APMA E4990 Project: Wine Master

Explore the wine world catered to your taste

Xiaojing Dong, Keran Li, Zhen Li, Zihan Yi

Columbia University in the City of New York
March 22, 2018

Project Goal

- Have you ever tasted some good wines by chance, or insisted on some particular wines for years, but don’t have another opportunity to try other wines just because you don’t know it?

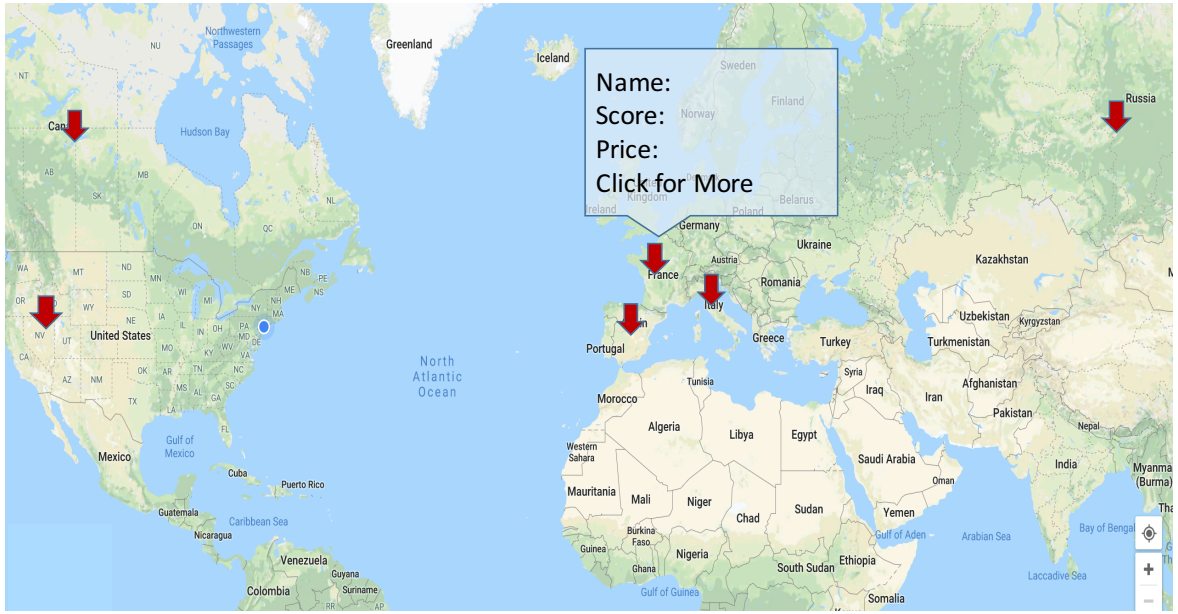
- Actually, you are very close to being a *Wine Master* with our upcoming webapp.

Project Goal

Upload a wine label
or type the key information

Provide its basic information

Recommendation similar wines for your choice



Data

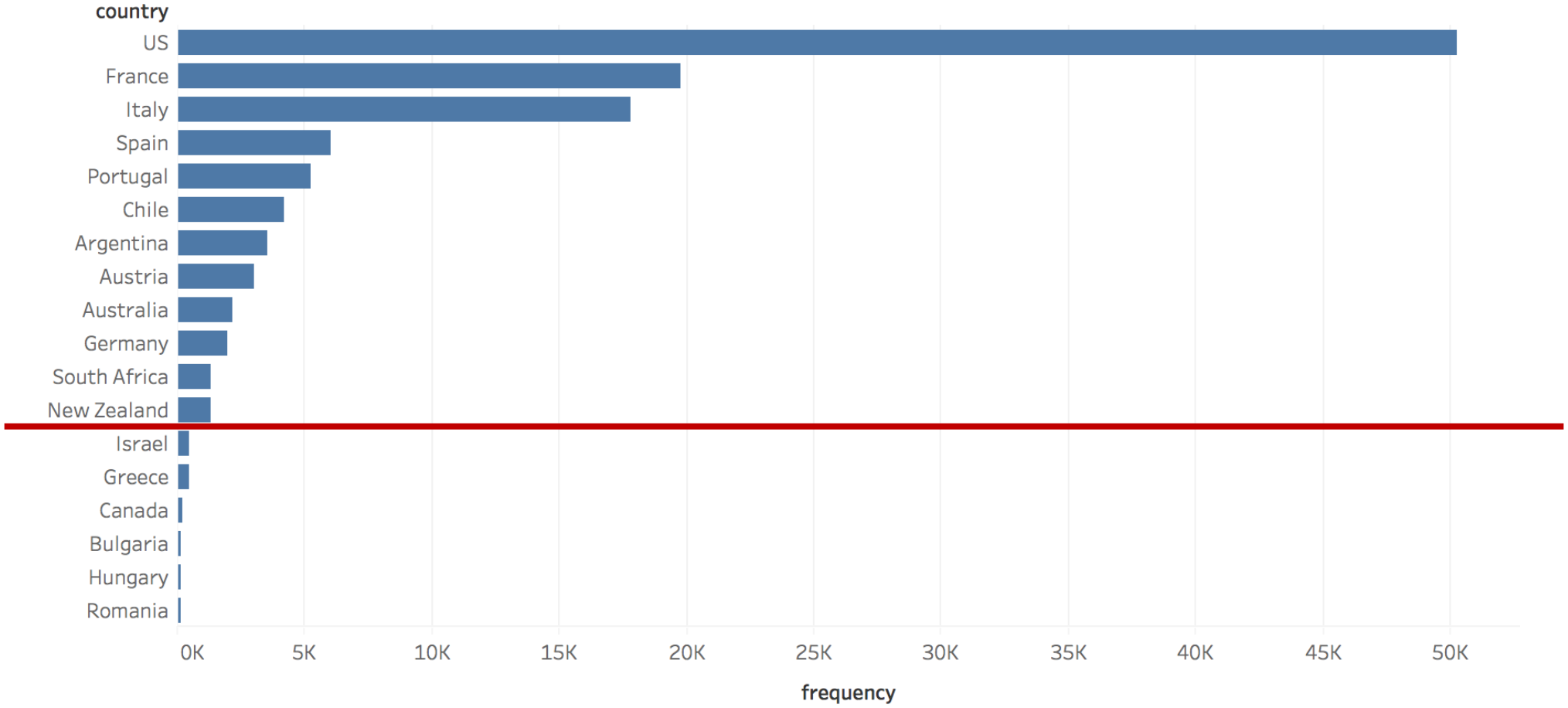
	country	description	designation	points	price	province	region_1	region_2	taster_name	title	variety	winery
0	Italy	Aromas include tropical fruit, broom, brimston...	Vulkà Bianco	87	NaN	Sicily & Sardinia	Etna	NaN	Kerin O'Keefe	Nicosia 2013 Vulkà Bianco (Etna)	White Blend	Nicosia
1	Portugal	This is ripe and fruity, a wine that is smooth...	Avidagos	87	15.0	Douro	NaN	NaN	Roger Voss	Quinta dos Avidagos 2011 Avidagos Red (Douro)	Portuguese Red	Quinta dos Avidagos
2	US	Tart and snappy, the flavors of lime flesh and...	NaN	87	14.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	Rainstorm 2013 Pinot Gris (Willamette Valley)	Pinot Gris	Rainstorm
3	US	Pineapple rind, lemon pith and orange blossom ...	Reserve Late Harvest	87	13.0	Michigan	Lake Michigan Shore	NaN	Alexander Peartree	St. Julian 2013 Reserve Late Harvest Riesling ...	Riesling	St. Julian
4	US	Much like the regular bottling from 2012, this...	Vintner's Reserve Wild Child Block	87	65.0	Oregon	Willamette Valley	Willamette Valley	Paul Gregutt	Sweet Cheeks 2012 Vintner's Reserve Wild Child...	Pinot Noir	Sweet Cheeks

Unique name

Wine Review data from Kaggle: https://www.kaggle.com/zynicide/wine-reviews

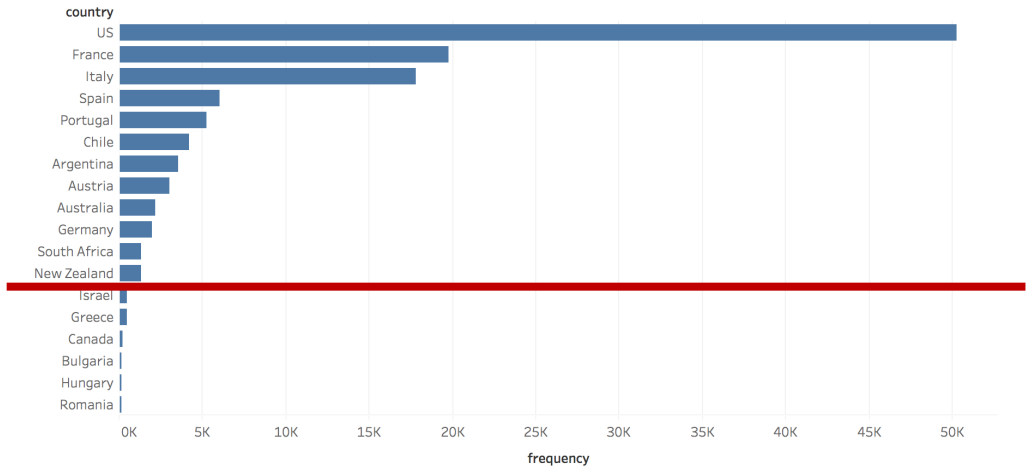
Data
Date Cleansing

1. Remove countries with frequency less than 1000



Data
Date Cleansing

1. Remove countries with frequency less than 1000
   - 12 countries remaining
   - 129,971 rows to 127,404 rows



2. Remove duplicate rows
3. Extract column *year* from column *title*

……

Method

1. Convert wine label image into machine-encoded text

- Technique: Optical Character Recognition (OCR)
- Python packages: pytesseract, tesserocr
  - Language involved:

Language	Country
English	US, Australia, New Zealand
French	France
German	Germany, Austria
Italian	Italy
Spanish	Spain, Argentina, Chile
Portuguese	Portugal
Afrikaans	South Africa

Method

2. Match the wine label to our data

- Extract useful information
- Correct the OCR errors

	country	description	designation	points	price	province	region_1	region_2	taster_name	title	variety	winery	year
125413	France	Firm tannins and great freshness, with a touch...	Carruades de Lafite	92	NaN	Bordeaux	Pauillac	NaN	Roger Voss	Château Lafite Rothschild 2008 Carruades de La...	Bordeaux-style Red Blend	Château Lafite Rothschild	2008

Method

3. Recommend similar wines

- Technique: Content-based Recommendation Engine
- Python packages: sklearn, NLTK

1 Apply TFIDF (term frequency–inverse document frequency) to vectorize wine *description*

Method

3. Recommend similar wines

- Technique: Content-based Recommendation Engine
- Python packages: sklearn, NLTK

1. Apply TFIDF (term frequency–inverse document frequency) to vectorize wine *description*
2. Select words with appropriate importance range (e.g. from 0.05 to 0.95)
3. Calculate the cosine similarity among all wines: $\cos(\theta) = \frac{V_1 \cdot V_2}{\|V_1\| \|V_2\|} = \frac{\sum_{i=1}^{n} V_{i1} V_{i2}}{\sqrt{\sum_{i=1}^{n} V_{i1}^2} \sqrt{\sum_{i=1}^{n} V_{i2}^2}}$
4. Select the top N similar wines from our dataset

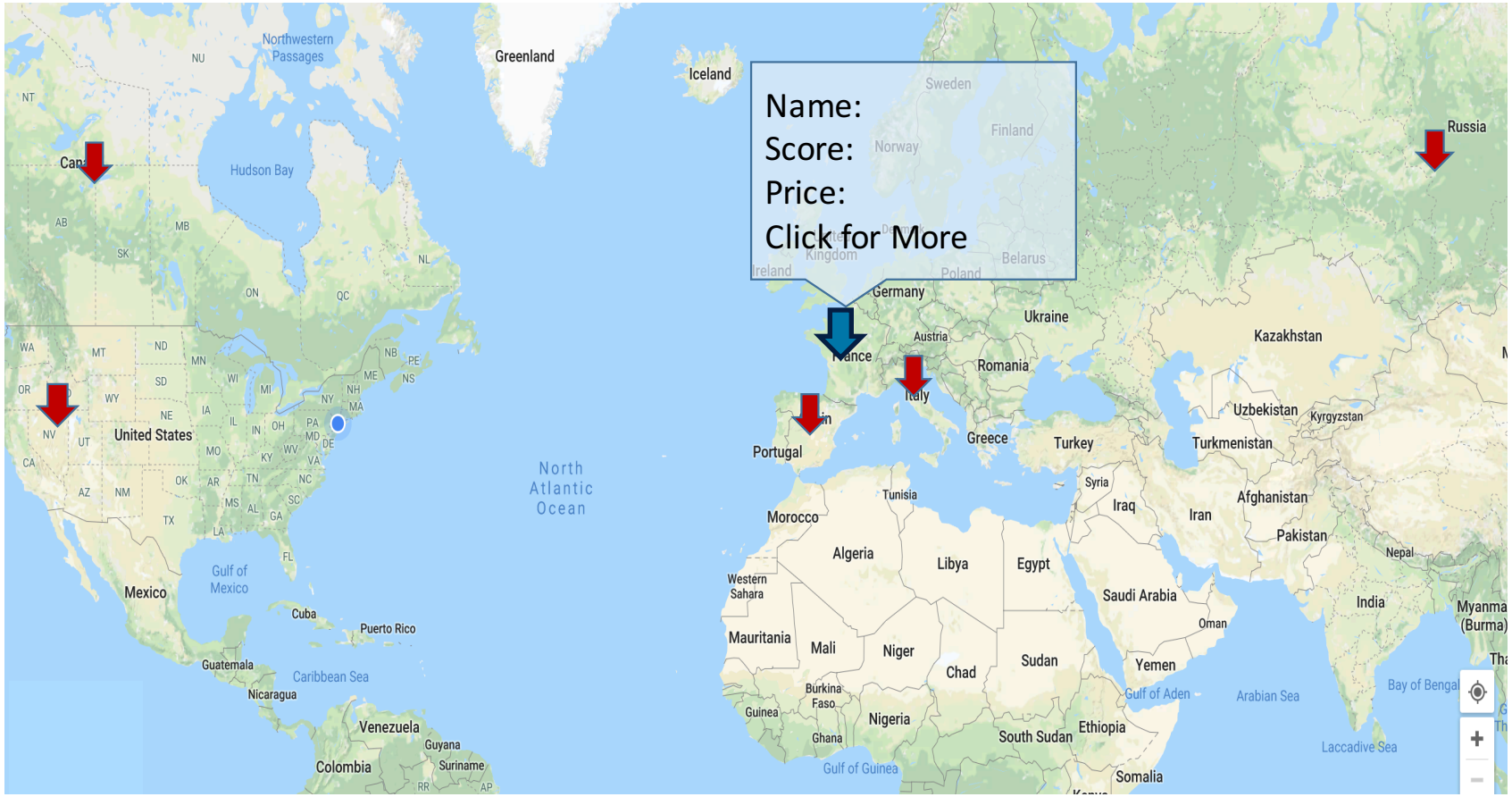
Method

3. Recommend similar wines

- Visualize the results
  - Blue arrow:

    uploaded wine
  - Red arrows:

    recommended wines



Potential Improvements

1. Validate the recommendation algorithm
   - Solution: set *like* and *dislike* buttons
2. Enrich our database
   - Solution: crowdsourcing
     - Allow users to update the wine information
     - Save the uploaded images for future use
3. Save user searching histories to implement better recommendation algorithm
4. Expand functions: e.g. comments, wine food pairing, and wine taste tips

Thank you!