Midterm Report

Spring 2020 CSCI 3390 Network Science

Group Members: Lizzy Hanley, Kieran Roth, Catriona Sullivan, Kevin Li

Friday, April 3rd, 2020

Implementation Progress

For our midterm progress report, our group will be sharing details on the progress we have made on our early-stage prototype. We have chosen to write our code in python because there are many powerful data analysis libraries available to us. Everyone in our group has worked with python before, and we all feel comfortable using it in this project.

The first task we tackled was Airbnb price mapping. The most recent Inside Airbnb web-scrape of the Airbnb site was done on March 13th, 2020, but to avoid the effects of COVID-19 on our data, we used data scrapped on February 12th, 2020. After reading in our CSV file as a pandas dataframe, we began to clean the data. There are three fields in the raw data we will be using for our mapping: latitude, longitude, and price. Both latitude and longitude were in a usable format, but price was not. For the price field, we had to get rid of the "$" symbol and the comma identifying the thousands, and we also had to convert the field to a numerical format. With the cleaned data, we then made a map to visualize where listings are located. Using a python package called folium (which lets us use OpenStreetMaps), we iterated over our dataframe for each pair of coordinates, and added a marker to the map. Finally we plotted a heatmap of the prices of each listing as shown in Figure 1 below. We are still trying to add a legend to the heatmap to improve readability.

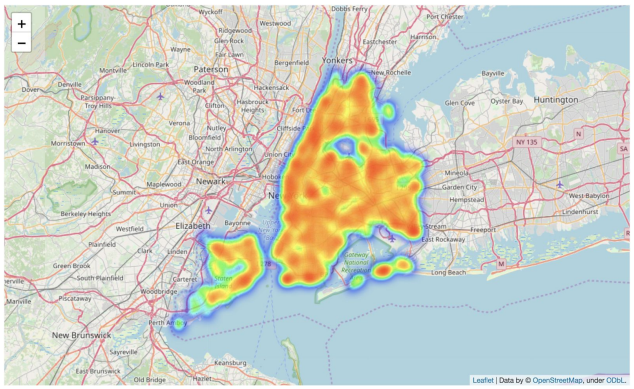

Figure 1: Airbnb price heatmap of NYC

Next we tackled mapping wealth distribution in a city. This proved to be a challenge since the wealth data we collected from DataUSA did not include sufficient location detail for our analysis. After researching for a new dataset, we decided to work with a publically available ArcGIS web map of 2019 Median Household Income in the United States. As stated in the map's description, this map shows the median household income in the U.S. in 2019 in a multiscale map by country, state, county, ZIP Code, tract, and block group. So far we have figured out how to bring this web map into python (see Figure 2 below), and we are now exploring how to convert our Airbnb price heatmap into an ArcGIS MapLayer to add on top of the 2019 Median Household Income web map. Our future work also includes figuring out how to compare the similarity between the two maps (ie. are the more expensive Airbnbs located in the more wealthy parts of a city).

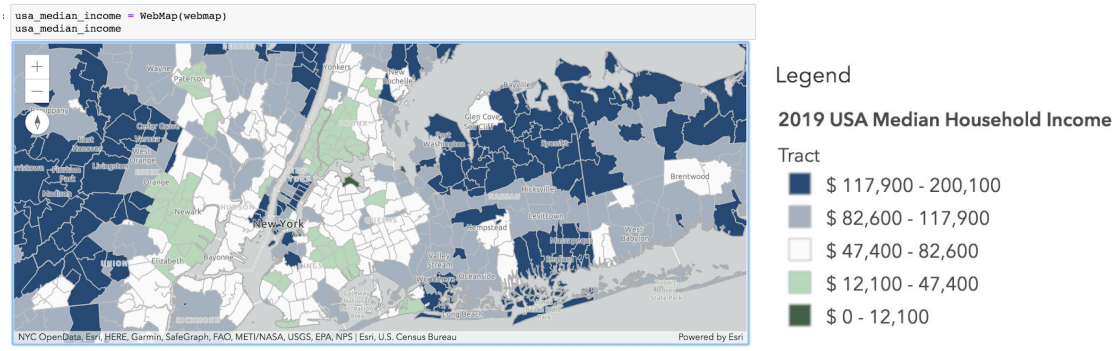

Figure 2a: Wealth distribution in NYC and map legend

Midterm Milestone

As highlighted in our initial project proposal, our midterm milestone was to finish the Airbnb price mapping and to begin the wealth distribution mapping. The section above explains that we were able to meet the milestone for Airbnb price mapping, but have run into challenges with the wealth distribution

mapping. Our original wealth dataset did not provide location data, therefore we had no way to create a map to visualize wealth distribution. Fortunately, we overcame this challenge by finding a different dataset to use, but is in a format none of our group has worked with before: an ArcGIS web map. ArcGIS is a cloud-based website which allows you to drop spreadsheets into MapViewer and display your coordinates on the map, including the data stored at those specific points. Learning how to work with this software is an issue we will have to address as we continue to work on our project. For example, an issue we encountered while implementing our Airbnb CSV file as a MapLayer was that all of our data points wouldn't load. We overcame this challenge by splitting up our data into multiple CSV files in order to work with smaller file sizes.

We are now using Esri's publicly available MapLayer of 2019 Median Household Income in the United States. Esri is a geographic information system company, and one of the largest distributors of geographic information software. The data shown in the MapLayer comes from Esri's 2019 Updated Demographic estimates which uses Census data to visualize wealth distribution across the United States. So far we have figured out to create a ArcGIS MapLayer to visualize the locations of Airbnbs in a city (see Figure 3 below), but we are still working to incorporate price into this visualization, and layer it on top of the existing wealth distribution map.

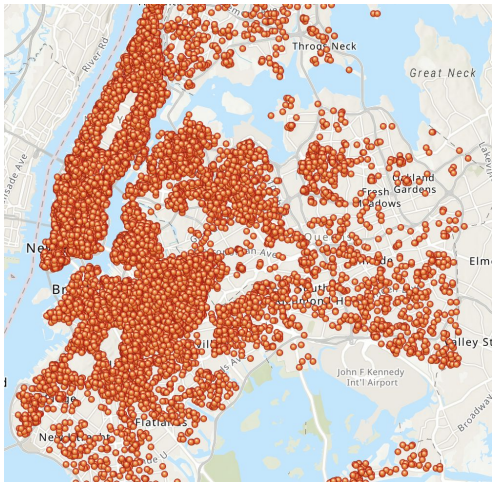

Figure 3: ArcGis MapLayer of Airbnb locations in NYC

Challenges & Solutions

As stated earlier, we used the folium python library to create a heatmap for our Airbnb price data. One of the challenges we faced with this task was learning how to use the library and implementing the code in a way that would work for our data. Once we got the heatmap to work, we realized that there wasn't enough information indicating how to read the map, such as a legend. Additionally, based on your comments on our initial report, we wanted to validate our criteria more thoroughly. This Harvard study seeks to measure inequality in cities and uses household income as one of the metrics for measuring wealth. After reviewing this source, we feel more confident that household income is a fair way to measure wealth in a city. Next we had to choose what type of income metric to use: median or average.

According to this article, Average Income in the USA by Family and Household, the "mean sums up all incomes and divides by the number of people reporting. The median income is the point where half the people make more and half make less." The mean, or average, income is usually higher, due to the few outliers who make an exorbitant amount of money. Therefore, in order to present the best representation of wealth and household incomes in the US, we chose to use median household income. For example, in 2014 the real median household income was $53,657, while the real mean household income was $73,298.

As we started to work on the wealth distribution part of the data, we also realized that our dataset was not suited to the type of visualization we wanted to create. It was very difficult to find a free dataset that included median household income along with latitude and longitude coordinates. There was also the issue of implementing ArcGIS within our python code. Esri has available MapLayers to use, but initially we were writing code for a web map rather than a layer, so we couldn't get the map to display.

The solutions to these challenges mostly involved reading more documentation and looking for datasets to fit the programs we wanted to use. For the folium heatmap issues, we were able to fix a lot of problems by looking over documentation and other resources. We looked up how to implement a legend, but we still couldn't get labels for the corresponding colors. There is still no clear solution for the comparison metric as we are trying to compare a python folium map to an ArcGIS map. As for the wealth distribution aspect of the project, we did some more searching and were able to find a better dataset that included both median income and latitude and longitude points. Instead of struggling to incorporate the ArcGIS code, we created a developer account instead to directly make the map from the website. Figure 3 shows our progress on creating an ArcGIS MapLayer that visualizes Airbnb locations in a city, but we haven't incorporated price into this map yet. Overall, we made some good progress on a lot of issues, but we still need to create a legend for our price heatmap, convert this heatmap to a ArcGIS MapLayer, and then compare our price map to our wealth distribution map.

Modifying our Final Report Milestone

In addition to the dataset and technical obstacles encountered while working through the project, circumstances with COVID-19 and time constraints added on to the workload of the project. Therefore, in order to maximize time and effectiveness, we want to modify our milestone goals. One possibility is to focus on improving our price and wealth visualizations. We would spend time learning how to use ArcGIS, and devote our time to producing a single ArcGIS web map that displays wealth distribution across the US and Airbnb price distribution in our three target cities. Because ArcGIS maps can be easily viewed in the browser, we would not need to make our own web map. We would include metrics to compare the price and wealth layer, and provide key takeaways for each city.

Another possible route is focusing more towards creating our own website to present our project. We would leverage our current progress in price and wealth mapping, and since the maps are interactive, a user can click a spot on each map in order to view the price of an Airbnb and the median income in that area. With only so much time in the semester left, we want to make the most out of our project by creating something interesting, while learning new web development skills. Focusing on this second option would help individuals in this group gain exposure to web technology and strengthen knowledge learned for those already familiar with it. The best case scenario would be being able to incorporate and execute both

options, but we want to keep this project feasible given the constraints we have now. Ultimately, we want to be flexible with our milestone goals and create a product that is both meaningful and obtainable. Please let us know your feedback on how you think it is best to proceed with our project!