CS513: Theory & Practice of Data Cleaning

Final Project Project Phase-I (Summer 2023)

University of Illinois, Urbana-Champaign

Team ID: Team 88

Ahmed Elfarra Ahmedse2@illinois.edu
Nishanth Alladi Nalladi2@illinois.edu
Adam Michalsky Adamwm3@illinois.edu

Motivation

In today's world, analytics are being leveraged throughout organizations to make critical decisions that can impact the success of a business. Analytics are not only used by executives, but also business leaders and analysts to help identify and implement operational efficiency in an organization. The decisions based on analytics are only as good as the data being displayed. A common saying in the industry is "Garbage in, garbage out." referring to the data used in analytics. In other words, if the data quality is poor when being leveraged for analytics, then the decisions made on the data are likely to be poor decisions.

Part of the course requirements for CS 513: Theory and Practice of Data Cleaning is to complete a group project that is focused on applying what we have learned throughout the course on a raw dataset so that the data can be utilized for use cases that we will identify below. This will provide hands-on experience since most (if not all) datasets require some form of data cleansing to be leveraged for analysis. In the sections that follow, we will describe the dataset, identify visible quality problems, the use cases we aim to fulfill, and a plan of how our group will distribute the workload.

Dataset

Name: Winery-Kaggle
Location: https://uofi.app.box.com/s/whvfh9jio38ck0m9gz58s31srx8iwg4i

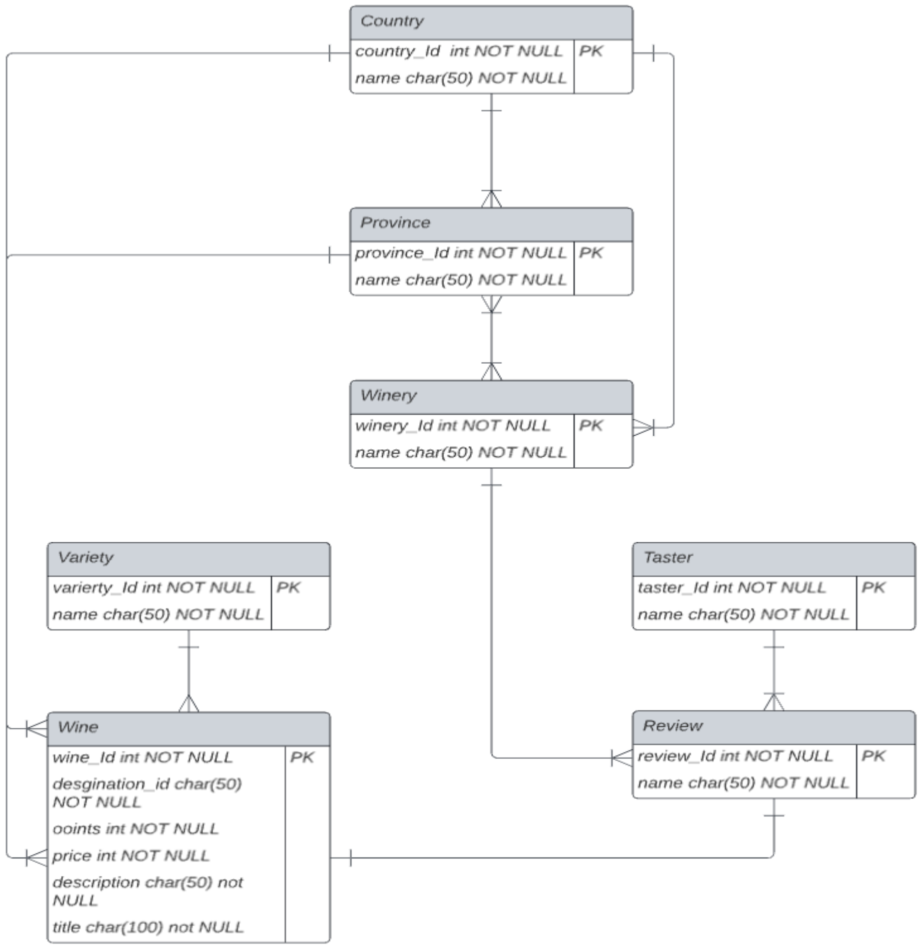
Exploratory Data Analytics (EDA) of the Dataset:

- **Total number of records:** 129970
- **Data diversity:** 706 different varieties of wine from 16,757 different wineries, 42 separate countries included in WK dataset. Over 51084 regions, 13643 tasters and 16032 reviews.
- **Age:** Data was uploaded to share on March 23, 2022.

Data Description

The dataset pool was already provided by course staff. Team 88 has chosen the dataset named "Winery-Kaggle" (WK). Unfortunately a dataset description was not provided so the following is a description of the data based on what can be observed in the file. The WK dataset has been provided as a single CSV file. The WK dataset contains 129970 wine reviews sourced from Twitter. Each review contains a review consisting of a description, rating on a scale of 1-100, taster (i.e, reviewer), and the taster's Twitter handle. Along with the review, characteristics of the reviewed wine are available which include the designation the producer gave the wine, price, title, variety, winery which produced the wine, and finally the location (country and province/region) where it was produced.

Entity Relationships



The following are the entities that were observed and were broken out in the entity diagram so we could understand the relationships. For phase II we plan on using a single file unless there is a clear benefit to breaking out the entities into their own files.

The main entities in the schema are:

- **Country**: Represents a country where wines are produced.
- **Province**: Represents a province or state within a country.
- **Review**: Represents the comments of the tasters reviews on the wines.
- **Winery**: Represents a winery or wine producer.
- **Variety**: Represents the type of grape used to make a wine.
- **Wine**: Represents a specific wine.
- **Taster**: Represents a wine taster who reviews wines.

The relationships depicted in the schema are:

- The **Country** entity has a one-to-many relationship with the Province entity, indicating that a country can have multiple provinces, but each province is associated with only one country.
- The **Province** entity has a one-to-many relationship with the winery entity, meaning that a province can have multiple wineries, but each winery is associated with only one province.
- The **Review** entity has a one-to-one relationship with the Wine entity, signifying that a review can have a single review for a single wine, but each wine is associated with only one review.
- The **Winery** entity has a one-to-many relationship with the Wine entity, indicating that a winery can produce multiple wines, but each wine is associated with only one winery.
- The **Variety** entity has a one-to-many relationship with the Wine entity, meaning that a grape variety can be used to produce multiple wines, but each wine is associated with only one variety.
- The **Wine** entity has a many-to-one relationship with the Variety entity, indicating that multiple wines can belong to the same variety, but each wine is associated with only one variety.
- The **Wine** entity has a one-to-many relationship with the Taster entity, signifying that a wine can be reviewed by multiple tasters, but each review is associated with only one wine.

Use Cases

U_0 - *No cleaning necessary*

Use Case: What varieties of wine are the highest rated in France?
Role: A tourist
Motivation: A tourist is traveling to popular regions that are known for their wines and would like to sample the highest rated wines.

Reasoning: Points for each review are available for almost all records. There is also a substantial subset of the data that is associated with France. A tourist would be able to figure out at least a top 10 list easily despite the inconsistencies in the country column.

U_1 - *Data cleaning is necessary and sufficient*

Use Case: Which wines does a taster prefer based on the points awarded and region it was cultivated in?
Role: A skeptical wine enthusiast based out of the United States
Motivation: A wine enthusiast is skeptical about the ratings given to some wine produced at their local winery. They suspect that each taster has a bias for preferred varieties of wine.
Reasoning: Reasoning for this use case is detailed below in the *Data Quality Problems* section below.

U_2 : *Data cleaning is not sufficient*

Use Case: What qualities/factors lead to higher rated wines?
Role: A wine producer/winery owner
Motivation: A wine producer would like to understand what factors lead to higher ratings in reviews so that they can improve their product.
Reasoning: Each review is an opinion of the taster and no context was provided on how points were awarded for each wine.

Data Quality Problems

Data cleaning is crucial to support the main use case U_1 by ensuring the accuracy, reliability, and interpretability of the dataset. Firstly, the absence of *taster_name* for some reviews limits the ability to analyze reviewers' biases. Reviews without an associated *taster_name* are considered unreliable and hold no value. Additionally, the lack of *taster_twitter_handle* prevents the wine skeptic from following reviewers whose opinions they agree with. Therefore, reviews without a valid twitter handle cannot be tracked.

Another data quality problem is the points column stored as a string, requiring a data type conversion to determine the highest-rated wines accurately. Incomplete data is also present in several fields, such as missing country, variety, designation, price, province, taster name, and taster twitter handle for certain reviews. By applying data cleaning techniques such as imputation or deletion, we can appropriately handle missing values and ensure that the dataset used for the main use case U_1 is complete and reliable.

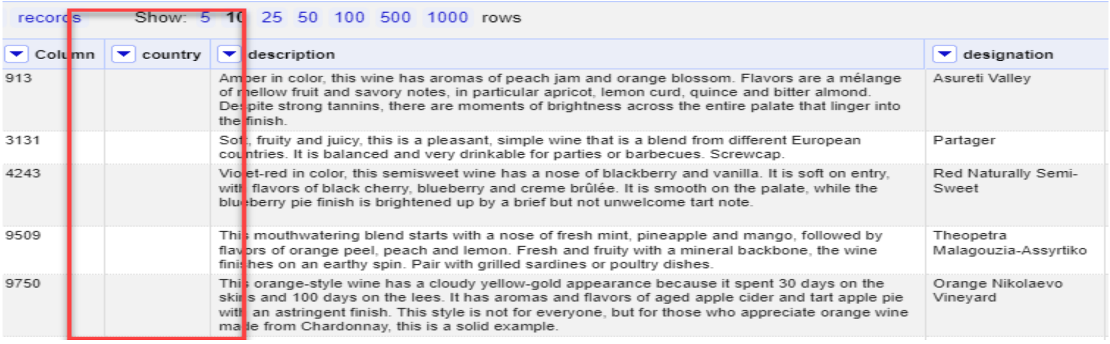
Along with the incomplete data, there is vague data. Vague data issues are issues where more information is required and items such as the absence of currency associated with the price field and a lack of explanation on how the score in the points column is calculated.

The unstructured nature of the description field limits the development of more meaningful use cases as it contains free text analysis by the taster. Lastly, there is a lack of temporal data, as the dataset does not indicate when the wines were reviewed despite being uploaded on March 23, 2022. Addressing these data quality problems is crucial to ensure the reliability, accuracy, and usefulness of the dataset for the main use case U_1 and other analytical purposes.

In general the dataset lacks proper data type enforcement, with all data points being stored as text and exhibits data inconsistencies. Inconsistent data entries, such as different spellings or variations of the same entity, can create challenges during analysis. Data cleaning involves harmonizing and standardizing the data to ensure consistent representation. This standardization enables grouping, aggregation, and comparison operations necessary for analysis, providing accurate and meaningful insights.

Overall, data cleaning plays a crucial role in supporting the main use case U_1 by ensuring data accuracy, handling missing values, resolving inconsistencies, and enhancing the overall quality and interpretability of the dataset. By cleaning the data, we can confidently perform analysis, build models, and draw reliable conclusions, enabling effective decision-making in various wine-related domains. Below are examples of the data quality issues we have encountered during phase I of our project.

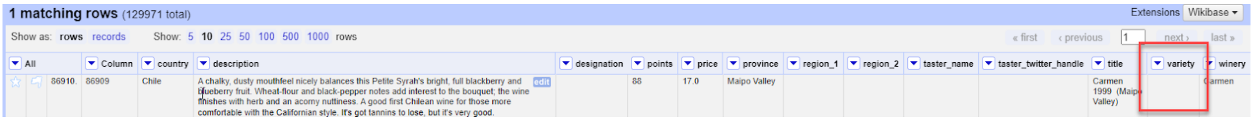
Example 1 - Missing Country:
ID: 913



records Show: 5 10 25 50 100 500 1000 rows

Column	country	description	designation
913		Amber in color, this wine has aromas of peach jam and orange blossom. Flavors are a mélange of mellow fruit and savory notes, in particular apricot, lemon curd, quince and bitter almond. Despite strong tannins, there are moments of brightness across the entire palate that linger into the finish.	Asureti Valley
3131		Soft, fruity and juicy, this is a pleasant, simple wine that is a blend from different European countries. It is balanced and very drinkable for parties or barbecues. Screwcap.	Partager
4243		Violet-red in color, this semisweet wine has a nose of blackberry and vanilla. It is soft on entry, with flavors of black cherry, blueberry and creme brûlée. It is smooth on the palate, while the blueberry pie finish is brightened up by a brief but not unwelcome tart note.	Red Naturally Semi-Sweet
9509		This mouthwatering blend starts with a nose of fresh mint, pineapple and mango, followed by flavors of orange peel, peach and lemon. Fresh and fruity with a mineral backbone, the wine finishes on an earthy spin. Pair with grilled sardines or poultry dishes.	Theopetra Malagouzia-Assyrtiko
9750		This orange-style wine has a cloudy yellow-gold appearance because it spent 30 days on the skins and 100 days on the lees. It has aromas and flavors of aged apple cider and tart apple pie with an astringent finish. This style is not for everyone, but for those who appreciate orange wine made from Chardonnay, this is a solid example.	Orange Nikolaevo Vineyard

Example 2 - Missing Variety:
ID: 86909



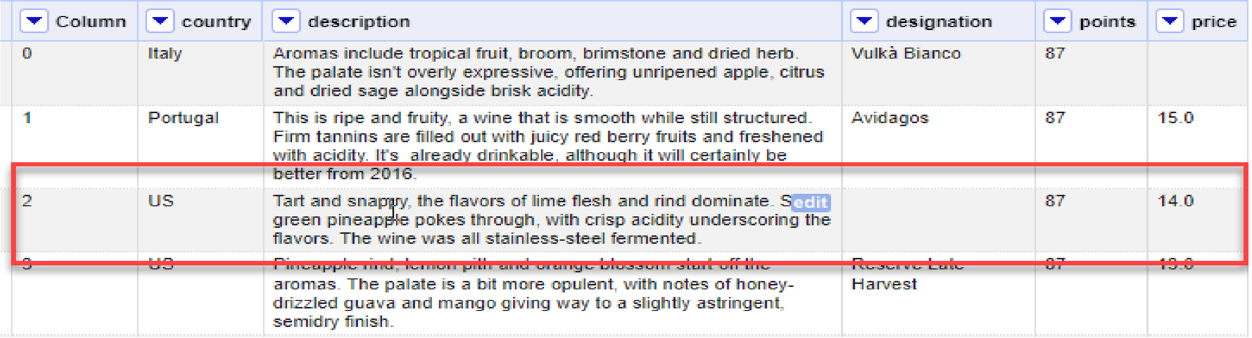
1 matching rows (129971 total)

Show as: rows records   Show: 5 10 25 50 100 500 1000 rows   « first ‹ previous 1 next › last »

All	Column	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_handle	title	variety	winery
86910.	86909	Chile	A chalky, dusty mouthfeel nicely balances this Petite Syrah's bright, full blackberry and blueberry fruit. Wheat-flour and black-pepper notes add interest to the bouquet; the wine finishes with herb and an acorny nuttiness. A good first Chilean wine for those more comfortable with the Californian style. It's got tannins to lose, but it's very good.		88	17.0	Maipo Valley					Carmen 1999 (Maipo Valley)		Carmen

Example 3 - Missing Designation:

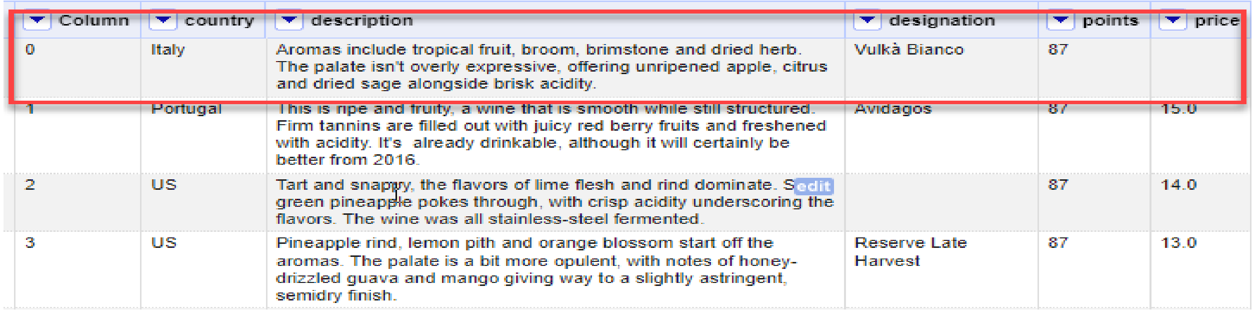
ID: 2



Column	country	description	designation	points	price
0	Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.	Vulkà Bianco	87	
1	Portugal	This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.	Avidagos	87	15.0
2	US	Tart and snappy, the flavors of lime flesh and rind dominate. S edit green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.		87	14.0
3	US	Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.	Reserve Late Harvest	87	13.0

Example 4 - Missing Price

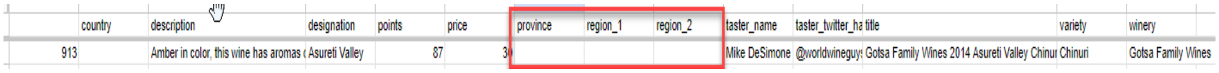
ID: 0



Column	country	description	designation	points	price
0	Italy	Aromas include tropical fruit, broom, brimstone and dried herb. The palate isn't overly expressive, offering unripened apple, citrus and dried sage alongside brisk acidity.	Vulkà Bianco	87	
1	Portugal	This is ripe and fruity, a wine that is smooth while still structured. Firm tannins are filled out with juicy red berry fruits and freshened with acidity. It's already drinkable, although it will certainly be better from 2016.	Avidagos	87	15.0
2	US	Tart and snappy, the flavors of lime flesh and rind dominate. S edit green pineapple pokes through, with crisp acidity underscoring the flavors. The wine was all stainless-steel fermented.		87	14.0
3	US	Pineapple rind, lemon pith and orange blossom start off the aromas. The palate is a bit more opulent, with notes of honey-drizzled guava and mango giving way to a slightly astringent, semidry finish.	Reserve Late Harvest	87	13.0

Example 5 - Missing Province/State

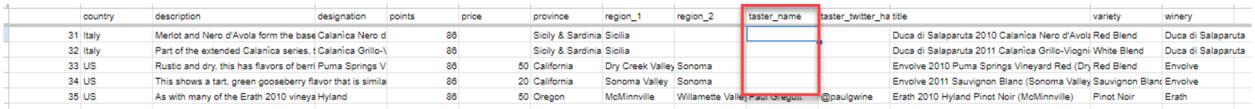
ID: 913



	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_ha	title	variety	winery
913		Amber in color, this wine has aromas (	Asureti Valley	87	3				Mike DeSimone	@worldwineguys	Gotsa Family Wines 2014 Asureti Valley Chinur	Chinuri	Gotsa Family Wines

Example 6 - Missing Taster

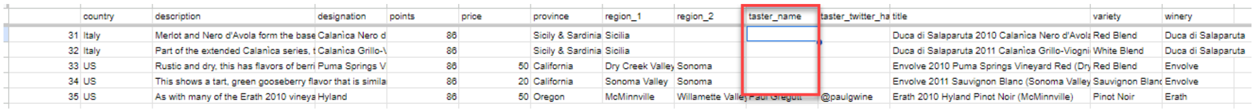
ID: 31, 32, 33, 34



	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_ha	title	variety	winery
31	Italy	Merlot and Nero d'Avola form the base	Calanìca Nero d	86		Sicily & Sardinia	Sicilia				Duca di Salaparuta 2010 Calanìca Nero d'Avola	Red Blend	Duca di Salaparuta
32	Italy	Part of the extended Calanìca series, t	Calanìca Grillo-\	86		Sicily & Sardinia	Sicilia				Duca di Salaparuta 2011 Calanìca Grillo-Viogni	White Blend	Duca di Salaparuta
33	US	Rustic and dry, this has flavors of berri	Puma Springs V	86	50	California	Dry Creek Valley	Sonoma			Envolve 2010 Puma Springs Vineyard Red (Dry	Red Blend	Envolve
34	US	This shows a tart, green gooseberry flavor that is simila		86	20	California	Sonoma Valley	Sonoma			Envolve 2011 Sauvignon Blanc (Sonoma Valley	Sauvignon Blanc	Envolve
35	US	As with many of the Erath 2010 vineya	Hyland	86	50	Oregon	McMinnville	Willamette Valley	Paul Gregutt	@paulgwine	Erath 2010 Hyland Pinot Noir (McMinnville)	Pinot Noir	Erath

Example 7 - Missing Twitter Handle

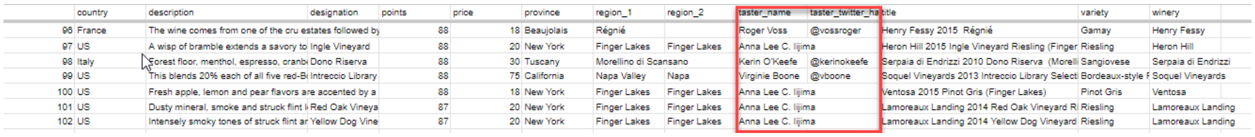
ID: 31, 32, 33, 34



	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_ha	title	variety	winery
31	Italy	Merlot and Nero d'Avola form the base	Calanìca Nero d	86		Sicily & Sardinia	Sicilia				Duca di Salaparuta 2010 Calanìca Nero d'Avola	Red Blend	Duca di Salaparuta
32	Italy	Part of the extended Calanìca series, t	Calanìca Grillo-\	86		Sicily & Sardinia	Sicilia				Duca di Salaparuta 2011 Calanìca Grillo-Viogni	White Blend	Duca di Salaparuta
33	US	Rustic and dry, this has flavors of berri	Puma Springs V	86	50	California	Dry Creek Valley	Sonoma			Envolve 2010 Puma Springs Vineyard Red (Dry	Red Blend	Envolve
34	US	This shows a tart, green gooseberry flavor that is simila		86	20	California	Sonoma Valley	Sonoma			Envolve 2011 Sauvignon Blanc (Sonoma Valley	Sauvignon Blanc	Envolve
35	US	As with many of the Erath 2010 vineya	Hyland	86	50	Oregon	McMinnville	Willamette Valley	Paul Gregutt	@paulgwine	Erath 2010 Hyland Pinot Noir (McMinnville)	Pinot Noir	Erath

Example 8 - Provided Taster with Missing Twitter Handle

ID: 97, 100, 101, 102



	country	description	designation	points	price	province	region_1	region_2	taster_name	taster_twitter_ha	title	variety	winery
96	France	The wine comes from one of the cru estates followed by		88	18	Beaujolais	Régnié		Roger Voss	@vossroger	Henry Fessy 2015 Régnié	Gamay	Henry Fessy
97	US	A wisp of bramble extends a savory to	Ingle Vineyard	88	20	New York	Finger Lakes	Finger Lakes	Anna Lee C. Iijima		Heron Hill 2015 Ingle Vineyard Riesling (Finger	Riesling	Heron Hill
98	Italy	Forest floor, menthol, espresso, cranb	Dono Riserva	88	30	Tuscany	Morellino di Scansano		Kerin O'Keefe	@kerinokeefe	Serpaia di Endrizzi 2010 Dono Riserva (Morelli	Sangiovese	Serpaia di Endrizzi
99	US	This blends 20% each of all five red-B	Intreccio Library	88	75	California	Napa Valley	Napa	Virginie Boone	@vboone	Soquel Vineyards 2013 Intreccio Library Selecti	Bordeaux-style F	Soquel Vineyards
100	US	Fresh apple, lemon and pear flavors are accented by a		88	18	New York	Finger Lakes	Finger Lakes	Anna Lee C. Iijima		Ventosa 2015 Pinot Gris (Finger Lakes)	Pinot Gris	Ventosa
101	US	Dusty mineral, smoke and struck flint l	Red Oak Vineya	87	20	New York	Finger Lakes	Finger Lakes	Anna Lee C. Iijima		Lamoreaux Landing 2014 Red Oak Vineyard Ri	Riesling	Lamoreaux Landing
102	US	Intensely smoky tones of struck flint ar	Yellow Dog Vine	87	20	New York	Finger Lakes	Finger Lakes	Anna Lee C. Iijima		Lamoreaux Landing 2014 Yellow Dog Vineyard	Riesling	Lamoreaux Landing

Phase II Initial Plan

Disclaimer: Assignments below are tentative and subject to change based on workload.

Step 1 (Finish By July 1) (Ahmed)

A. With our updated knowledge of the dataset, read through our ERD and update as necessary (Ahmed)

B. Update the Relationships section as necessary (Ahmed)

Step 2 (Finish by July 6) (Nishanth)

A. Using OpenRefine, look through data entries and find small and large scale inconsistencies manually. We've listed some preliminary quality problems we've found in Step 3, but we are not sure they are expansive, so the work will be to refine those steps further.

B. Use the cluster feature to clean columns

C. Using Python, find libraries that can create data similar to text facets in OpenRefine – OR will be too slow for such a large dataset

Step 3: (Finish by July 10) (Adam,Ahmed)

A. Clean inconsistencies found in OpenRefine using knowledge from Step 2

    a. Fix the designation column

        i. Create 1 column per semantic group since there are some semantic variations within designation – There are wine names, years, places, what look to be brand names, vineyards, single characters, and non-word character strings

    b. Cluster the following columns (and whichever new ones are created from the designation column)

        i. region_1

        ii. region_2

        iii. Province

        iv. Designation

        v. Title

        vi. Variety

    c. Clean the title column: it is of the format "Title (Province/region_1/region_2)" and the data in parenthesis is already captured in another column so it is unnecessary to have twice, it can be made to "Title"

    d. Clean the region_1 and region_2 columns: they are the same for some rows, region_2 should be blank when that is the case

    e. Delete the “Column” column: it is unnecessary, remove it
    f. Format data the same
        i. Make everything title case
        ii. Use the correct accents wherever necessary
        iii. Make numeric columns into numbers

B. Clean the rest of the inconsistencies we can’t clean in OR in Python

Step 4 (Finish by July 14) (Nishanth)

A. Print unique values of designation, province, region_1, region_2, title, variety before and after cleaning. Verify that we see improvement and that the final result is clean enough for our use case. This makes sure…
    a. We dont have different values for what is supposed to be the same value (clustering)
    b. Our values are reasonable and understandable by humans
    c. Each column is a semantic group

B. For a given entry, make sure no two columns have the same value. This makes sure…
    a. There is no redundant data

Step 5 (Adam, Nishanth, Ahmed) (July 26)

A. Show the OpenRefine log from Step 3 in a understandable way (Adam)
B. Show a before and after “highlights” section – biggest changes and modifications (Nishanth)
C. Using Python, quantify percentage of dataset affected, along with percentage of “new” data that was created (Ahmed)
D. Create a easily viewable dashboard or scrollable report of the changes in a visual form (Group)