



POLITÉCNICA

UNIVERSIDAD
POLITÉCNICA
DE MADRID

Data Visualization Final Project

New York Yellow Taxi

January 2025



Federico Castejón
Julián López
Aritz Martín
Unai Zuazo

Table of Contents

1. Introduction.....	3
2. Problem characterization in the application domain.....	3
3. Data and task abstractions.....	4
Task Abstractions.....	4
Data abstractions.....	4
Dataset Filtering and Cleaning.....	6
4. Interaction and visual encoding.....	6
Idiom 1 - Passenger Count by Day.....	7
Idiom 2 - Trip Distance by Passenger Count.....	7
Idiom 3 - Geographical Heatmap of Pickups.....	8
Idiom 4 - Trip Counts Across Two Days of the Week.....	9
Idiom 5 - Correlation Analysis.....	9
Idiom 6 - Tip Amount by Neighborhood.....	10
5. Algorithmic implementation.....	11
Idiom 1 - Passenger Count by Day.....	11
Idiom 2 - Trip Distance by Passenger Count.....	11
Idiom 3 - Geographical Heatmap of Pickups.....	11
Idiom 4 - Trip Counts Across Two Days of the Week.....	12
Idiom 5 - Correlation Analysis.....	12
Idiom 6 - Tip Amount by Neighborhood.....	12
6. Instructions for using the application.....	13

1. Introduction

Data visualization is a powerful tool for uncovering insights, identifying trends, and supporting decision-making processes. Over the years, the development of visualization best practices and standards has played a key role in turning complex datasets into actionable knowledge. These techniques are particularly well-suited for addressing challenges faced by policymakers and city planners, enabling the design of more effective policies and optimal decisions for the collective wellbeing.

In this project, we leverage data visualization to explore and share new knowledge about the intricate dynamics of New York City's taxi ecosystem. Taxis are a crucial component of the city's transportation network, and understanding their patterns can offer valuable insights into urban mobility, congestion, and economic activity. Our goal is to investigate these patterns and use visualization tools to uncover trends, correlations, anomalies, and outliers that can inform better policy and planning decisions.

This study will be structured as follows: we will begin by presenting the scope of the problem and the specific objectives we aim to achieve. Next, we will discuss the types of tasks and questions we seek to answer, as well as the datasets used in our analysis. In the third section, we will elaborate on the visual encodings and techniques that proved most effective in answering the posed questions. Finally, we will briefly present the interactive application we developed to facilitate deeper exploration of the findings.

2. Problem characterization in the application domain

Taxis play a vital role in urban transportation, offering a reliable and efficient means of travel. In the United States, particularly in New York City, taxis are used significantly more than in many European countries, reaching staggering numbers of more than 400.000 taxi rides per day in 2015 (from TLC database). Their ubiquity and accessibility make them an integral part of daily commuting and tourism alike, making it really important to explore the best way to distribute them around the city for example, to avoid excessive traffic and organise them in the best way possible.

The New York City (NYC) Yellow Taxi dataset provides a detailed record of millions of taxi trips taken across the city. This dataset includes information on trip times, distances, pickup and drop-off locations, passenger counts, payment types, and fares. Its richness makes it an invaluable resource for analyzing transportation patterns and understanding urban mobility.

The primary objective of this analysis is to uncover insights about taxi usage patterns in NYC. Specifically, this report aims to:

- Identify temporal and spatial demand trends.
- Explore correlations between trip characteristics.
- Analyze tipping behaviors to understand passenger-driver interactions.

To achieve these objectives, we utilized the R programming language and the Shiny framework to create interactive visualizations and analyze the dataset. The report is structured to walk through the data preprocessing steps, key visualizations, and the insights derived from them. Finally, we discuss the findings and limitations.

3. Data and task abstractions

Task Abstractions

The purpose of this visualization project is to provide insights into taxi usage patterns in NYC to uncover new knowledge and generate actionable hypotheses. The information derived from the visualizations can assist transportation planners, policymakers, and businesses in making data-driven decisions to optimize operations and improve customer experience.

The analysis process involves **exploration**, as the user knows the data to focus on (pickup locations, trip durations, fares, and tips) and seeks to identify relationships and trends. This exploration provides a comprehensive understanding of patterns in taxi demand and customer behavior.

Specifically, the questions addressed in this project include:

1. How do the number and distribution of taxi pickups vary by time and location?
2. What are the common trip durations, and how are they distributed?
3. How do users behave in terms of passenger counts? Does the behaviour change over time?
4. How do different trip metrics, such as fare, distance, and tips, correlate with one another?
5. What are the tipping patterns of clients?

By answering these questions, the visualizations offer actionable insights for understanding urban mobility and enhancing decision-making processes in NYC's taxi industry.

Data abstractions

New York City (NYC) Taxi & Limousine Commission (TLC) keeps data from all its cabs, and it is freely available to download from its official website. Now, the TLC primarily keeps and manages data for yellow and green taxis and for-hire vehicles.

In the used dataset, we only considered the Yellow Taxis Data, for the month of January 2015. The reason for not taking a more recent database is that instead of providing the pickup & dropoff coordinates, the TLC has divided the NYC into regions and indexed those regions, and in the CSV files, they have provided these indices. This means we are no longer provided the exact location of pickups and dropoffs, but rather an estimation. This is sad because one of our goals was to examine the effect on cab use by ridehailing apps (Uber, Lyft, etc.). However, when coordinates were provided in datasets by the TLC, these companies were not really popular yet.

The dataset includes the following features.

Field Name	Description
VendorID	A code indicating the TPEP provider that provided the record. 1. Creative Mobile Technologies; 2. VeriFone Inc.
tpep_pickup_datetime	The date and time when the meter was engaged.
tpep_dropoff_datetime	The date and time when the meter was disengaged.
Passenger_count	The number of passengers in the vehicle. This is a driver-entered value.
Trip_distance	The elapsed trip distance in miles reported by the taximeter.
Pickup_longitude	Longitude where the meter was engaged.
Pickup_latitude	Latitude where the meter was engaged.
RateCodeID	The final rate code in effect at the end of the trip. 1. Standard rate; 2. JFK; 3. Newark; 4. Nassau or Westchester; 5. Negotiated fare; 6. Group ride
Store_and_fwd_flag	This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. Y= store and forward trip; N= not a store and forward trip
Dropoff_longitude	Longitude where the meter was disengaged.
Dropoff_latitude	Latitude where the meter was disengaged.
Payment_type	A numeric code signifying how the passenger paid for the trip. 1.Credit card; 2. Cash; 3. No charge; 4. Dispute; 5. Unknown; 6. Voided trip
Fare_amount	The time-and-distance fare calculated by the meter.
Extra	Miscellaneous extras and surcharges. Currently, this only includes the \$0.50 and \$1 rush hour and overnight charges.

Field Name	Description
MTA_tax	0.50 MTA tax that is automatically triggered based on the metered rate in use.
Improvement_surcharge	0.30 improvement surcharge assessed trips at the flag drop. the improvement surcharge began being levied in 2015.
Tip_amount	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.
Tolls_amount	Total amount of all tolls paid in trip.
Total_amount	The total amount charged to passengers. Does not include cash tips.

The dataset consists of almost 13 million entries, each of them representing a registered taxi trip (avg. 400.000+ trips per day). This is obviously far more than we either need nor can process at once. Attempting to use all at once would result in a slow and inefficient app, with no visible benefits. That is why we will only be using 200.000 entries for our visualizations, which should also be overkill, but far more manageable.

Dataset Filtering and Cleaning

In order to be able to use this dataset, it had to be cleaned and filtered first. It was already quite clean from the start, but some changes still had to be made. For example, we had to address errors in the coordinates. The features *Pickup_longitude*, *Pickup_latitude*, *Dropoff_longitude* and *Dropoff_latitude* sometimes showed coordinates that were very far away from New York, sometimes not even in the USA. These are obviously errors and had to be removed.

We also incorporated the origin and destination neighborhoods into the dataset. By using the exact coordinates of each pickup and dropoff, we identified their locations and mapped them to the corresponding neighborhoods in NYC. This was achieved using a shapefile provided by NYC Open Data ([Borough Boundaries | NYC Open Data](#)), which defines the boundaries of each neighborhood through polygons in a GeoJSON format. This enables neighborhood analysis, identifying high-demand areas.

For visualization purposes, another feature was added, *Trip Duration*. As its name suggests, we used the columns *tpep_pickup_datetime* and *tpep_dropoff_datetime* to compute the time duration of the ride (dropoff time - pickup time).

4. Interaction and visual encoding

This section aims to explain the objective of each visualization and its visual arrangement, together with how the user can interact with the interface.

Idiom 1 - Passenger Count by Day

This visualization shows the **percentage of trips** by **passenger count** (1-6) across each **day of the week**. We can observe that 1 passenger (single riders) consistently dominates (e.g., 74.4% on Monday and 65.9% on Sunday).

The proportion of shared rides (more than 2 passengers) increases slightly on weekends (Saturday and Sunday), potentially due to social activities or group outings. On weekdays, shared rides are less common, with one-passenger trips making up around 73-74%.

Displaying this information can be useful to taxi companies to adjust **fleet availability** or **target promotions** for shared rides, especially on weekends when demand for group travel may be higher. Furthermore, travelers could benefit from understanding how common solo vs. group taxi usage is and how the **patterns** shift throughout the week.

It is also interesting to check the **behavioral insights by day**, identifying variations in ride-sharing behavior across the week, such as increased group rides on weekends, which may correlate with social activities.

Idiom	Normalized Stacked Bar Charts
Data	Quantitative variable: "percentage of trips", "Number of passengers" Qualitative variables: "Day of the Week"
Encode	X-axis: Day of the Week, Y-axis: Percentage of Trips divided with different colors for each class of passenger_count
Task	Visualize the proportion of Trips By Passenger count for each day.
Interaction	In this case there is no interaction, since this visualization is correlated with the next one.

Idiom 2 - Trip Distance by Passenger Count

This visualization displays statistics such as the **number of trips**, **average trip distance**, **average fare amount**, **total amount**, and **trip duration** categorized by **passenger count** for a selected day. In addition, it can be observed the **average trip distance (in miles)** for different passenger counts on Monday.

This representation provides insights into **trip patterns** by passenger count and day of the week. It can help taxi operators and ride-sharing services optimize route planning, pricing models, and group trip promotions.

Since the visualization allows selecting different days, companies can compare passenger behavior across days to optimize for weekday versus weekend travel patterns and find trends like longer trips for specific passenger counts (e.g., groups of 4), which can help focus on promoting those trip types for higher revenue.

Idiom	Bar Charts
Data	Quantitative variables: "avg_trip_distance", "number_passengers"
Encode	X-axis: Number of passengers, Y-axis: Average Trip Distance
Task	Explore the trend of the average distance traveled by each group of passenger's for each day.
Interaction	Select the day of the week.

Idiom 3 - Geographical Heatmap of Pickups

To visualize the distribution of taxi pickups in New York City, a **geographical map with a heatmap** has been chosen. The heatmap uses latitude and longitude values to represent the locations of taxi pickups, with the intensity of color showing the density of pickups in each area. This approach is effective for identifying areas with higher demand for taxis, such as busy neighborhoods or transit hubs, which can provide useful information for optimizing taxi services.

Given the dynamic nature of taxi demand, the heatmap includes options for filtering data by **day of the week** and **range of hours of the day**. This allows users to identify patterns specific to weekdays, weekends, or particular times, such as rush hours or late-night periods. For example, users can examine how demand shifts between business hours and leisure hours or compare weekdays to weekends to observe behavioral trends. To improve the interactivity and usability of the visualization, the map supports **zooming and panning**, enabling users to focus on specific regions.

This visualization has been created to support the analyst in understanding geographical patterns within the city to improve resource allocation and driver navigation strategies based on the observed patterns.

Idiom	Geographical Map
Data	Latitude and longitude of pickup locations
Encode	Spatial position to represent the geographic location of pickups; color gradient to encode the density of pickups
Task	Display the geographical distribution of taxi pickups across New York City, highlighting areas with the highest demand

Interaction	The user can select a specific day of the week and time range to filter data. Additionally, the user can zoom in and out to focus on specific areas.
--------------------	--

Idiom 4 - Trip Counts Across Two Days of the Week

To compare the amount of taxi trips across different days, a **bar chart** has been chosen. The X-axis represents the **hour of the day**, while the Y-axis shows the **number of trips** for each hour.

Users can **select two days of the week for comparison**. For example, comparing trips on Tuesday and Saturday (see the next figure) can demonstrate how taxi usage is different between a weekday and a weekend. This visualization is also useful for detecting peak hours, such as early morning commutes or late-night rides.

Additionally, the user can also filter the date of observations based on a **specific date range**, giving a more granular view. This functionality can help understand seasonal variations, such as differences in trends in January and August.

Idiom	Bar Chart
Data	Time of day (hour), day of the week and number of trips
Encode	X-axis: Time of day (hour), Y-axis: Number of trips
Task	Compare trip volume for two selected days and detect peak usage periods
Interaction	Select two days of the week, filter by date range

Idiom 5 - Correlation Analysis

To explore relationships between key trip metrics, a **scatter plot** has been chosen. Users can select any two quantitative variables, such as **trip distance**, **total fare**, **tip amount**, or **trip duration**, to plot on the X and Y axes. This visualization helps identify potential correlations, such as longer trips leading to higher fares or tips.

To further enhance the analysis, users can choose to overlay a **regression line** on the scatter plot. This allows for easy identification of trends in the data, making it simpler to interpret relationships between the selected variables.

On top of that, a correlation heatmap matrix was included, as it offers a general insight into which variables may be more correlated than others.

This visualization is particularly valuable for uncovering operational insights, such as how trip duration changes with its distance, and for building predictive models to improve service efficiency, as factors like traffic may greatly disturb this relation.

Idiom	Scatter Plot, Correlation Matrix
Data	Quantitative variables: "trip_distance", "total_amount", "tip_amount", "trip_duration"
Encode	X-axis: First variable, Y-axis: Second variable Matrix with pairwise correlation
Task	Explore relationships between trip metrics and identify trends. Observe which variables show high correlation
Interaction	Select X and Y variables, update scatter plot

Idiom 6 - Tip Amount by Neighborhood

A radar chart was chosen to analyze tipping behavior across NYC. The chart uses the neighborhood columns we created (explained in the [Dataset Filtering and Cleaning section](#)). With this information, we were able to see which neighborhood each taxi ride came from and in turn, observe how generous the tip is. This visualization helps identify hotspots for tipping, such as areas with higher-income passengers or tourist-heavy locations.

Users can select how many neighborhoods they want to see plotted in the chart, as all may not be of interest. However, at least three neighborhoods have to be selected, as it makes no sense to make out a polygon of less than three dots.

This visualization is particularly useful for taxi drivers and organisations, as taking taxi calls from those origins can help inform strategies for improving driver earnings by targeting specific passenger demographics.

Idiom	Radar Chart
Data	Pickup neighborhood and tip amount
Encode	Neighborhood names in the perimeter, mean tip amounts in the axes
Task	Identify tipping hotspots
Interaction	Filter for neighborhoods of interest

5. Algorithmic implementation

Idiom 1 - Passenger Count by Day

To create this visualization of a Normalized Stacked Bar Chart using the *ggplot* library in R. First, a preprocessing of the data is made by extracting the names of the days in English. Then, a new column is created (*trip_duration*) by calculating the difference between the pick-up and drop-off time of the taxi. Continuing with this task, there is data removal of the null values and abnormal values (negative trip duration or zero passengers). To ensure the equal rows for all days, we define seed to balance the dataset with the *sample_n* function. Finally, we ordered the dataset from Monday to Sunday.

For the plot, the data is mutated to show the percentage of the number of passengers for each day. Then, the plot is created with the *day_of_week* variable on the x-axis and the stacked percentage of passengers. We also use the *scale_fill_brewer* function to provide diverging and qualitative color schemes from ColorBrewer to distinguish between the types of the feature *passenger_count*.

Idiom 2 - Trip Distance by Passenger Count

For a more detailed visualization of the number of passengers on each trip, we decided to add this implementation. The preprocessing followed is the same as the idiom before. After this part, we developed the UI interface, with a *sidebarLayout* function to plot the different parts of the visualization. On the one hand, to know the information about each day, there is a *sidebarPanel* and the *mainPanel* shows the table and plot output.

Then, to show the interesting aggregated statistics, it is calculated the average for the trip distance, fare amount, total amount and the trip duration. These values are stored in a new variable which is displayed in a table for each class of the feature *passenger_count* with the *reactable* function of the library *reactable* library in R.

The bar chart, it is calculated the average trip distance for each type of passenger (1,2,3...), ensuring all passenger counts are represented with the *tidyr* library. Finally, the bar plot is created with the *passenger count* on the x-axis and the average distance on the y-axis.

Idiom 3 - Geographical Heatmap of Pickups

This implementation creates a dynamic heatmap visualization using the *leaflet* library in R. Data is filtered by inputs by the user, such as the day of the week and time range. The map uses a *CartoDB.Positron* layer displays a base map, which dynamically adjusts its center and zoom based on data bounds, defaulting to New York City if no data matches. The heatmap is generated with *addHeatmap*, using longitude and latitude points. The heatmap

employs a color palette that varies from cooler to warmer tones, representing the density of pickups on each point.

The UI shows a dropdown menu (`selectInput`) to choose the day and a slider (`sliderInput`) to set a time range. The heatmap is displayed through *leafletOutput*.

Idiom 4 - Trip Counts Across Two Days of the Week

To implement this bar plot, the first step was to filter the data to include only the entries within the date range specified by the user. The *hour()* function was then used to extract the hour from each timestamp, and the *wday()* function extracted the day of the week.

Next, the data was grouped by the hour of the day using *group_by()*, and the total number of trips for each hour was calculated with a summarization operation. This process was repeated for both selected weekdays to enable comparison.

Finally, to plot the idiom proposed, a bar plot was created using *geom_bar()* to display the hourly trip counts side by side.

Idiom 5 - Correlation Analysis

For the scatter plot, the first step is to add an additional layer of outlier detection, as any dot too far away from the rest would make the whole plot look bad. For this, we removed any values that were less than/ exceeded the 0.01st/ 0.99th quantile.

After this, we used the *ggplot()*, *geom_point()* and *geom_smooth()* functions from the *ggplot2* package to plot the scatter plot and regression line, respectively, using the variables from the user input.

This plot is not dynamically updated, as there are a lot of points to be plotted and a regression line to be drawn. Dynamical updates would lead to unwanted plots, which take some time.

For the heatmap correlation matrix, the first step was to compute the correlation matrix using the command *cor()*. Following, using the *ggplot()* and *melt()* functions from the *ggplot2* package, we could plot the heatmap correlation matrix.

Idiom 6 - Tip Amount by Neighborhood

We wanted a way to compare the tipping habits between different regions, but the dataset only offered geographical coordinates. That is why, as described in the [Dataset Filtering and Cleaning section](#), we added an additional column that mapped every row in the dataset to one of five neighborhoods: Manhattan, Queens, Brooklyn, Bronx and Staten Island. Once we had done this we used the function *radarchart()* from the *fmsb* package to plot the chart. As inputs we used the specified neighborhoods chosen by the user and a mean of all the tip amounts found in the dataset for the 'height' in the chart.

The whole chart is dynamically updated, so it is not necessary to press an 'update' button. We did this because it still remains really fast, as the means are calculated straight away, so changing input parameters just means we have to update the plot with no computing at all.

6. Instructions for using the application

To get started, click the following link: <https://unai.shinyapps.io/taxisnycshiny/>. The **home page** provides an introduction to the topic, along with information about the group number and student names. Each tab presents a different visualization.

NYC Taxi Data Visualization

HomeDatabase PreviewPassenger Count by DayTrip Distance by Passenger CountPickups HeatmapTrip Counts Across Two Days of the WeekCorrelation AnalysisTip Amount by Pickup Neighborhood


Welcome to the NYC Taxi Data Visualization App!

New York City's taxi system is a dynamic and essential aspect of urban mobility. With more than 400,000 trips daily in 2015, these iconic yellow cabs serve as one of the most frequently used modes of transportation. The vastness and complexity of the city's traffic require a well-organized and efficient approach, making the analysis of taxi data a powerful tool for optimizing operations. Our app gives you the ability to explore and visualize various aspects of NYC's taxi data, offering insights into trip statistics, distribution patterns, and correlations. Whether you're a researcher, data enthusiast, or someone interested in urban transportation, this app provides an interactive and intuitive interface for uncovering key trends and making data-driven decisions.

Here are some of the key features you can explore:

- **Database Preview:** Preview the dataset with variable descriptions for easy understanding.
- **Passenger Proportion by Day of the Week:** View the distribution of passengers across the week in a bar chart.
- **Average Trip Distance by Passenger Count:** Analyze how passenger count affects the average trip distance, based on the day of the week.
- **Heat Pickups Heatmap:** Visualize the density of taxi pickups across NYC.
- **Comparing Trip Counts Across Two Days of the Week:** See a comparison of the number of trips for any two days.
- **Correlation Analysis:** Explore relationships between key variables, including scatter plots and correlation matrices.
- **Radial Chart of Mean Tips by Neighborhood:** Get insights into tipping patterns across NYC's neighborhoods.

The app allows you to interact with the data to draw your own conclusions, making it an excellent tool for anyone interested in exploring the intersection of urban transportation and data science.



This app has been developed using the Shiny package for R.
Group: 14 of the 'Data Visualization' subject in Universidad Politécnica de Madrid.
Group Participants:

- Federico Castellón
- Julián López
- Anja Martín
- Unai Zubizarreta

Home page

In general, within each visualization panel, you'll find an example question along with a hypothetical answer. Below that, the user interface allows you to select the variables and filters to explore the data and generate answers to those questions.

There is a tab that allows the user to **preview the database**, in order to see specific rows that the user can filter by data range. The user can also limit the number of rows displayed.

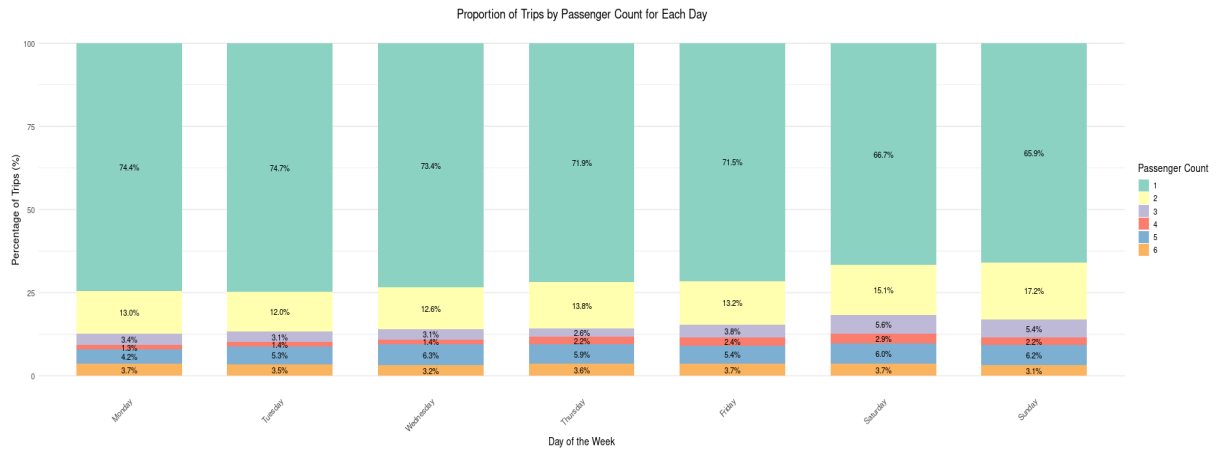
Dataset Preview		Variable Descriptions																
Number of rows to display:		Select Date Range:																
10		2015-01-01to2015-01-31																
VendorID	time_pickup_datetime	time_dropoff_datetime	passenger_count	trip_distance	pickup_longitude	pickup_latitude	RateCodeID	store_and_fwd_flag	dropoff_longitude	dropoff_latitude	payment_type	fare_amount	extra	mta_tax	tip			
2	2015-01-15T19:05:39Z	2015-01-15T19:23:42Z	1	1.59	-73.993896484375	40.7501106262207	1	N	-73.9747848510742	40.750617980957	1	12	1	0.5				
1	2015-01-10T20:33:38Z	2015-01-10T20:53:28Z	1	3.3	-74.0016479492188	40.7242431640625	1	N	-73.9944152832031	40.7591094970703	1	14.5	0.5	0.5				
1	2015-01-10T20:33:38Z	2015-01-10T20:43:41Z	1	1.8	-73.9633407592773	40.8027877807617	1	N	-73.9518203735352	40.8244132995605	2	9.5	0.5	0.5				
1	2015-01-10T20:33:39Z	2015-01-10T20:35:31Z	1	0.5	-74.0090866088867	40.7138175964355	1	N	-74.0043258666992	40.7199859619141	2	3.5	0.5	0.5				
1	2015-01-10T20:33:39Z	2015-01-10T20:52:58Z	1	3	-73.9711761474609	40.7624282836914	1	N	-74.0041809082031	40.7426528930664	2	15	0.5	0.5				

Database preview

In **Idiom 1 - Passenger Count by Day**, the user can visualize the proportion of passenger count for each day of the week.

How common is it for people to share taxis? Does this depend on the day of the week?

By analyzing the data, it seems like on weekdays its more rare for people to share taxis than on weekends. This is likely due to gropus of friends going to and returning from parties.



Idiom 1 - Passenger Count by Day

In **Idiom 2: Trip Distance by Passenger Count**, the user can select a day of the week to obtain aggregated statistics of a specific day by comparing different numbers of passengers.

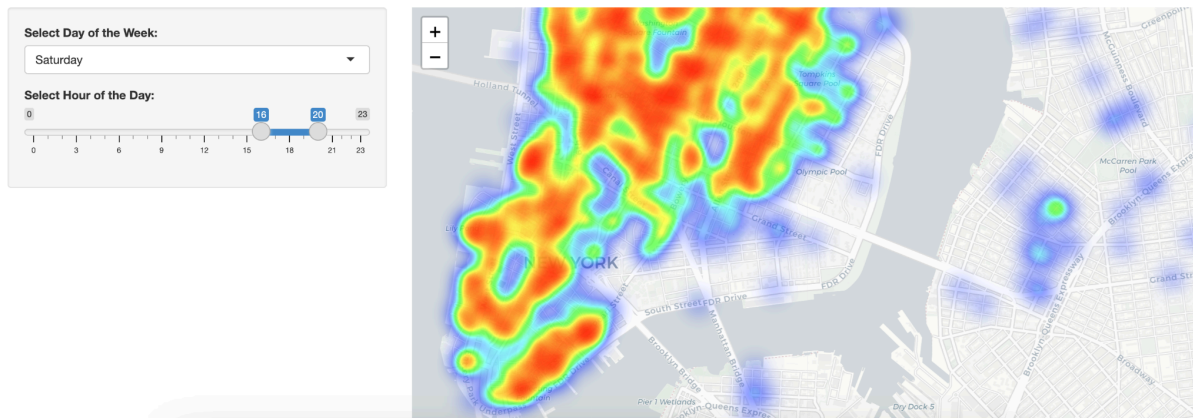


Idiom 2 - Trip Distance by Passenger Count

In **Idiom 3: Geographical Heatmap of Pickups**, the user can select a day of the week and select a specific hours range, so that the heatmap displays the number of pickups within the filtered data, e.g., check the most visited zones on Fridays from 19 p.m. to 23 p.m.

Where do most pick ups take place? Are there different patterns at different times or on different days?

From the heatmap we can observe that most pickups take place in Manhattan. There are also some focuses at important places like both airports (JFK and LaGuardia). This seems to be the pattern across days and time.

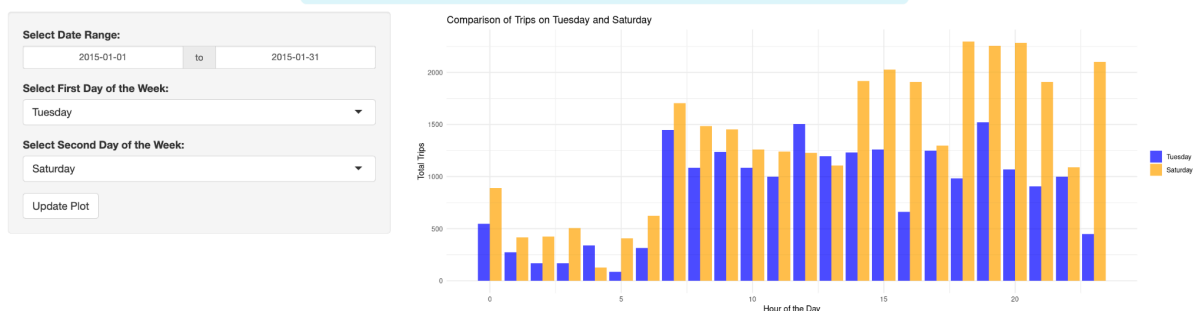


Idiom 3 - Geographic Heatmap of the Pickups

In **Idiom 4: Trip Counts Across Two Days of the Week**, the user can filter data within a specific date range. Then, the user has to select two different days of the week to compare the temporal trends of the number of taxi trips at each hour of the day.

Are taxis specially demanded on some days of the week?

By comparing different days of the week in the barplots one can observe that Fridays and Saturdays have considerably more trips than other days of the week.

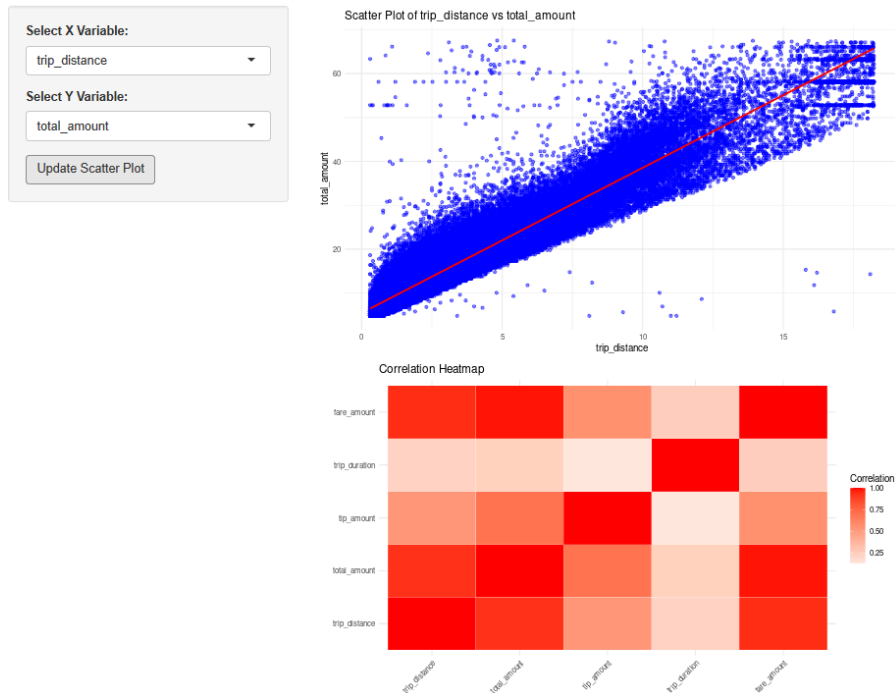


Idiom 4 - Trip Counts Across Two Days of the Week

In **Idiom 5: Correlation Analysis**, the page shows a scatter plot between the two variable inputs from the input box on the left. Below is a heatmap correlation matrix that will give an overview of the correlation value of each pair of variables. For example, if the matrix shows a high correlation between two variables, one may want to plot them in the scatter plot to visualize it. The regression line is a tool to help the user identify the correlation (positive incline means positive correlation).

Which variables are highly correlated? How do they compare?

The most correlated variables are fare_amount, total_amount and trip_distance. This makes sense, because the longer the trip, the more expensive it is. Surprisingly, trip_duration and trip_distance do not seem too correlated to each other. This is likely due to heterocedastity caused by traffic jams, leading to high variances in trip duration.

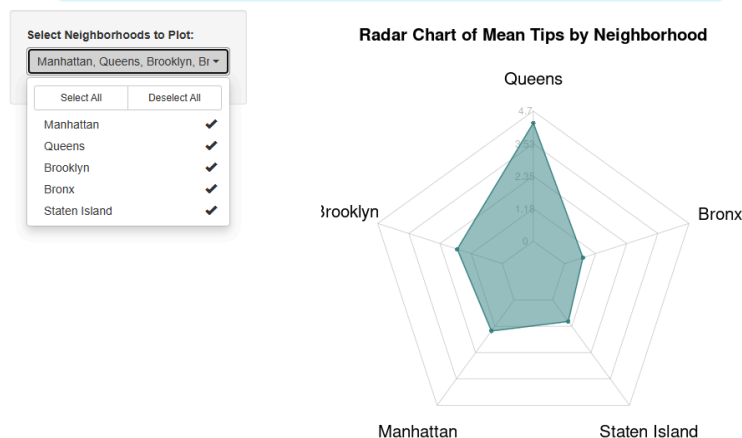


Idiom 5 - Correlation Analysis

In **Idiom 6: Tip Amount by Neighborhood**, the interface shows a deployable list of the neighborhoods explored in the dataset. By selecting at least three of them, a radar chart appears, showing the mean tip amount in a specific neighborhood. the 'pointier' the polygon is in the direction of a variable, the higher the expected tip amount.

Where do clients leave the biggest tips?

It seems like the biggest tips come from clients from Queens. This could be because it is far away, so the trips are more expensive in general. However, the most likely hypothesis is that Queens has very wealthy parts such as Forest Hills, Bayside, Douglaston, and select areas of Astoria. Therefore, clients have a lot of money and are more generous. This is backed by the fact that the most humble neighborhood, the Bronx, leaves the least tips.



Idiom 6 - Tip Amount by Neighborhood