**Taxi Data NYC Per Month**

Data Available Per Month since 2011 <https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page>

https://docs.microsoft.com/en-us/azure/open-datasets/dataset-taxi-yellow?tabs=azureml-opendatasets

Each Month has 2+ Million Records

* doLocationId
* endLat
* endLon
* Extra //Miscellaneous extras and surcharges. Currently, this only includes the $0.50 and $1 rush hour and overnight charges.
* fareAmount
* improvementSurcharge
* mtaTax
* passengerCount
* paymentType
* puLocationId
* puMonth
* puYear
* rateCodeId
* startLat
* startLon
* storeAndFwdFlag //This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, also known as “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.
* tipAmount
* tollsAmount
* totalAmount
* **tpepDropoffDateTime**
* **tpepPickupDateTime**
* tripDistance
* vendorID
* vendorID

Data Cleaning:

* Phase where we need to explore each and every feature that we have and clean it out by checking the extremes and averages (this could be clearly visualized using box plot).
* Filling missing values.

Possible Queries

* Measure change of a parameter(s) throughout the day based on dropoff/pickup time and location.  
   E.G. measuring how tip values change throughout a day at different locations.
* Measuring change of a parameter(s) through the months/years.  
   E.G. How tip values have changed since 2011.
* **Maybe focus on parameters pre and post the introduction of Uber, Lyft, etc.?**
* Heatmap of driver locations throughout the day
* For each location, we can see the relation between tip, payment method and fare amount to further help drivers choose where they operate
* **We can pick holidays across a calendar year and then show a trend of average tourists in the city, or pick out specific days in a month when the city has the most people (This looks really great! - Deepak)**
* We can use a clustering algorithm to derive the clusters across the regions and post average pickup in each region by month or time window (say before k days).
* We can display the trend of how the pickup rate is correlated with various times of the day like in the morning - high, lunchtime it falls down and rises up at evening.
* Linear regression forecast at each cluster for every k time bins, which will give a good understanding to the driver where he needs to go.
* We can display the frequency of payments modes of passengers, which will be helpful for banks or other entities.
* We can also get the total metered tax paid by the drivers across different regions, which will give a view on the actual profit made by the drivers.
* We can display the relation between trip-distance and weekdays ⇾ trip distance on weekends must be higher than the other working days.
* Take in average prices of taxis in an area over a certain amount of time to draw conclusions about things such as inflation - K
* Could determine what days and times of day taxis are being used most so that drivers can figure out when they could maximize their profits - K
* **Could potentially average out the number of passengers(or other parameters) per month from the years 2019-2021 to see the effects that COVID may have had on people taking a taxi. - K**
* Track the counts of different payment methods over time used by passengers for taxis to determine the effects of the introduction of online/mobile electronic payments. - K
* Trend query with x-axis as time (5 trends)

Potential Problems

* Do we have too much data for the scope of the assignment? 2.5 million records for January 2022, and 3 million records for February 2022. Finding trends month by month could be a lot of data, should we focus on only 1 month and analyzing data based on time of day? ⇾ we need to sample data in such a way that you have data for all the months and the tuples must sum across 1 million. 1 million should be fine, I think.
* Can we focus on adjoining areas with concentrated data and remove the outliers to filter 2mil records per month, so we have limited data to work with

⇾ We will be running data-cleaning functions on top of relations, so it should filter the outliers.

Query List :

Heatmap for total rides over time for different locations in NYC

Effect of Covid-19 on revenue generated or trips for NYC cabs over time :p

Effect of Uber, Lyft on cab rides from 2011 onwards and 2015 onwards

Taking holidays across a calendar year, observe the trend for trips, tips and overall fare maybe?!