Rosé All Day

Wine Recommender, Interactive Dashboard, & Taster Trading Cards

CS5010 Final Project, August 2020

Nikki Aaron, Bev Dobrenz, Amanda West, Joseph Wysocki

Introduction

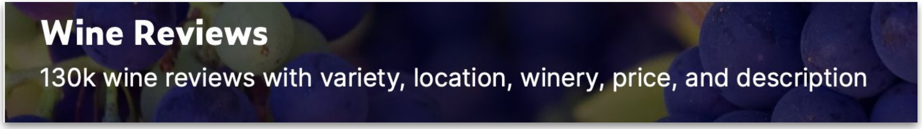
Goal: Create a system that takes input from the user regarding their taste preferences and returns a list of recommended wines.

Dataset: WineEnthusiast Review Data from Kaggle

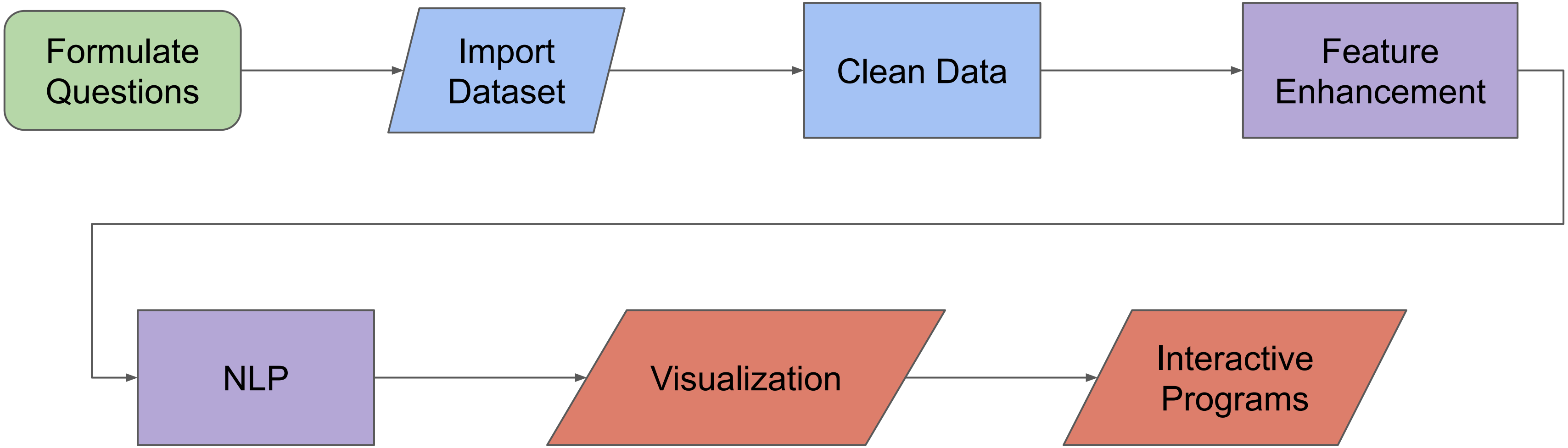
- 130k wine reviews with variables such as variety, location, winery, price, and description

Results: Users can:

- Receive wine recommendations using a ***Terminal Input Wine Recommender***
- Visualize and interact with updated wine review data using a ***Tableau Dashboard***
- Obtain information about favorite wine tasters from ***Taster Review Profiles***



Project Flow Chart

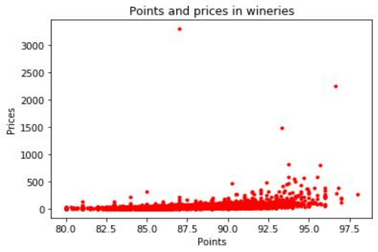


Data Cleaning/Initial Investigation

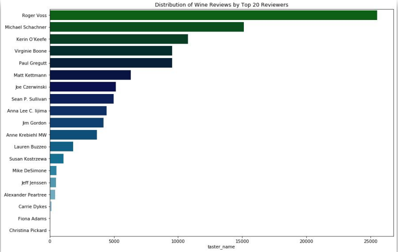
- Wine Review Data from Kaggle
  - Winemag-data-130k-v2.csv with 130,000 rows of wine reviews
  - Variables include:
    - Country, Description, Points, Price, Province, Region_1, Region_2, Taster_Name, Taster_Twitter_Handle, Title, Variety, Winery
- Created summary plots for initial investigation
  - Excluded missing (NaN) values: 96,400 rows to perform analysis

```
[5 rows x 7 columns]
country           63
points             0
price           8996
province          63
taster_name    26244
variety            1
winery             0
dtype: int64
```

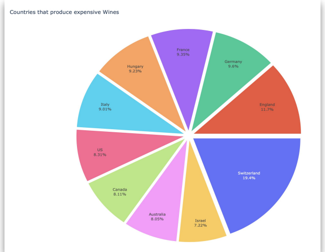
Are there any missing values?



Is there a relationship between price and points?



What taster writes the most reviews?

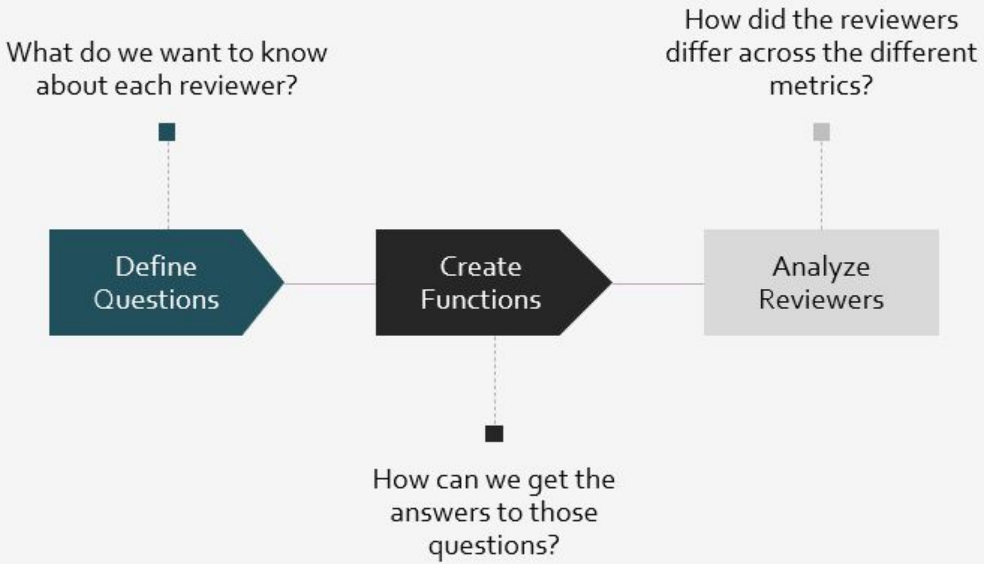


What Country has the most expensive wine?

Wine Reviewer Profiles

Process

Our process for analyzing the reviewers was quite straightforward. We defined questions that we wanted to explore, created functions to answer those questions, and then analyzed those metrics across each reviewer



Output

We applied our functions to each reviewer and stored the results in a dictionary. Next, we created a new function that returned the key metrics for a given reviewer. We also compiled this information in a new data frame. The results are shown to the right.

Results of get_info() function

```
get_info('Lauren Buzzeo')
```

```
Twitter handle:  @laurbuzz
Number of Reviews:  1711
Average Score:  87.57
Highest Score:  95
Lowest Score:  81
Perfect Score Percent:  0.0
Diversity Percent:  10.0
Word Usage:  Wordy
Status:  Experienced
Scoring Style:  Generous
Review Count by Country:  {'South Africa': 905, 'Israel': 198, 'France': 586,
'US': 19, 'Canada': 1, 'Portugal': 1, 'Spain': 1}
Average Price of Wine Reviewed:  24.49
```

Reviewer Analysis Data Frame

	Twitter handle	Number of Reviews	Average Score	Highest Score	Lowest Score	Perfect Score Percent	Diversity Percent	Word Usage	Status	Scoring Style	Review Count by Country	Average Price of Wine Reviewed
Paul Gregutt	@paulgwine	9494	89.09	100	80	0.0211	7	Average	Power Reviewer	Generous	{'US': 9268, 'France': 34, 'Canada': 184, 'Ita...	33.65
Michael Schachner	@wineschach	14941	86.91	98	80	0	16	Average	Power Reviewer	Tough	{'Argentina': 3752, 'Chile': 4280, 'Spain': 65...	25.23
Roger Voss	@vossroger	20167	88.61	100	80	0.0496	9	Curt	Power Reviewer	Generous	{'Austria': 831, 'Portugal': 4842, 'France': 1...	38.65
Lauren Buzzeo	@laurbuzz	1711	87.57	95	81	0	10	Wordy	Experienced	Generous	{'South Africa': 905, 'Israel': 198, 'France':...	24.49
Joe Czerwinski	@JoeCz	5006	88.54	100	80	0.02	16	Average	Power Reviewer	Generous	{'New Zealand': 1270, 'US': 102, 'France': 112...	35.2

Trading Cards

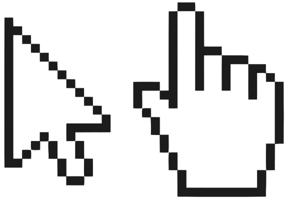
Last, for some extra fun, we created a fake brand of trading cards – known as "Somm Stars" – and featured each reviewer. The cards organize the information gathered through our work in Python and present it in a friendly format.

An Example

Roger Voss – the King of Reviewers

Wine Variety Analysis

- Incorporated Text Mining & NLP Components (Python NLTK library)
  - **Text mining** is the process of deriving high quality information from text.
    - Goal is to turn the text into data for analysis via application of NLP.
  - **NLP** - **N**atural **L**anguage **P**rocessing
    - *"Natural language processing (NLP) is a branch of artificial intelligence that helps computers understand, interpret and manipulate human language."* - *SAS*



Text mining ⇒ information

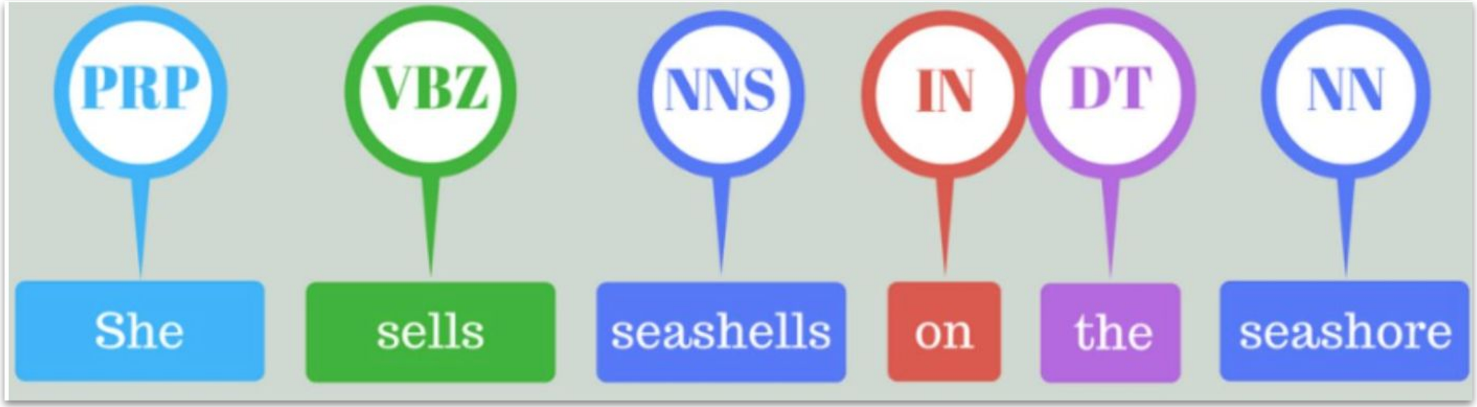
NLP ⇒ knowledge

Wine Variety Analysis

- *...Continued*
  - **Tokenization** is the process of breaking strings into tokens of small structures or units
    - *“She sells seashells by the seashore*”
    - ⇒ [“She”, “sells”, “seashells”, “by”, “the”, “seashore”]
  - **Stop words** are common “connector” words. These words do not provide any meaning and are best removed.
    - *examples*: “the”, “a”, “at”, “for”, “above”, “on”, “is”,“all”

Wine Variety Analysis

- *...Continued*
    - **Part-of-speech** understands and matches parts of speech to each tokenized word in a text (adjectives, adverbs, nouns, verbs, etc).



Source: Dhilip Subramanian

Word Dictionary Creation for Descriptions

- Applying NLP to Wine Tasting Reviews

  a. Identified the **column of interest**. ⇒ “The happy jumping dog”

  b. Used **tokenization** to break up paragraph into a list of words. ⇒ [“The”, “happy”, “jumping”, “dog”]

  c. Filtered out **stop words** from each list of text. ⇒ [“happy”, “jumping”, “dog”]

  d. Appended the **part-of-speech** to each word. ⇒ [[“happy”, ADJ], [“jumping”, V], [“dog”, N]]

  e. Filtered parts of speech for only nouns and adjectives. ⇒ [[“happy”, ADJ], [“dog”, N]]

* ADJ = Adjective, N = Noun, V = Verb

Wine Variety Analysis

- Creating Wine Flavor Word Dictionary
    a. “Fruity”, “Earthy”, “Floral”, “Bitter” examples of how one might describe a wine.
    b. Created word lists for each category ⇒ [Fruity: “apple”, “pear”, “juicy”]
    c. If descriptor list had enough matches, considered wine apart of that category.
        - Wine could live in more than one category (i.e. “sweet” and “earthy”)

Flavor Profile Types		
Sweetness	**Flavor**	**Body**
Sweet	Savory	Light-Bodied
Dry	Fruity	Full-Bodied
	Earthy	
	Bitter	
	Floral	

Wine Type Analysis

1. Identified the **columns of interest** --
   title, designator, grape variety, description

2. Matched title and designator to **list of color words** in English, French, Italian, Spanish and German

3. Matched grape variety to list of **grape variety colors**

4. Matched description to **list of color words**

5. Count **weighted matches** and assign color with the **highest count**

6. Add “Blend” and “Sparkling” designations

Wine Type Analysis

Red => [noir, rotwein, rosso, rouge]

White => [blanc, bianco, bianca, weißwein, weis]

Rosé => [rosato, rosado, rosat, roséwein, roséfine]

Sparkling => [champagne, bubbles, bubbly, brut, bruto, sekt, schaumwein, effervescent, spumante, scintillante]

Red grapes => [cabernet, malbec]

White grapes => [sauvignon blanc, chardonnay, riesling]

Ambiguous grapes => [pinot noir (blanc), chardonnay (noir), pinot gris (mendoza)]

Sample Recommendation

Cavas Hill NV 1887 Rosado Sparkling

(Sparkling Rosé Blend from Spain)

$13 | 82 Points | Sweet, Light-Bodied

Red in color, with berry and apple aromas, this is a sweet blend, with a light body and nose tingling effervescence.

Wine Recommender Terminal App

- Collect user preferences
  a. Wine Type
  b. Flavor Profile
  c. Taster Score
  d. Price Range
  e. Origin Country

Pandas Filtering

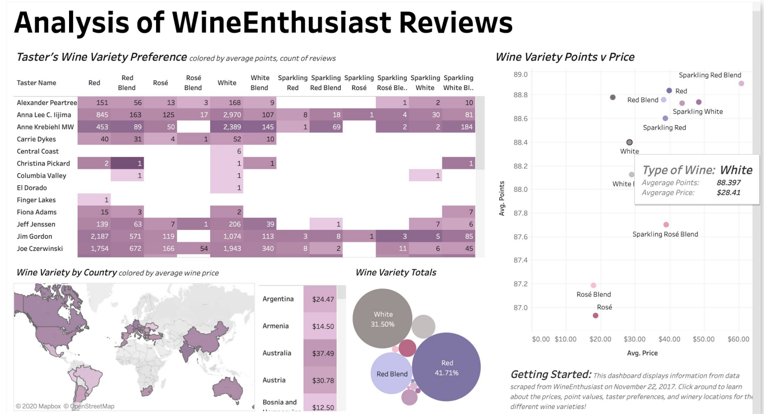
- Display 10 Results
  a. Wine Title
  b. Wine Type
  c. Flavor Profile
  d. Taster Score
  e. Price
  f. Origin Country
  g. Description

Tableau Dashboard

Tableau is a **data visualization** tool that makes it easy to create **interactive visual analytics** in the form of dashboards for both **technical and non-technical end users.**

```python
print('Number of country list in data:',df['country'].nunique())
plt.figure(figsize=(14,10))
cnt = df['country'].value_counts().to_frame()[0:20]
#plt.xscale('log')
sns.barplot(x= cnt['country'], y =cnt.index, data=cnt, palette='ocean',orient='h')
plt.title('Distribution of Wine Reviews by Top 20 Countrie');
#%%

#look at the tasters — are they biased?

#average mark by a taster
tasters = df[ ["taster_name", "points"] ].groupby(by="taster_name").mean()[::-1]
print(tasters)

#standard deviation by taster
print( df[ ["taster_name", "points"] ].groupby(by="taster_name").std()[::-1] )

cnt = df['taster_name'].value_counts().to_frame()[0:20]
print(cnt)
#%%
```

```
Nikki's Top 10 Wine Recommendations
----------------------

Marco Abella 2012 El Perer Carignan:
This single-vineyard Carignan is living proof of how impre
ssive the variety can be, especially when hailing from a t
op year like 2012. Schisty aromas of blackberry, chocolate
 and mocha are smooth and alluring. This feels focused, pu
re and juicy, while dense blackberry, mocha and baking spi
ce flavors tail off with exotic notes of Malabar pepper, t
oast and the finest black coffee. Drink through 2023.

Garcia Figuero 2012 Tinus:
Extremely ripe aromas of prune, blackberry, cassis and cin
namon announce a jammy loaded palate that weighs nothing s
hort of a ton. The flavor profile is a black-fruit bonanza
 accented by graphite and blackened toast. Drink this mons
ter wine from a dry, hot vintage from 2019-2034.

Vega Sicilia NV Unico Reserva Especial:
This blend of the 1996, '98 and '02 vintages is mature and
 browning in hue. The nose is superripe, with prune, brand
ied cherry, tobacco and molasses notes. A soft, creamy pal
ate holds vanilla, tobacco, baking spice, prune and raisin
```



Python

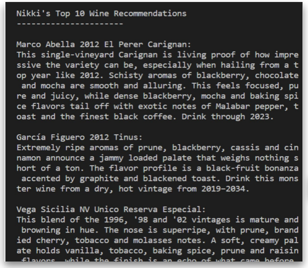
Putting it All Together: Sample Use Case

User is traveling to Spain and wants to learn about Spanish wines before their trip!

```
print('Number of country list in data:',df['country'].nunique())
plt.figure(figsize=(14,10))
cnt = df['country'].value_counts().to_frame()[0:20]
#plt.xscale('log')
sns.barplot(x= cnt['country'], y =cnt.index, data=cnt, palette='ocean',orient='h')
plt.title('Distribution of Wine Reviews by Top 20 Countrie');

#%%

#look at the tasters -- are they biased?

#average mark by a taster
tasters = df[ ["taster_name", "points"] ].groupby(by="taster_name").mean()[::-1]
print(tasters)

#standard deviation by taster
print( df[ ["taster_name", "points"] ].groupby(by="taster_name").std()[::-1] )


cnt = df['taster_name'].value_counts().to_frame()[0:20]
print(cnt)

#%%
```

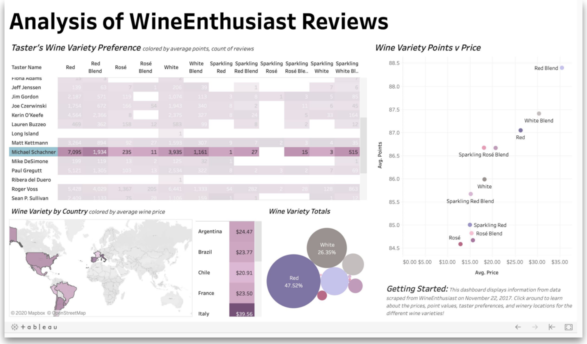
Data Scientists perform initial analysis, data transformations, and create a Wine Recommender

```
Is there a particular country of origin you are interested
 in? Enter the number for one of these options:
0: No Preference
1: US (37505 wines)
2: Argentina (3752 wines)
3: Austria (2790 wines)
4: Portugal (4869 wines)
5: South Africa (1218 wines)
6: New Zealand (1270 wines)
7: Israel (484 wines)
8: Chile (4303 wines)
9: France (17518 wines)
10: Spain (6508 wines)
11: Italy (10121 wines)
12: Australia (2007 wines)
13: Germany (2093 wines)
14: Hungary (144 wines)
15: England (69 wines)
16: Canada (253 wines)
```



```
Nikki's Top 10 Wine Recommendations
---------------------

Marco Abella 2012 El Perer Carignan:
This single-vineyard Carignan is living proof of how impre
ssive the variety can be, especially when hailing from a t
op year like 2012. Schisty aromas of blackberry, chocolate
 and mocha are smooth and alluring. This feels focused, pu
re and juicy, while dense blackberry, mocha and baking spi
ce flavors tail off with exotic notes of Malabar pepper, t
oast and the finest black coffee. Drink through 2023.

García Figuero 2012 Tinus:
Extremely ripe aromas of prune, blackberry, cassis and cin
namon announce a jammy loaded palate that weighs nothing s
hort of a ton. The flavor profile is a black-fruit bonanza
 accented by graphite and blackened toast. Drink this mons
ter wine from a dry, hot vintage from 2019-2034.

Vega Sicilia NV Unico Reserva Especial:
This blend of the 1996, '98 and '02 vintages is mature and
 browning in hue. The nose is superripe, with prune, brand
ied cherry, tobacco and molasses notes. A soft, creamy pal
ate holds vanilla, tobacco, baking spice, prune and raisin
```

User specifies wine preferences and receives a Top Ten List of wine recommendations



User uses Tableau Dashboard to dive deeper into their Top Ten List of wine recommendations

User combines information from Wine Recommender and Tableau Dashboard to select a Taster Trading Card!

Wine Recommender: Use Case

User specifies a preference for Spanish wines and receives a Top Ten list of wine recommendations for wines from Spain.

- *Where can the user go to learn more about Spanish wines?*

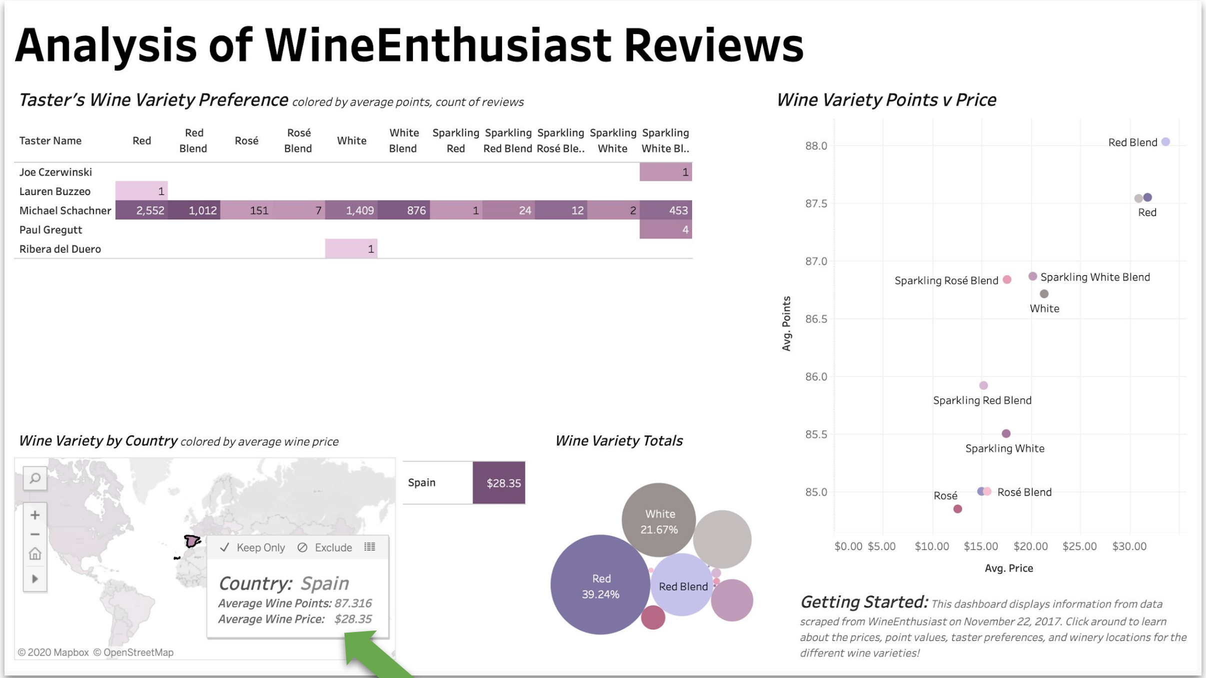
```
Is there a particular country of origin you are interested
 in? Enter the number for one of these options:
0: No Preference
1: US (37505 wines)
2: Argentina (3752 wines)
3: Austria (2790 wines)
4: Portugal (4869 wines)
5: South Africa (1218 wines)
6: New Zealand (1270 wines)
7: Israel (484 wines)
8: Chile (4303 wines)
9: France (17518 wines)
10: Spain (6508 wines)
11: Italy (10121 wines)
12: Australia (2007 wines)
13: Germany (2093 wines)
14: Hungary (144 wines)
15: England (69 wines)
16: Canada (253 wines)
```

Original user input, user selects Spain

```
Nikki's Top 10 Wine Recommendations
----------------------

Marco Abella 2012 El Perer Carignan:
This single-vineyard Carignan is living proof of how impre
ssive the variety can be, especially when hailing from a t
op year like 2012. Schisty aromas of blackberry, chocolate
 and mocha are smooth and alluring. This feels focused, pu
re and juicy, while dense blackberry, mocha and baking spi
ce flavors tail off with exotic notes of Malabar pepper, t
oast and the finest black coffee. Drink through 2023.

García Figuero 2012 Tinus:
Extremely ripe aromas of prune, blackberry, cassis and cin
namon announce a jammy loaded palate that weighs nothing s
hort of a ton. The flavor profile is a black-fruit bonanza
 accented by graphite and blackened toast. Drink this mons
ter wine from a dry, hot vintage from 2019-2034.

Vega Sicilia NV Unico Reserva Especial:
This blend of the 1996, '98 and '02 vintages is mature and
 browning in hue. The nose is superripe, with prune, brand
ied cherry, tobacco and molasses notes. A soft, creamy pal
ate holds vanilla, tobacco, baking spice, prune and raisin
 flavors, while the finish is an echo of what came before
```

Output, user receives Spanish wine recommendations

Tableau Dashboard: Use Case



Original user input, user selects Spain

User Input:

User selects Spain from the Map to display information about wine from Spain.

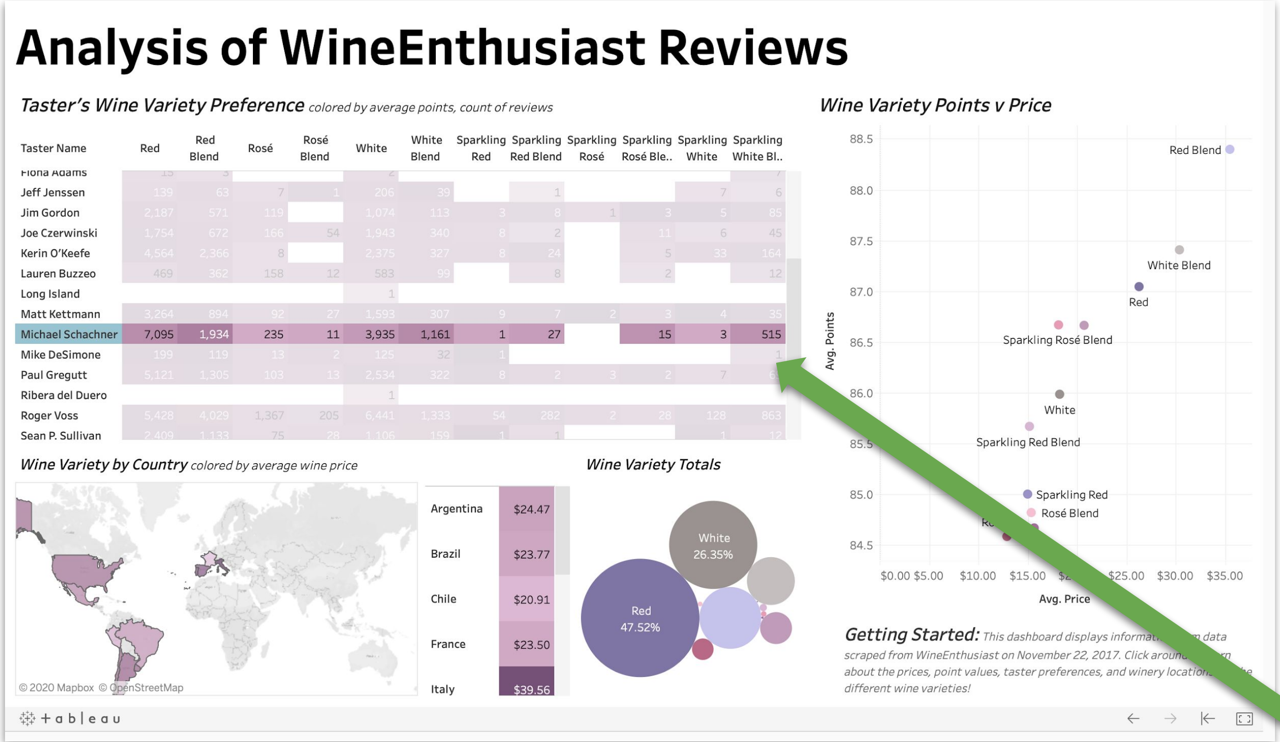
User Insights:

Spain has 9 varieties, with Red, White, and Red Blend accounting for over half of Spain's total wine.

The most expensive wines from Spain are Red Blends. Red Blends also have, on average, the most points.

Five tasters have reviewed wine from Spain, with Michael Schachner having the most reviews -- ***does Michael try wines from other countries?***

Tableau Dashboard: Use Case Continued



Original user input, user selects Michael Schachner

User Input:

User selects Michael Schachner: does Michael try wines from countries other than Spain?

User Insights:

Michael Schachner tries all different varieties of wine from countries all over the world.

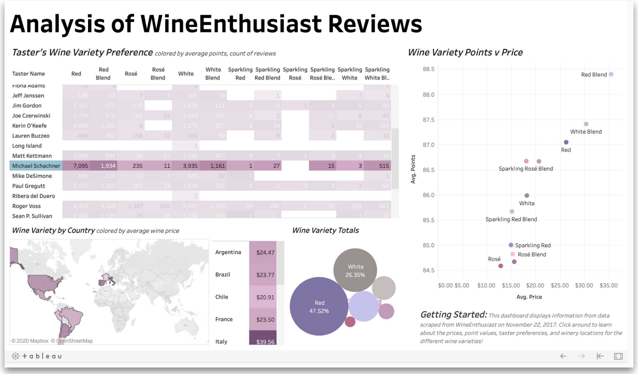
Michael Schachner gives the most points to Red Blends and mostly reviews Red and White Wines.

Where can the user go to learn more about Michael Schachner?

Taster Profile: Use Case

After filtering the Tableau Dashboard to show information about Michael Schachner, a user can learn even more about Michael by getting a copy of his taster's trading card!

...



Conclusion

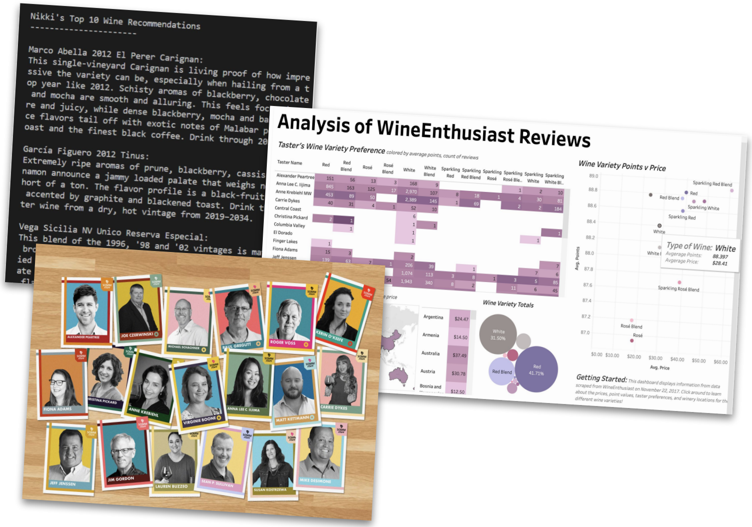
Goal: Create a system that takes input from the user regarding their taste preferences and returns a list of recommended wines.

Results:

- Terminal Input Wine Recommender
- WineEnthusiast Tableau Dashboard
- Taster Profiles

Future Steps:

- Web scrape updated wine reviews from Wine Enthusiast



Thank you!