PROCESS BOOK

UBER-MOVES

Overview and Motivation

Most people see everyday traffic on their daily routine but often have any clue about the overall traffic. This project provides a visual exploration one can use to find out where and when people hire cabs mostly. This project helps Uber driver to see and visualize pickups and place himself in a better place and time for maximum pickups during his work hours.

Related Work

The maps that we did using latitude and longitude was great. Apart from that, the presentation about data visualization @ uber by Shan He was inspirational. Here is the link of that presentation:

https://www.youtube.com/watch?v=nLy3OQYsXWA

Questions

This project is about showing when and where uber pickups were made in New York City for a given month (Eg: April 2014). The primary questions that this project answers –

1. Which places have more pickups?
2. What are the peak timings in a day where most pickups were?
3. Is Uber busy on weekdays or weekends?
4. How busy is uber on the wee hours of a weekday or weekend?
5. For each pickup, tooltip provides the exact location and pickup time on mouse-hover.
6. This helps a uber driver place himself in a better position to get maximum pickups.

This helps city traffic controllers visualize the movement of FHV(for-hire vehicles) in the city.

Data

The data is collected from *data.world* website.

The data has all the uber pickup information in New York City for a given month (Eg: April 2014). The total number of pickups for just one month is close to 600,000!

The data is in the following format:

- Date/Time : The date and time of the Uber pickup
- Lat : The latitude of the Uber pickup
- Lon : The longitude of the Uber pickup
- Base : The TLC (Taxi & Limousine Commission) base company code affiliated with the Uber pickup

Link to the database:
https://data.world/data-society/uber-pickups-in-nyc

Data Processing

The dataset includes about 600,000 items. It is very large for a serverless website. We will clean up the dataset and save those data in grids to support streaming-based data loading. Each small grid is a single JSON file. We will only load the data that will appear in our current viewport.

We will build another derived dataset, sampling each hour's total pickup count in a uniformly distributed grid system into one single file. When the viewport of the map is in a large scale, we will use it to display a heat-map.

The data is for a complete month of April 2014, about 600,000 pickups. Mapping these many pickups didn’t seem to be a good idea on a small map on the screen. Sampling seemed to be a good idea. We have sampled 10,000 pickups as follows:

1. Wrote python scripts to group pickups by a key that comprises of Date and Hour. Eg: All uber pickups on 04/01/2014 in 9th hour (9 am) will be keyed '04/01/2014 09'. We took the hour as part of the key as well, since the hour of the day also plays an important role in deciding how many trips would be made in prime hours.
2. Every such key is given a weight percentage based on the total number of pickups for that key vs the total trips in April. That is, weightage is given by the total number of trips per day/per hour versus the total trips in April.
3. Finally, sampling will be done based on weight. We have used Python dataframe sample method for sampling.
4. We added two more columns to our dataset for ease of filtering data. The columns are
    a. Numbers between 1 and 7 to denote Day of the Week (1-Mon, 2-Tue...7-Sun)
    b. Numbers between 0 and 23 to denote Hour of the Day.

So the final version is a sample of 10,000 pickups, with two additional fields used for filtering.

To retrieve category data of specific geographic position, we will try to do some data preprocessing based on Here Location Services' API. It provides an API can return category information with inputting latitude and longitude.
https://developer.here.com/documentation/places/topics/quick-start-find-text-string.html

Exploratory Data Analysis

We use a heat map as our main visualization on the New York City map, as displayed in *Figure 1* below, to look at the density of pick-up points. This design helps users quickly identify busy or less busy areas.

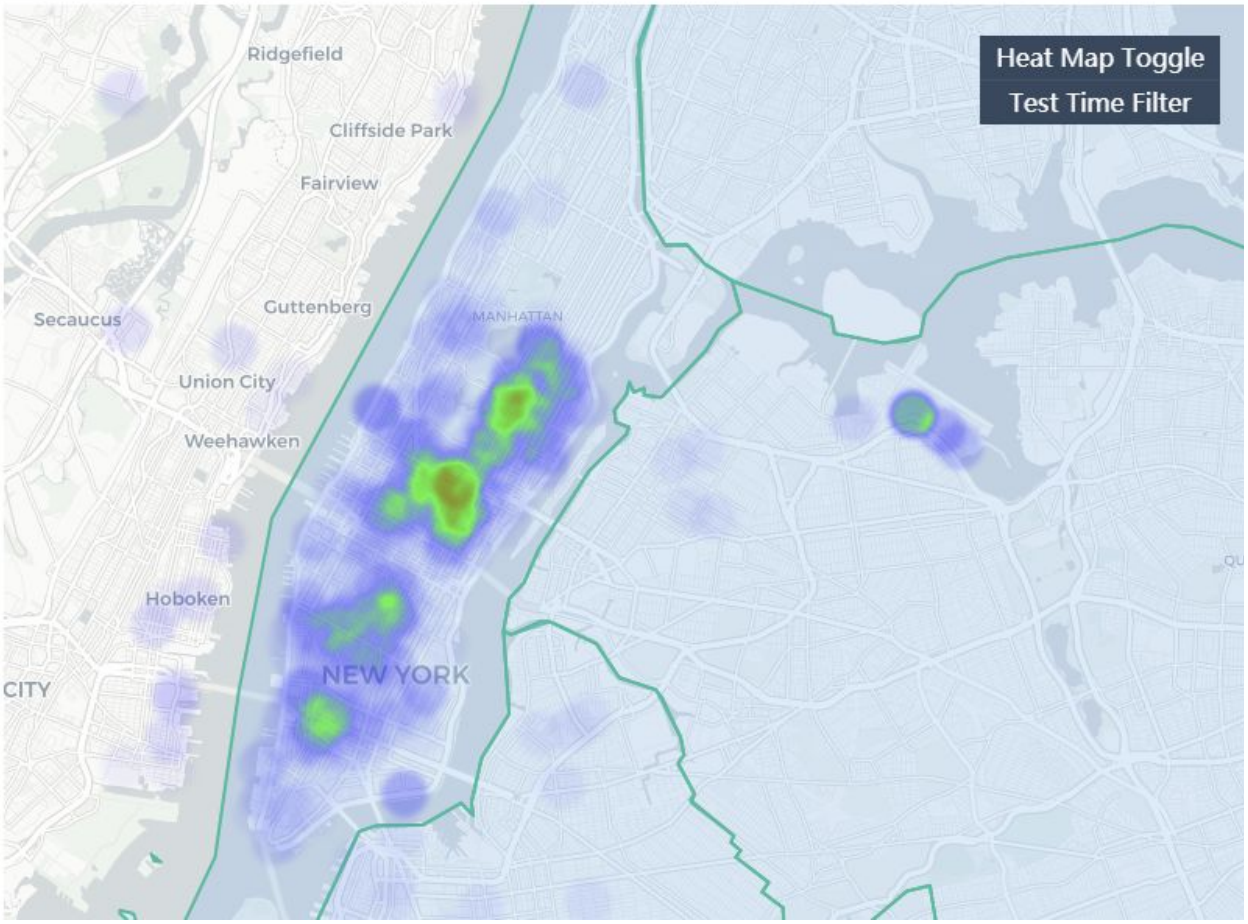

Figure 1. *a small sample of data displayed as heat map*

There will be a time and date sliders, as seen in *Figure 2* below. This option provides users the freedom to choose a timeframe and a range of days in the week to explore and test their hypothesis.

- Users can use brushing on either Time or Days of the week. The uber pickup points will show up or disappear on the map based on brushing area. A click anywhere on the Time or Day of the Week axis will undo brushing.
- The header on top of the map shows the selected Days and Time from brushing. By default it will be “Days: All Days Time: All time.”

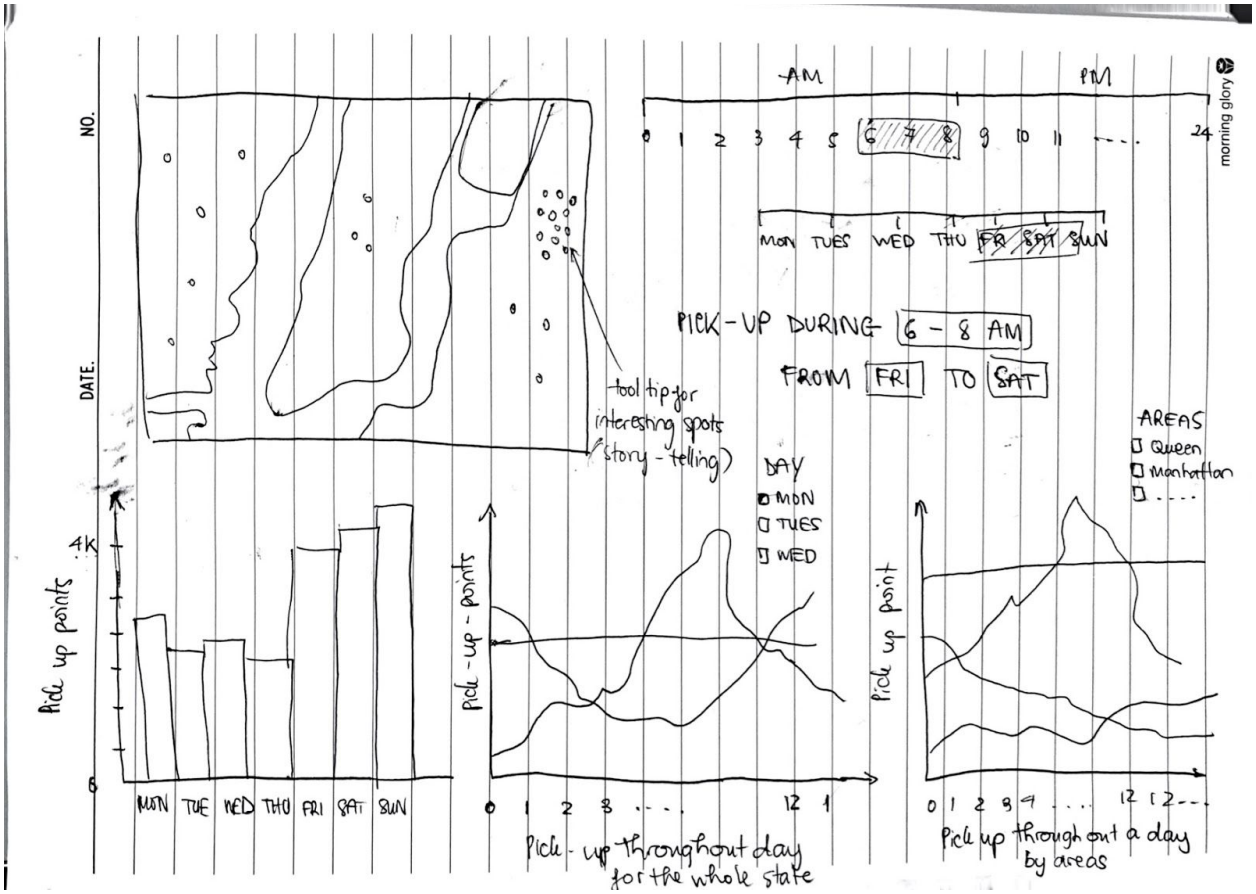

Figure 2. *a mock-up of our visualization layout*

Along with the heat map, there will be an option for plot chart visualization as shown in *Figure 3* below.

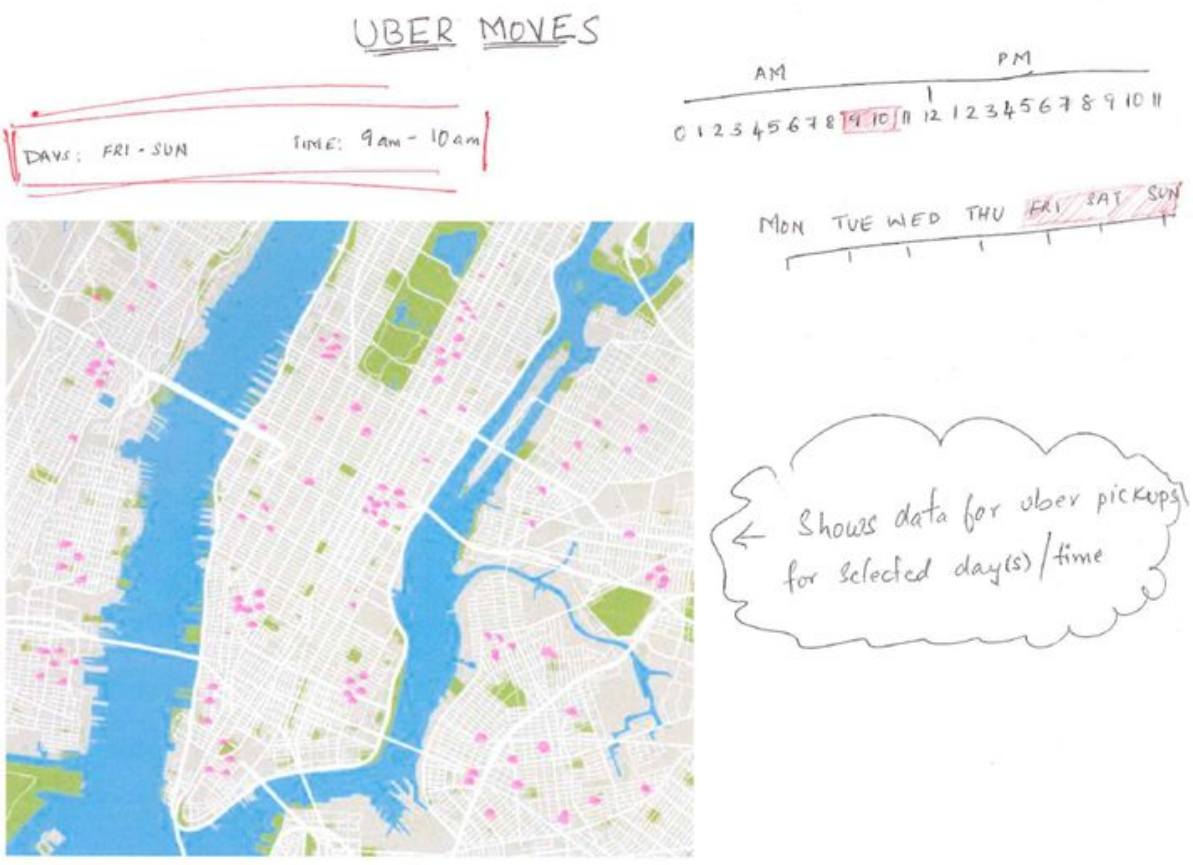

Figure 3. *plot chart visualization*

For any type of map visualization method, an area of interest can be zoomed in by selecting a region.

For the plot map, a tooltip will show the details of an uber pickup point on mouse hover. Tooltip will show all the details of all points if there are multiple pickups from the same location, as shown in *Figure 4* below.

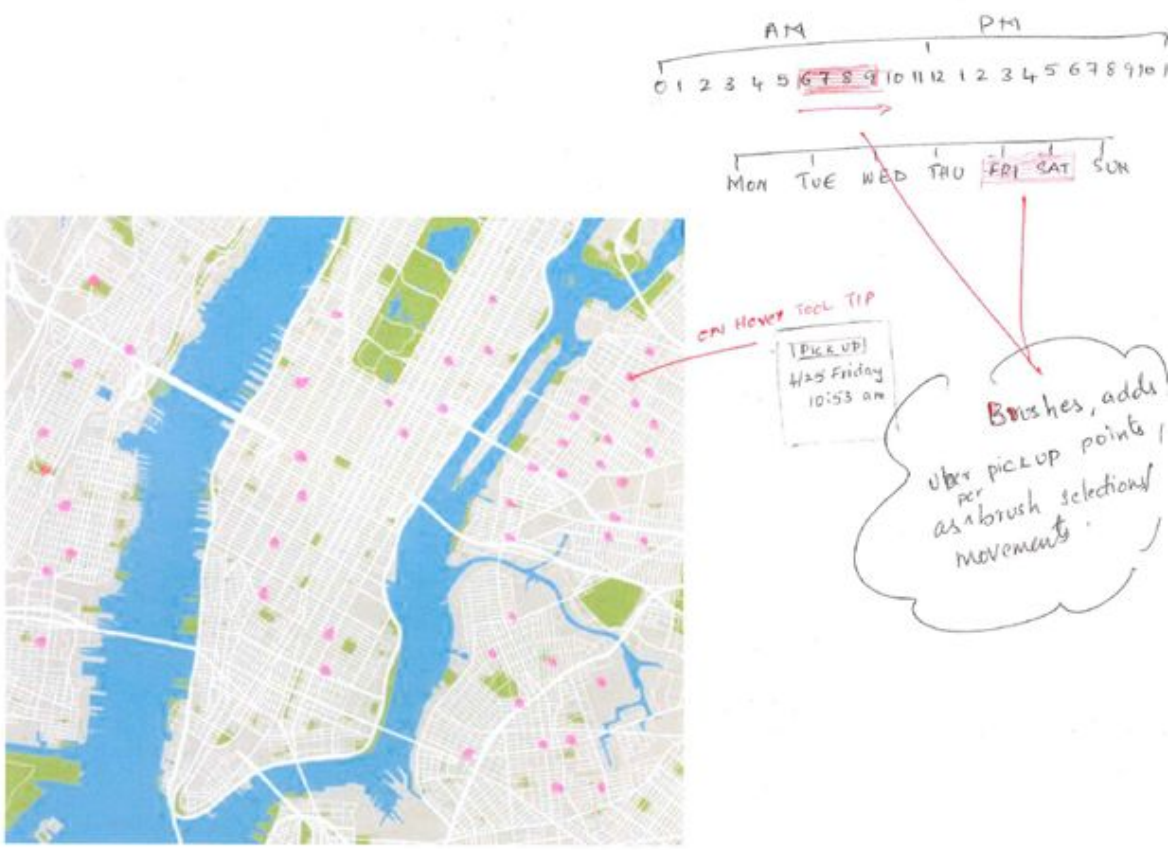

Figure 4. *a tool tip mock-up for plot chart*

For story telling, there will be a tool tip highlighting something interesting in the visualization, as shown in *Figure 5* below.

Figure 5. *a tool tip highlighting interesting fact*

As a third option for visualization, users can view data as a bubble chart on the map, as shown in the first screenshot from *Figure 6* below. This option allows users to easily get the precise number of pick-up points based on location.

The following dynamic charts will be displayed to show relationships of the data based on user's selection of date and time:

- A bar chart or line chart to show pick-up frequency throughout the week. It can show which date is most busy during the week. For this chart, users can customize the timeframe, and the chart will be updated accordingly.
- Pick-up throughout the day for the whole week. (Example: the second chart in *Figure 2*)
- Pick-up throughout the day by areas (Example: the third chart in *Figure 2*). Users can customize the day and time range.

Design Evolution

We have considered multiple visualizations designs. There were some good ones that we stick with, but also some that did not work out well. Below is a list of all the visualizations we have considered and our final decisions in chronological order:

- **Plot map**
  Plot map is the initial thought for our design because it makes sense to visualize the density of pick-up points. However, this is not a great choice to display a busy pick-up area. There will be a lot of overlapping points in that case. Therefore, we chose to have it as a secondary option to view the data.
- **Heat map**
  Since we ran into a problem with plotting busy areas, heat map becomes a perfect solution. Users can easily spot a high-density area, and there will be no overlapping issue compared to the plot map. Therefore, we chose this as our main visualization design.
- **Street view**
  Street view is fascinating and cool but it is complicated and hard to implement. The bigger issue is that it is not efficient to visualize the density of pick-up points via a street view. Therefore, we eliminated this option from our final design.
- **3D map**
  Similar to street view, 3D map is complex and unnecessary to provide insights for users. Thus, we eliminated this visualization from our final design.
- **Bar chart**

Bar charts are great to compare data. Therefore, we add them to our design to show more helpful insights. Users can customize the time and date range for these bar charts accordingly.

- **Line chart**

  Line chart is perfect for comparing changes over the same period of time for more than one group. Thus, we chose it to display the pick-up trend over a customized period.

- **Bubble chart**

  For both plot map and heat map, users won't be able to easily identify the precise pick-up count for a location. Therefore, we decided to implement an additional visualization option using bubble chart.

Implementation

Currently, we implemented an interactive map with three different rendering mode, circles, heatmap, clustered circles. It is based on Maptalk library.

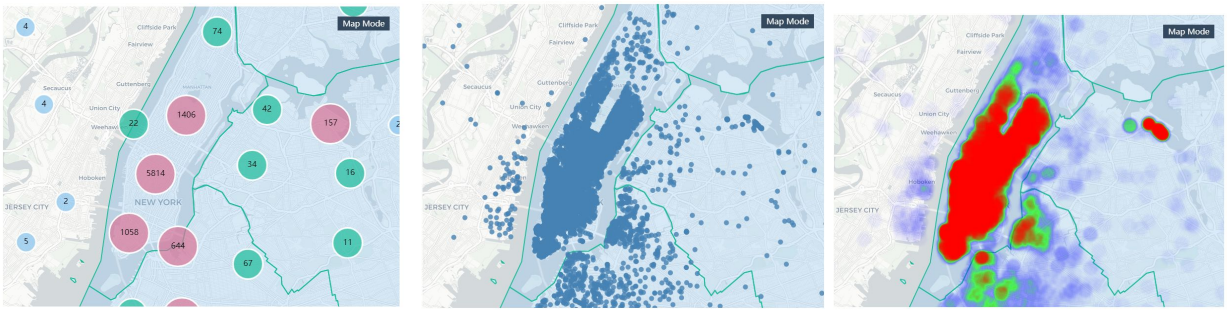

Figure 6. *Screenshots of three different visualization map options*

Other components are based on D3. Below are the details:

1. Day Time Brush Filter, a brush slider which is used for filtering by time in hours and minutes or weekdays.

Figure 7. *Screenshot of time slider*

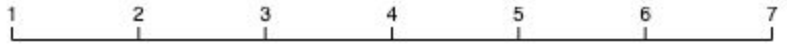

Figure 8. *Screenshot of date slider (from 1-Sunday to 7-Saturday)*

Future work: histogram, storying telling, user guides, etc.

Evaluation

So far, we can find the density from 16pm to 22pm is significantly bigger than other period.

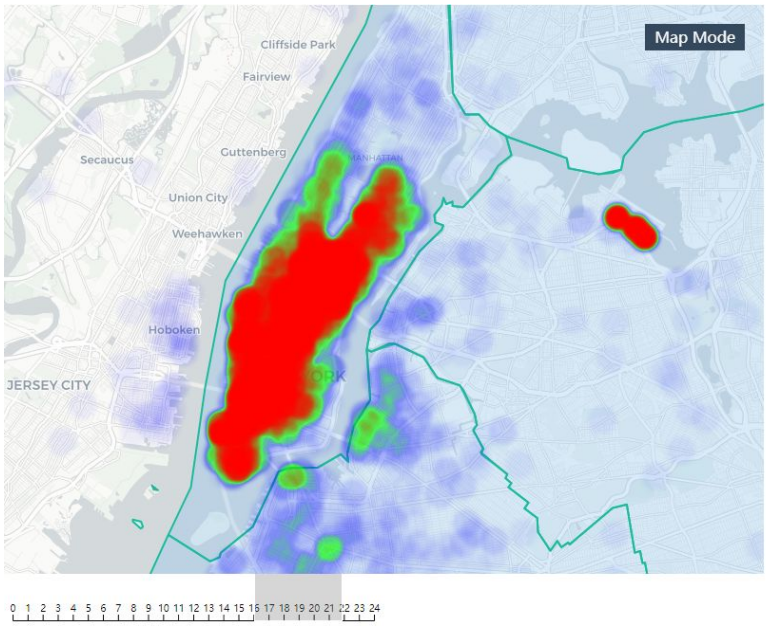

Figure 9. *Screenshot of heat map visualization*

What's more, if we observe this map in clustered mode. We can find that the Diamond district has the biggest pickup frequency. It is not obvious enough because users should manually choose a specific perspective. Maybe in the future, we can add some storytelling feature to show the maximum value in the map.

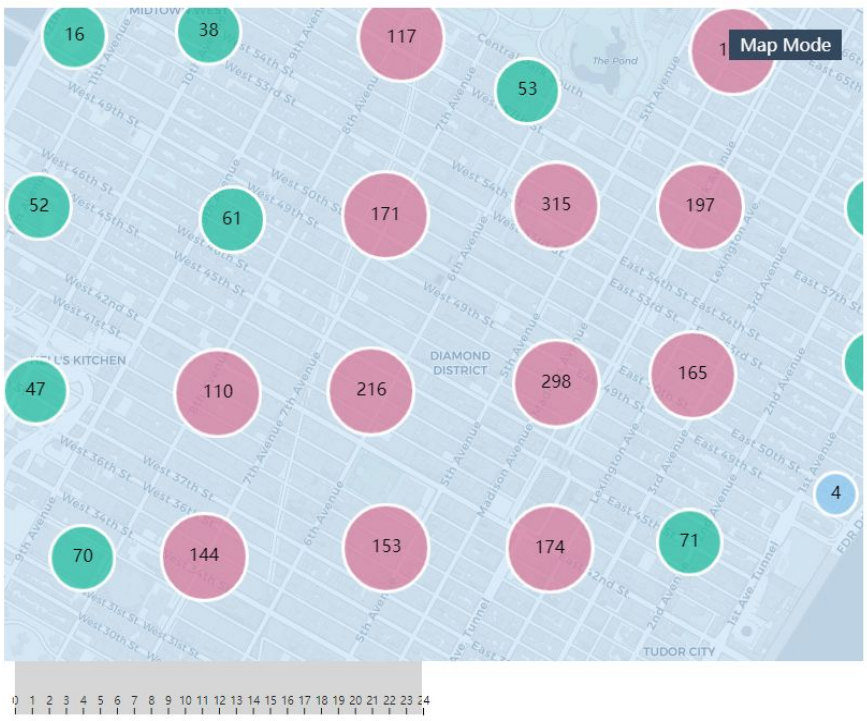

Figure 10. *Screenshot of bubble chart visualization*