

**Team Tomahawks Project 1**

This document contains our project objective for the following:

* Team Efforts
* Project Proposal
* Data Exploration
* Data Cleaning Processes

**## Team Effort =** (.group({Rob Chesser : Michael Bell : Adam Durar: Regina Foster})).style

**\* [ ]** Describe the core message or hypothesis for your project.

Topic Selection

The group decided on the topic of “Wine” in general with curiosity to know more about particular aspects of the wine industry such as: consumer preference, flavor ranking, production locations, and industry sales to assess average cost per bottle in relation to preference, quality, and winery location.

**## Project Proposal**

**\*[ ]** Describe the questions you and your group found interesting, and what motivated you to answer them

Concept on “How to Analyze Topic”

Each team member posed a question as to their interest on the topic of “Wine” then conducted research via internet sites for data available on their question about “wine”.

|  |  |
| --- | --- |
| **Original Questions:**   1. What country produces the best wine?   1a. For US producers, what state has the best wine?  1b. If you are going to choose a wine by county/state, which has your best chance  of being a “good”wine?  1c. Where does Missouri rate?   1. Which wine variety do ‘Consumers most prefer’? 2. Which state sells the most wine? 3. What is the average price per bottle?   (Average High Rank and Average Low Rank)  5. Of the top 5 higest possible wine scores (96-100) from WineEnthusiast, what are the top 10 overall wines that offer the best "bang for your buck" based on rating vs. price? | **Data Found**  1a. Wine Scores by State and Country  2a. Wine Variety Ranked by Sales Volume  3a. Wine Producers by State  3c. Annual Wine Sales by Industry  4a. Annual Wine Production by Industry  5. WineEnthusiast Wine Ratings |
| **Revised Questions:**   1. Which state produces the highest scored wine?   1b. If you are going to choose a wine by county/state, which has your best chance of being a “good” wine?  1c. Where does Missouri rate?   1. Which state produces the ‘most’ wine? 2. How much wine is sold each year?   3b. Which state sells the most wine per year?   1. What would the average price of wine be   based on the data available?  5. Of the top 5 higest possible wine scores (96-100) from WineEnthusiast, what are the top 10 overall wines that offer the best "bang for your buck" based on rating vs. price?  5(b). What are the top "Bang for Buck" (rating - price) wines from WineEnthusiast that scored ratings between 97-100? | **Reason for Change**   1. Consumer preference data not found feasible for use in timeframe allotted for completion of Project 1 – scores reflect taste ratings   \*1a. ‘State Wine Score’ data offers a top ranking by states   1. Consumer preference data not found to determine a classification of “best” thus analysis of volume data was implemented 2. Data found contained the ‘industry of wine sales’ without defined ‘state sales’.   \*3b. ‘Wine Producers per State’ will be used to get an average of ‘state sales’ based on the percentage of ‘producers per state’ divided by ‘annual sales’ and production volumes.   1. Data found can formulate ‘Average Price per Bottle’ based on ‘U.S. Wine Production’ and ‘U.S Wine Sales’ volumes   5. Wanted to also look at the best scoring “Bang For Buck” wines that had ratings above 96 as my original top 10 list only included wines with a 96 rating. |

**## Finding Data**

**\* [ ]** Summarize where and how you found the data you used to answer these questions

Date collected was gathered from the: Kaggle, Inc a dataset repository site, Government Statistics sites – TTB (Tax and Trade Bureau) and the Census Bureau- NAICS (North American Industry Classification System , and industry marketing data sites - The Nelson Group, The Wine Institute and the OECD iLibrary.

Request for access to data via API keys or downloadable documents were made to industry marketing data sites which are paid access databases and/or member publication sites. Inquires were not responded to within the timeframe of Project 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Question #** | **FileName** | **SourceName** | **URL** |
| 1 | winemag\_data\_v2.csv | Kaggle/Firenze11 | <https://www.kaggle.com/lezhili/most-common-wine-scores>  <https://www.kaggle.com/zynicide/wine-reviews> |
| 2 | top\_wine\_variety\_by\_sales\_volume.csv | Wine Institute | <https://www.wineinstitute.org/resources/pressroom/06242019> |
| 3 | [q2-bonded-wine-producers-by-state-1999-june-2019.csv](http://localhost:8888/edit/KUProject/tomahawks/Data/q2-bonded-wine-producers-by-state-1999-june-2019.csv)  #\*\*State Liquor Sales Not Feasible\*\*  [wine\_sales\_2018.csv](http://localhost:8888/edit/KUProject/tomahawks/Data/wine_sales_2018.csv)  wine\_sales\_2018\_cali\_only .csv | TTB Tableau data source  NAICS - Manufacturing and Trade Inventories and Sales report  Wine Institute | <https://ttb.gov/wine/wine-stats.shtml>    <https://www.census.gov/mtis/index.html>  <https://www.wineinstitute.org/resources/pressroom/06242019> |
| 4 | Wine\_Production\_By\_State\_2017.csv | TTB Tableau data source | <https://www.ttb.gov/statistics/2017/final17wine.pdf> |

**Conclusion of Data Exploration and Cleaning**

The original topic we chose was broad, then streamlined based on data available to compare wine scores and number of producers by states in comparison to sales by variety type demonstrating the bang for your buck bottle of wine.

Data retrieved from Kaggle/Firenze11, Wine magazine – ~129K entries presented challenges. There were many duplicate entries which required removal of unnecessary columns such as twitter handle, taster name, title, description, blanks rows, and drop null.

The second set of data was revised and the number samples per country/state were not consistent. Several scores were listed as “America” with wine names spread across columns. Two sets of data were created in dataframes for world data and US specific data. The cleaned data was used to calculate the average wine score for each country producing wine and graphed by country vs wine score. US data was extracted from the World data and the average wine score for each state was compared and graphed.

Clean up of a .csv file downloaded from the U. S. Census Bureau statistics for licenced wine producers in the US required removal of text headers and footers along with specific data extraction for plotting. The final data set demonstrated 'California' ranking highest with 5, 217 producers, Wyoming ranked second with 1, 171 and Oregon ranked third with 758 wine producers in 2018 .

Data entry of sales reported by the Wine Institute, Jun 24, 2019 press release ‘California Wine Sales in U.S. Market Hit $40 Billion in 2018’ presented two sources of data to gather Wine Variety Sales and National Wine Sales by State. Data was prepared uniform for plotting.

\*\*Please add any remaining details and conclusions and remove this message before submission\*\*

See Jupyter Notebook : Final Analysis