DATA VISUALIZATION PROJECT

1. SUMMARY:

It is commonly said that “Wine gets better with age “. In order to answer the question: what makes a wine great, I did an exploratory data analysis on wine ratings dataset. This included effects of 11 different chemical properties. Some interesting relations were observed in this study. This visualization focuses on one such relationship. D3.js is used for this project.

Relation: Effect of volatile acidity, citrus acid & alcohol on red wine quality.

2. VISUALIZATION DESCRIPTION:

A scatterplot is selected as the most effective graph type for this particular relation. The motivation for this decision was the following graph created in the data analysis:

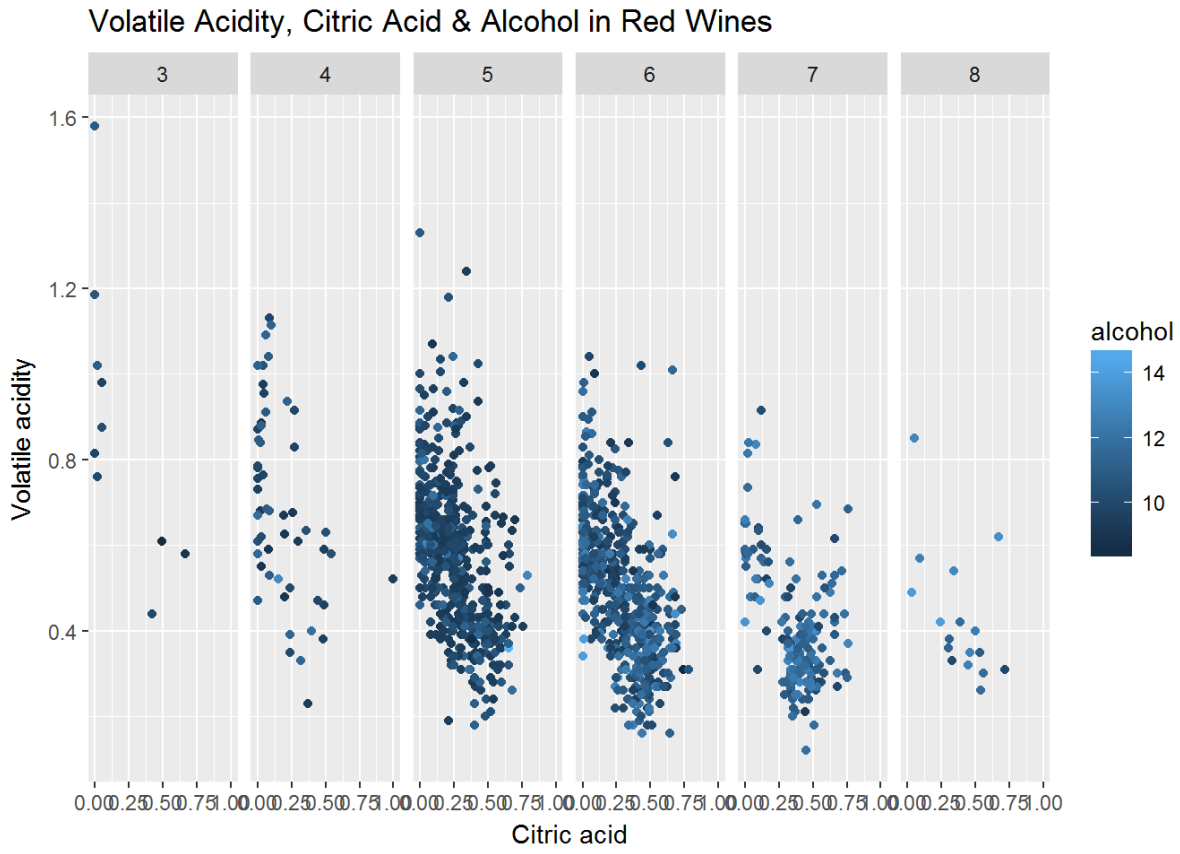

Figure 1: Graph obtained from data analysis

The objective for this project was to capture similar relation, but with some interaction. From above image, we can see that both these acids are negatively correlated to each other. (cor: -0.552) Upon further inspection of these two properties, we come to know that citric acid brings freshness while volatile acid (primarily acetic acid) provides unpleasant taste to wines. Some kind of trend between alcohol and ratings is also seen. (More alcohol in higher quality wine)

For creating such visualization, a basic scatterplot was created first. ('project1.html') Feedbacks from people were taken as to how to 'evolve' this into interactive visualization on general level. After each feedback, the plot was updated and finally an interactive plot is achieved. ('project_final.html') The next section will follow through the entire development process for the same.

3. DESIGN PROCESS:

A. Basic plot (Initial Design):

A basic scatterplot is constructed with volatile acidity on y axis and citric acid on x axis. The data points are represented as circles with their radius indicating alcohol level. The image for this plot ('project1.html') is given below:

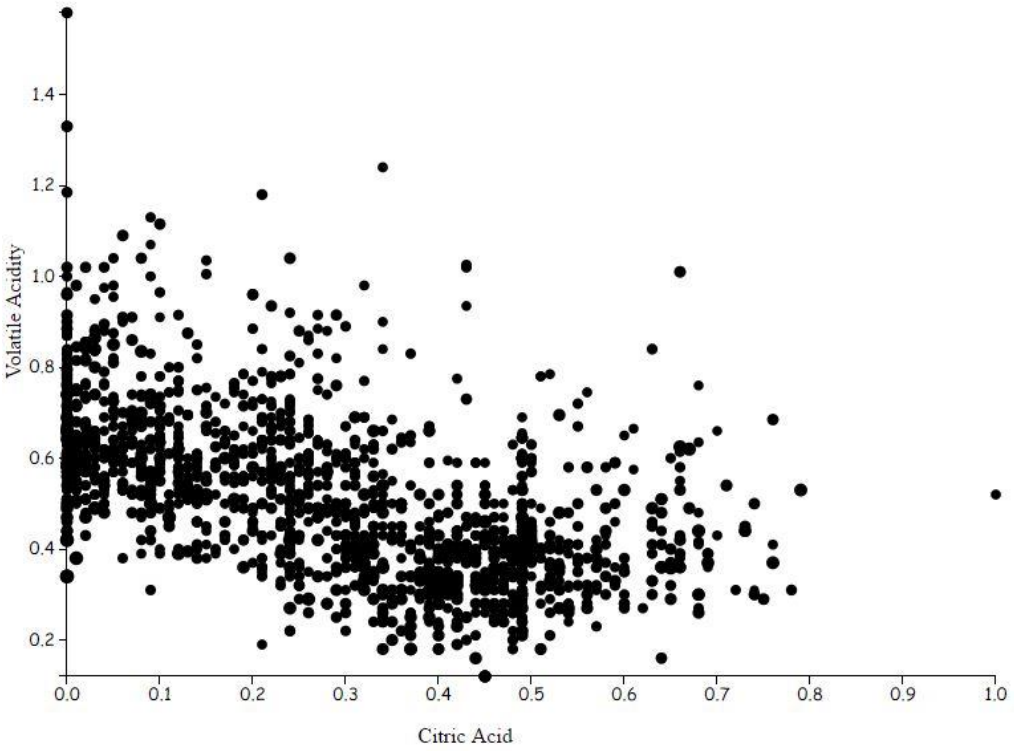

Figure 2: Basic Level Plot

B. Second Plot (Mid-Level Design):

Based on this plot, first and second feedback was received. The summary of this observation is given below. Plot constructed after improvement is shown at the end. ('project2.html')

Feedback 1:

Since the data contains values for alcohol level between 8 to 14, the radius scale defined by this factor does not add any substantial visual effect. The alcohol values represent almost the same radius size for each data point.

Action Taken:

Scatter plot for volatile acidity vs citric acid is constructed, keeping radius size constant for each data point.

Feedback 2:

As the radius size is not indicative of alcohol level, colour fill property can be utilized to incorporate this factor. Colour plot will bring more effect to this visualization. Also, the plot has large number of data points. Hence, overlapping is taking place. Using opacity for these colour circles will be better.

Action Taken:

Continuous colour scale is implemented to represent alcohol level factor. The reason for continuous scale as opposed to discrete one is that the values of alcohol are not discrete (like 8,9). So, intermediate values (like 9.2, 10.6) can be represented accurately. (Relatively) Opacity of 0.8 factor is introduced for better visualization.

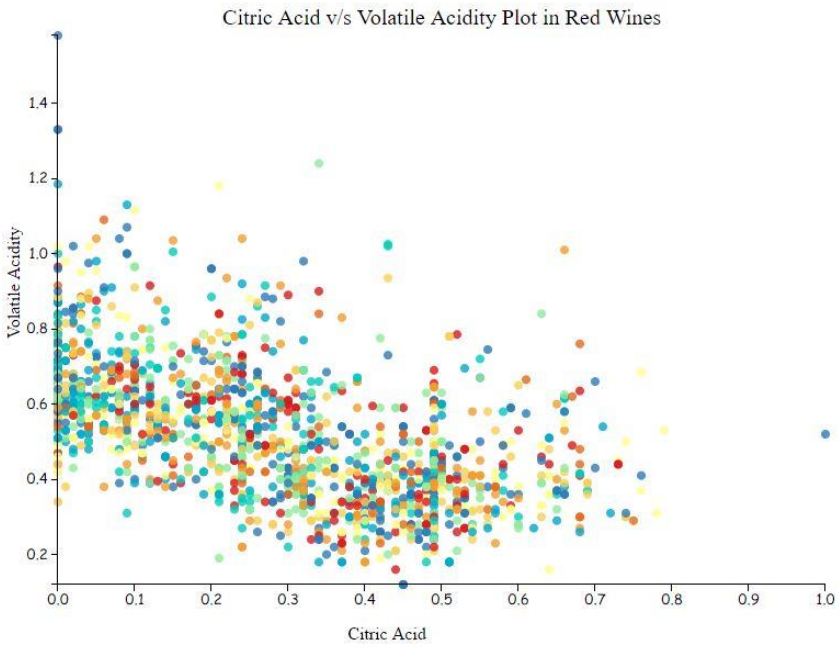

Figure 3: Mid-Level Plot

C. Third Plot (High-Level Design):

The second plot provides some level of understanding between the relation of citric acid, volatile acidity and alcohol, as found in the data analysis. For this plot, feedback 3 was received. Plot based on this observation is given below.

Feedback 3:

Similar to the first plot, some kind of facet wrap can be applied which will divide the data into different quality groups. This will provide clear observation of data according to the quality and some kind of interactivity to the user.

Action Taken:

In order to group the data in different quality, a filter function is incorporated which selects the data points according to required quality value. These different groups can be viewed by clicking on appropriate buttons. So, data set for each quality group can be viewed and toggled separately.

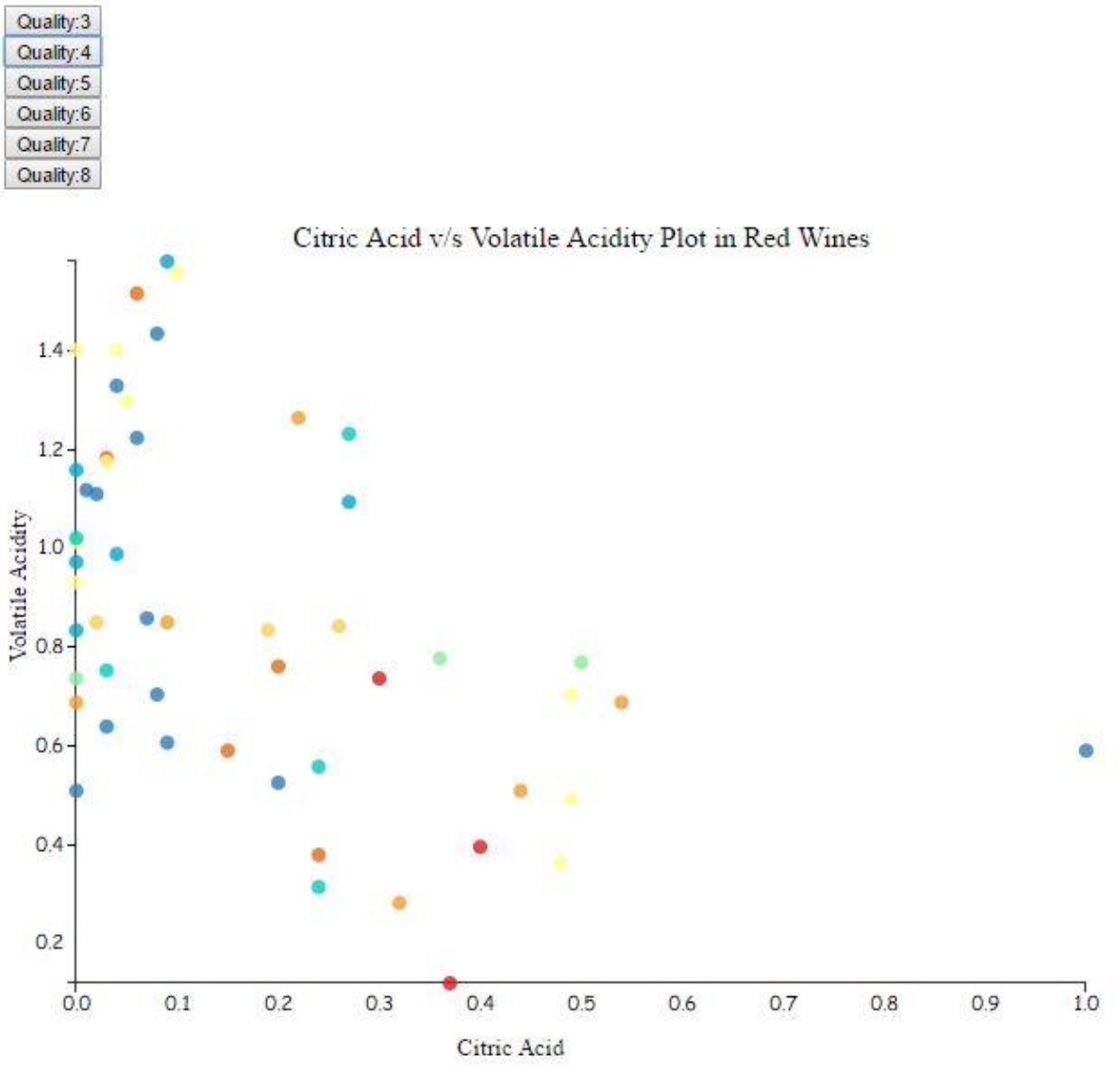

Figure 4: High Level Plot for Quality 4

D. Fourth Plot (Final Design):

Some final modifications are done to make the visualization more eye pleasing. ('project_final.html') Minor improvements are carried out. These are explained below.

Modifications:

1. Position of buttons is changed to bottom of the plot.
2. Dynamic title is added, which changes according to the button selected. This provides the user idea about which group he is viewing.
3. A continuous colour gradient legend is added with range of values for alcohol level.
4. It was observed that as the button was clicked, the range for the x and y axis changed accordingly. To compare the groups relatively, another function to set the domain of these axes is defined. So, when the page is loaded initially, the range for axis are defined for whole data individually. This ensures that the user can see all groups and compare them relatively.
5. Transition effect is introduced when toggling from one group to other so that the data points move smoothly to new location when a button is clicked.
6. Description for axis and alcohol factor is updated by including appropriate units.

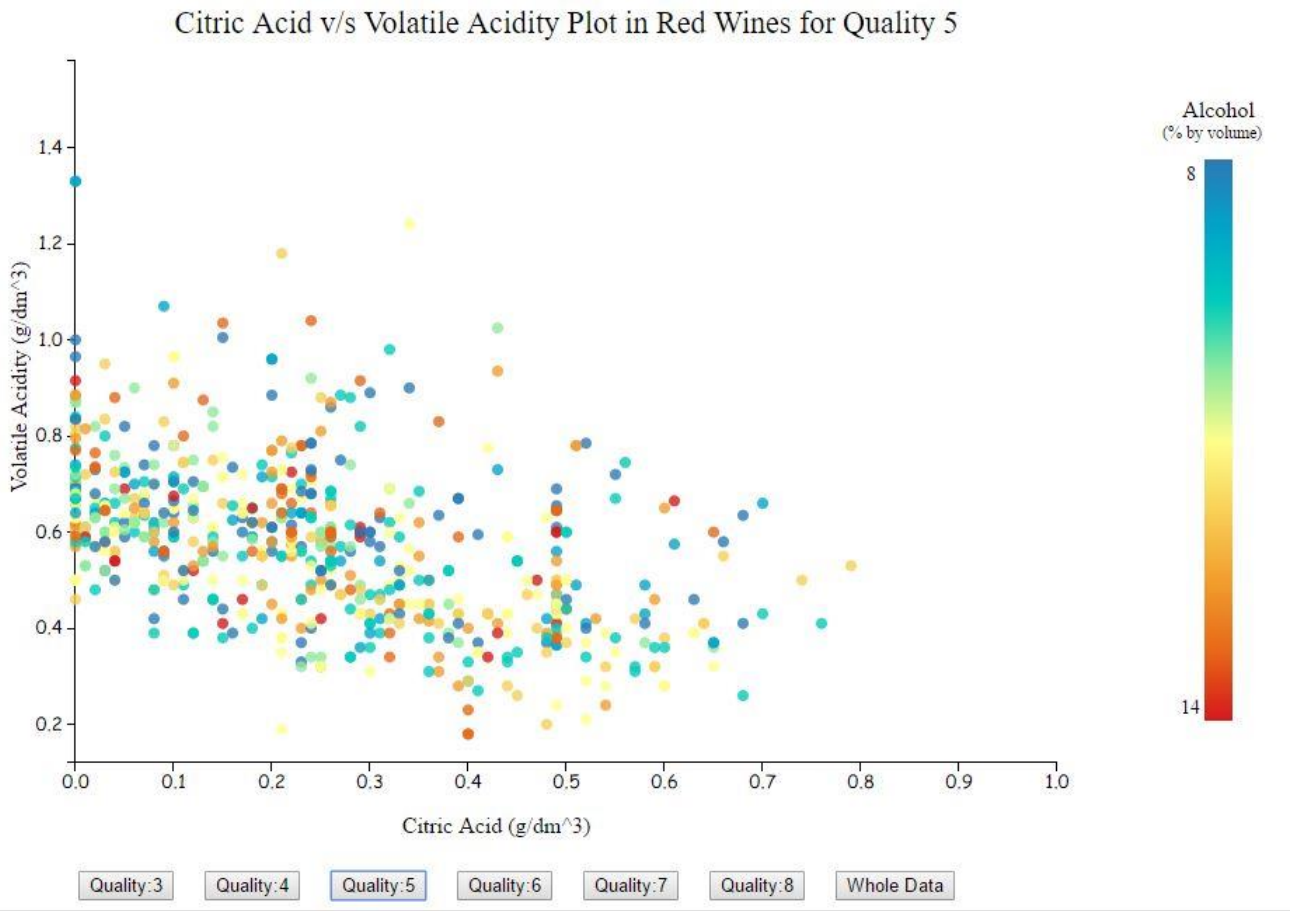

Figure 5: Final Visualization

4. FINAL COMMENTS:

As mentioned in the course, a good visualization must consist of two parts:

- Author driven plot
- Interactive part for user to explore

For the author driven part, a colour scatter plot is made. In this plot, one clearly sees how citric acid and volatile acidity are negatively correlated. Another observation is that wines with high amount of alcohol (red colour) prefer relatively more citrus acid and less volatile acidity. This relation is developed and conveyed to the user in the effective way.

The interactive part of the visualization consists of toggling the scatter plot with specific groups. (quality/rating of wine) **Quality range is 1-10. (1: Bad & 10: Excellent)** The data set contains values for ratings from 3 to 8 only. The user can explore further some relations between different quality groups and make their own observations. Some of the trends which can be explored are for example:

- Relation between quality and alcohol
- Disproportionate amount of data points between quality groups

Project report for the analysis of red wines using R software is also included as reference. ('project.html') Code for this visualization can be inspected using appropriate editor. Required comments are included in the code.

5. REFERENCES:

- Red wine exploratory data analysis project using R
- "Introduction to D3" by Curran Kelleher (Video Lecture)
- "Boost D3.js charts with SVG gradients" by Nadieh Bremer (Online Blog)