Guided Interactive Visualization of NYC Taxi Trips During COVID-19

Jerry Zhang

MIT

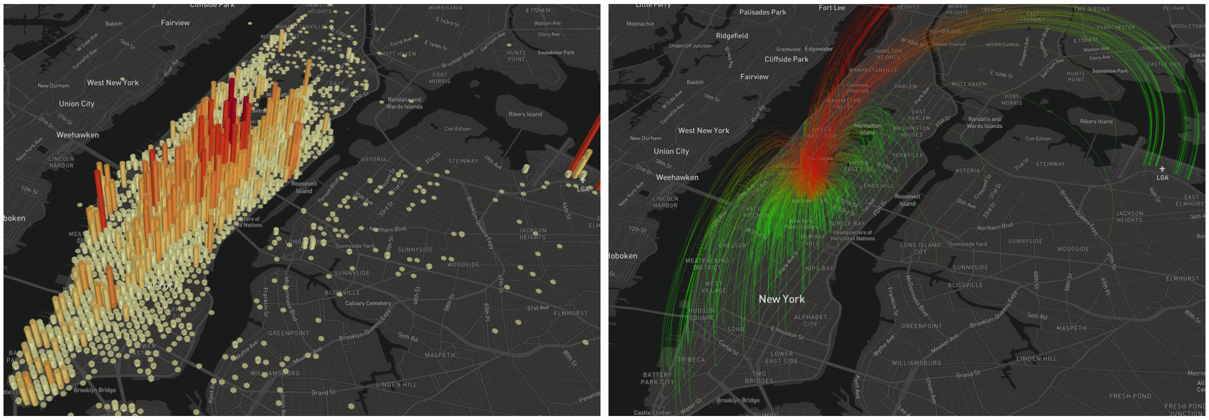

Figure 1: **Left**: Geographically binned histograms aggregates the total drop-offs in New York City and surrounding areas. **Right**: A focused location brushing of all trips where drop-offs are located at the southern edge of Central Park.

Abstract

While amidst the travel restrictions presented by the COVID-19 pandemic and subsequent lockdowns, the New York City taxi services avoided a collapse by providing a socially isolated transport service for riders. The ridership trends captured by these COVID-time trips demonstrate the changes in the priorities of the services' riders. These trends provide key insights on popular COVID-time destinations. The visualization and analysis are achieved through the translation of NYC Taxi and Limousine Commission (TLC) data into displayable information [4]. Coupled with the geographic nature of the data, an interactive slippy map built on the Mapbox and WebGL framework facilitates both individualized and aggregated views of the data to provide an intuitive and quantitative understanding of COVID-time taxi ridership trends in New York City [6].[1]

Index Terms: Human-centered computing—Visualization—Visualization techniques—WebGL; Human-centered computing—Visualization—Visualization design methods

1 Introduction

As the COVID-19 pandemic gripped the world for most of 2020, industries across entire sectors began to feel the economic impacts of the virus. The businesses of America's largest urban area, New York City, were no exception. With stay-at-home orders and business closures in areas like dining, night life, and retail, New York's unemployment rates have hit unprecedented highs, and the city faces a long road to economic recovery [1]. But an often forgotten piece of New York's challenging commercial situation comes not from traditional brick and mortar businesses, but from the transportation industry that previously drove the bustling metropolis. A prime example of this sector's struggles can be seen in the NYC taxi industry that, by some reports, has experienced declines in ridership in excess of 80% [2]. Yet with these dismal reports predicting that the pandemic travel restriction was the final nail in the coffin for the taxi industry, the industry found a new niche during these times of travel restrictions; as a low-contact method of direct transport, the taxis acted as the preferred choice of transportation for medium and long distances within New York City. The ridership trends also shifted, resulting in new popular and socially distant destinations to emerge. The motivation to observe and understand these trends drove the development and design of the interactive visualization.

2 Motivating Questions

The main questions that the visualization is set out to answer are

1. Which taxi destinations were the most popular during the initial phase of the COVID-19 pandemic?

2. Of the riders who go to these popular places, where do their trips originate from?

[1] Deployed interactive visualization at: `https://6859-sp21.github.io/final-project-travel-nyc/`

3. How do these popular destinations change over the course of a day?

These questions drove the data processing and design decisions of the resulting interactive visualization.

3 Dataset

The main dataset of interest for this project is provided by the Taxi and Limousine Commission (TLC) of the government of New York City. [4] The dataset spans the years 2009 to 2020 and contains trip information of the four main services that TLC provides: yellow taxi, green taxi, for-hire vehicles, and high volume for-hire vehicles. For the purpose of this visualization, the focus will be on the Yellow Taxi data between March 2020 and June 2020: the initial months of the pandemic hitting the US and the initial stages of lockdown in New York City. These trips are the majority of the ridership within Manhattan, and provide the most amount of detail in the data.

The dimensions of the Yellow Taxi data include

- Locations (Pickup and Drop-off)
- Timestamps (Pickup and Drop-off)
- Distance Traveled
- Duration of trip
- Fare of trip
- Passenger count

For the purposes of answering the central questions, the main dimensions of interest are the locations, timestamps, duration, and distances of taxi trips.

4 Related Works

4.1 Related Data Analysis

Taxi trips are a common source of data to analyze for ridership trends. Specifically, analysis during COVID-19 has been done in regards to the taxis services of different cities. In [3], Nian and their group analyzed taxi travel in Chongqing, China during the time of the pandemic in the region. In particular, the researchers used GIS and a spatial information to construct network models for analysis of ridership.

The results presented by Nian provide keen insights into how the taxi ridership changed as a whole due to COVID-19 within the city of Chongqing and how the favored destinations shifted as the pandemic progressed. However, Nian places a lot of emphasis on the quantitative aspect of the dataset, which removes the individual and human aspect that individualized taxi trip data provides. As one of the central motivating questions is to understand how individual trips collectively make up the aggregate popular destinations, a deliberate choice was made to ensure these aspects of the dataset were apparent in the visualization.

4.2 Related Visualization

Visualizations of the NYC Taxi dataset are quite prevalent, examples such as [7] tracks the journey of single taxis during their day throughout the city, or [5] which focuses specifically on the tips riders give in relation to other aspects of their rides such as destination and duration. However, while each of these contain some aspect of answering the central questions, neither visualization provides both the aggregate and individualized view of the dataset.

5 Methods

5.1 Data Preprocessing

The scale of the dataset was too large to be reasonably analyze all at once on a normal laptop, so the dataset was parsed and condensed through a distributed processing cluster into a file size that could easily be read and accessed by the resulting visualization. Each trip entry was also processed to ensure that the trip did not have missing information and had at least a trip endpoint within New York City.

As the filtering and aggregation methods of the visualization need to be responsive to the inputs of the user, a few techniques were implemented.

5.1.1 Geographical Bounding Boxes

As the dataset is mostly geographical, the preprocessing step organizes the trips based on geographical area. Shorter distanced trips were grouped together as the location brushing filter could more efficiently evaluate such data. These created natural geographical bounding boxes that allowed for the visualization processes to quickly ”short circuit” filter certain sections of the data.

5.1.2 Zoom Level Detail

Given the geographical nature of the dataset, densely populated trip data often results in overlapping marks that do not provide the user with any more information. These areas become even more obfuscated when a viewer zooms out and decrease the markers’ relative sizes and distances. Thus, certain trip endpoints were determined to be visually irrelevant at certain zoom levels of the map. This provided a performance boost at rendering time as less render and animation requests would be needed when rendering dense data plots on the map.

5.2 Render-Time Processing

5.2.1 Data Caching

When users interact with the visualization, they are likely to make small adjustments to filters, or would like to return to an initial setting. To decrease the filter processing time for these particular instances, these data and filter states are cached within the visualization code in order to optimize the users’ experience for when they want to restart their interactions with the visualization.

6 Visualization Design

The design of the visualization is driven by the central questions. The nature of the question lends itself to explore both individual and cumulative data measures. This led to the two main modes of interacting with the data in the resulting visualization: location brushing and geographical histograms.

6.1 Arcs and Brushing

To explore individual trips and to analyze the locations by which it originate and terminate, each trip is designated by a visual arc from the source to the destination, as shown in Figure 1. These trip arcs can then be explored in more detail by using a location brushing filter on the map to isolate certain trip endpoints within the vicinity of the viewer's cursor. This method of interaction and visual encoding helps answer the question of where users travel from or travel to.

6.2 Geographic Binned Histograms

To help answer the aggregate questions of where popular COVID-time destinations are, a geographical aggregation can allow for a better view than individual points can. This method facilitates a view of the larger taxi trends than individual trips, and viewers can determine which areas see more pickups and which areas see more drop-offs.

6.3 Autoplay

As the data is naturally time dependent, the time filter is given a autoplay function. This allows the viewer to interact with other aspects of the visualization as the time filter progresses. This adds a dimension to the visualization, and when coupled with the aforementioned brushing and aggregation views, the progressing filter can help viewers answer questions regarding how trip locations and destinations change throughout the day.

This particular feature was highly requested during demos and tests of the initial two visualization techniques as the smooth progression of the time filter can provide a better understanding of travel trends than manual scrubbing of the time filter.

6.4 Guided Questions

To guide the user through the different aspects of the visualization, tools are individually introduced, and guiding questions are given along with them to allow the viewers to familiarize themselves with what each type of visualization method is informing them of. These questions also build the data narrative for the viewers and to provide one perspective of what the visualization can reveal about the taxi trip trends during COVID-19 pandemic.

7 Discussion

Through user testing and revisions, the implementation of the final visualization offers enough complexity and interactions for the viewer to be able to gain a solid understanding of the ridership trends and popular destinations of COVID-time taxi trips while also not overwhelming the viewer with options and quantitative information overload. This balance was fine-tuned through user feedback from the deploy minimum viable product (MVP) as well as through iterated testing after the MVP.

The implementation of the guided question and slow-ramp introduction to the available interactions facilitated the viewer friendliness of the final visualization. The questions also enabled the viewers to gain an understanding of the goals of the visualization and subsequently resulted better viewer understanding of the data present in the visualization.

8 Future Works

The next steps for the visualization comes in both optimizing the data display pipeline and the breadth of data covered by the visualization.

A current limitation to the visualization is the data processing required by the filtering of the data. The performance bottleneck becomes apparent when the automatic time filter is used to analyze the change over time of a large span of data. Optimizations on this include further preprocessing and data binning as well as reduced filter checks to decrease computational complexity.

This version of the visualization is to give the viewer an easily digestible yet interactively dense visualization to understand the taxi trips taking in NYC during COVID-19. This can be expanded upon to span larger time periods of pre-COVID and eventually post-COVID trips. Furthermore, the extra dimensions of the data such as type of cab, distance traveled, and other metrics that might help understand the taxi trips can be incorporated into this type of visualization. The difficulty in developing those aspects is to ensure that additional views of the data does not increase the complexity for the viewer.

References

[1] P. Mcgeehan. Why n.y.c.'s economic recovery may lag the rest of the country's, Oct 2020.

[2] D. Naresh. 'we're on the brink of utter collapse.' yellow cabdrivers in new york struggle to stay alive as the pandemic rages on, Jan 2021.

[3] G. Nian, B. Peng, D. J. Sun, W. Ma, B. Peng, and T. Huang. Impact of COVID-19 on urban mobility during post-epidemic period in megacities: From the perspectives of taxi travel and social vitality. *Sustainability (Switzerland)*, 12(19):7954, sep 2020. doi: 10.3390/SU12197954

[4] NYCTaxi and LimousineCommission. Tlc trip record data.

[5] OmniSci. Data visualization example / demo: Nyc taxi ride data.

[6] Uber. deck.gl: Webgl2 powered geospatial visualization layers, 2016.

[7] C. Whong. Nyc taxis: A day in the life.