	Sangiovese	Abbadia Ardenga
27608	Italy	You'll need to swirl the glass a few times to ...	Praepositus	95	Northeastern Italy	Kerin O'Keefe	@kerinokeefe	Abbazia di Novacella 2015 Praepositus Kerner (...	Kerner	Abbazia di Novacella
33841	US	Sourced from old-vine Bacchus and Weinbau frui...	Reserve	97	Washington	Paul Gregutt	@paulgwine	Abeja 2007 Reserve Cabernet Sauvignon (Columbi...	Cabernet Sauvignon	Abeja
99317	US	This splendid, resonant, beautifully detailed ...	Gran Moraine Vineyard	95	Oregon	Paul Gregutt	@paulgwine	Aberrant Cellars 2014 Gran Moraine Vineyard Pi...	Pinot Noir	Aberrant Cellars
83367	US	The best of a great flight of single-vineyard ...	Nicholas Vineyard	95	Oregon	Paul Gregutt	@paulgwine	Adelsheim 2009 Nicholas Vineyard Pinot Noir (C...	Pinot Noir	Adelsheim
...	...	...	...	...	...	...	...	...	...	...
126238	US	Pure Cabernet from some of Washington's oldest...	Old Vines	95	Washington	Paul Gregutt	@paulgwine	Woodward Canyon 2007 Old Vines Cabernet Sauvig...	Cabernet Sauvignon	Woodward Canyon
122571	US	Here the high alcohol does not obscure the lay...	Artist Series #18	95	Washington	Paul Gregutt	@paulgwine	Woodward Canyon 2009 Artist Series #18 Caberne...	Cabernet Sauvignon	Woodward Canyon

---

2.2 Database tables, fields and evaluation against normal forms

Tables and fields

1. Varieties {
   - **variety_id** - int [primary key]
     - Unique identifier for each distinct variety
   - **variety_name** - varchar(35) [not null]

    - Name of variety
}

Evaluation of Varieties Table:

1NF	2NF	3NF
It has a primary key (`variety_id`) and all columns contain atomic values.	There is only one candidate key (`variety_id`), and no non-prime attribute is dependent on only a portion of the candidate key.	There are no transitive dependencies, and every non-prime attribute is dependent on the primary key.

2. Locations {
- **location_id** - int [primary key]
    - Unique identifier for each distinct location where the wine is from (specific to designation)
- **country_name** - varchar(15)
    - Country within which province is located
- **province_name** - varchar(35)
    - Province within which winery is located
- **winery_name -** varchar(55)
    - Winery within which the designation is located
- **designation_name** - varchar(90)
    - Name of designation
}

Evaluation of the Locations Table:

1NF	2NF	3NF
All columns contain atomic values.	Single candidate key (`location_id`), and no non-prime attribute is dependent on only a portion of the candidate key.	No transitive dependencies, and every non-prime attribute is dependent on the primary key.

3. Tasters {
- **taster_id** - int [primary key]
    - Unique identifier for each distinct Taster (author of the reviews)
- **taster_name** - varchar(20)
    - Name of Taster (author of review)
- **taster_twitter_handle** - varchar(20)
    - Twitter handle of Taster
}

Evaluation of the Tasters Table:

1NF	2NF	3NF

All columns contain atomic values.	Single candidate key (`taster_id`), and no non-prime attribute is dependent on only a portion of the candidate key.	No transitive dependencies, and every non-prime attribute is dependent on the primary key.

4. Reviews {
- **review_id** - int [primary key]
  - Unique identifier for each distinct review: specific to title
- **title -** varchar(140)
  - Title of the review
- **description** - text
  - Content of the review
- **points** - int
  - Points awarded for the wine by the Taster (author of the review)
- **location_id** - int [foreign key references > Locations.location_id]
  - Identifier that links to the Locations table on where the wine in the review is from
- **variety_id** - int [foreign key references > Varieties.variety_id]
  - Identifier that links to the Varieties table on the variety of the wine in the review
- **taster_id** - int [foreign key references > Tasters.taster_id]
  - Identifier that links to the Tasters table on the taster who wrote the review

}

Evaluation of the Reviews Table:

1NF	2NF	3NF
All columns contain atomic values.	Single candidate key (`taster_id`), and no non-prime attribute is dependent on only a portion of the candidate key.	No transitive dependencies, and every non-prime attribute is dependent on the primary key.

Overall evaluation and justification:

- At the outset, I explored a more granular approach, contemplating distinct tables for countries, provinces, designations, wineries, and wines. However, I encountered challenges during data ingestion due to the intricate structure of the denormalised table. When joining the 'Reviews' table, multiple entries for identical values surfaced, stemming from various reviews of the same wine by different tasters, published in different years, each with different points, vintages, and prices. After several iterations, I found that maintaining 'Reviews' as the key table, with auxiliary tables encapsulating its crucial components, worked most effectively.
- If I were to refine this approach, I would include a 'Wines' table to consolidate distinguishing wine characteristics, being variety, designation and vintage. This would enable multiple reviews or prices to be linked to a single unique 'wine_id.'
- However, considering the data scale-down that excluded prices and repetitive reviews, the chosen level of normalisation (3NF) seemed most pragmatic and aligned with the dataset's characteristics.

3. Creation of database

3.1 Build the database structure in MySQL

Create Data Tables

- The tables which have no dependencies shall be created first, followed by those with less dependencies and finally the one that has the most dependencies.
  - From the relational schema, start from tables at the edges and then move inwards.
- To drop a table, do the reverse. Drop the table which has the highest dependencies, followed by less and finally those that have no dependencies.
  - Start from the inner most and then move outwards towards the edges.

Step 1: Iteratively add a `CREATE TABLE SQL` scripts in the following order: `Locations, Tasters, Varieties and Reviews.`

```
////////////////////////////////////////////////////////////////////
USE wine_reviews;

DROP TABLE IF EXISTS Reviews;
DROP TABLE IF EXISTS Varieties;
DROP TABLE IF EXISTS Tasters;
DROP TABLE IF EXISTS Locations;

-- Create Locations table
CREATE TABLE Locations (
  location_id int AUTO_INCREMENT PRIMARY KEY,
  country_name varchar(15) NOT NULL,
  province_name varchar(35) NOT NULL,
  winery_name varchar(55) NOT NULL,
  designation_name varchar(90) NOT NULL
);

-- Create Tasters table
CREATE TABLE Tasters (
  taster_id int AUTO_INCREMENT PRIMARY KEY,
  taster_name varchar(20) NOT NULL,
  taster_twitter_handle varchar(20) NOT NULL
);

-- Create Varieties table
CREATE TABLE Varieties (
  variety_id int AUTO_INCREMENT PRIMARY KEY,
  variety_name varchar(35) NOT NULL
);

-- Create Reviews table
CREATE TABLE Reviews (
  review_id int AUTO_INCREMENT PRIMARY KEY,
  title varchar(140) NOT NULL,
  description text NOT NULL,
  points int NOT NULL,
  location_id int,
  variety_id int,
  taster_id int,
  FOREIGN KEY (location_id) REFERENCES Locations (location_id),
  FOREIGN KEY (variety_id) REFERENCES Varieties (variety_id),
  FOREIGN KEY (taster_id) REFERENCES Tasters (taster_id)
);
////////////////////////////////////////////////////////////////////
```

Step 2: Run `create-tables.sql` script. This will create tables specified in the SQL script.

```
!mysql -t < /home/coder/project/wine_reviews/scripts/create-tables.sql
```

Step 3: Verify if the tables are created correctly.

Step 4: Iterate back to **Step 1** until all the tables are created.

3.2 Enter instance data

Load Denormalised Data (required for Data Ingestion)

I loaded the denormalised data into the `wine_reviews` database, which will then be used for data ingestion to the tables created previously.

Step 1: Create a SQL script which contains:

- Create `denormalised` table for temporarily storing our denormalised data.
- Load the denormalised data into the `denormalised` table.

```
////////////////////////////////////////////////////////////////////
USE wine_reviews;
DROP TABLE IF EXISTS denormalised;

-- Create denormalised table
CREATE TABLE denormalised (
    country VARCHAR(15),
    description TEXT,
    designation VARCHAR(90),
    points INT,
    province VARCHAR(35),
    taster_name VARCHAR(20),
    taster_twitter_handle VARCHAR(20),
    title VARCHAR(140),
    variety VARCHAR(35),
    winery VARCHAR(55)
);

LOAD DATA INFILE '/home/coder/project/wine_reviews/data/winemag-records.csv'
INTO TABLE denormalised
FIELDS TERMINATED BY ','
ENCLOSED BY '"'
LINES TERMINATED BY '\n'
IGNORE 1 LINES
;
////////////////////////////////////////////////////////////////////
```

Step 2: Run `load-dnorm_data.sql` script. This will create tables specified in the SQL script, load the CSV data into the database and then pivot it into a tall table.

```
!mysql -t < /home/coder/project/wine_reviews/scripts/load-dnorm-data.sql
```

Step 3: Verify if the data is correctly loaded and then pivoted as intended.

Ingest Data into the Normalised Tables

Data was ingested into tables one at a time, iteratively while also verifying it after the ingestion.

- The tables which have no dependencies are ingested first, followed by those with less dependencies and finally the one that has the most dependencies.
- Deletion of tables will be in reverse order - Starting with the table with highest dependencies, followed by less and finally those that have no dependencies.

Step 1: Before writing the data insertion script, I tried out the script to see if it is creating the data as required for the tables. Data was inserted in the following order of tables: `Varieties`, `Tasters`, `Locations and Reviews`.

Step 2: Create a SQL script using `INSERT INTO` which contains data insertion to all the normalised tables.

```
////////////////////////////////////////////////////////////////////
USE wine_reviews;
DELETE FROM Reviews;
DELETE FROM Tasters;
DELETE FROM Locations;
DELETE FROM Varieties;

-- Insert data into Varieties table
INSERT INTO Varieties (variety_name)
SELECT DISTINCT variety
FROM denormalised;

-- Insert data into Locations table
INSERT INTO Locations (country_name, province_name, winery_name,
designation_name)
SELECT DISTINCT country, province, winery, designation
FROM denormalised;

-- Insert data into Tasters table
INSERT INTO Tasters (taster_name, taster_twitter_handle)
SELECT DISTINCT taster_name, taster_twitter_handle
FROM denormalised;

-- Insert data into Reviews table
INSERT INTO Reviews (title, description, points, location_id, variety_id,
taster_id)
SELECT
  title,
  description,
  points,
  l.location_id,
  v.variety_id,
  t.taster_id
FROM
  denormalised d
  JOIN Locations l ON d.country = l.country_name
    AND d.province = l.province_name
    AND d.winery = l.winery_name
    AND d.designation = l.designation_name
  JOIN Varieties v ON d.variety = v.variety_name
  JOIN Tasters t ON d.taster_name = t.taster_name;
////////////////////////////////////////////////////////////////////
```

Step 3: Run `ingest-data.sql` script. This will create tables specified in the SQL script, load the CSV data into the database and then pivot it into a tall table.

```
!mysql -t < /home/coder/project/wine_reviews/scripts/ingest-data.sql
```

Step 4: Verify if the data is correctly ingested into the table as intended.

Step 5: Iterate back to **Step 1** until data is ingested to all the tables.

3.3 Reflect on how well the database reflects the data

- **Elements that work well:**
  - The database works well in accurately reflecting reviews with the highest points, providing comprehensive access to all data related to specific reviews within this dataset.
  - The structured layout enables easy retrieval of information on individual reviews, contributing to a user-friendly experience.

- **Elements to be improved upon:**
  - The reliance on the structure centred around reviews creates a dependency, particularly on the title of each review. As a result, access to distinct rows is contingent on the review title. The streamlined data, containing only unique titles with respective descriptions and points, restricts user access to individual components of the dataset, such as wines individually.

  - The encapsulation of data related to the origin of wines in a single table (Locations) limits the variability in data access. This constraint essentially ties back to the reviews, impeding more diverse ways of exploring the data.

  - The absence of certain components, such as price or different tasters' descriptions/ ratings of the same wine, imposes limitations on accessing comprehensive knowledge about wines. This lack of data facets hinders a holistic understanding of wine characteristics and taster perspectives.

3.4 List SQL commands

***Question 1**: What are the reviews with the maximum points?*

- Identify the Highest Rated Wines:
  - Retrieve details of wines with the highest points, including location, variety, taster, and description.

```
////////////////////////////////////////////////////////////////////
SELECT
  r.title, r.description, r.points, l.country_name, l.province_name,
l.winery_name, l.designation_name, v.variety_name, t.taster_name,
t.taster_twitter_handle
FROM Reviews r
JOIN Locations l ON r.location_id = l.location_id
JOIN Varieties v ON r.variety_id = v.variety_id
JOIN Tasters t ON r.taster_id = t.taster_id
WHERE r.points = (SELECT MAX(points) FROM Reviews);
////////////////////////////////////////////////////////////////////
```

***Question 2**: What are the most popular and highly rated varieties?*

- List Unique Varieties with Review Counts and Highest Points:
  - Identify the most commonly occurring unique wine varieties, their review counts, and the highest points per variety.

```
////////////////////////////////////////////////////////////////////////
SELECT
  v.variety_name,
  COUNT(*) AS variety_count,
  MAX(r.points) AS highest_points
FROM Varieties v
JOIN Reviews r ON v.variety_id = r.variety_id
GROUP BY v.variety_name
HAVING variety_count > 10
ORDER BY variety_count DESC;
////////////////////////////////////////////////////////////////////////
```

***Question 3**: Where do most highly rated wines come from?*

- Find the Highest Rated Wine by Country:
  - Identify the highest-rated wine for each country, including details like province, title, and points.

```
////////////////////////////////////////////////////////////////////////
WITH RankedReviews AS (
  SELECT
    l.country_name,
    l.province_name,
    r.title AS highest_rated_wine,
    r.points,
    ROW_NUMBER() OVER (PARTITION BY l.country_name ORDER BY r.points DESC) AS
rnk
  FROM
    Locations l
    JOIN Reviews r ON l.location_id = r.location_id
)
SELECT
  country_name AS Country,
  province_name AS Province,
  highest_rated_wine,
  points
FROM RankedReviews
WHERE rnk = 1
ORDER BY points DESC;
////////////////////////////////////////////////////////////////////////
```

***Question 4**: Which wineries produce the most highly rated wines?*

- List the Top 20 Most Popular Wineries:
  - Identify the top 20 most popular wineries based on review count and average points, including location details.

```
////////////////////////////////////////////////////////////////////////
SELECT
  l.winery_name AS Winery,
  CONCAT(l.country_name, ", ", l.province_name) AS Location,
  COUNT(r.review_id) AS review_count,
  AVG(r.points) AS avg_points
FROM Locations l
JOIN Reviews r ON l.location_id = r.location_id
GROUP BY l.winery_name, location
ORDER BY review_count DESC, avg_points DESC
LIMIT 20;
////////////////////////////////////////////////////////////////////////
```

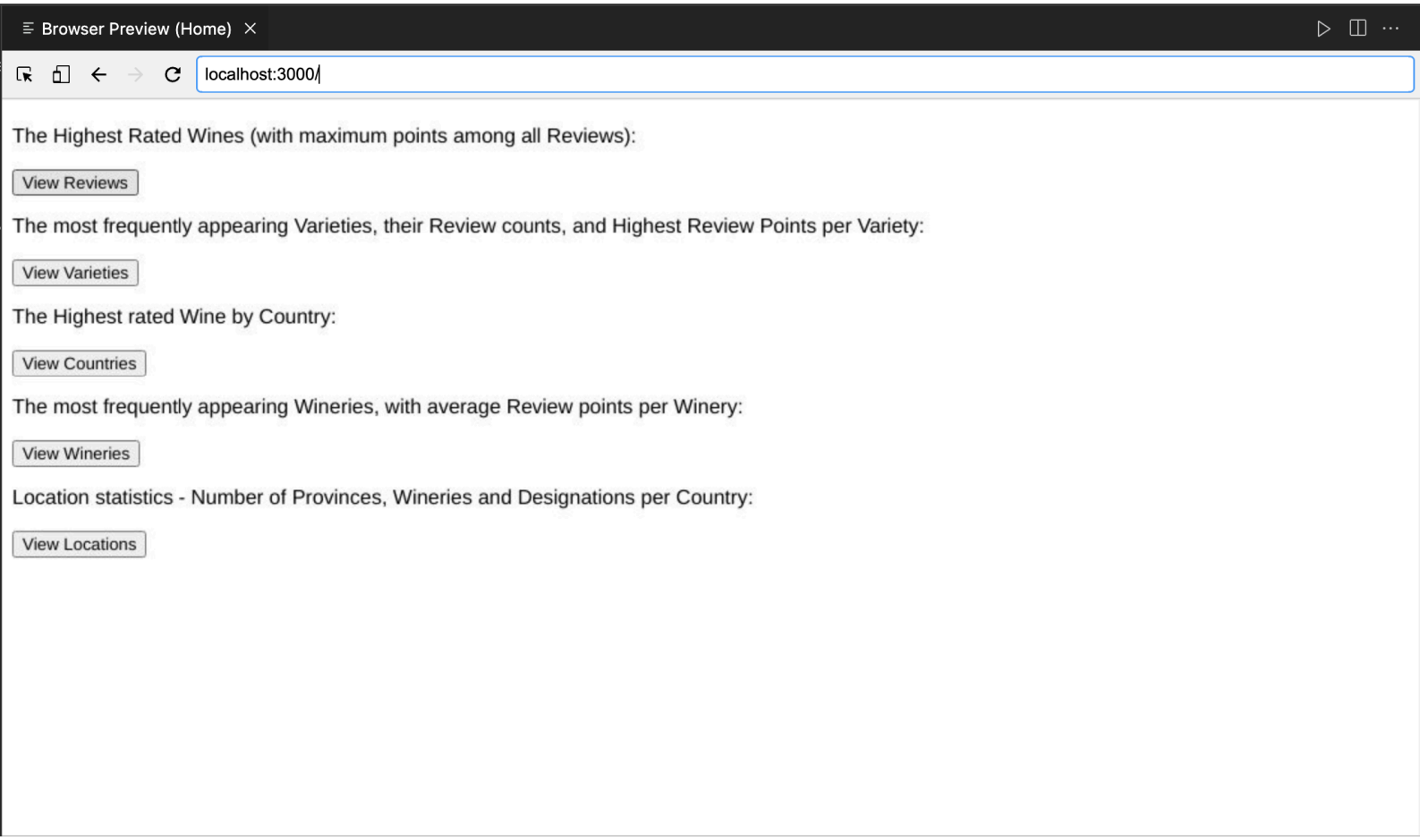
***Question 5**: What are the locations producing the most wine (of points > 95)?*

- Location Statistics:
  - Retrieve statistics for each country, including the number of provinces, wineries, and designations.

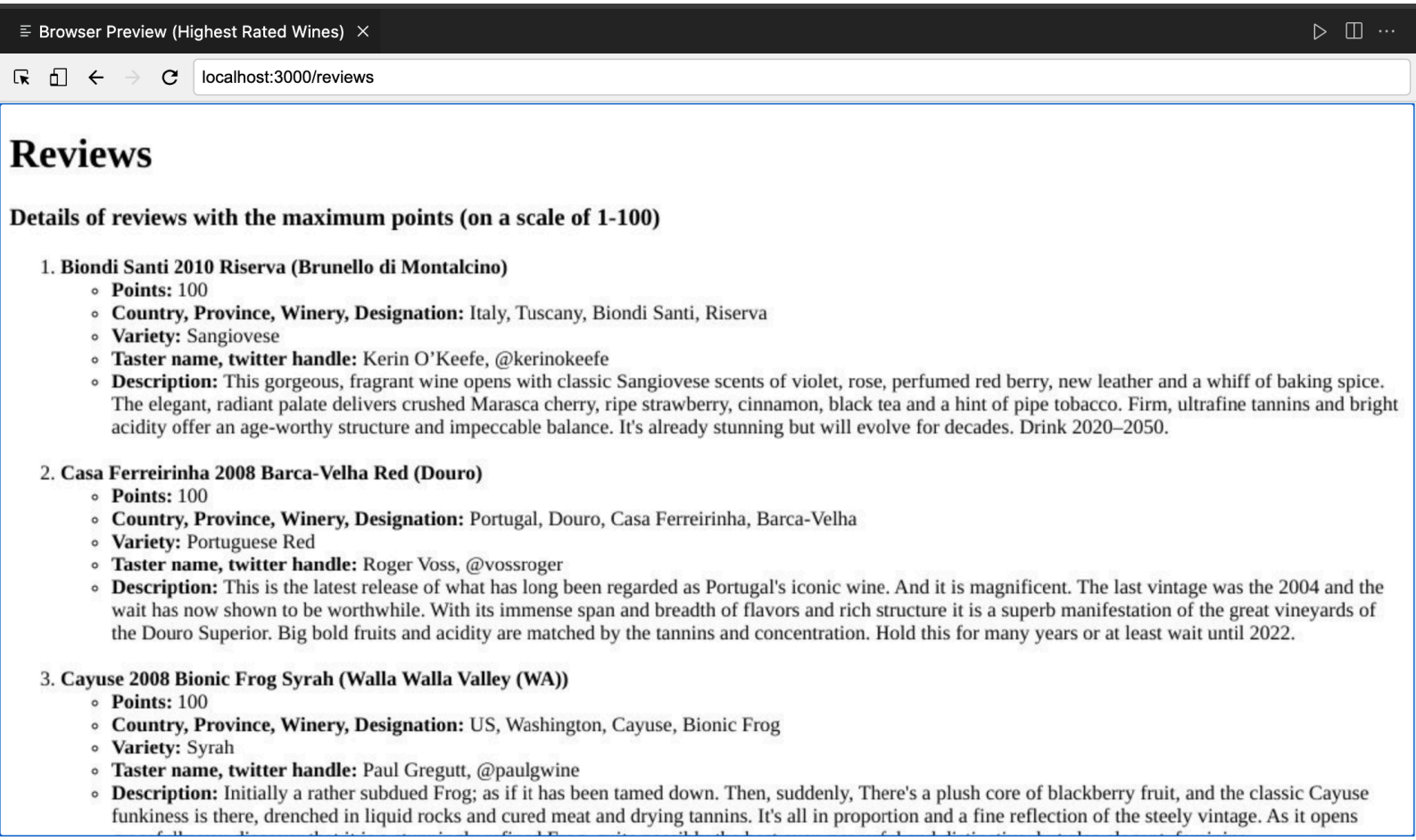
```
////////////////////////////////////////////////////////////////////////
SELECT
  l.country_name AS Country,
  COUNT(DISTINCT l.province_name) AS num_provinces,
  COUNT(DISTINCT l.winery_name) AS num_wineries,
  COUNT(DISTINCT l.designation_name) AS num_designations
FROM Locations l
LEFT JOIN Reviews r ON l.location_id = r.location_id
GROUP BY l.country_name
ORDER BY num_designations DESC;
////////////////////////////////////////////////////////////////////////
```

4. Web application

Main Page:



Click on “View Reviews”:



Reviews

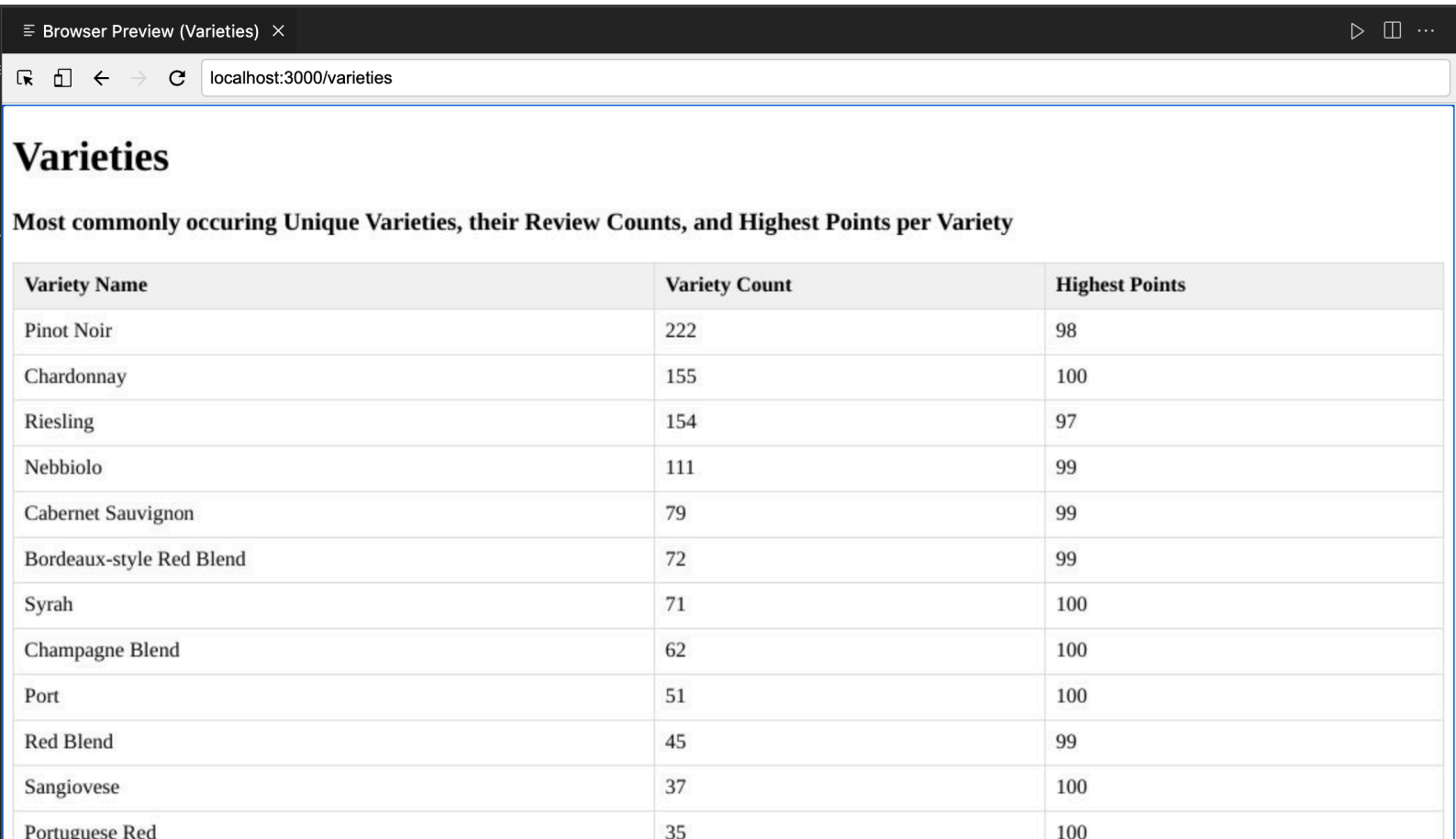
Details of reviews with the maximum points (on a scale of 1-100)

1. **Biondi Santi 2010 Riserva (Brunello di Montalcino)**
   - **Points:** 100
   - **Country, Province, Winery, Designation:** Italy, Tuscany, Biondi Santi, Riserva
   - **Variety:** Sangiovese
   - **Taster name, twitter handle:** Kerin O’Keefe, @kerinokeefe
   - **Description:** This gorgeous, fragrant wine opens with classic Sangiovese scents of violet, rose, perfumed red berry, new leather and a whiff of baking spice. The elegant, radiant palate delivers crushed Marasca cherry, ripe strawberry, cinnamon, black tea and a hint of pipe tobacco. Firm, ultrafine tannins and bright acidity offer an age-worthy structure and impeccable balance. It's already stunning but will evolve for decades. Drink 2020–2050.

2. **Casa Ferreirinha 2008 Barca-Velha Red (Douro)**
   - **Points:** 100
   - **Country, Province, Winery, Designation:** Portugal, Douro, Casa Ferreirinha, Barca-Velha
   - **Variety:** Portuguese Red
   - **Taster name, twitter handle:** Roger Voss, @vossroger
   - **Description:** This is the latest release of what has long been regarded as Portugal's iconic wine. And it is magnificent. The last vintage was the 2004 and the wait has now shown to be worthwhile. With its immense span and breadth of flavors and rich structure it is a superb manifestation of the great vineyards of the Douro Superior. Big bold fruits and acidity are matched by the tannins and concentration. Hold this for many years or at least wait until 2022.

3. **Cayuse 2008 Bionic Frog Syrah (Walla Walla Valley (WA))**
   - **Points:** 100
   - **Country, Province, Winery, Designation:** US, Washington, Cayuse, Bionic Frog
   - **Variety:** Syrah
   - **Taster name, twitter handle:** Paul Gregutt, @paulgwine
   - **Description:** Initially a rather subdued Frog; as if it has been tamed down. Then, suddenly, There's a plush core of blackberry fruit, and the classic Cayuse funkiness is there, drenched in liquid rocks and cured meat and drying tannins. It's all in proportion and a fine reflection of the steely vintage. As it opens

Click on ”View Varieties”:

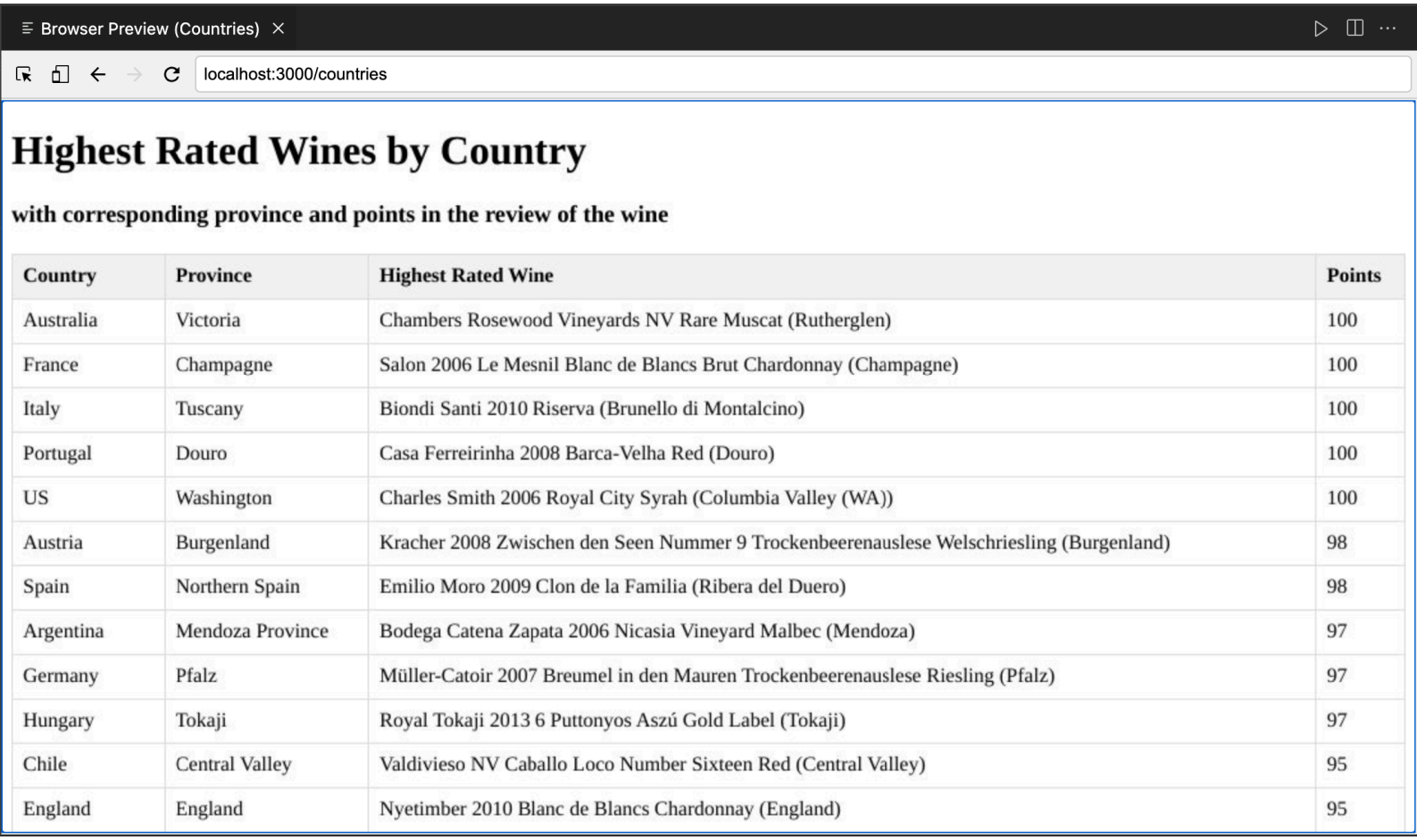


Varieties

Most commonly occuring Unique Varieties, their Review Counts, and Highest Points per Variety

Variety Name	Variety Count	Highest Points
Pinot Noir	222	98
Chardonnay	155	100
Riesling	154	97
Nebbiolo	111	99
Cabernet Sauvignon	79	99
Bordeaux-style Red Blend	72	99
Syrah	71	100
Champagne Blend	62	100
Port	51	100
Red Blend	45	99
Sangiovese	37	100
Portuguese Red	35	100

Click on ”View Countries”:

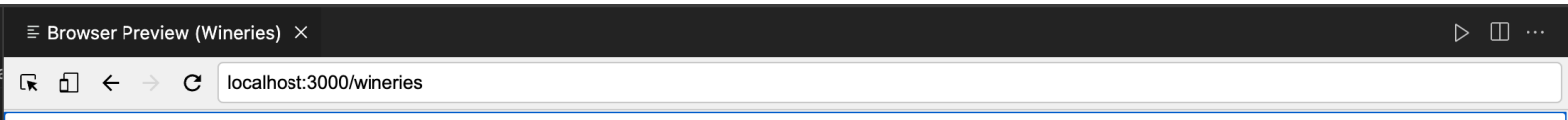


Highest Rated Wines by Country

with corresponding province and points in the review of the wine

Country	Province	Highest Rated Wine	Points
Australia	Victoria	Chambers Rosewood Vineyards NV Rare Muscat (Rutherglen)	100
France	Champagne	Salon 2006 Le Mesnil Blanc de Blancs Brut Chardonnay (Champagne)	100
Italy	Tuscany	Biondi Santi 2010 Riserva (Brunello di Montalcino)	100
Portugal	Douro	Casa Ferreirinha 2008 Barca-Velha Red (Douro)	100
US	Washington	Charles Smith 2006 Royal City Syrah (Columbia Valley (WA))	100
Austria	Burgenland	Kracher 2008 Zwischen den Seen Nummer 9 Trockenbeerenauslese Welschriesling (Burgenland)	98
Spain	Northern Spain	Emilio Moro 2009 Clon de la Familia (Ribera del Duero)	98
Argentina	Mendoza Province	Bodega Catena Zapata 2006 Nicasia Vineyard Malbec (Mendoza)	97
Germany	Pfalz	Müller-Catoir 2007 Breumel in den Mauren Trockenbeerenauslese Riesling (Pfalz)	97
Hungary	Tokaji	Royal Tokaji 2013 6 Puttonyos Aszú Gold Label (Tokaji)	97
Chile	Central Valley	Valdivieso NV Caballo Loco Number Sixteen Red (Central Valley)	95
England	England	Nyetimber 2010 Blanc de Blancs Chardonnay (England)	95

Click on “View Wineries”:

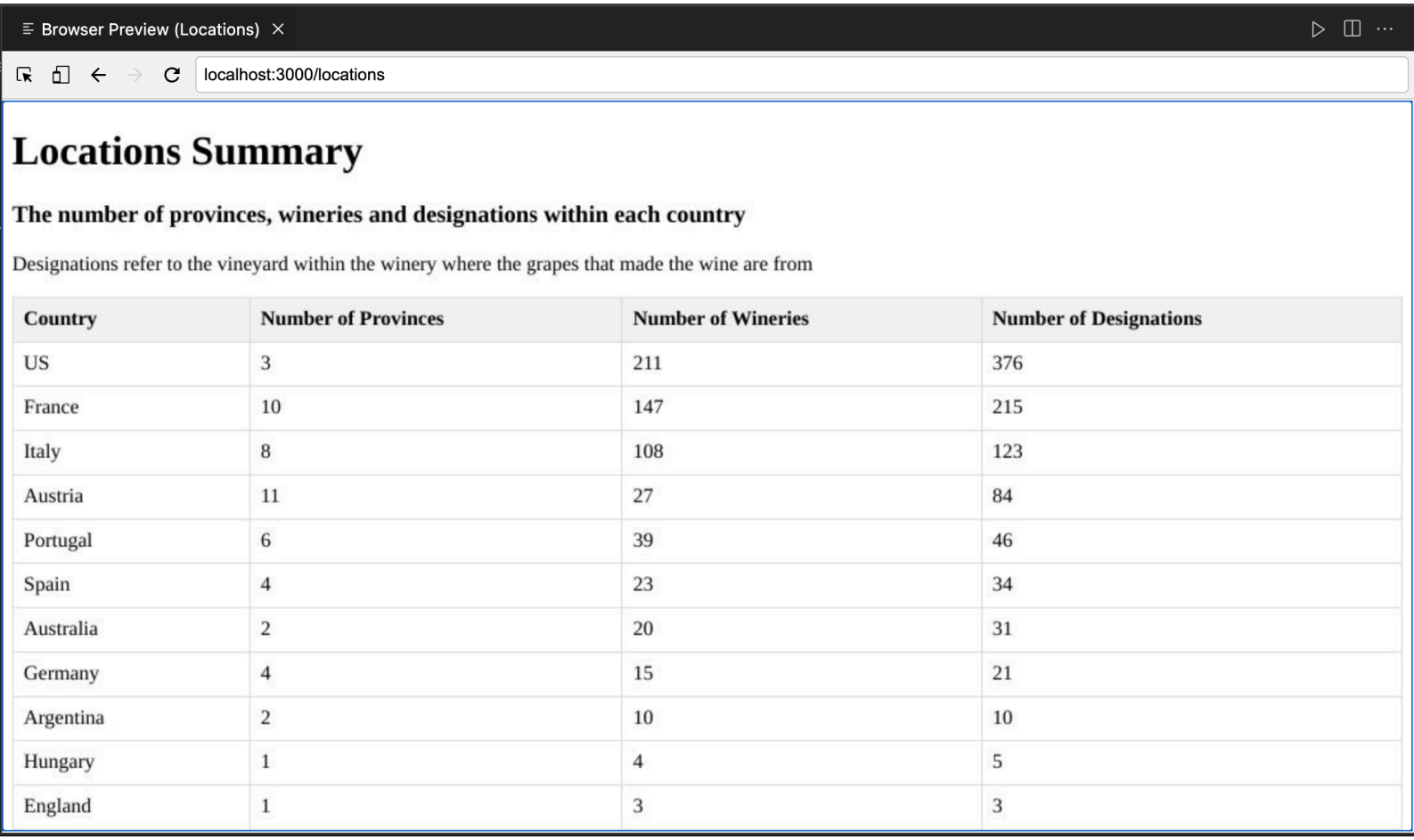


Top 20 Wineries

Most commonly occurring Wineries, their respective locations and average points of reviews

Winery	Location	Review Count	Average Points
Williams Selyem	US,California	28	95.8571
Domaine Zind-Humbrecht	France,Alsace	20	95.5
Cayuse	US,Washington	15	96.1333
Kracher	Austria,Burgenland	14	96.1429
Louis Roederer	France,Champagne	14	95.7857
Emmerich Knoll	Austria,Wachau	14	95.4286
F X Pichler	Austria,Wachau	14	95.1429
Cayuse	US,Oregon	12	96.75
Wayfarer	US,California	12	96.1667
Bründlmayer	Austria,Kamptal	12	95.3333
Lynmar	US,California	12	95.3333
Franz Hirtzberger	Austria,Wachau	11	95.1818

Click on “View Locations”:



Locations Summary

The number of provinces, wineries and designations within each country

Designations refer to the vineyard within the winery where the grapes that made the wine are from

Country	Number of Provinces	Number of Wineries	Number of Designations
US	3	211	376
France	10	147	215
Italy	8	108	123
Austria	11	27	84
Portugal	6	39	46
Spain	4	23	34
Australia	2	20	31
Germany	4	15	21
Argentina	2	10	10
Hungary	1	4	5
England	1	3	3