

**# Guidelines for Project 1**

This document contains our project objective for the following:

* Team Efforts
* Project Proposal
* Data Exploration
* Data Cleaning Processes
* Data Analysis in Jupyter Notebook as code comments and summary after image

**## 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, highest rated wine with the lowest price (best value), 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) | **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 |
| **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. What wines give you the best overall value based on their rating vs. their price? | **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. Created metric to analyze value for the top rated wines (96-100 ratings) by taking the rating minus the price (called this “Bank For Buck Score”) |

**## 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\_all\_row.csv  winemag\_reviews\_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> |

,

**## Data Cleanup**

**\* [ ] Describe the data exploration and clean-up process (accompanied by your Jupyter Notebook)**

**Plot 1 – Wines Scores by State and Country**

1. Removed unnecessary columns such as twitter handle, taster name, title, description
2. Removed blank rows via dropna
3. Two sets of data were created in dataframes for world data and US specific data.
4. Analysis
5. The cleaned data was used to calculate the average wine score for each country producing wine and graphed by country vs wine score.
6. US data was extracted from the World data and the average wine score for each state was compared and graphed.
7. Issues/concerns
8. We had duplicate entries. The second set of data was revised
9. The numbers samples per country/state were not consistent
10. Several scores were listed as “America” instead of individual states.
11. Specific wine label names were spread across two columns

**Plot 2 – Wine Variety by Sales Volume**

1. Add problems with coding

**Plot 3 – Wines Sales by State Producers**

1. Add problems with coding

**Plot 4 – Average Price of Wine per Bottle**

1. Add problems with coding

**Conclusion of Data Exploration and Cleaning**

The original topic we chose was broad, then streamlined based on the availability of data found feasible for usage within the timeline for project completion.

\*\*\*Elaboration needed

Jupyter Notebook for Final Analysis