CS513 Phase-I Report: Wine Reviews

Team18

Yunfei Ouyang Sicheng Jiang

yunfeio2@illinois.edu sj62@illinois.edu

July 9, 2023

1. Dataset Chosen

Winery-Kaggle

Content: 130k wine reviews with variety, location, winery, price, and description

Source: https://www.kaggle.com/datasets/zynicide/wine-reviews

2. Description of Dataset

a. Database schema

```
CREATE TABLE wine_reviews (
     id INT NOT NULL PRIMARY KEY,
     country VARCHAR(255),
     description VARCHAR(255),
     designation VARCHAR(255),
     points INT,
     price INT,
     province VARCHAR(255),
     region_1 VARCHAR(255),
     region_2 VARCHAR(255),
     taster_name VARCHAR(255),
     taster_twitter_handle VARCHAR(255),
     title  VARCHAR(255),
     variety VARCHAR(255),
     winery VARCHAR(255)
);
```

b. a narrative description of the dataset covering structure and content

The dataset "wine_reviews" is designed to store information about wine reviews. It consists of the following columns:

1. id: An integer value that serves as a unique identifier for each wine review.
2. country: A string indicating the country of origin for the wine.
3. description: A brief textual description provided by a sommelier, describing various aspects of the wine such as taste, smell, appearance, and texture.
4. designation: This field specifies the specific vineyard within the winery where the grapes used to make the wine are grown.
5. points: An integer representing the rating assigned to the wine by WineEnthusiast on a scale of 1 to 100. It is mentioned that only wines with a score of 80 or higher have reviews posted.
6. price: The cost associated with purchasing a bottle of the wine.
7. province: The province or state from which the wine originates.
8. region_1: This column captures the wine growing area within a province or state. For example, it could be "Napa" if the wine is from the Napa Valley.
9. region_2: Sometimes, a more specific sub-region is specified within a wine growing area. However, this field may be left blank in some cases.
10. taster_name: The name of the sommelier or wine taster who provided the review.
11. taster_twitter_handle: The Twitter handle of the sommelier or wine taster.
12. title: The title of the wine review.
13. variety: The type of grapes used in making the wine. For instance, it could be "Pinot Noir."
14. winery: The name of the winery that produced the wine.

3. Use Cases

a. “Zero cleaning” use case U0: data cleaning is not necessary

We want to investigate on all countries of origin that contains at least one wine with the highest rating among all others.

b. “Main” use case U1: data cleaning is necessary and sufficient

We want to investigate on taster's name and twitter handle of the individual who reviewed the most wines.

c. “Never enough” use case U2 : data cleaning is not sufficient

Since our user likes to drink French wines, we would like to receive recommendations for wines (wine name) from the Deutz winery.

4. Data Quality Problems

a. List obvious data quality problems with evidence (examples and/or screenshots)

1. There are empty values in designation, price, region_1, region_2, taster_name and taster_twitter_handle.
   screenshot:

	country	descriptio	designatic	points	price	province	region_1	region_2
0	Italy	Aromas in	Vulk?? Bia	87		Sicily & Sa	Etna	
1	Portugal	This is ripe	Avidagos	87	15	Douro		
2	US	Tart and s	nappy, the	87	14	Oregon	Willamett	Willamett
3	US	Pineapple	Reserve L	87	13	Michigan	Lake Michigan Shore	
4	US	Much like	Vintner's I	87	65	Oregon	Willamett	Willamett
5	Spain	Blackberry	Ars In Vitr	87	15	Northern	Navarra	
6	Italy	Here's a b	Belsito	87	16	Sicily & Sa	Vittoria	
7	France	This dry a	nd restrain	87	24	Alsace	Alsace	
8	Germany	Savory dri	Shine	87	12	Rheinhessen		
9	France	This has g	Les Nature	87	27	Alsace	Alsace	
10	US	Soft, supp	Mountain	87	19	California	Napa Vall	Napa
11	France	This is a d	ry wine, ve	87	30	Alsace	Alsace	
12	US	Slightly re	duced, thi	87	34	California	Alexande	Sonoma
13	Italy	This is dor	Rosso	87		Sicily & Sa	Etna	
14	US	Building o	n 150 year	87	12	California	Central Cc	Central Cc
15	Germany	Zesty orar	Devon	87	24	Mosel		

2. In the dataset, all points fall within the range of 80 to 100, which limits the differentiation between wines with similar ratings. It is worth noting that the data was specifically crawled for wines with ratings above 80 points. To facilitate better comparison and enable more effective queries in the future, it is desirable to normalize the points to a 0-10 scale. This normalization will allow for a standardized rating system.

b. Explain why / how data cleaning is necessary to support the main use case U1

To investigate the taster's name and Twitter handle of the individual who reviewed the most wines, data cleaning is necessary to support the main use case.

1. Standardization of Names and Handles: Clustering helps ensure consistency in the taster's names and Twitter handles throughout the dataset. It addresses potential variations, misspellings or formatting issues, making it easier to identify and analyze the taster who reviewed the most wines.
   Without 1, a taster may be identified as a different person in different reviews because of the lack of consistency in tasters' names and Twitter handles.

2. Handling Missing or Incomplete Data: The dataset may contain missing or incomplete information for certain tasters' names or Twitter handles, as well as the destination or price which makes the review incomplete. Moreover duplicate entries for the same

taster could distort the analysis. Data cleaning ensures that we focus on accurate and unique data.

Without 2, we may encounter NULL data on tasters' names and Twitter handles. This can lead to difficulties in determing who reviewed the most wines.

There are additional data cleaning techniques that we are going to deploy, such as normalizing the points range, check consistency for all other entries, etc. However, these techniques won't effect the U1 analysis,The above two points illustrate why data cleaning is necessary to support U1.

5. Initial Plan for Phase-II

a. Below is a possible plan, listing typical data cleaning workflow steps. In your Plan for Phase-II, fill in additional details for the project steps as needed. In particular, include who of your team members will be responsible for which steps, and list the timeline that you are setting yourselves!

■ S1: Review (and update if necessary) our use case description and dataset description

1. We will review our case and dataset description prior to our work in Phase-II

■ S2: Profile D to identify DQ problems: How do you plan to do it? What tools are you going to use?

1. We will use Open Refine to identify the rows that contain empty values in taster’s name or taster’s twitter handles.
2. We will use Open Refine to identify rows that the same taster’s name are spelled differently.
3. We will use Open Refine to check if there are extra spaces in the taster’s name and twitter handles columns.
4. We will use Python to check if there are two different tasters who have the same twitter handle (IC violations).

■ S3: Perform DC “proper”: How are you going to do it? What tools do you plan to use? Who does what?

1. We will use Open Refine to remove the rows that contain empty values in taster’s name or taster’s twitter handles.
2. We will use Open Refine to cluster the rows that the same taster’s name are spelled differently.
3. We will use Open Refine to remove any extra spaces in the taster’s name and twitter handles columns.
4. We will use Python to normalize the points of each wine to a 0-10 scale.
5. We will use Python to ensure each taster’s twitter handle is only associated with one taster’s name.

■ S4: Data quality checking: is D' really "cleaner" than D?

1. Use Open Refine to find the rows that contain empty values in taster's name and taster's twitter handle in D'. The result should be an empty set.
2. Use Open Refine to ensure that all points range from 0 to 10 in D'.
3. Use Python to find different tasters who have the same twitter handle in D'. The result should be an empty set.
4. Use Python to find the taster name and the taster's twitter handle of the taster who tastes the most wines.

■ S5: Document and quantify change

1. columns and cells changed
2. IC violations detected: before vs after