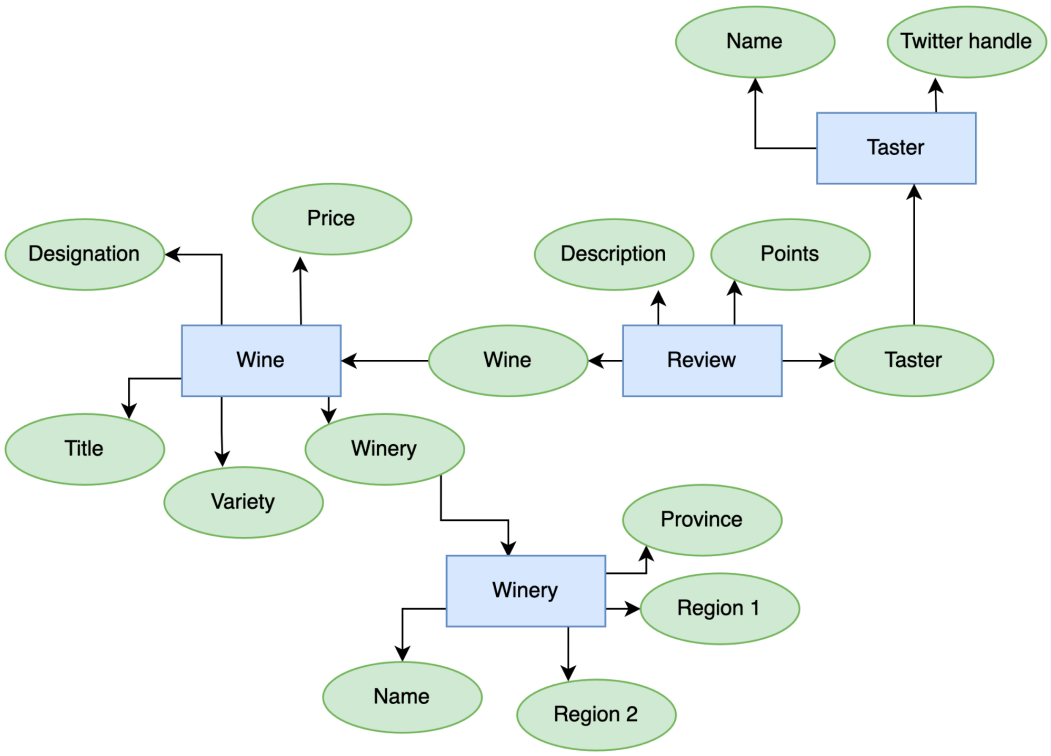
CS 513: Data Cleaning
Summer 2022
Project Phase 1
Tyler Davis, Liz Jackson, Austin Wang

1. **Identify a dataset:** We are using the winery dataset from Kaggle, which is one of the sample datasets. This dataset includes the following information:
    a. Information about the wines themselves, such as their title, the winery at which they were produced, their variety, and their price.
    b. Geographic information about where the wines were produced, such as their country, province, and region.
    c. Information about the reviewers who tasted these wines; specifically their names and Twitter handles.
    d. Reviewers' assessments of these wines in free-text and numeric form.

2. **Describe use cases:**
    a. **Target use case (U_1), in which data cleaning is necessary and sufficient for data analysis:** After we clean this dataset, we will use the data to evaluate the words in wines' reviews and observe how certain words correlate with wine rating and price. For example, we will evaluate which words tend to appear in the highest-scoring and highest-price wines and which terms correspond to the lowest-scoring and lowest-price wines. We assume that the wine bottle sizes are standardized. We also want to examine the relationship between wine harvest year, price, and rating, which will require extracting year data from the names of wines.
    b. **Minor use case (U_0) in which the data requires no cleaning:** The Country column in the dataset contains discrete country values as is and the Price column only contains numeric values with assumed consistent currencies in all rows. As a result, the existing dataset allows us to evaluate which countries produce the cheapest wines and which countries produce the most expensive wines.
    c. **Minor use case (U_2) in which data cleaning is not sufficient to produce the analysis:** This dataset does not include the date on which the reviewer tasted and reviewed the wine. As a result, we are unable to determine the age of the wine when its rating and price were documented in the review, and we therefore cannot know whether wine age correlates with ratings and price.

3. **Describe the dataset:** The ER diagram below illustrates this dataset:



(Diagram link)

4. **Obvious data quality problems:** The following data quality problems require data cleaning for us to do our intended data analysis:
    a. **No discrete year column:** Many wines' titles contain their harvest year in the title (such as "Louis M. Martini ***2012*** Cabernet Sauvignon (Alexander Valley)" or "Envolve ***2010*** Puma Springs Vineyard Red (Dry Creek Valley)") but the wines' harvest year is not stored as its own discrete column. As a result, we cannot evaluate a correlation between wines' harvest years and their prices and ratings until we extract the year into its own column.
    b. **Locations in wines' titles:** Many wines contain their provinces in parentheses at the end of their titles (such as "Lava Cap 2010 Battonage Chardonnay ***(El Dorado)***"). If a wine does not have an associated region, its region is at the end of its title (such as "Vigneti Le Monde 2010 Sauvignon ***(Friuli Grave)***"). This parenthetical location information does not provide additional information about the wines (the information is contained in other columns in the dataset) and it may prevent us from correlating two identically-titled wines if they have different location information appended to the end of their titles. We can remove the parenthetical location information in wines' titles without losing value.
    c. **Free-text comments:** Each review contains free-text comments and these comments are difficult to report on without additional cleaning. By evaluating the

most common words (excluding stopwords) in the corpus of comments, we can better understand which words tend to indicate the highest-rated wines.

d. **Missing price data**: Some reviews do not contain wine prices, which makes it impossible to include those reviews in our analyses that involve wine prices.

5. **Devise an initial plan:**
    a. **S_1, description of dataset D and matching use case U_1:** See questions 1 and 2.
    b. **S_2, profiling of D to identify the quality problems P that need to be addressed to support U_1:** See question 4.
    c. **S_3, performing the data cleaning process using one or more tools to address the problems P:** We will first use Python to do string cleanup and analysis and OpenRefine for remaining data cleanup. We will use YesWorkflow to document our workflow and for provenance purposes. Specifically:
        i. We will use OpenRefine to perform basic text cleaning on all string columns and numerical formatting cleaning on numeric columns.
        ii. We will use regex functions in Python to find four-digit integers within wine titles to determine wines' years and will extract this information into its own column.
        iii. We will use string-splitting functions in Python to remove parenthetical information at the end of wines' titles.
        iv. We will use Python to extract words from wines' descriptions and use the Natural Language Toolkit library to stem and normalize those words and filter stopwords. The results will then be split into two new tables foreign keyed to the original dataset, enabling analysis of the most frequent words in wines' descriptions.
        v. We will use OpenRefine to do any remaining cleaning on columns with discrete values (such as Province and Winery) to ensure that information is consistent across entries.
    d. **S_4, checking that your new dataset D′ is an improved version of D, e.g., by documenting that certain problems P are now absent and that U_1 is now supported:** We will do the following to confirm that D' is an improved version of D:
        i. Validate that wine year column contains only four-digit integers greater than 1900 and less than or equal to 2022
        ii. Validate that wine titles *do not* end with a parenthetical note about their location origin
        iii. Validate that term usage can be aggregated per and across reviews.
    e. **S_5, documenting the types and amount of changes that have been executed on D to obtain D′:** To obtain D', we made the following changes:
        i. Removed parenthetical location information from wines' titles.
        ii. Created a new column for wine year and pulled the wine year into this column.

iii. Cluster and clean text columns. Ensure all words representing the same entity have the same spelling.
iv. Clean numerical columns for any formatting issues, i.e., inclusion of a string or a leading zero.
v. Extract words and frequency from review text, using Python's NLTK library to both filter stopwords and to stem words for normalization purposes. These words will then be processed into a set of normalized tables:
    1. The first table is a mapping between reviews and words. Each row consists of three columns, with each row representing the frequency of a single term within a single review:
        a. Word Foreign Key (PK on second table)
        b. Review Foreign Key (PK on main table)
        c. Frequency (within a given review)
    2. The second table is the normalized table for words extracted from review text. We would use this table to query by word rather than querying words by review. Each row consists of three columns:
        a. Word Primary Key
        b. Word (stemmed using NLTK)
    3. Original review description will be left as is.

6. **Assignment of tasks:**
    a. Word splitting and normalization: Tyler
    b. Year extraction and removal of parenthetical notes in titles: Liz
    c. General column cleaning / OpenRefine work: Austin