Writeup

The Data

The data represented in our project is distilled from two separate Kaggle data sets provided below. They are both NY State data sets that are updated and maintained by Kaggle itself. The first data set, The Brand Label data set, had the License Class Description, the Product Description, and whether it was imported or domestic, among other data which we decided not to use. The Active Liquor License data set had the latitude and longitude information that we used for the NY State map overlay, among other data which, again, was not used. Both data sets had a Wholesale License Number which was what we used to stitch the data sets together. We filtered out any data that did not have any coordinate information because we considered it an incomplete data record. Additionally we decided to combine a few similar classes (such as “wine specialties” and “wine-low alcohol” into just “wine”) and delete a few that only had less than 50 entries for clarity sake. These two alterations allowed us to cut our alcohol categories roughly in half (25 to 14). Finally, we used an external JSON data set in order to draw the NY state outline and its counties. The GeoJson data for New York state was open source code and freely available on Github. Since the Active Liquor License data set had information about latitude, longitude, and whether it was domestic or imported, we could easily work with this data set with GeoJson data. We plotted domestic and imported Alcohol License distribution across the state using two different colored dots. Since some of the data points had the same coordinate values, some circles overlapped together. In order to correct for this we added some jitter to the location of the circles and decreased the opacity of the circles. These two factors allowed us to decouple many of the circles and show a greater density on the map.

Data Set 1: https://www.kaggle.com/new-york-state/nys-liquor-authority-brand-label-and-wholesaler

Data Set 2:
https://www.kaggle.com/new-york-state/nys-liquor-authority-active-licenses-and-permits

Json data:
https://github.com/deldersveld/topojson/tree/master/countries/us-states

Design Rationale:

The first visualization we made were bar charts that showed how many products of each class were registered in the state of NY. Since the low was in the teens and the high was in the thousands, we decided to use a logarithmic scale in order to present the data. We thought it would be interesting to divide the data by origin of the product (Domestic or Imported) and originally had these bars stacked on top of each other. This presented a problem however. Since it was a log scale, unless you were paying attention to this fact, a bar that was half one color and half the other might indicate that there was a 50/50 split when this was not the case at all. To prevent this, we decided to go with a clustered bar chart instead so that every bar started from 0. In order to ensure that people understood that there were two variables being displayed, we decided to use different colors for the Domestic and Imported products. This is true across the entirety of the project. We wanted to make sure we used high contrasting colors so it was easy to see, even when really bunched up like in the maps. We also wanted to make sure that the colors were color-blind friendly so that our visualization could be viewed effectively by everyone. Another thing to note is the class labels. At first the labels were really small so that everything could fit between the hashes but we decided that the small print might be inaccessible to some. Instead, we decided to stager the labels so that they could be bigger (and therefore longer) without overlapping and becoming illegible.

The other form of visualizations we made were state maps. We used an external JSON file to draw the map and then mapped our coordinate data onto the map. This is a very intuitive way to present this data because it allows users to see the connection between License Class and its distribution. We made sure to choose a GeoJson file that also contained counties boundaries for our final version because the map was very confusing without them drawn in. We thought it would be clearer and more intuitive for users to see with the counties boundary and were ultimately very pleased with the results.

The first map represents overall Alcohol License distribution across the state of New York state. We plotted two different colors for domestic and imported, making sure that they were the same colors which were used in the bar charts. We also added a jitter to the pixel location to prevent circles from clustering together at the same point and making it hard to distinguish. After projection of the location, we used Math.floor to get the exact pixel in map. Then we changed the capacity of the circle to 0.3, so we could easily compare the distribution situation over the state. Secondly, we selected two popular alcohol licenses: Whiskey and Beer & Lager and plotted the distribution of these two licenses across the state. In this way we were able to make a comparison of the two types of popular alcohol to the general.

The Story:

Manufacturers and wholesalers must register their establishments according to the New York State Alcohol Beverage Control Law, so the dataset is accurate and complete. From the first bar chart and map of Distribution of Alcohol Licenses across NY State, we discovered that most of the alcohol in New York State was of domestic origin. Much of the imported alcohol licenses were located around the greater New York City area. For the second graph, we discovered that, to no surprise, all of the Irish and Canadian Whiskey was imported, as one would expect. What we did not suspect however, was how unevenly spread out the Whiskey licenses were, being mainly distributed in New York City than other parts of NYS. For the third graph, we found that the Beer & Lager License Class had more domestic than imported Alcohol Licenses. We also found that domestic Beer & Lager License Class was almost uniformly distributed over the NYS. But the imported Beer & Lager License Class was concentrated in New York City.

Work Distribution:

The overall work distribution was essentially 50/50. Noah did the initial splicing and cleaning of the data in order to get one usable data set to begin working with. He also did the 3 bar charts whereas Xinyue did the 3 state scatterplots. We both worked on the initial design ideas and various written pieces We spent about 3 hours deciding upon the initial idea and sketching initial visualizations and then spent about 7 hours each making our respective graphs. Noah then spent about an hour combining all of the work together and styling it. The writeup portion was split in half as well and took about an hour for each person. The hardest part about making the bar charts was getting the scale correct. This is in terms of making sure everything looked good on the log scale but also ensuring that there was enough room to read the labels and that it didn't seem overly cluttered. Xinyue spent 5 hours searching for the NYS map data and 3 hours on learning make maps. The hardest part about making the state maps was learning how to make a map and interact with the data as we didn't learn anything about map when working on this project. This included learning steps in processing the geojson data and also how to project related data onto the map.