INFO3300 P2: Final Writeup

Adam Liu, April Ye, Cindy Zhang

Wine Reviews Analysis

a) Data Description

For this project, we have four datasets. We used two wine reviews datasets on wine that we found on Kaggle (https://www.kaggle.com/zynicide/wine-reviews). Specifically, we have 150k entries of data on the country, designation, points, price, province, region, variety, winery, and description on different wines, and a separate 130k entries of wine reviews with the taster name and taster's twitter handle data. These two datasets share some of the same fields but different selections of wine and wine reviews. We have another dataset associating every country with its respective country code (https://github.com/d3/d3.github.com/blob/master/world-110m.v1.json?fbclid=IwAR1DGwGLtL6W9wEF4wQPsGq5fbSHO_Tzp5-t1V7jmN5kcWPtIHQYXo76CPs). And lastly, we have a dataset mapping the shape of each country on a world map, the world_110m.json file that was given in INFO3300 homework 5. We filtered out data that had empty values for the following fields: descriptions, prices, variety names, provinces, taster names, and taster twitter handles. We also filtered out repetitive entries on the same wine. Our rationale for filtering out repetitive entries for the same wine is to make our graphs more easily interpretable, given the large amount of data points. When working with plotting data for both datasets, we used the same country code in each dataset for association.

b) Visual Design Rationale

Graph 1:

[Overview]

Our first graph is a circular graph mapping the top 10 most popular wine varieties on the right side to their regions on the left side. When hovering over a wine variety on the right side of the circle, the user can see the lines that are drawn connecting the variety to all the regions that grow this specific grape. Each wine variety has a corresponding shade of red, matching the color of the wine. An image of the wine bottle is also shown on the right side. We chose to have

the wine varieties appear on the right side of the circle because the focus of this graph is the variety, and the right side of the circle is in the center of the web page.

The purpose of the graph is to give a brief introduction and educate the audience about the different wine varieties in our data visualization. At first glance, the audience can judge from the density of the lines and learn about how widely a specific wine variety grows. They can also see which region is suitable for growing many different types of grapes.

[Trade-offs]

One tradeoff of the circular graph is that a few of the colors are similar because wines generally tend to have similar shades. However, we expect the user to hover over each wine variety which will remove any of the other similarly colored lines that might hinder the user's ability to see the origins of the wine. We thought it was more important to have the user really associate the arc colors with the wines' true colors rather than have very discernable colors that don't relate to the wines at all.

Another tradeoff of the circular graph is that it is more difficult to read the detailed information (which regions have a particular wine variety), but it is easier to get a good overall understanding of the distribution of wine varieties and regions.

Also, we chose to use the mouseover interactive element rather than mouse click elements because we felt that it would make the user's experience much more fluid and smooth. It would also allow the user to just get a quick preview of the wines before taking a look at the more data-heavy graphs afterwards.

[Mapping]

Data on wine varieties are represented by the colored arcs with labels on the right half of the circle. Data on regions are represented by points with labels on the left half of the circle. The marks are the arcs representing wine varieties, points representing regions in the world, and lines connecting the two. The channels are the colors indicating the different varieties of wines and the position of the lines leading to the different regions.

Graph 2:

[Overview]

Our second graph utilizes a scatterplot graph to show the rating plotted against the for the different reviewed bottles of the top 10 most popular wine varieties. When filtering via

clicking on each button (that has the name of its corresponding wine variety filter), all the points respective to this specific wine variety are shown on the graph. When hovering over any point on the scatterplot, a more detailed description of the wine is shown, including the corresponding reviewer, rating, price, variety, and region.

We used a scatterplot to be able to visualize the distribution of wine reviews relative to its price and rating, filterable by a specific type of wine. We chose to color code the different points on the graph according to their unique wine variety. This specific color encoding is actually carried from the first graph to tie in an element of uniformity across our entire data visualization; that way, the audience can be able to associate a specific wine variety and its features/performance with its unique color throughout our visualization. On hover over a specific data point, the point will dilate and become a much lighter whiter color for the user to be able to more easily pinpoint which point they are viewing against our visualization's dark background. This matching whiter color is used for the hover card that displays more details related to the hovered data point.

[Trade-offs]

We made the choice to only include the top ten most popular wine varieties due to the sheer overwhelming amount of data points and few possibilities of ratings–as a result, this scatterplot graph initially displayed repetitive vertical lines that were made from the opaqueness and density of the plotted circles. We realized that especially for a scatterplot, these vertical lines would be very difficult to interpret and also hard to hover over to learn more about distinct points. Though this made our graph a lot more visually interpretable, there are some more data points from less popular wine varieties that could affect the presented visualization in terms of outliers and opaqueness of points (from point density).

Additionally, this graph's y axis tick labels range from $5 to $2k. Though there are currently no data points displayed that represent wines with the price tag of $2k, we did want to illustrate the fact that our dataset contained wines within this range and the possibility of a much more expensive wine. Our data displayed currently only encompasses the top 10 wine varieties, in which no wines cost $2k. This was a design choice and trade-off that we chose in order to represent the outlier data points of less popular wines.

We also debated over changing the hover color of the filtering buttons according to the color code of their respective wine variety. However, doing so made the visualization look a bit loud and tacky, so we decided on a sleeker gray to make the buttons more uniform and

embrace a more low-key appearance. We wanted the focus to be on the scatterplot graph, so we wanted to introduce as little distraction towards the graph as possible.

[Mapping]

We used a log scale for the prices to account for the natural distribution of the prices of wines. Wine rating is plotted on the x axis, and prices on the y axis; data was plotted with respect to their wine rating and price, so that each point represented the review of a specific wine. Each button represented a top 10 (according to popularity) wine variety. We used circle points as our marks, and color, position, and opacity as our visual channel.

Graph 3:

[Overview]

Our third graph is a world map with a slider. On the slider, there are 10 ticks representing the top 10 most popular wine varieties. On the map, the corresponding countries where the grapes are grown are highlighted. The audience can learn more about each wine variety and get a better sense of where in the world they came from by looking at the map.

[Trade-offs]

One tradeoff of the world map visualization was the way our slider was labeled. We really wanted to emphasize the tick that the user was currently on. In order to show a large difference, we made the text of the other wine varieties small and printed out the selected wine variety on the left of the world map in a large font with a bright magenta color to match with the highlighted countries. We felt that this added another aspect to our data visualization because it makes a direct color connection between the wine variety and the countries and also makes the wine variety stand out to the user.

We made a choice to color countries instead of specific towns/provinces that grow the grapes because we believe that the information would be interpreted much easier on a world map. We also decided to cut Antarctica out of the visualization because no grapes grow there, so the user would not miss out on any crucial information. In addition, this allowed for a much better viewing experience because the entire world and slider would fit on the screen much better.

Another tradeoff we made was the bright magenta coloring of the countries. Although the previous two graphs had similar, more mild colors to imitate the actual colors of the wines, we

decided to make the coloring of this graph much more bright and vibrant. This would ensure that even smaller countries/land masses would be able to stand out enough for the user to notice. With more mild colors, it's very easy to miss some of the countries with smaller areas of land.

[Mapping]

The ten wine varieties are shown as ticks on the slider. The magenta color of the countries on the map indicate whether the grapes are grown there or not. At each tick on the slider, a different set of countries will be highlighted with a magenta color which indicates that the grapes are grown in that country. The marks are the countries, and the channels are the colors and the positioning of the countries.

c) Interactive Elements and Design Rationale

Graph 1:

The circle graph acts as an introduction to our entire data visualization, so we thought it would be ideal to use a mouseover/hovering interactive element. We considered using mouse clicks but felt that hovering over the wine varieties offered the user a much smoother and easier experience and allows the user to just get a quick preview of the wines before moving onto our second graph. When hovering over a particular wine variety, an image of the wine appears and the right side of the web page so that we can allow the user to get a mental image and establish a connection between the arcs of the circle and the wines that they represent. In addition, we know that the nature of the graph we made makes it very difficult to see and trace all the lines in the middle of the circle. Therefore, when the user hovers over a wine variety, we removed the lines of all other wine varieties on the left side, so the user can focus on the regions of the wine that the user chose to hover over.

In addition, we added a small transition to enlarge the arcs as the user hovers over them. This gives the user feedback as they hover over the different wine varieties, so they can make sure that they're hovering over the intended arc.

In order to indicate to the user that we want them to hover over the wine variety arcs on the right side of the circle, we added a caption to instruct the user to "hover over the wine varieties on the right to see their origins."

Graph 2:

We wanted to use a scatterplot to display our data points to showcase correlation between a wine's rating and price. To introduce interactivity, we decided that it would be really interesting to be able to hover over a data point plotted on the graph and be able to see more details about it (reviewer, rating, province of origin, and price). We thought that this hovering functionality humanizes our data visualization because each data point correlated to a specific reviewer, making our data portrayal a story consisting of different people's opinions and critiques rather than simply a plotting of wine rating against prices. In order to prompt the user to hover over the data points, we displayed a small caption to do so underneath the header of this second graph.

Due to the sheer amount of data points, we also decided to allow users to interactively view data points based on the wine variety they selected. We represented the top 10 most popular wine varieties (that we filtered through and calculated programmatically) via buttons that the user can interact with and see the change in data plotting on the scatterplot graph. This way, the user will be able to see the distribution of not only all wine reviews as a whole, but also the unique skews and performance of a certain wine variety. They will also then be able to more easily view a specific wine and distinguish its general position in terms of price and rating. Since the buttons changed on hover and that it is usually intuitive for users to click on buttons for web interaction, we decided that a second prompting caption (that was used for hovering) would be redundant.

Graph 3:

To visualize the wines from the wine reviews geographically, we used a world map. We wanted to see the countries specifically from where each wine could be sourced from. With a world map, it would be a lot easier to correlate a wine's origination with different parts of the globe, and would be easier to notice patterns of sunlight or lack there of depending on the country of origination's position relative to the equator. Countries that the wine is sourced from are shaded in a bright magenta.

We thought that it would be very useful and interesting to learn where each of the top 10 grape varieties are sourced from. To do so, we introduced an interactive slider with tick marks corresponding to each of the grape varieties. Upon sliding through the slider (which snapped to the nearest tick mark), the world map's magenta coloring of countries will change according to the list of countries that the specific grape variety is from. This level of interactivity would allow the user to be able to differentiate which countries correspond to a specific grape variety. We

also displayed the current grape variety the slider was slided to in bright magenta lettering under the world map that changes as the user interacts with the slider. We thought it might be helpful to hide the tick label on the slider when it was toggled to in order to emphasize the user's current position on the slider. Lastly, to make it more intuitive for users to toggle through the slider instead of dragging, we included a brief caption under our graph header to prompt the user to click through the slider to find out more of each grape variety.

d) The Story

Overall:

In order to connect our entire data visualization across graphs, we used the same colors for each wine variety. This also ensured that the user would be able to associate each type of grape with its true color and give an overall sense of unity to the whole data visualization. We also analyzed the same top 10 most popular wine varieties and displayed data corresponding to these wine varieties throughout our data visualization.

Graph 1:

This visualization gives us a holistic view of the top ten wine varieties, where they come from, and what they look like. When we first started creating this graph, we didn't realize these different types of grapes could come from so many different places and environments around the world. While some grapes do have more limited options for regions to grow in, we are able to see that these grapes are still able to adapt and survive in many different spots around the world. We also found it surprising that the colors of the wines were so different, varying from shades of yellow to extremely dark blues. With this visualization, we wanted to give the viewer a mental image of the varying grapes and wines that exist and convey that there are numerous different types of grapes that come from so many different regions of the world.

Graph 2:

Though we initially thought that wines would usually be in the same price range, our visualization shows us that different wines could range from a wide span of price and also garner a wide range of ratings. For example, Chardonnay ranges from the lowest price of $7 to the highest of $800, with ratings from 80 to 99 on a scale of 80 to 100. Looking at our visualization, we do indeed see a general correlation between a wine's price and rating–the more expensive the wine, the higher the rating. Wine varieties such as Sauvignon Blanc and

Chardonnay clearly illustrate this correlation. It was also interesting to see that most of the reviews comprised of the same reviewers. While we initially thought that with 130,000 reviews the ratings would be diversified, hovering over data points on our scatterplot does seem to showcase that the same reviewers are extremely enthusiastic about rating wines, and tend to monopolize the ratings for a specific wine variety. This could raise question of potential bias in the ratings, which the user can witness and contemplate upon hovering over data points in a certain wine variety category in the scatterplot.

Graph 3:

While we initially thought that each wine variety would be sourced from a distinct country or general geographic location, our world map visualization shows us that at least for the top 10 most popular wine varieties, the countries of origins have many overlaps. The most often recurring countries/regions are North America, Chile, South Africa, and Europe. This makes sense upon reflection, since many of the wines most known to us today are usually sourced from North America and Europe. Because we had decided to focus only on the top 10 most popular wine varieties, our visualization tells users that the most popular wines are usually from these repeating areas. This could be very helpful for our audience, so that in the future, they are more aware of where the most popular wines are from and can either visit these wineries to taste test in person or purchase wines from these general areas.

Team Contributions:

Adam:

I helped brainstorm different ideas for the visualizations and met with the group for TA meetings and group work sessions. I implemented the circle graph visualization which included designing the arcs and making sure the labels and points followed the same path. I had a difficult time figuring out how to match each point to the arcs and make the hovering interactive feature work with those matchings. After completing the circle visualization, I helped develop the logic to properly filter the data for the world map visualization and match each wine variety to the appropriate set of countries. I started right before spring break and spent around 30 hours working on this project.

April:

I contributed towards the brainstorming and planning of all 3 visualizations. I also met up alongside my teammates for every TA check-in meeting. My main contribution was creating the second visualization and helping initially set up and later further implement the slider for the third visualization. The zooming implementation (that was later remove due to too much interactivity that would make the second graph hard to navigate) and styling the buttons took a weirdly unnecessary but nonetheless a huge chunk of time. I also struggled a lot with figuring out how to display the data in the scatterplot due to the immense amount of data points. Towards the end of the project, figuring out how to style the slider took several hours on the last day of the project submission as well. I started during spring break, and overall took around 30+ hours.

Cindy:

I contributed towards the brainstorming and planning of all 3 visualizations. I also met up alongside my teammates for the TA check-in meetings. I worked alongside Adam to get the world map working and filtering logic. In addition, I had to find a country code dataset so that we could match the wine variety countries to the world map. I also worked with April for the slider implementation.