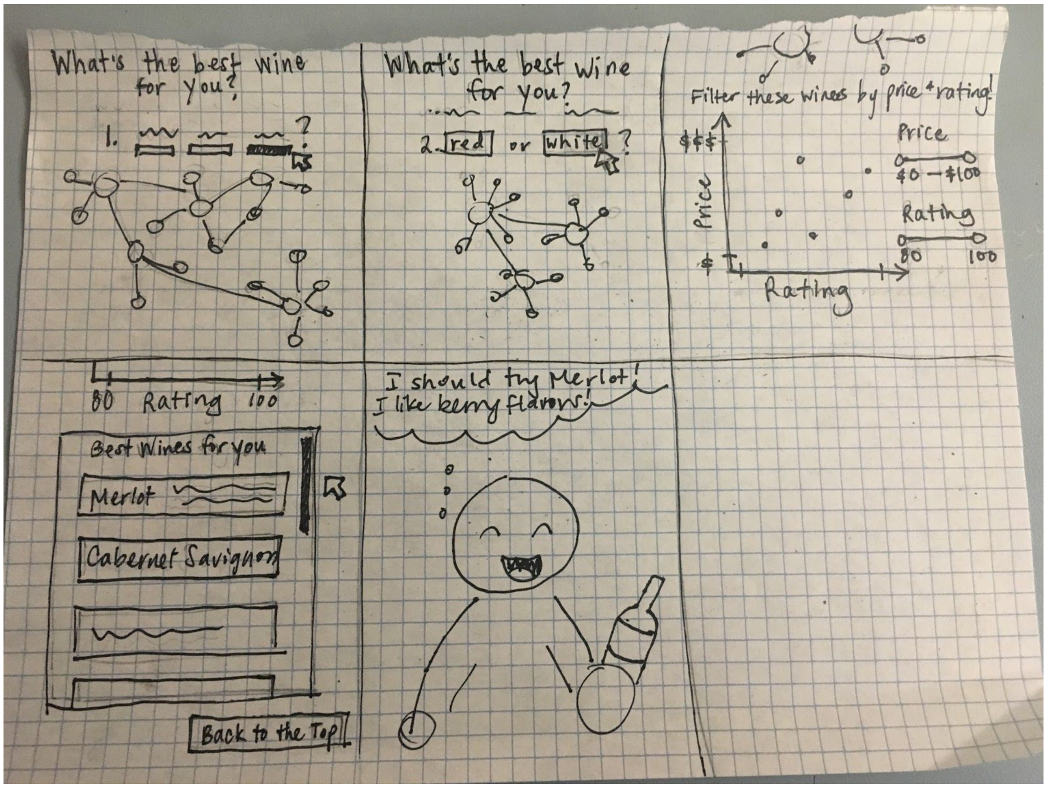
Charles Yu, Michelle Yang, Jonathan Lee
INFO 4310 A3 - Design Doc
Due: 03/15/2018

Describe the data you chose and identify specific insights/use cases for that data that will align with your chosen interactions.

The dataset we chose is a dataset of 150K wine reviews that we found via Kaggle. The format of data entries is quite simple: Each row entry is a collection of data about a specific bottle of wine including its varietal, country of origin, region, producer, price, a rating (as determined by Wine Enthusiast, a wine magazine), and a summary of the taster's impressions of wine.

To explore the data, we used Python Jupyter notebooks with the Pandas package. For initial exploration insights, we used Kaggle user amukho33's previous explorations of the dataset to guide us in our initial examinations. Like amukho33's, we decided to filtered out wines in the dataset that did not stem from one of the top-12 wine-sending countries. We also filtered out wines which were not within the top-100 most popular wine varietals by bottle count in our data set. In our exploratory analysis looked at relations between price and rating, which countries were sending out the most wines, and finally, what words were most salient in describing certain varietals of wine. Eventually, we settled on looking at the textual features of wine varietals and whether or not they uncovered any interesting insights. To do this, for each varietal, we used a Count-Vectorizer on all the of the description texts of that varietal and obtained a set of 100-top features (filtering out stop-words). We then created a matrix of varietals by varietals in which each cell (i, j) denotes the similarity of the wine varietal i and wine varietal j based on a simple jaccard similarity between the varietals' features. When we looked at the most similar varietals for a given query, we were surprised by how accurately these reported similarities aligned with how wine reviewers actually suggested to be similar (based on qualitative Google searches). We used this wine similarity matrix as the driving inspiration for our network of varietal nodes. To clean our data for this metaphor, we set a threshold of a jaccard similarity score of 0.45 (empirically chosen) if there is to be an edge between two varietals.

Provide storyboards that outline the interactions you will design for your dataset and justify why you are using those particular interactions (done in step 1)


What's the best wine for you?
1. ?
What's the best wine for you?
2. red or white ?
Filter these wines by price + rating!
$$$
Price
$
Rating
Price
$0 — $100
Rating
80
100
80 Rating 100
Best Wines for you
Merlot
Cabernet Savignon
Back to the Top
I should try Merlot! I like berry flavors!

Above is our storyboard. It read from left to right, top to bottom.

1. When the user first arrives at our data visualization, the data visualization would show the network of wine varietals and have a question for the user to answer. The question might prompt the user to type in or click some flavors that he/she likes. Once those are selected, the complete question fades out or up, the nodes that do not fit the criteria fade disappear, and the next question shows up.
2. The next question would be “Do you prefer red or white wine?” The user can click on “red” or “white” and the same animations occurs.
3. Once these questions about the network are completed, the visualization will automatically scroll down to a rating v. price scatterplot that contains all the wines still remaining from the above network. The user can use sliders to adjust their price range and rating.
4. As the user is adjusting the sliders, a box will show up below the scatterplot with all the wines that are best suited for your preferences based on what you filtered out in the network and the scatterplot. It would be displayed similar to search results, where the user can scroll through the wines. Each wine would be displayed in its own cell, which contains extra information about it.

5. The goal of the data visualization is to give users a resource where they can discover which wines best suit their tastes. Additionally, we hope the user can gain a high level sense of what kinds of wines are more similar to others (the more embedded nodes), and which ones are quite different (the outer, less connected nodes).

Briefly describe your final interactive visualization application.
Our final data visualization shows the user a network of wine varietals, all connected based on jaccard similarity of feature words used by wine tasters to describe varietals. In our data analysis, we found that red and white varietals naturally clustered into two large groups, and we chose color metaphors to align with this observation. Red wine varietals form their own network and these nodes are colored red. White wine varietals form its own network and are colored pale green. And wines that don't fall under either category are pink and dispersed outside of the networks to express their uniqueness. We specifically chose these colors to make our network of nodes look like bunches of grapes, a tongue-in-cheek visual metaphor.

Users are able to filter and narrow down the varietals shown via the filtering search bar by typing in keywords that describe their taste preferences. The user can then click reset to start over and reset their preference filters.

Hovering over any of the nodes will highlight the varietals that are most similar to the varietal of that node. It also causes all other non-related nodes to fade to grey to further increase the salience of the highlight. Clicking on any of the varietals will then reveal a new set of filters for price range and rating preferences. Once these are filled in (or not if they are not important to the user) the user can click "Pick a Bottle" to be given a bottle suggestion that best fits all of the user's preferences.

Step back and think about issues or trade-offs associated with the interactions you developed, and how you might alleviate those (or whether they are unavoidable).
We originally had the network of wine varieties interactable via dragging the points around to let the user unravel the red wines from white wines in case the network generated had the networks overlapping. The tradeoff that arises from this design is that it assumes the user would understand that the network points are clickable and draggable. We might have been able to alleviate this through the codebase by making sure that the networks generated would never overlap.

Another area that could have been better executed is how the visualization treats the action of clicking on individual nodes. More specifically, a user's intention of clicking may be as a means to drag the nodes around. However, our visualization currently treats that click the same as a click by itself which keeps the highlighted network of similar wine varieties highlighted and reveals additional filters for users to input their price and points preferences. We could have alleviated these issues by dividing those two functionalities into different modes. For example, there could be a toggle that makes it clear how a click will be treated differently. One side of the toggle could indicate it's in a mode that allows the user to move the network around. The other side of the toggle could switch to a mode that indicates its used to narrow down on the wine bottle you want. That way the user's expectations align more accurately with the actual functionality of the visualization.

Finally, we keep the price and rating filters hidden until the user clicks on a variety of wine. This was intentional because we wanted to keep the user's attention focused on one task at a time (narrowing down the wine varieties based on taste preferences before considering the price and rating range). Our hypothesis was that users would not care about price or rating if the wine variety didn't even offer a bottle that would fit their taste preferences. There are two main trade offs with this approach. The first is that it assumes all users care about wine attributes the same way when in fact some people might not care about taste and care more about price and rating of a wine. We could have better addressed this by letting the user adjust all of the filters at the same time before picking a variety that fits their taste preference. The other tradeoff of our approach is that the appearance of filters are not as salient. As a result, they could be overlooked. We could have alleviated this by either putting them all upfront (as just proposed) or adding an animation or increasing the contrast of these filters to better highlight the appearance of these new features upon click.

Briefly outline the development process of your tool.

- Brainstorming session to bounce ideas
- Data exploration of initial dataset and follow up brainstorm while interactively looking at data
- Cleaning the data in Python Notebooks for export to d3
- Building the basic wine varietal network in d3
- Adding animation features & filtering features to the network
- Experimenting with visual and design choices and getting feedback from friends
- Finalizing designs & features

Explain how your visualization/interactions changed between storyboarding and final implementation. Comment on any trade-offs or design choices you had to make while developing.

There were a few features and interactions that were explored in our storyboarding and early work that didn't make it to the final implementation. One such feature was producing a scatter plot that plotted wine bottles under the same variety by price and rating. Although we initially thought this would bring a lot of value, after developing it, it really just caused a lot of issues both in interaction and significance than what was worth. In many cases, bottles would overlap, making them extremely difficult to select. We attempted to address this issue by implementing a jiggle function. We even considered listing out wines bottles that shared a similar price/rating ratio with that of the bottle selected, but realized it would just be easier to have the user filter those fields (which is what we ended up designing instead). In terms of significance, we thought we would find a lot of interesting outliers or trends, but from extensively playing with the visualization, there simply wasn't much there. Wines under the same varietal often times behaved the same.

Another design decision that was made while developing that differed from our early design ideas was how the default network would appear. Initially we were thinking of just displaying all of the wine varietals at once with no organization. However, after much consideration, we realized that most users would likely begin their filtering process by the type of wine (red or white). Realizing that, we decided to change the organization of the default network to already have the different types of wines visually separated into their own respective groups. This saves the user a step while also still giving them the freedom to decide whether wine type matters to them as all of the wines are still displayed.

Identify how work was broken down in the group and explain each group member's contributions to the project. Give a rough breakdown of how much time you spent developing and which parts of the project took the most time.

- As a Team:
    - Finding workable datasets and brainstorming different ideas around them
    - Deciding on the data set and idea
    - Determining the purpose and goal of our data visualization
    - Discussing all design decisions together as they came up
    - Helped fix bugs
- Charles
    - Filtered the original data set
    - Developed the network data visualization and interactivity (click, drag, hover)
    - Implemented the functions that organize the nodes by wine type networks (red, white, other)
    - Implemented the taste/descriptor filter bar
    - Implemented "Pick a Wine" functionality
- Michelle
    - Oversaw the visual/user interface design aspect of the project
        - Layout
        - Colors
        - Typeface
    - Tested and implemented different CSS frameworks (we implemented Bootstrap ultimately)
    - Drew storyboards
- Jonathan
    - Implemented functions to filter the varietals after user selects a node
    - Implemented the scatter plot of the wines that fall under the selected varietal (later scrapped)
    - Helped Charles implement the second set of filters (min price, max price, min rating)

The entire project took roughly a weeks worth of focused work to complete, though we spread out the work more evenly throughout the time given. Surprisingly, the visual design implementation (CSS) and fixing functional bugs took the most time.