New York City Taxis:

Usage Statistics: Data Exploration & Insights

Data Details:

Description:

This is data from the New York City Taxi & Limousine Commission (TLC) which gives data records about trips taken (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml).

There are three kinds of taxis in New York, for which data and dictionaries are available:

1. Yellow Cabs
2. Green Cabs (or borough cabs)
3. FHV (For Hire Vehicles)

Range:

The data is available from Jan 2009 to Jun 2017 - data for yellow cabs from Jan 2009 to July 2017 (19 parameters), data for green cabs from Aug 2013 to June 2017 (20 parameters), and data for FHV from Jan 2015 to Jun 2017 (3 parameters).

For exploration purpose of this project, we retrieved data from year of 2015 for both Yellow Cabs and Green Cabs. The dataset contains more than 146 million records with 21 parameters and has a total size of about 25 gigabytes (GB).

Parameters:

The main parameters that would be used in the project are the following::

1. Vendor ID
2. Date Time Stamps - Pickup and Dropoff
3. Trip Distance
4. LatLong Stamps - Pickup and Dropoff

5. Fare Amount (along with breakup of tips, tolls, surcharges and taxes, etc.)

The total list of parameters in the data (21 parameters) and their data types are given below:

Data Label	Data Type
trip_distance	double
pickup_longitude	double
pickup_latitude	double
dropoff_longitude	double
dropoff_latitude	double
fare_amount	double
extra	double
mta_tax	double
tip_amount	double
tolls_amount	double
improvement_surcharge	double
total_amount	double
ehail_fee	double

Data Label	Data Type
vendorid	int
passenger_count	int
ratecodeid	int
payment_type	int
trip_type	int

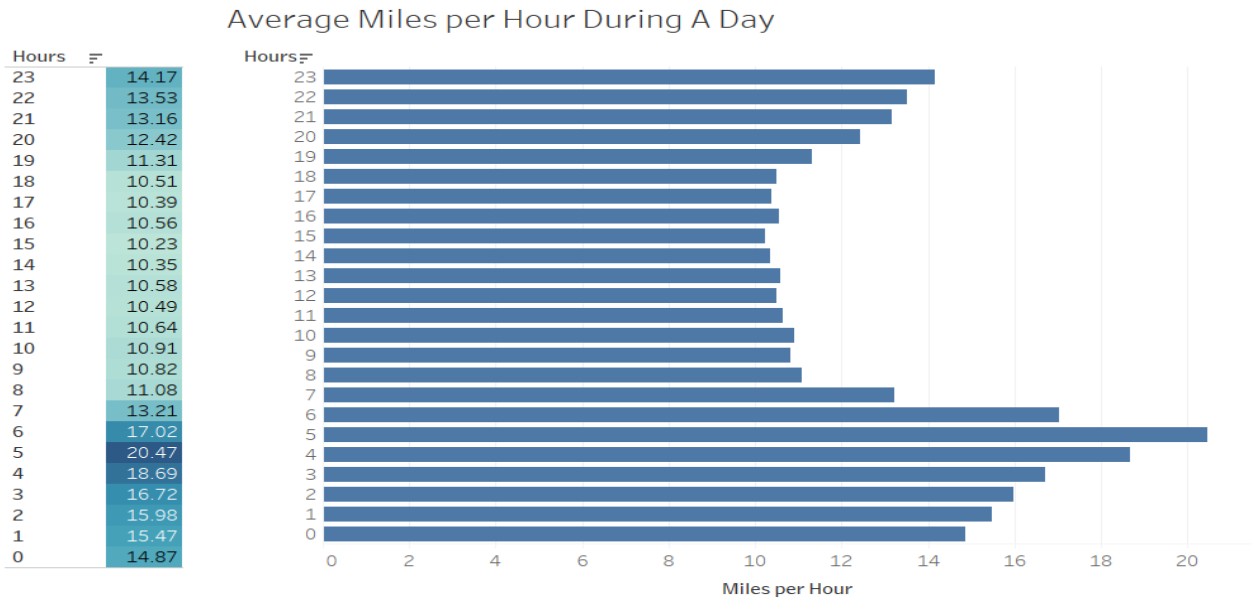
Data Label	Data Type
lpep_pickup_datetime	timestamp
lpep_dropoff_datetime	timestamp

Data Label	Data Type
store_and_fwd_flag	string

Hadoop Tools Used:

- **Interface:** Hue, AWS Athena
- **Framework:** AWS EMR, Presto
- **Languages:** Hive, Pig, R, SQL
- **Storage:** HDFS, AWS S3
- **Others:** Tableau

Findings: What is the average speed of a cab?



What we wanted to check:

- What is the average speed that a cab in NYC travels?
- Does this change with the time of day? If so, how?

How did we do this?

- There is a timestamp field in the parameters provided, which is in the yyyy/mm/dd hh:mm:ss format.
- There is also a total distance travelled parameter.
- Using these parameters, average speed was calculated.

- This was plotted against hour when this ride was taken.

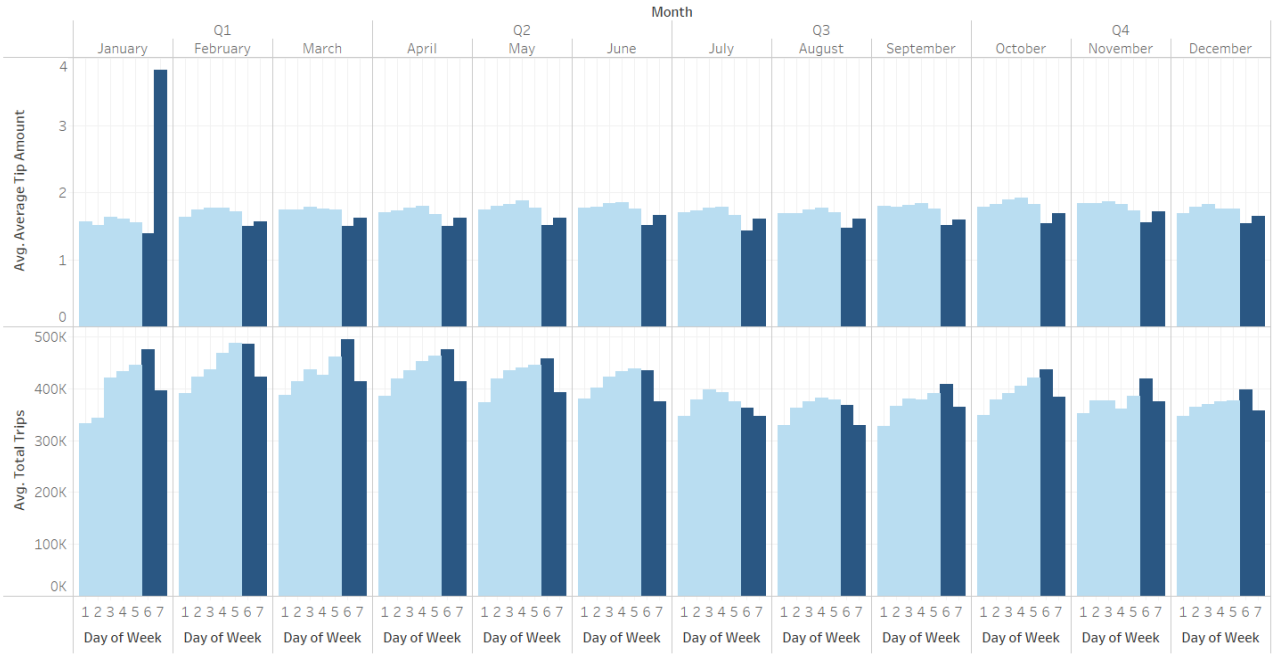
What we found

- Average speed is >13mph between 9pm and 7am and uniformly below 11mph for the rest of the day.
- Highest speed is between 5am and 6am – 21mph.

What we inferred

- NYC has a lot of traffic all through the day!
- The best time to go for a peaceful drive in NYC is early in the morning!

Findings: How do people tip cabbies in NYC?



What we wanted to check:

- What is the average trend of tipping?
- Does this change over months, or day of the week?

How did we do this?

- There is a tip field in the parameters provided.
- There is a timestamp for the trip start, which is in the yyyy/mm/dd hh:mm:ss format.
- Using these parameters, average tip was calculated.
- This was plotted against day, week, month.

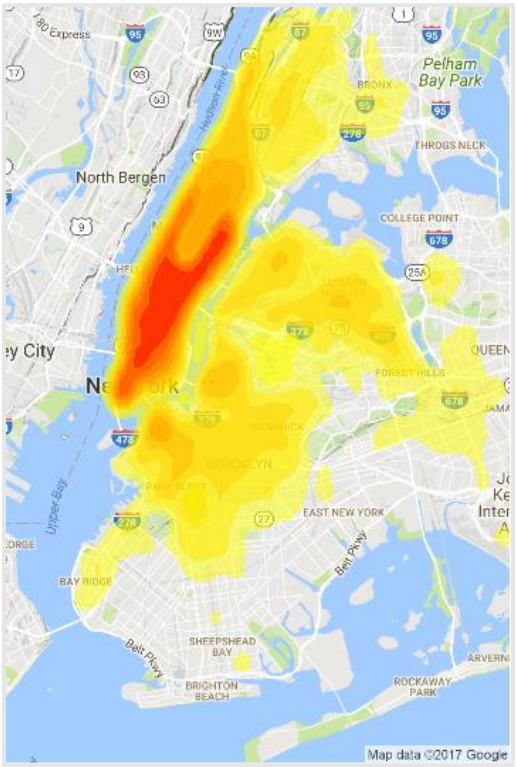
What we found

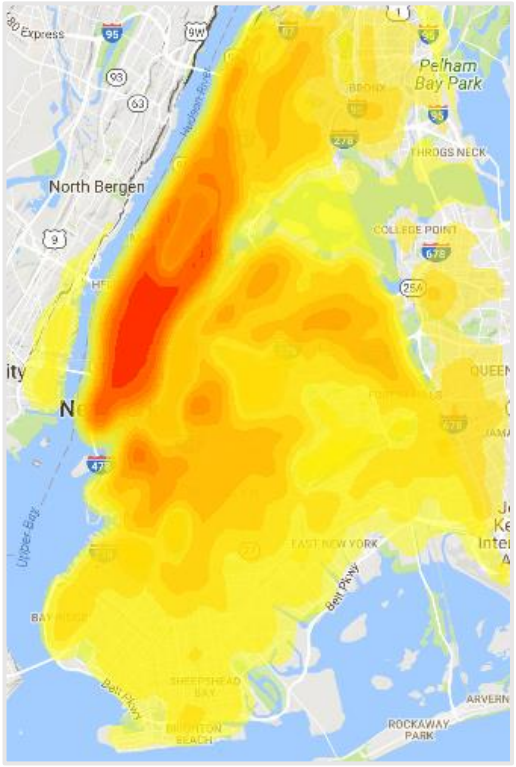
- Average tip is about $1.7.
- Average number of rides / day shows an increasing trend through the week (lowest on Monday, highest on Saturday).
- Weirdly, average tips are lower on Saturday!

What we inferred

- Cabbies shouldn't expect to make money through tips on Saturdays even though it's busy!
- Best time for cabbie's day off? Monday!

Findings: Where are maximum pickups and drops?





What we wanted to check:

- Which location has maximum pickups?
- Which locations have maximum drops?
- Is there a trend?

How did we do this?

- There were latlong values given for both pickup location and dropoff locations.
- Heatmap created using maps in R.

What we found

- Maximum pickups and drop-offs are both in Manhattan.
- However, drop-offs extend to a much wider area – even in the outer boroughs.

What we inferred

- A lot of people go from Manhattan to the outer boroughs. These could be people coming to Manhattan for work in the day.
- As a cabbie, you want to stick as close to Manhattan as possible to get long rides!

Findings: Is there a place where a higher fare is guaranteed?

What we wanted to check:

- What is the trend of total cost?
- Is there a place where you can get a fare that will definitely be higher than average?

How did we do this?

- Latlong values were given for pickup location.
- This was captured along with the cost of the trip.
- Scatter Plot created using maps in Tableau.

What we found

- There's no specific trend here – most places where pickups happen are pretty close to average fare.
- There are some in the outer boroughs that give a higher fare, but that is expected.

What we inferred

- Cabbies should not stick to one place in the hopes of getting a higher fare.
- Most fares, from most places are pretty close to average.

Learnings:

- A csv file of 1.8GB can take ages to download / upload.

- Given a dataset size of 25GB, we couldn't download and work on the all the data locally (or in most cases, even view it completely locally). Hence, most of the work we did was done on AWS.

- AWS accounts (and hence S3 buckets) can be shared between people through IAM. So not everyone needed to upload the data into their AWS account.

- AWS has a tool for pretty much everything you can think of - from storage (S3) to detection, definition and alteration of schemas (Glue), to Analytics (RedShift), to Visualization (QuickSight), and everything else.