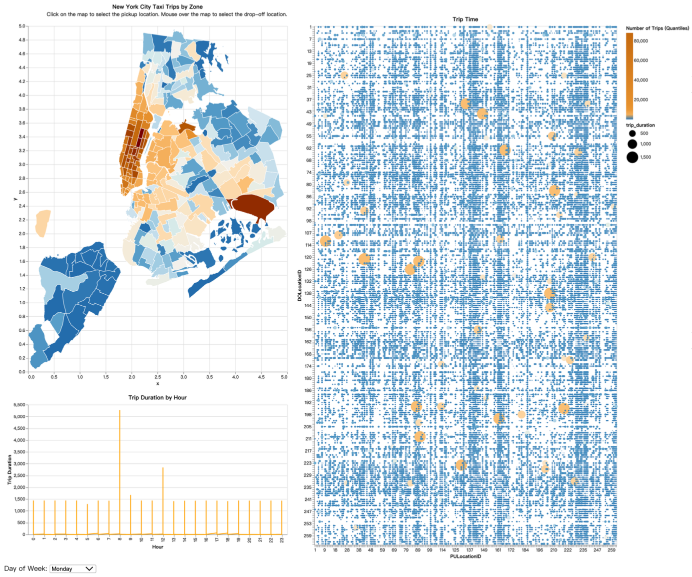
Interactive Visualization of New York City Taxi Trip Data

Tianhao Zhang
tz2518
tz2518@nyu.edu



Click on the map to select the pickup location. Mouse over the map to select the drop-off location.

Day of Week: Monday

Figure 1. Interactive visualization of New York City taxi trips data, showcasing the three main components: a map of New York City zones, a scatterplot of trip durations based on pickup and drop-off locations, and a line chart displaying trip durations by hour.

1. Introduction

In recent years, the growing availability of large-scale datasets has lead to the demand of data visualization tools that allow users to gain insights into complex patterns and relationships. Among these datasets, urban transportation data, such as taxi trips, have attracted considerable attention due to their potential applications in urban planning, traffic management, and public policy [2]. Expecially for New York city, it is a bustling metropolis and taxi plays an essentiial role in the city system [4].

In this project, an interactive data visualization tool is presented that enables users to explore and analyze New York City taxi trip data. The tool combines geospatial, temporal, and trip duration information to offer a comprehensive view of taxi trip patterns across the city. As illustrated in Figure 1, the primary visualizations in the tool include a New York city map of taxi trip origins and destinations by zipcode zone, a scatter plot of trip duration with respect to pickup and drop-off locations, and a line chart of trip duration by hour of the day.

The primary objective of this project is to help users identify trends and relationships in taxi trip data, such as the most popular pickup, peak travel times, and variations in trip duration. To achieve this goal, the visualization tool incorporates user-driven interactions, allowing users to filter and explore the data based on their own trips [3]. By providing an intuitive and engaging interface, the tool aims to empower users to derive actionable insights from the taxi trip data that can make it easier for people to take taxi and inform urban planning and policy decisions.

In the following sections, the functions of the visualization tool will be illustrated, and the implementation details will be discussed, including data processing, transformation, and visualization techniques used to create this interactive tool.

2. Demonstration

This section demonstrates the functionality and interactions of the interactive data visualization tool. The tool consists of three main components: a choropleth map, a scatter plot, and a line chart, as shown in Figure 1. The choropleth map represents the number of taxi trips originating and ending in each zone, the scatter plot illustrates trip durations based on pickup and drop-off locations, and the line chart displays trip durations by hour of the day.

2.1. Map

The map displays New York City's zones, with each zone color-coded based on the number of taxi trips originating in that zone. When a user clicks on or hovers over a zone, the selected zone retains its color, while the unselected zones turn gray. Users can interact with the map in the following ways:

- Click on a zone to select the pickup location for further analysis. The selected zone will retain its color, and unselected zones will turn gray.

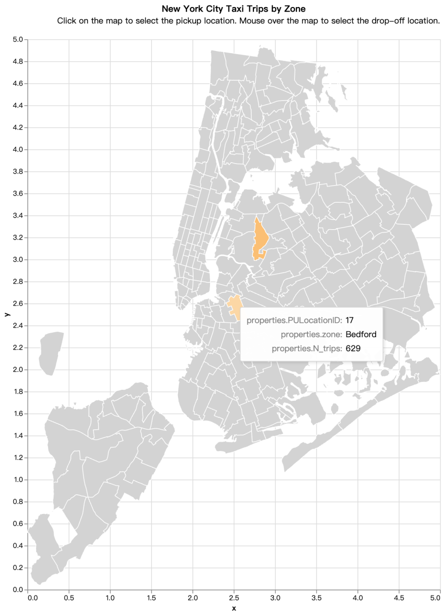

New York City Taxi Trips by Zone
Click on the map to select the pickup location. Mouse over the map to select the drop-off location.
properties.PULocationID: 17
properties.zone: Bedford
properties.N_trips: 629
y
x
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
2.8
3.0
3.2
3.4
3.6
3.8
4.0
4.2
4.4
4.6
4.8
5.0
0.5
1.5
2.5
3.5
4.5

Figure 2. An example of map zones selection

- Hover over a zone to select the drop-off location and view additional details, such as the zone name and the number of trips.

The selected zone will retain its color, and unselected zones will turn gray.

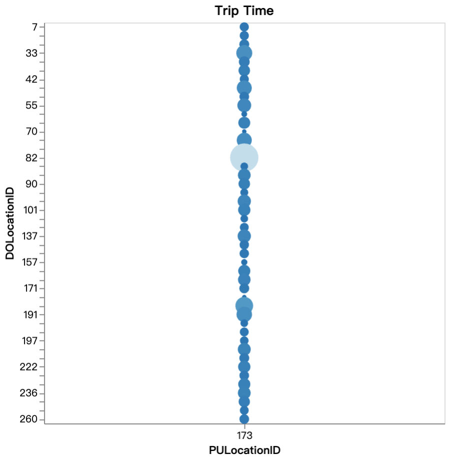


Figure 3. line chart example

2.2. Scatter Plot

As shown in Figure 3, The scatter plot represents trip durations as circles, with the x-axis corresponding to the pickup location and the y-axis to the drop-off location. The size and color of each circle indicate the duration of the trip. Users can interact with the scatter plot by selecting specific pickup locations on the map. The scatter plot will display trips originating from the selected pickup zone but will not respond to the selected drop-off locations.

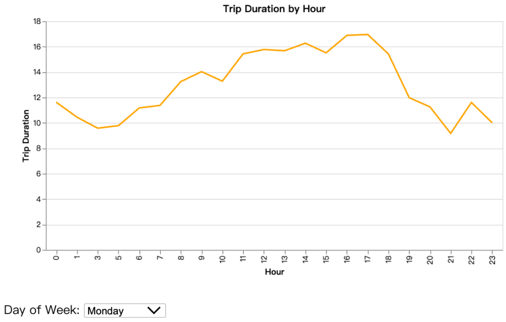


Figure 4. line chart example

2.3. Line Chart

The line chart as Figure 4 displays trip durations by the hour of the day, with the x-axis representing the hour and the y-axis representing the trip duration. Users can further filter the data by selecting a specific day of the week using the dropdown menu provided. The line chart updates dynamically based on the user's selections in the choropleth map, allowing users to explore trip durations for specific pickup and drop-off locations and days of the week.

Together, these interactive components empower users to explore and analyze New York City taxi trip data, uncovering trends and relationships that can inform urban planning and policy decisions.

3. Implementation Details

This section discusses the key steps and techniques involved in the implementation of the interactive data visualization tool, including data processing, visualization techniques, and interactions.

3.1. Data Processing

The raw taxi trip data was preprocessed, merged with New York City's geographic information, and aggregated to create a suitable format for visualization. The following steps were performed:

- Data cleaning: Missing values, outliers, and invalid data points were removed or imputed.

- Feature extraction: Relevant features, such as pickup and drop-off locations, trip duration, and hour of the day, were extracted from the raw data.

- Merging geographic information: The taxi trip data was merged with New York City's geographic information to enable the visualization of data on the map.

- Data aggregation: The data was aggregated using different methods to create three separate dataframes, each corresponding to the choropleth map, scatter plot, and line chart.

3.2. Visualization Techniques

The visualization tool was implemented using the Altair library, a declarative statistical visualization library for Python. The following techniques were used to create the interactive components of the tool:

- Choropleth map: The GeoJSON data format was used to represent New York City's zones. The map was created using Altair's 'mark_geoshape()' method, with colors representing the number of trips based on a quantile scale.

- Scatter plot: Altair's 'mark_circle()' method was used to represent trip durations, with the size and color of circles indicating the duration of the trip. User interactions, such as filtering by pickup and drop-off locations, were implemented using Altair's selection mechanisms.

- Line chart: Altair's 'mark_line()' method was used to create a line chart of trip durations by hour of the day. The chart was dynamically updated based on the user's selections in the choropleth map and scatter plot.

3.3. Interactions

The selection method of Altair enables users to explore and filter the data based on their interests [1].In the implementation, Altair's selection mechanisms, such as selection_single(), were used to implement these interactions, allowing users to select specific zones, pickup and drop-off locations, and days of the week.

The following selections were implemented in the code:

- Day of week selection: A single selection binding to a dropdown menu, allows users to select a specific day of the week. This selection filters the data displayed in the line chart.

- Pick-up location selection: A single selection is applied to the choropleth map, allowing users to click on a zone to select a specific pickup location (PULocationID). This selection filters the data displayed in the scatter plot and line chart.

- Drop-off location selection: A single selection applied to the choropleth map, allowing users to hover over a zone to select a specific drop-off location (DOLocationID). This selection filters the data displayed in the scatter plot and line chart.

By combining data processing, transformation, and visualization techniques, the interactive data visualization tool provides users with an intuitive and engaging interface to explore and analyze New York City taxi trip data, uncovering trends and relationships that can inform urban planning and policy decisions [5].

References

[1] Abha Belorkar, Sharath Chandra Guntuku, Shubhangi Hora, and Anshu Kumar. *Interactive Data Visualization with Python: Present your data as an effective and compelling story*. Packt Publishing Ltd, 2020. 3

[2] Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Cláudio T. Silva. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2149–2158, 2013. 1

[3] Petra Isenberg, Tobias Isenberg, Tobias Hesselmann, Bongshin Lee, Ulrich von Zadow, and Anthony Tang. Data visualization on interactive surfaces: A research agenda. *IEEE Computer Graphics and Applications*, 33(2):16–24, 2013. 1

[4] Xiaorui Jiang, Chunyi Zheng, Ya Tian, and Ronghua Liang. Large-scale taxi o/d visual analytics for understanding metropolitan human movement patterns. *Journal of Visualization*, 18:185–200, 2015. 1

[5] Fabio Miranda, Harish Doraiswamy, Marcos Lage, Kai Zhao, Bruno Gonçalves, Luc Wilson, Mondrian Hsieh, and Cláudio T. Silva. Urban pulse: Capturing the rhythm of cities. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):791–800, 2017. 3