Information Visualization

Fall 2016

NYC Taxi Analysis

Heejong Kim / hk2451@nyu.edu / hk2451
Hai Lan / hai.lan@nyu.edu / hl2892
Aayushi Acharya / aa4999@nyu.edu / aa4999
Himanshu Kumawat / hk1953@nyu.edu / hk1953

GitHub
https://github.com/NYU-CS6313-Fall16/NYC-Taxi-10

Video
https://vimeo.com/196902125

Live Demo
http://54.174.39.79/

What is the problem you want to solve and who has this problem?

New york city, is the most densely populated major city in the United States. Taxis using in New York city is more frequent than all others cities ever in US. Our project is focus on helping taxi drivers in New York City to easily find the “target spots” they want to pick up passengers.

Taxi drivers, especially in crowded city like New York, are facing intense competition every day. Every taxi drivers are hoping to minimize the empty-load time during their work hours and also, get more money earned. Fortunately, NYC Taxi and Limousine Commission (TLC) provide records of all historical taxi data in past few years. With that data, taxi drivers no longer need to find passengers in blind. However, simply showing that data to taxi drivers directly is not a wise way because those data is complex to analyze and it is difficult for drivers to find their interested information from those data directly.

Therefore, developing an effective visual analytic tool to interactively mine and visualize those data to those taxi drivers with their own needs will be helpful for them to work more efficiency. Considering a real scenario: a taxi driver just drop off a passenger and wish to find another passenger immediately, where he/she should go? He/She can simply check on mobile browser and see from current location and current time, which area usually shows the highest taxi demanding. Also, if he/she prefer to have a high tip ratio or prefer to receive some cash in next trip, he/she can also consider the high-tip area or cash-preferred area.

On the other hand, it can also optimize passengers’ satisfactory as well because no passenger prefer to wait a super long time until get a taxi. Taxi company can also get benefits from this to better assign their taxi services.

What are the driving analytical questions you want to be able to answer with your visualization?

Since our project aims to help taxi drivers in NYC to find the target spots, we want to answer the questions that taxi drivers might have.

- **Where is the highest taxi demanding area in Manhattan in different time intervals?** Taxi demanding will change at different place at different time. For example in rush hour, it is possible that more potential passengers may take taxi from residence places to their working places. Counting the pickups in history of each building block with 1 hour as time interval can help to provide the answer. With this analysis and visualization, taxi drivers can easily find in specific time interval, where exactly is the best place near them to pick up passengers.
- **Where and when to find the places those passengers prefer to pay highest tips?** Taxi drivers usually want to earn more for each trip. However, tips amount are not always affected by total amount. By calculating the average tip percentage of total amount and mapping results to map can help drivers find out in which area and time, people prefer to pay higher tips.

- **Where and when is the place that passengers usually pay with cash rather than credit cards?**
  Sometimes, drivers prefer to get cash directly to avoid withdraw from ATM again. Sometimes, they just prefer to get money with credit cards for some safety reasons. By analyzing the credit card/cash ratio in spatiotemporal can help to find the place that fit for those drivers' need.
- **Is there relationship between higher demanding areas and tips (or cash/credit ratio)?**
  In addition to finding the best place for taxi demands, tips or cash, we want to find the relationship between taxi demands and other attributes. Understanding the relationships could be helpful for taxi drivers to choose a place to go.

What does your data look like? Where does it come from? What real-world phenomena does it capture?

Our dataset are yellow cab records that provided by NYC Taxi & Limousine Commission. The first week of April 2016 (April 4th to April 10th) data are used in this project. Those data includes variety different attributes like pickup/drop-off coordinate data, tips amount, total amount, trip distances etc. Besides, shape-files of Manhattan in building blocks scale and neighborhood scale will be used as well by acquired from mappluto NYC building footprints data.

Following are detailed data which are used in this project:

Attribute Name	Attribute Type	Meaning	Values	Derived?
Pickup_time	Quantitative	The date and time when the meter was engaged.	MM/DD/YY HH:MM	No
Pickup_longitude	Quantitative	Longitude where the meter was engaged.	Signed degrees DDD.dddd	No
Pickup_latitude	Quantitative	Latitude where the meter was engaged.	Signed degrees DDD.dddd	No
Building blocks ID	Categorical	Id of each building block	Integer	No
Building blocks	Categorical	Boundary of all building blocks in Manhattan	Geo-polygon	No
Time interval	Ordinal	Time interval to	1 hour	Yes, divide a

		count taxi usage in each building block		week into 168 hours
Number of Taxi pick-ups per building block and hour	Quantitative	The number of taxi pick-ups in different building block in each hour	Integer	Yes, mapping coordinate points of pickups into building block shapefile and count
Number of Taxi pick-ups per neighborhood and hour	Quantitative	The number of taxi pick-ups in different neighborhood in each hour	Integer	Yes, mapping coordinate points of pickups into neighborhood shapefile and count
Total taxi fare	Quantitative	The total amount charged to passengers. Does not include cash tips	$00.00	No
Tip amount	Quantitative	Tip amount – This field is automatically populated for credit card tips. Cash tips are not included.	$00.00	No
Payment type	Categorical	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	Categorical 1= Credit card 2= Cash 3= No charge 4= Dispute 5= Unknown 6= Voided trip	No

Those data are pre-processed before visualizing. By using raw data, the first step is to mapping each pick-up point to each building blocks in NYC. There are total over 40000 building blocks in manhattan and 3 Gb taxi records data. To successfully do this mapping (spatial-join), We used Apache Spark on cluster to filter those data first. After this step, we can link each pick up record a building block id. Then we can use building block id as a key and applied 1 hours time interval as well to do “word count”. By doing this, we can know in each hour and each building block, how many pickups happened. With the similar technologies, we calculate the average tip ratio and average cash using ratio in each building block in each hour.

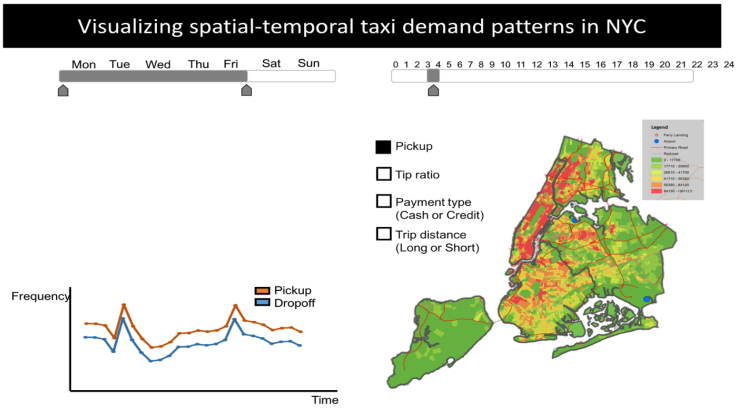
What have others done to solve this or related problems?

There has been many attempts to analyze the NYC Taxi data. NYC Taxi & Limousine Commission (TLC) analyzed the taxi data and provided TLC Factbook in 2014 and 2016 [1,2]. The factbook contains the analysis of trip trends, fares, drivers and app usage [1,2]. Since the data were collected by Taxicab & Livery Passenger Enhancement Programs (TPEP/LPEP) to improve the riding experience of passengers, the TLC Factbook depicts an overall trend of taxi data. We focused on analyzing taxi data from taxi drivers, and our projects is a specific version of the Factbook. The Factbook analyzed the taxi data for each year, but we focused on one week.

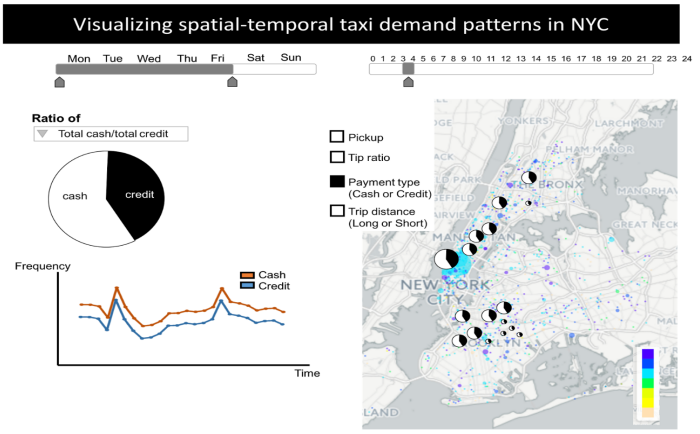
For the visualization, Chris Whong created “a day in the life of a NYC taxi”. The project visualizes the data of one random NYC yellow taxi on a single day in 2013 showing all pick up and drop points [3]. Juan Francisco Saldarriaga programmed two interesting visualization of map which are “New York City Taxi Activity” and “Taxi!”. Those visualizations map the trip data for 24 hours [4,5]. These two projects are related to our project since they visualized the data in spatial format. We also used map for the visualization.

There is another project entitled “Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance”. This project visualize the data of NYC Taxi and Uber Trips with some explanation about few topics [6]. It relates to “NYC late night taxi index”, “Specific area taxi pickups”, and “Cash and credit NYC taxi payment” by comparing NYC taxi and uber, taking our project one step further and considering seven years of data [6].

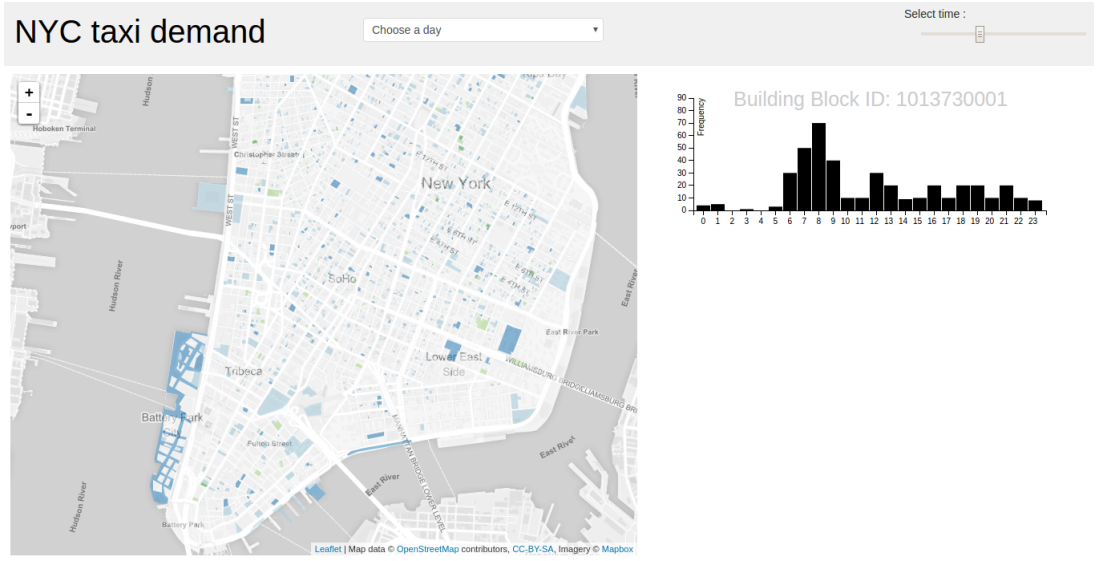
Design Iterations



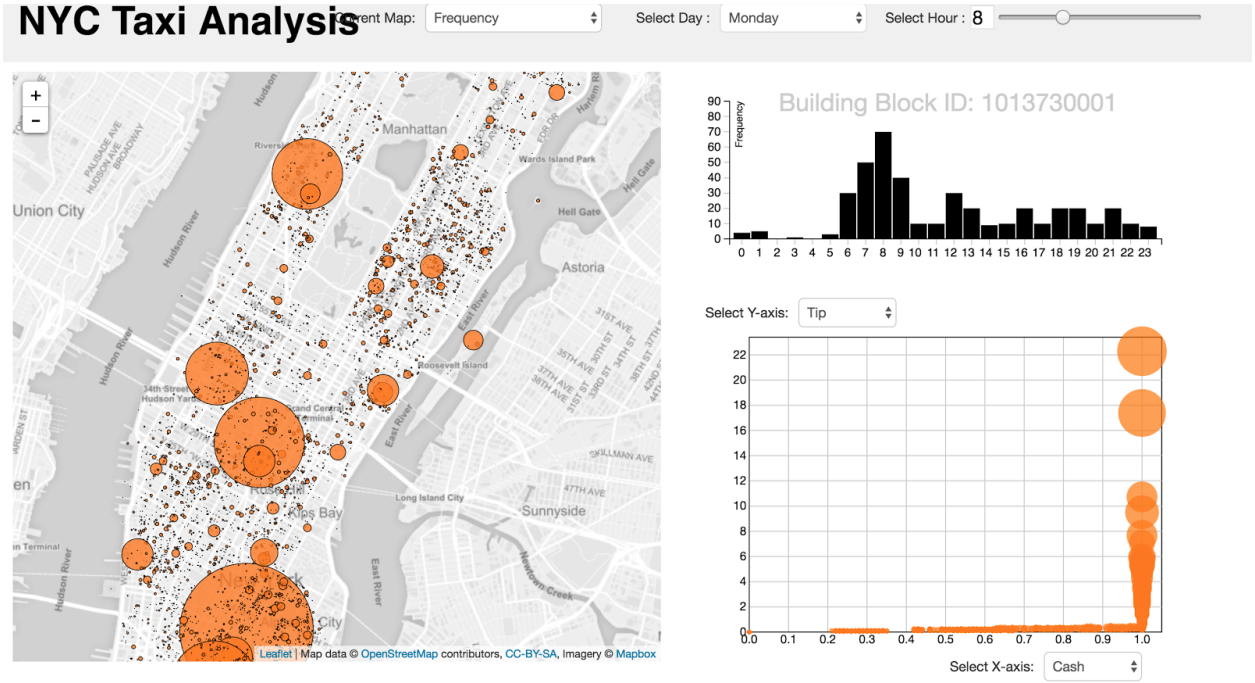
- This was a part of our original mock up and it showed a time and day slider with 4 checkboxes for different views of visualization.
- For the first option of Pickup Frequency, the primary view was a choropleth map with the denser colors showing more pickups in that area.
- Once a block was selected on the map, the graph on the left showed showed the frequencies for that building block across the parameters chosen in the top header.



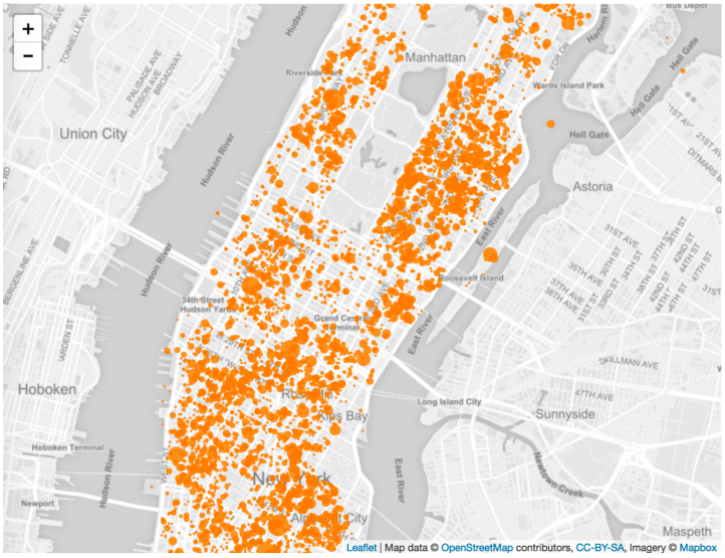
- Just like the above, the header parameters remain the same of choosing the day and time. But, with the check box of payment type selected.
- Hence, this shows small pie charts on the Geo-map showing the ratio of cash to credit card chosen as the payment method.
- A chart for the same building block would be shown just with payment types as the two lines.



- This was the first mock-up after some coding work on the project. On the left, there is the choropleth map.
- The blue building blocks show varying intensities with respect to pickup frequencies, but this did not give a clear picture. Since the blocks were at a very granular level, the differences were not evidently seen.
- The header has a drop down to choose the day and a slider to choose the period of one hour.
- We chose a bar chart since we had only one parameter to show on the Y axis and were using only the “pickup” data now, so there was no need of two lines on the “payment type” chart, either.

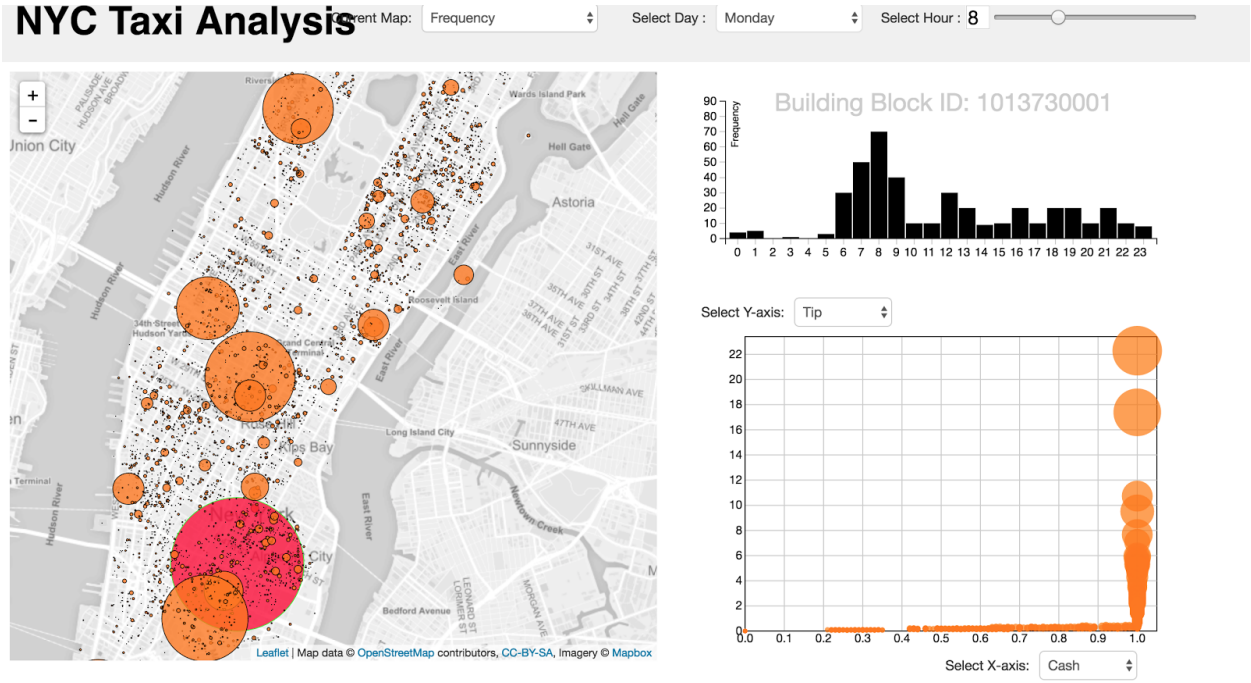


- Since the choropleth did not work out, we decided to change the visual mark to circles (after feedback from the professor in the class).
- Circles' radius are proportional to the number of pickups.
- This made it easier to visualize certain areas with higher pickups and choose it or zoom in if necessary.
- We were taught a guideline of 'Eyes beats memory', and hence decided to incorporate everything in that one page and not going ahead with the check-boxes design.



- This is additional map that we've tried. Each dot represents building block.
- Circle's radius are proportional to the number of pickups but in log scale.
- We didn't chose this map since it is hard to see the difference between building blocks.

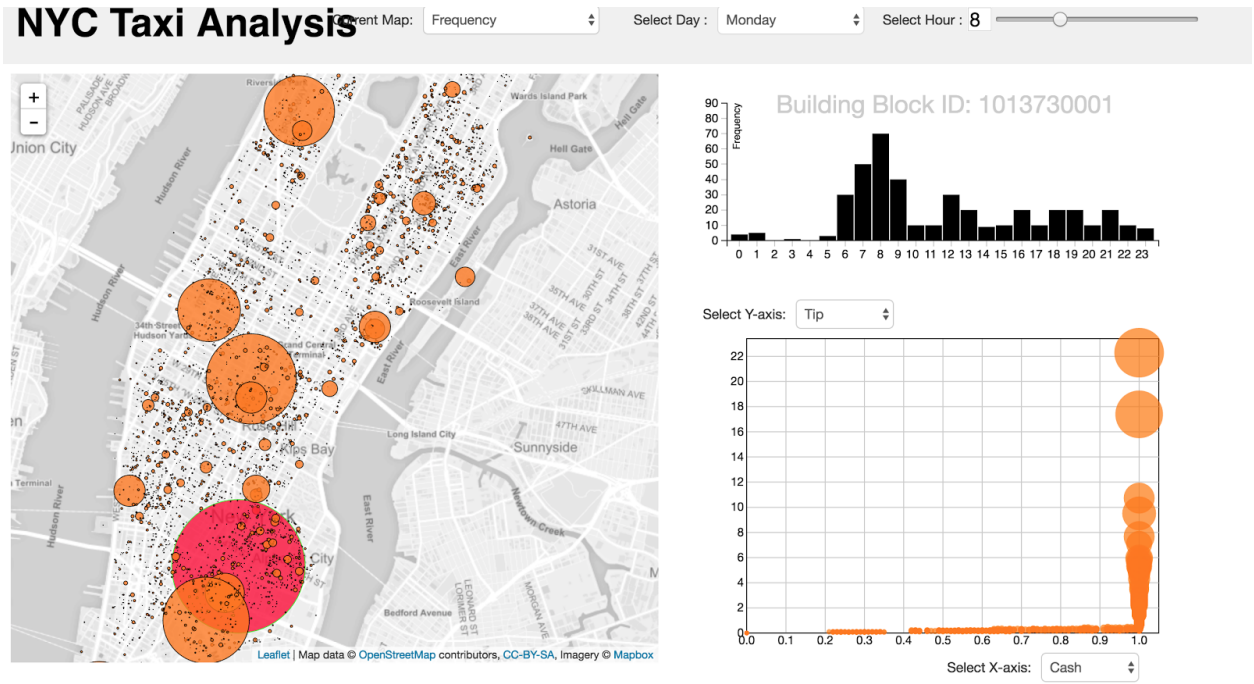
Final Visualization



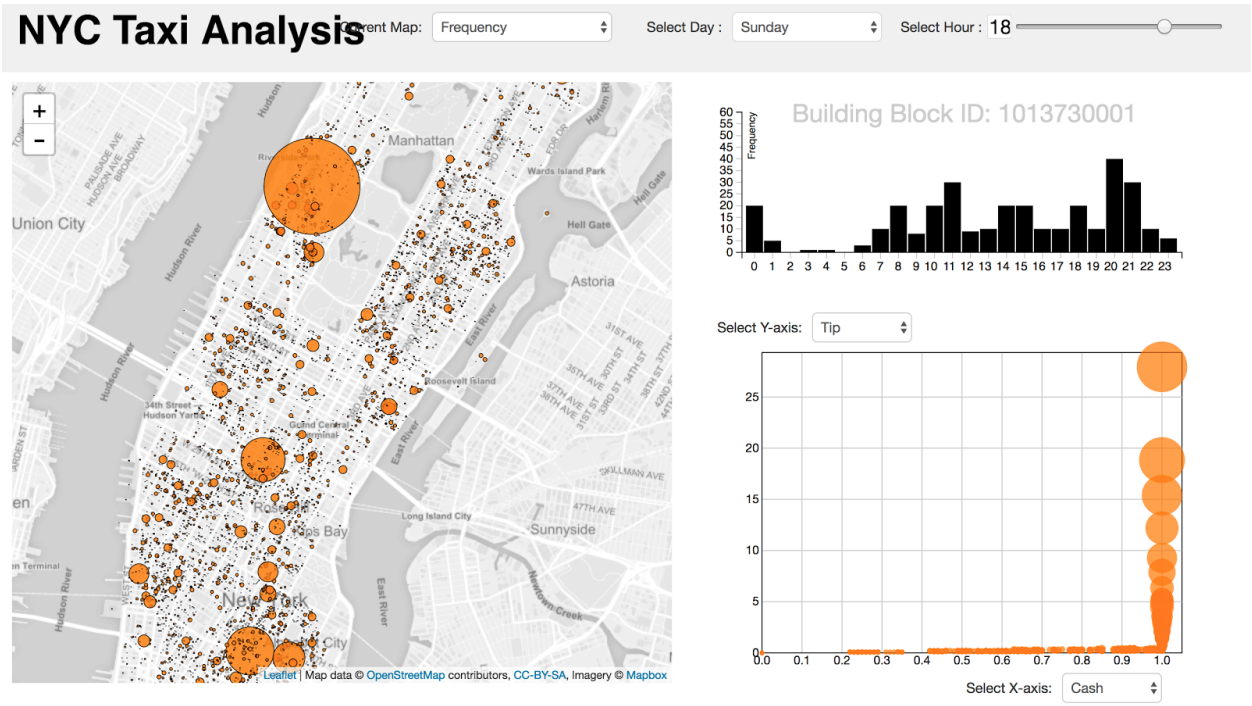
- The above picture shows our final visualization. There are 4 elements to it: The header, Manhattan Map, bar chart, scatter plot.
- Let us start with the **header**. The “current map” drop down shows which attribute to display on the map (which is related to the scatter plot). The three options in the drop down are Frequency, Tip and Cash.
- The day drop down has options from Monday to Sunday. Time can be selected by sliding the bar and the respective hour will be displayed in the small box next to the slider.
- On the left we have the Manhattan **map** with circles representing pickup frequency with its size being proportional to the number of pickups.
- Each circle when clicked, will change color as shown, get highlighted and send the respective building block to the bar chart.
- The **bar chart** overall shows the Frequency data/trend for one particular day (selected in the drop down) across the 24 hours for the particular building block selected.
- The chart shall keep on dynamically changing when different blocks are selected to give an overall trend of the entire day.
- The **scatter plot** shows the correlation between different attributes. The parameters could be changed by choosing the option in the drop downs given on both axes.
- The attributes include Cash, Tip and Frequency. This is used so that the comparison between different parameters could be done to find if there existed any logical relationships.
- The size of the bubbles on the scatter plot relate to the bubbles on the map with their sizes. For the above picture shown, the largest circle on the scatter plot is also the largest bubble on the map (selected one).

- The scatter plot for the above example could be read as; the circles around 1.0 on the X axis means they are definitively paid in cash and highest tip was 22% of the normal amount due.

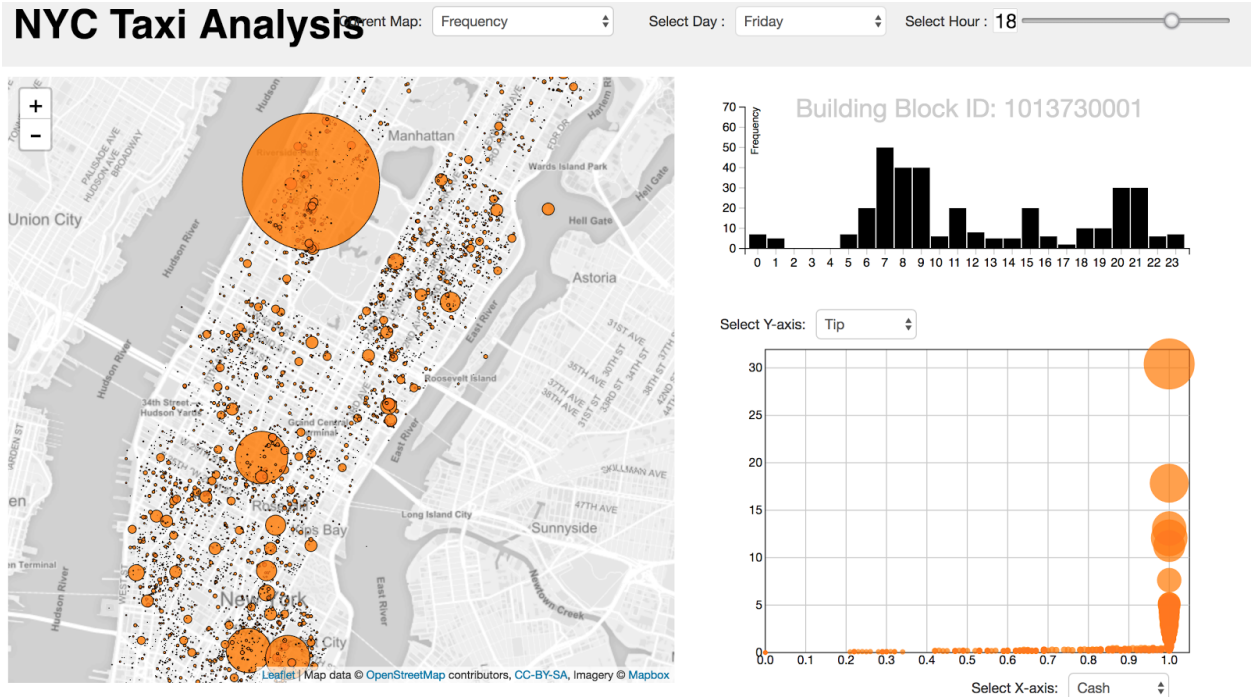
Findings



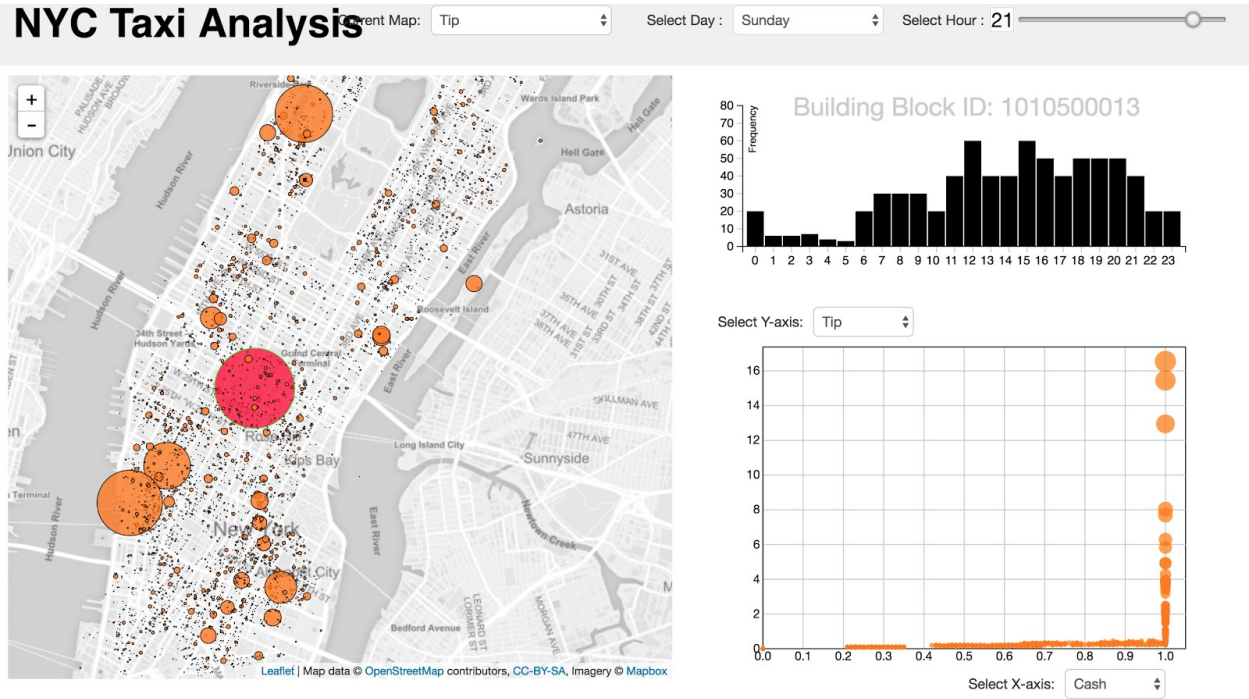
- Overall, relating to the primary question that our project would want to answer, as expected, morning times had a much higher frequency rate for pickups than the other times across the weekdays (normal office hours)



- Across the weekends, a “single” prime time was less obvious, since they were evenly distributed. However, overall, there were fewer pickups in the morning and much more in the evenings as shown in the picture above.
- What was interesting though, were the locations of the highest frequency (shown with larger bubbles).
- More or less they always remained the same; South point of Manhattan (South Ferry stop, New York Stock Exchange, Battery Park), Midtown Manhattan, Lower West Side.



- The above picture shows the pickup frequencies for Friday evening time. Since the area with the largest bubble is Upper West Side, we believe the reason for the same would be the famous museums like American Museum of Natural History and Central Park itself which are famous tourist spots.
- This gives a very clear picture to the taxi driver about which places to go to for a higher chance of pickups.



- This was one way of of helping the taxi drivers get more pickups. To earn more money, the drivers will of course want to get a larger tip. Since tips depend on a lot of random factors, one cannot say that if it is a longer trip in distance, it will surely conclude with more tip.
- The above map shows the “Tip” selected in the dropdown in the header and hence map shows the ratio of tips on Sunday at 9 pm
- Since Midtown Manhattan consists of major clubs and places to hang out, highest tips are given when pickups are in the areas shown on the map.
- The selected area also has many restaurants and tourist attractions.
- As we can see in the scatter plot, the tips are dealt with in cash.
- So if the driver is looking to make some extra cash, he can make sure to go to places with overall higher tip ratio (more bubbles towards the top) and at times with pickup frequency high by using the bar chart!

In general, the things we realized after creating the entire application are:

- It is a good visualization when there are a lot of questions since it means the user tried to figure out the right things and extract information from the visualization provided.
- Sometimes it can be necessary to change the visualization even after creating a primary view because how you expected to visualize might be different from how it actually looks.
- The rules/guidelines taught in this course with respect to the data marks and visual channels should be kept in mind because it helps in effectively choosing the right type of charts at least for the primary view

Limitations and Future Works

- Overview of the visualization is missing. What this means is that a general summary of this data could be first given.
- The limitation in this aspect is that the user can only see data specific to a particular day and time.
- This could be changed to reflect data on a monthly basis just to see the generic trends if a taxi driver has no idea where to start from.
- After a general idea is given, the user could choose to go one level deeper in the hierarchy to the granular details.
- This would include choosing a particular time and day to decide about certain factors (as the project focuses on) and know which spot would be the most beneficial for the driver.
- So, the next step in the project would be to add more data for an overall view and not make it specific to the one week of data that we chose.
- Additionally, we can bind the coordinates data of the map with neighborhood details, so that the building blocks will be described in more detail/neighborhood names when hovered over.
- This will make it easier for the driver since he/she will get an idea about the preferable area he/she should be in and will be more familiar with those neighborhood names.

References

[1] Taxi, N. Y. C., & Limousine Commision. (2014). Taxicab Factbook. New York Taxi and Limousine Com http://www.nyc.gov/html/tlc/downloads/pdf/2014_taxicab_fact_book.pdf.

[2] Taxi, N. Y. C., & Limousine Commision. (2016). Taxicab Factbook. New York Taxi and Limousine Com http://www.nyc.gov/html/tlc/downloads/pdf/2016_tlc_factbook.pdf.

[3] Whong, C. (2014). NYC Taxis: A Day in the Life. http://chriswhong.github.io/nyctaxi/

[4] Saldarriaga, J. (2012). New York City Taxi Activity. https://vimeo.com/35433719.

[5] Saldarriaga, J. (2011). Taxi!. https://vimeo.com/31298658.

[6] Schneider, T. (2015). Analyzing 1.1 Billion NYC Taxi and Uber Trips, with a Vengeance.