# The Wranglers (Team28)


(1) Naveen Kumar Conjeevaram Baskaran (email id: nc42@illinois.edu)
(2) Alvin Do (email id: alvindo2@illinois.edu)
(3) Hemantika Dasgupta (email id: hd8@illinois.edu)


## 1. Dataset Chosen

We opted for the Winery-Kaggle dataset (winemag-data-130k-v2.csv).

## 2. Description of Dataset

The winery-Kaggle dataset D contains about 130K records of wine reviews from around the world. This data was scraped from the Wine Enthusiast (https://www.wineenthusiast.com/ - a popular site dedicated to wine, spirits, and the culture surrounding them) website during the week of June 15th, 2017.

The csv file originally consisted of 14 columns as mentioned below:

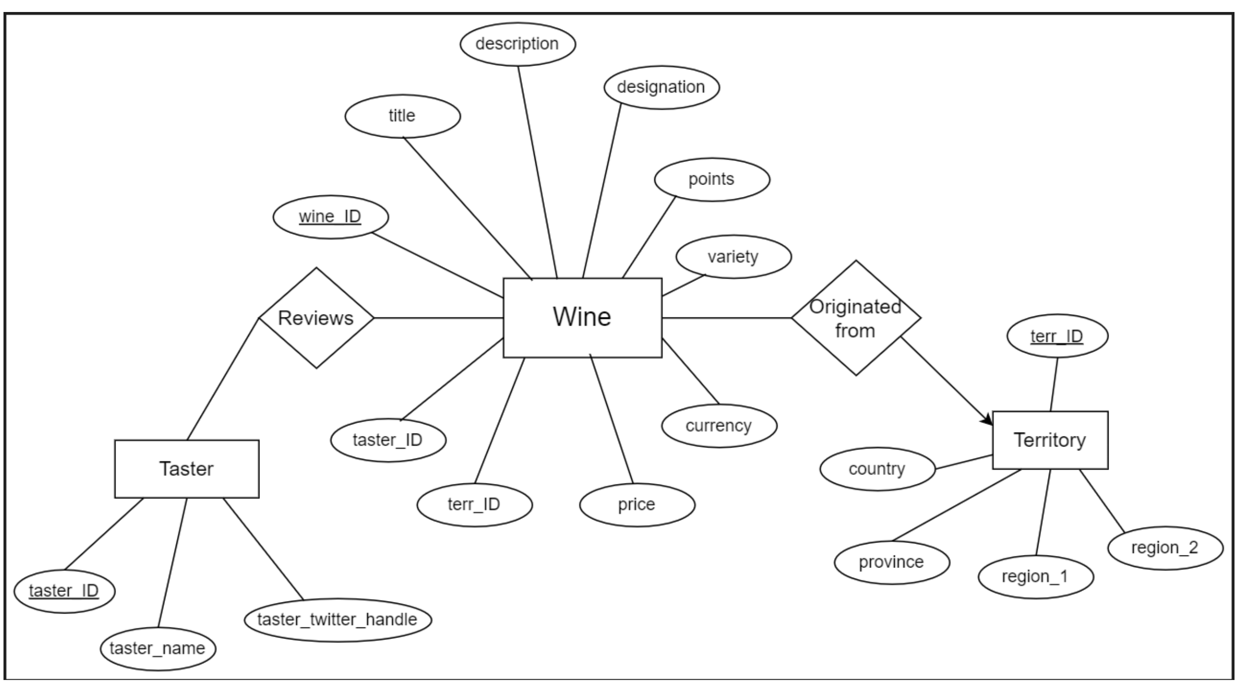
i. *Wine_ID: A unique ID for each wine review record in the csv dataset.*
ii. *Country: the country from which the wine was originated (ex: US).*
iii. *Description: text describing the specific wine's taste, smell, look, feel, color, specialty, etc.*
iv. *Designation: the vineyard within the winery where the grapes that made the wine are from.*
v. *Points: number of points provided for the wine on a scale of 1-100 by the tasters tasting the wine.*
vi. *Price: the cost for one bottle of the wine in local currency (ex: USD).*
vii. *Province: the province or state within the country that originated the wine (ex: Washington).*
viii. *Region 1: the area within the province or state (ex: Snipes Mountain).*
ix. *Region 2: sometimes there are more specific regions specified within a wine growing area (ex:Columbia Valley inside of Snipes Mountain), but this value can sometimes be blank.*
x. *Taster Name: name of the person who tasted and reviewed the wine.*
xi. *Taster Twitter Handle: Twitter handle identifier for the person who tasted and reviewed the wine.*
xii. *Title: name of the wine along with the territory which could be country, province or region if the province contains text "other".*
xiii. *Variety: the type of grapes used to make the wine (ie Pinot Noir).*
xiv. *Winery: the winery that produced the wine.*

Additionally, some extra fields(as shown below) were added by us to provide more context around the data(currency) & to efficiently create a relationship model for this dataset(with additional IDs).

xv. *Currency: The currency that's used to quote the wine price*
xvi. *Taster_ID: A unique identifier for the taster*
xvii. *Terr_ID: A unique identifier for the territory*

The dataset can be illustrated using the below ER diagram. The ER diagram can below be effectively transformed into relational tables, presenting a coherent and logical structure. There are 3 entities in the diagram below – Taster, Wine and Territory connected by the relationships – 'Reviews' and 'Originated from'.  As we could see from the data, multiple tasters can review multiple wines and provides the ratings. Therefore, we have a many-to-many relationship between the 'Taster' & 'Wine' entities. Many different wines can originate from a specific territory, hence we have a many-to-one relationship between the wine and territory entities.

We noticed that there's some information about the tasters and a lot of details of the wine in this dataset. Hence, we can organize the taster details in a separate 'Taster' table having the associated ID, name & twitter handle. The territory details of the wine specifically outlining the country, province, region can be moved to a separate 'Territory' table too.  Lastly, all the wine attributes (title, description, designation, points, price, currency, variety etc.) can be put in a table of its own. The primary keys taster_ID and terr_ID from the Taster & Territory entities respectively are included in the 'Wine' entity as foreign keys for easier association of relational data.



## 3. Use Cases

a. <u>“Main” use case U1: data cleaning is necessary and sufficient</u>
   Our target use case is that of a ‘wedding wine selection where the wedding is being held in the US and the guests would prefer European wine’. As we know, weddings are a grand affair and drinks play a crucial role. Hence, in order to arrange an assortment of wines that should delight the guests and yet be within the wedding budget, we need to look at the points, price, title, description, winery and territory details(country, province, region_1, region_2). Barring the ‘points’ data, every other data point mentioned above needs to be cleaned to some extent to make it ‘fit-for-purpose’ for the wedding wine analysis.

b. “<u>Zero cleaning” use case U0: data cleaning is not necessary</u>

   If we were to just give a cursory look at this dataset to get a list of the wineries where they produce highly-rated wines and the wine variety, then just by looking at the ‘points’, ‘winery’ and ‘variety’ columns, we can draw an easy conclusion. The data in these 3 columns is pretty comprehensive and clean. There are no null values in ‘points’ and ‘winery’ columns and the ‘variety’ field has only 1 NULL value(out of the 130K records and can be easily researched), hence the data is good enough to use as is.

c. <u>“Never enough” use case U2 : data cleaning is not sufficient</u>

   If we want to contact the tasters online using the information in this dataset, it would never fulfill this need as we noticed that a lot of records are missing the taster names and their twitter handles. No amount of manipulation in these 2 fields can point us to the correct individual, hence the data cleaning on these 2 columns will never be enough.

## 4. Data Quality Problems

One look at the dataset and we noticed the syntactic errors of various special characters and symbols that has corrupted a lot of data. Almost all text columns like description, province, region_1, region_2, title etc. are riddled with these symbols. Even the accented characters in some names(e.g. Saint-estèphe changed to Saint-EstÃ¨phe) have taken a different form in the csv data. Some data in the dataset also have the syntactic errors of white spaces. For e.g., an entry “Nicosia 2013 VulkÃ  Bianco  (Etna)” in title column has 2 consecutive white spaces before the parenthesis. Morever, we noticed a lot of semantic errors as well. There are a lot of missing data in columns like price, country and the regions. In our DB schema, we would define the price field to not have NULL values, hence few records in the price field would fail this NULL constraint. Additionally, we also noticed duplicate records in this dataset.

For us to determine a moderately-budgeted wine from European countries, we definitely need to clean up the price and territory information. Also, having duplicate information wouldn't help as if we were to even consider some average values, the additional duplicate rows would skew this calculation. Reading through the descriptions and getting the appropriate winery information would help with the analysis, hence some extent of cleaning is definitely required for our primary use case U1.

Some screenshots have been provided below that highlight these errors:

➔ Country fields NULL and special characters highlighted in red:

| 1 | | country | description |
|---|---|---|---|
| 915 | 913 | | Amber in color, this wine has aromas of peach jam and orange blossom. Flavors are a mÃ©lange of mellow fruit and savory notes, in particular apricot, lemon curd, qu |
| 3133 | 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. Screwc |
| 4245 | 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 brÃ»lÃ©e. It is smooth |
| 9511 | 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 |
| 9752 | 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 cid |
| 11152 | 11150 | | A blend of 85% Melnik, 10% Grenache Noir and 5% Petit Verdot, this wine has aromas of saddle leather, cassis and dark chocolate. In the mouth there are flavors of che |
| 11350 | 11348 | | Light and fruity, this is a wine that has some sweetness, a gentle series of white fruits, acidity and lemon zest at the end. Screwcap. |
| 14032 | 14030 | | This Furmint, grown in marl soils, has aromas of brioche, apple tart and caramelized pear and flavors of lemon-lime juice, green apple and lime blossom before a crisp |
| 16002 | 16000 | | Jumpy, jammy aromas of foxy black fruits are slightly volatile. This has a narrow mouthfeel that's high in acidity. Blackened berry flavors are spicy and peppery, while th |
| 16751 | 16749 | | Winemaker: Bartho Eksteen. This wooded Sauvy shows incredible restraint, harmony and balance. It opens with attractive notes of ripe gooseberry, apple flesh and gre |
| 18077 | 18075 | | Delicate white flowers and a spin of lemon peel on the nose start this refined white. The bright fruit on the palate is tropical and exotic, but the minerality gives it lift. Fu |
| 26487 | 26485 | | This wine has aromas of black berry, dried red raspberry and blueberry. In the mouth there are flavors of blueberry juice and black plum. The finish is slightly jammy ye |
| 26488 | 26486 | | Aromas of green apple and white flowers prepare the palate for a wine that fills the mouth with soft fruit flavors, in particular apple, grapefruit and pineapple. The soft |
| 26491 | 26489 | | Balanced aromas of green herbs and citrus zest are fresh and vital but not aggressive. A lemony twinge of acidity lends snap to the palate, while this tastes of pithy brin |
| 27824 | 27822 | | This is a reasonably rich, concentrated example of Pinot Gris, but one that seems a bit repressed aromatically. It hints rather than suggests at fig, pear and melon, befor |
| 36114 | 36112 | | An interesting blend of indigenous Bulgarian and uncommon French varieties, this red wine is made from equal parts Melnik, MourvÃ¨dre and Marselan. It offers arom |
| 38242 | 38240 | | Subdued citrus and pear notes on the nose find fuller expression on the rounded, very textured palate that almost seems to have a pleasingly bitter edge of earth and l |
| 38900 | 38898 | | Scents of clover, stem, green herb and red currant lead into a complex wine that adds sassafras highlights as it glides into a gentle finish. Fermented with wild yeasts an |
| 44676 | 44674 | | Crisp apple freshness almost tips into full citrus zestiness on the nose of this subtle, graceful wine. Very slight spritz tickles the palate on a textured and dry palate. It's l |
| 44852 | 44850 | | This blend of Gamay and Prokupe has aromas of tart cherry and Mediterranean herbs. Flavors recall cherry, black plum and stewed prune. |
| 44853 | 44851 | | This wine has aromas of honeysuckle and lemon curd, with flavors of lemon-meringue pie and lemon pith. The finish is clean and crisp. |
| 45249 | 45247 | | Just a whiff of citrus shows on the restrained nose. The palate is very crisp, even lean, disclosing slight spritz. Some riper fruit waits at the core but you have to savor th |
| 45404 | 45402 | | Basic cherry aromas turn more earthy and soupy with air. A crisp pulling palate holds flavors of tomato and spiced plum, while this is mild, with nothing much to speak |
| 46354 | 46352 | | A dark color and rich, jammy, baked aromas of charred oak and lush berries are the opening to this blend of 72% Cabernet Sauvignon, 20% CarmenÃ¨re and 8% Caberr |

winemag-data-130k-v2 Sheet1 +

Ready 63 of 129971 records found Accessibility: Unavailable Count: 1 100%

➔ Duplicate records in the data:

| 1 | | country | description | designation | points | price | provinc | region_ | region_ | taster_name | taster_twitter_hand | title | variety | winery |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3211 | **3209** | **US** | **Creamy, lush a** | **Sonoma Brut** | **90** | **22** | **California** | **Sonoma C** | **Sonoma** | **Virginie Boone** | **@vboone** | **Gloria Ferrer NV** | **Sparkling Blend** | **Gloria Ferrer** |
| 4401 | 4399 | US | Made predomi | Sonoma Brut | 88 | 22 | California | Sonoma C | Sonoma | Virginie Boone | @vboone | Gloria Ferrer NV | Sparkling Blend | Gloria Ferrer |
| 27775 | 27773 | US | A wonderfully ( | Sonoma Brut | 90 | 20 | California | Sonoma C | Sonoma | | | Gloria Ferrer NV | Sparkling Blend | Gloria Ferrer |
| 63181 | **63179** | **US** | **Made from mo** | **Sonoma Brut** | **92** | **22** | **California** | **Sonoma C** | **Sonoma** | **Virginie Boone** | **@vboone** | **Gloria Ferrer NV** | **Sparkling Blend** | **Gloria Ferrer** |
| 65586 | **65584** | **US** | **Made from mo** | **Sonoma Brut** | **92** | **22** | **California** | **Sonoma C** | **Sonoma** | **Virginie Boone** | **@vboone** | **Gloria Ferrer NV** | **Sparkling Blend** | **Gloria Ferrer** |
| 81565 | 81563 | US | This wine show | Sonoma Brut | 89 | 24 | California | Sonoma C | Sonoma | | | Gloria Ferrer NV | Sparkling Blend | Gloria Ferrer |
| 90691 | **90689** | **US** | **Creamy, lush a** | **Sonoma Brut** | **90** | **22** | **California** | **Sonoma C** | **Sonoma** | **Virginie Boone** | **@vboone** | **Gloria Ferrer NV** | **Sparkling Blend** | **Gloria Ferrer** |
| 94323 | 94321 | US | A bit rough anc | Sonoma Brut | 85 | 20 | California | Sonoma C | Sonoma | | | Gloria Ferrer NV | Sparkling Blend | Gloria Ferrer |
| 100740 | 100738 | US | A good, dry an | Sonoma Brut | 89 | 20 | California | Sonoma C | Sonoma | | | Gloria Ferrer NV | Sparkling Blend | Gloria Ferrer |
| 109003 | 109001 | US | From almost 9: | Sonoma Brut | 88 | 22 | California | Sonoma C | Sonoma | Virginie Boone | @vboone | Gloria Ferrer NV | Sparkling Blend | Gloria Ferrer |
| 122210 | 122208 | US | This bubbly is r | Sonoma Brut | 86 | 20 | California | Sonoma C | Sonoma | | | Gloria Ferrer NV | Sparkling Blend | Gloria Ferrer |
| 129973 | | | | | | | | | | | | | | |
| 129974 | | | | | | | | | | | | | | |

➔ Price field having NULL values, extra white spaces and junk characters in text fields

| | | country | description | points | price | province | region_1 | region_ | title |
|---|---|---|---|---|---|---|---|---|---|
| 2 | 0 | Italy | Aromas include tropical fruit, broo | 87 | | Sicily & Sardin | Etna | | Nicosia 2013 VulkÃ  Bianco  (Etna) |
| 15 | 13 | Italy | This is dominated by oak and oak-( | 87 | | Sicily & Sardin | Etna | | Masseria Setteporte 2012 Rosso  (Etna) |
| 32 | 30 | France | Red cherry fruit comes laced with l | 86 | | Beaujolais | Beaujolais-Villages | | Domaine de la Madone 2012 Nouveau  (Beaujolais-Villages) |
| 33 | 31 | Italy | Merlot and Nero d'Avola form the | 86 | | Sicily & Sardin | Sicilia | | Duca di Salaparuta 2010 CalanÃ¬ca Nero d'Avola-Merlot Red (Sicilia) |
| 34 | 32 | Italy | Part of the extended CalanÃ¬ca se | 86 | | Sicily & Sardin | Sicilia | | Duca di Salaparuta 2011 CalanÃ¬ca Grillo-Viognier White (Sicilia) |
| 52 | 50 | Italy | This blend of Nero d'Avola and Syr | 86 | | Sicily & Sardin | Sicilia | | Viticultori Associati Canicatti 2008 Scialo Red (Sicilia) |
| 56 | 54 | Italy | A blend of Nero d'Avola and Nerell | 85 | | Sicily & Sardin | Sicilia | | Corvo 2010 Rosso Red (Sicilia) |
| 81 | 79 | Portugal | Grown on the sandy soil of Tejo, th | 86 | | Tejo | | | Adega Cooperativa do Cartaxo 2014 BridÃ£o Touriga Nacional (Tejo) |
| 139 | 137 | South Afri | This is great Chenin Blanc, wood fe | 90 | | Walker Bay | | | Beaumont 2005 Hope Marguerite Chenin Blanc (Walker Bay) |
| 161 | 159 | Italy | Intense aromas of ripe red berry, r | 91 | | Tuscany | Brunello di Montalcino | | Castello Romitorio 2011 Filo di Seta  (Brunello di Montalcino) |
| 165 | 163 | France | Produced from vineyards donated | 91 | | Beaujolais | Moulin-Ã -Vent | | Collin-Bourisset 2011 Hospices Civils de RomanÃ¨che Thurins  (Moulin-Ã -Vent) |
| 184 | 182 | Italy | Loaded with bold, ripe fruit, exotic | 88 | | Tuscany | Brunello di Montalcino | | Abbadia Ardenga 2003 M. Vigna  (Brunello di Montalcino) |
| 196 | 194 | Italy | Here's a traditional Chianti Classicc | 87 | | Tuscany | Chianti Classico | | Campomaggio 2005  Chianti Classico |
| 202 | 200 | Italy | Aromas of mature black-skinned b | 90 | | Tuscany | Brunello di Montalcino | | Tenuta di Sesta 2011 Riserva  (Brunello di Montalcino) |
| 224 | 222 | Italy | Chopped herb, forest floor, leather | 90 | | Tuscany | Brunello di Montalcino | | Talenti 2011 Trentennale  (Brunello di Montalcino) |
| 225 | 223 | Italy | Bright and creamy, this savory whi | 90 | | Northeastern | Alto Adige | | Terlan 2015 Pinot Bianco (Alto Adige) |
| 287 | 285 | Austria | This is a very aromatic wine that's | 92 | | Kremstal | | | Salomon-Undhof 2011 Steiner KÃ¶gl Erste Lage Riesling (Kremstal) |
| 290 | 288 | Austria | While it feels rich and round, this is | 92 | | Kremstal | | | Josef Schmid 2011 Kremser Gebling Erste Lage GrÃ¼ner Veltliner (Kremstal) |
| 292 | 290 | France | This is a wine that has great potent | 92 | | Bordeaux | Saint-EstÃ¨phe | | ChÃ¢teau Lafon-Rochet 2011  Saint-EstÃ¨phe |
| 293 | 291 | Italy | This powerful Sagrantino opens wi | 92 | | Central Italy | Sagrantino di Montefalco | | Dionigi 2006  Sagrantino di Montefalco |

## 5. Initial Plan for Phase-II

➔ The team discussed these data issues and came up with a draft plan as stated in the below steps:

- **Step 1(S1):** We'll use the winery-Kaggle dataset D that has a good collection of wine data and reviews to use for our target use case U1 of 'wedding wine selection where the wedding is being held in the US and the guests would prefer European wine'.

- **Step 2(S2):** We require the points, price, title, winery, description and territory details(country, province, region_1, region_2) from the dataset to get an idea of good-rated wine within a budget and a certain country we would like it from. We don't require the taster name and the twitter handle for this analysis. Alongside from the visual inspection & analysis we performed on the dataset, we are also planning to use 'Regex' to identify the syntactic errors and erroneous data patterns and 'Datalog' to identify the integrity constraint violations.

- **Step 3(S3):** To help with this arduous task of data cleaning, we are planning to use (i) **Regex** to identify the syntactic errors (special symbols, characters, white spaces), (ii) **OpenRefine** to clean the syntactic errors to transform the dataset. Since OpenRefine allows use of GREL, Regex and Python languages, we intend to leverage these options to perform our data cleaning. (iii) **Datalog** to check for integrity constraints like duplicate data & NULL values (iv) **SQLite** to fix semantic errors and update the records with values that can be determined. As the weeks goes by and we learn about other tools such as YesWorkflow, we would also explore possibilities to use this in our data cleaning exercise.

- **Step 4(S4):** Once the new dataset is ready, we want to compare this with the original dataset. We plan to use SQLite queries to do this 'before & after' comparison. For e.g., we plan to write queries for checking counts of NULL values and verify if duplicates have been removed from the dataset.

- **Step 5(S5):** As we understand, documentation of the steps will greatly help with this assignment evaluation. Hence, we are thinking of using the OpenRefine JSON log to capture the cleaning steps that will occur in OpenRefine tool and for the rest of the items, we can prepare a document summarizing the steps and capturing screenshots of the results.

➔ Steps S1, S2, S4 and S5 will be a joint effort as we have a weekly meeting cadence and we would address the work during that time. We plan to dedicate 2-3 hrs per week till the week of July 24 and finish the project.

➔ For the tasks listed in step S3, Alvin will be responsible for the Regex and Datalog work; Naveen will be responsible for the OpenRefine tool usage and Hemantika will own the SQLite work.