NYC Ride Hailing Optimization – Fare Amount, Pickup Points and Collision Predictive Analytics

Amol Wadwalkar, Jagannadh Commuri, Promodh N Ravichandran,
Sriram Ayyagari, Suresh Kadavath

Introduction

Currently, taxi drivers rely on historical industry knowledge of the area in which they work to attempt to maximize profit and reduce risk of collision. Uber drivers are enabled with tools/interfaces that provide them with data-driven insights that Yellow and Green taxi drivers lack. We have analyzed NYC Taxi, Collision & Uber datasets for New York City to understand the relationship between pickup frequency, traffic, and fare amount. Visualization is provided for drivers to interact with and identify locations for increasing pickups volume while accounting for the current likelihood of getting into an accident. The size and feature richness of NYC Taxi, Collision and Pickup points, and digital experience that we could bring leveraging the data have motivated us to work on this project.

Project Website can be accessed via:
http://ride-sharing-project.s3-website-us-east-1.amazonaws.com/

Problem definition

Few cities provide an online infrastructure to hail a ride, but the technology is the bare minimum and provides no useful information which will help taxi drivers to make decisions. This project aims to address the issues by leveraging data and providing insights into the traditional taxi drivers.

We believe that our project is better than the state of the art because currently there are not many options/digital technology tools available for traditional taxi drivers that would help them with a data-driven solution to increase their efficiency and to maximize their profits by helping identify the demands in pickup locations. The user interface provides features, such as heatmaps for visualizing pickup locations, accident-prone zones, demand for pickup locations, and variations of these patterns based on average daily historical data (day-of-the-week and hour-of-the-day).

Survey

Our survey consists of multiple papers related to the study of traffic, weather, collision patterns leveraging Uber, NYC datasets. These papers were helpful as we are working with similar datasets and these papers establish a sample framework for visualizing data. This is beneficial from a high-level understanding to help frame our choice of different analyses to perform. Some of the salient features from these papers are regression analysis with data from many sources, correlating collisions to taxi pickup points for a given day (day-of-the-week) and location, insights about computational processing of large datasets, computational aspects and the interactive usability of the data. For more details of the survey - (Project Survey)

Proposed method

Intuition

We found that there are opportunities to provide better customer experience by leveraging the features that are available in large data sets. Based on what we learned in this course, we thought it would be a good use-case to work on this project and come up with innovations such as:

- Create a better user experience by consolidating different heatmaps into a single view.
- Provide better predictability of fare amount, pickup demand and collision probability for pickup zones.
- Provide a web tool-tip feature indicating historical and predicted values of above metrics for a selected location/zone on the map.

Description of the approach/design

Our solution design and approach include data collection, data cleansing, data crunching, data extraction, and exposing the data metrics through REST-APIs. User experience is created through web interfaces using D3, JavaScript, H3 and MapBox-APIs. The solution consists of multi-tier architecture utilizing AWS services and Azure ML services.

1) Data were extracted in a consumable format and then stored into AWS S3 buckets for different visualization use cases such as traffic, collision heatmaps, prediction of price, demand for pickup zones.

2) AWS EMR/Hadoop/Pig services to process the datasets. During data processing additional attributes such as Fare amount, the Trip distance was introduced to predict the price.

3) AWS RDS service is used as a MySQL database to store the processed data.

4) Leveraged H3 algorithm (https://eng.uber.com/h3/) to map latitude-longitude coordinates into NYC Zones. Zone information is required to visualize the data for better customer experience. Resolution value of 10 was used during Zone Calculations.

5) User experience is dictated by the JavaScript framework by utilizing the data that is extracted for various use-cases as mentioned above. Users are provided with choices to focus on different map views such as a number of pickups and collisions by selecting the checkboxes. When both the choices are selected, the combined value is derived by giving positive weight to pickups and negative weight to collisions for a better visual representation.

6) APIs deployed in AWS Lambda (with Node.js) is used to connect to data stored in AWS S3 for historical values.

7) Azure ML models that we developed to predict average fare amount per mile at any given pickup location within NYC (specifically for Yellow and Green Taxi data sets as Uber data set doesn't contain fare amounts). Ref#

8) Based on the data elements like fatalities, injuries, we categorized collision data into 3 categories.

- R - when there is a fatality.
- O – when there are injuries to passengers.
- Y – when there are no injuries.

We used these categories in training the model. We used a classification model to predict the category of a collision.

9) Additional 2 models have been developed to predict collision-prone zones and number of pickup counts for a given zone.

a) For collision, Multiclass Decision Jungle model from Azure ML studio has been chosen. This model can be used to predict a target that has multiple values.

b) For fare amount, and pickup counts we used Fast Forest Quantile Regressions (A decision tree-based quantile estimator) from the ML studio. It is supervised learning that can predict values for a specified number of quantiles, unlike linear regression that predicts the value a single estimate, *mean.*

These models include features such as trip distance, day-of-the-week, and hour-of-the-day to predict the fare amount category. Models were experimented using 2 training data sets – collision data with 50-50% train-test split, and the fare data with 90-10% train-test split with cross-validation. Some of the details about the approach and operation that we used are described below.

Data Collection

Data collection hinges upon 5 datasets ranging between April-Sept 2014 and Jan-Jun 2015 related to Uber NYC pickup & drop off locations, NYC Yellow and Green taxi dataset. Important attributes selected were datetime, location (latitude-longitude), fare amount, the trip distance for our analysis. Additional but important data attributes such as dataset qualifier (uber/yellow/green taxi indicator), and day-of-the-week are also introduced. For Collision data, we introduced a category to indicate the severity of the collision (R/O/Y). The raw data consisted of approximately 180 million records and a total size of 15 GB.
The following data has been used for analysis and prediction:

- Trip data from NYC TLC (New York City Taxi and Limousine Commission)
- Collision data from NYC Open Data
- Uber trip data from five-thirty-eight

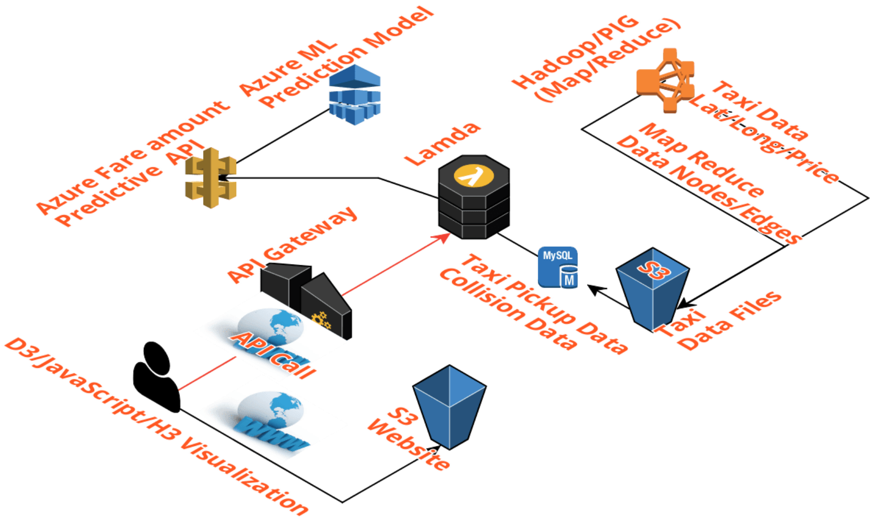
Data Cleansing

The team selected AWS EMR, Hadoop, and Pig for processing the data as detailed below:

- Feature selection
    - Source qualifier (‘u’ for uber, ’g’ for a green taxi, or ‘y’ for a yellow taxi)
    - Datetime
    - Day of the week (‘Mon, ’Tue’, etc.)
    - Fare Amount
    - Trip Distance
    - Pickup Location (Lat/Lon)
- Eliminated null or empty records for the selected attributes.
- Filtered distinct records
- Formatted the data

Using this processing framework, about 180 million original yellow, green, and uber trip records for the periods April-Sep 2014 and Jan-June 2015 have been parsed and calculated. For the uber dataset, the Fare amount and trip distance are set to 0.0 while extracting the data but not intending to use it as part of our prediction model.

Reference Architecture Diagram


Azure ML Prediction Model
Hadoop/PIG (Map/Reduce)
Taxi Data Lat/Long/Price
Azure Fare amount Predictive API
Lamda
Map Reduce Data Nodes/Edges
API Gateway
MySQL
M
S3
Taxi Pickup Data Collision Data
Taxi Data Files
API Call
D3/JavaScript/H3 Visualization
WWW
S3 Website

Analysis

The team reviewed different concepts for the project work based on ideas from team members and public datasets that are available. The team came up with concepts such as - NYC government published public data sets such as “Medallion Vehicles”, “RideAustin”, “Impaired Driving Death Rate“, “NYC taxi information”, “NYPD vehicle collision information”, “NYC Central park weather history information”. Uber also published datasets for 2014 and 2015 with basic data elements. After analyzing the different datasets, the team agreed to focus on opportunities that would provide valuable information and interactive digital experience. So we focused on three data sets, TLC Trip Record Data(Ref#), NYPD vehicle collision information and Uber NYC Taxi trips(2014 & 2015). We couldn’t leverage weather data into consideration for predicting fare amounts and collision factors as we didn’t have enough time before submitting the project.

The data we considered from the datasets consists of records with data elements such as Taxi Information (pickup location/date time/fare amount/drop-off location/fare amount/tax/tip/trip distance/travel time etc). Uber Data (pickup location/date time), Collision Information (incident location/injured people count/people got killed count/pedestrians involved/date time/nature of accident etc.) We considered below data features for our analysis to come up with a correlation between pickup location/date time/collision counts & nature of the collision and fare amount. Data dictionaries for each of the datasets were attached in the package for reference (Ref#).

Since the datasets are very big, consists of around 180 million records (15gb of data) with all data in CSV format, we considered Hadoop/pig scripts to extract data and to cleanse. We have written pig scripts and created ESM Hadoop node clusters and cleansed data. Since initially were running these scripts and it was turning out expensive to run in EMR, so we created Hadoop clusters and started running pig scripts locally (there was performance lag with running pig scripts locally with 1st big data processing was taking around 3-4 hours).

Final data after extracting the key data elements required for our project consists of pickup location (latitude/longitude), date time (day of the week and hour), average fare amount per mile for given weekday (average of aggregation of fare amount and aggregation of trip distance for any given day of the week – Mon, Tue, Wed etc).

Our initial plan was to leverage any mapping technology such as google maps to present the pickup location and use any visualization tools such as D3 to present the traffic network patterns. After further analysis of different technologies available team agreed to use AWS RDS (MySQL) to store the extracted data to present historical information, and Azure ML studio to develop prediction models for fare amount, collision-prone zones, high demand for pickup location and H3 visualization algorithm to present grid cells on NYC map.

Algorithms

H3 (Hexagonal Hierarchical Spatial Index Ref#), which was open sourced by Uber, for visualization algorithm to transform Lat-Long to identifiable hexagonal grid cells across the map of NYC has been chosen for grid layouts. Integration of various datasets like collision/ TLC data with similar columns is also done. Some of the columns are merged/transformed (group by "day of the week")
H3 results are rendered using API from Azure ML Studio for Multiclass Decision Jungle and Fast Forest Quantile Regression and underlying work in Azure ML Studio is illustrated in diagrams.
We looked at different ML algorithms for our prediction model but we chose to leverage Multiclass Decision Jungle algorithm for collision-prone zones with the classification of the zones based on the type of accidents and severity of the accidents with different criteria such as pedestrians or people involved in the collision, injuries/deaths involved. We have used Fast Forest Quantile Regression algorithm for predicting average fare amount/mile and demand for a pickup location for any given location, day of the week and hour of the day.

Some of the reasons for selecting Fast Forest Quantile Regression and Multiclass Decision Jungles algorithms as mentioned in Microsoft Azure ML reference documentation would fall into the criteria that we selected in our project are given below.

Forest Quantile Regression (Ref#)

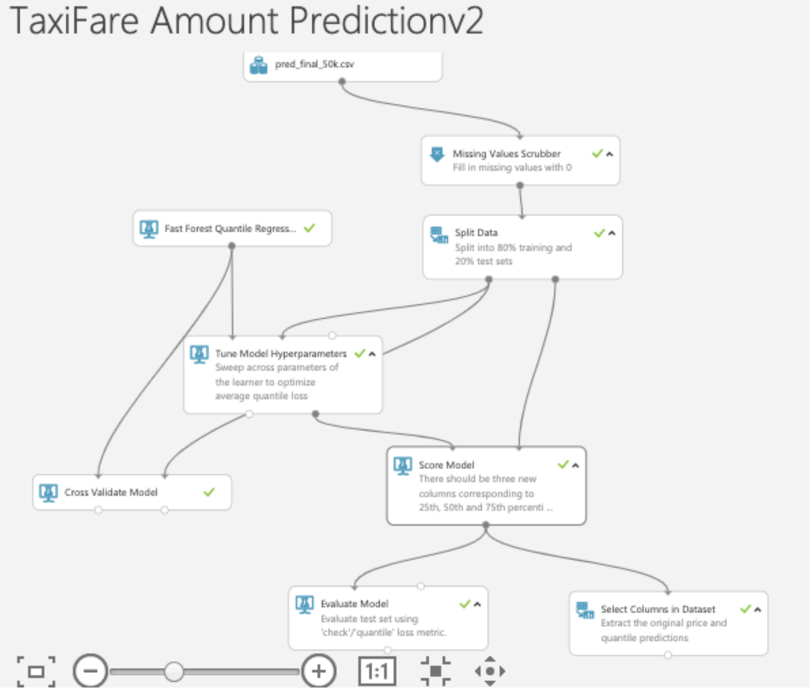
- Predicting prices
- Discovering predictive relationships in cases where there is only a weak relationship between variables

Multiclass Decision Jungle (Ref#)

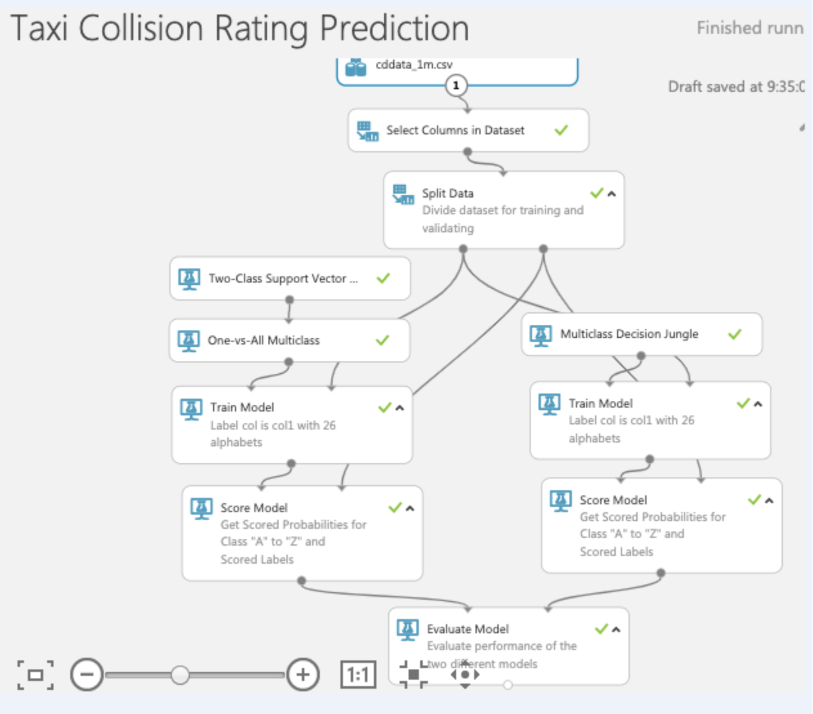
- Since some of the features that we selected day of the week, hour, collision indicators are non-linear and multi-class decision jungle can handle data with the varied distribution.
- They perform integrated feature selection and classification and are resilient in the presence of noisy features.

Prediction Model Details/Diagrams

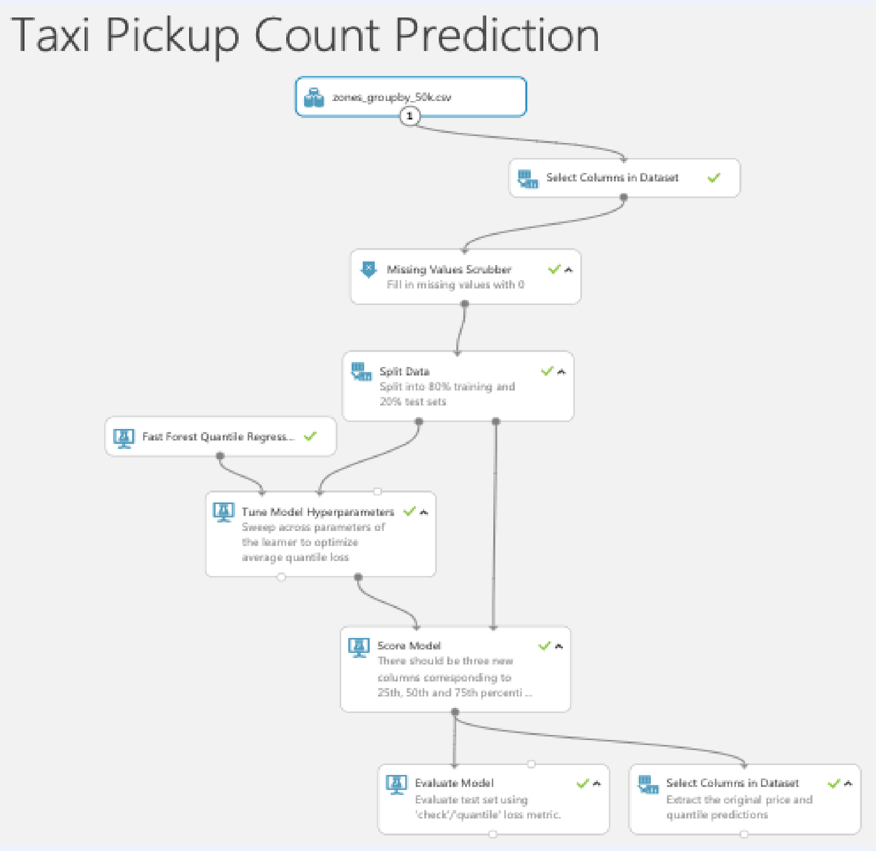
Azure ML Studio Snapshot for Fare Prediction


TaxiFare Amount Predictionv2
pred_final_50k.csv
Missing Values Scrubber
Fill in missing values with 0
Fast Forest Quantile Regress...
Split Data
Split into 80% training and 20% test sets
Tune Model Hyperparameters
Sweep across parameters of the learner to optimize average quantile loss
Cross Validate Model
Score Model
There should be three new columns corresponding to 25th, 50th and 75th percenti ...
Evaluate Model
Evaluate test set using 'check'/'quantile' loss metric.
Select Columns in Dataset
Extract the original price and quantile predictions
1:1

Taxi Collision Rating Prediction


Taxi Collision Rating Prediction
Finished runn
Draft saved at 9:35:(
cddata_1m.csv
1
Select Columns in Dataset
Split Data
Divide dataset for training and validating
Two-Class Support Vector ...
One-vs-All Multiclass
Multiclass Decision Jungle
Train Model
Label col is col1 with 26 alphabets
Train Model
Label col is col1 with 26 alphabets
Score Model
Get Scored Probabilities for Class "A" to "Z" and Scored Labels
Score Model
Get Scored Probabilities for Class "A" to "Z" and Scored Labels
Evaluate Model
Evaluate performance of the two different models
1:1

Taxi Pickup Count Prediction



Taxi Pickup Count Prediction

zones_groupby_50k.csv

Select Columns in Dataset ✓

Missing Values Scrubber ✓
Fill in missing values with 0

Split Data ✓
Split into 80% training and 20% test sets

Fast Forest Quantile Regress... ✓

Tune Model Hyperparameters ✓
Sweep across parameters of the learner to optimize average quantile loss

Score Model ✓
There should be three new columns corresponding to 25th, 50th and 75th percenti ...

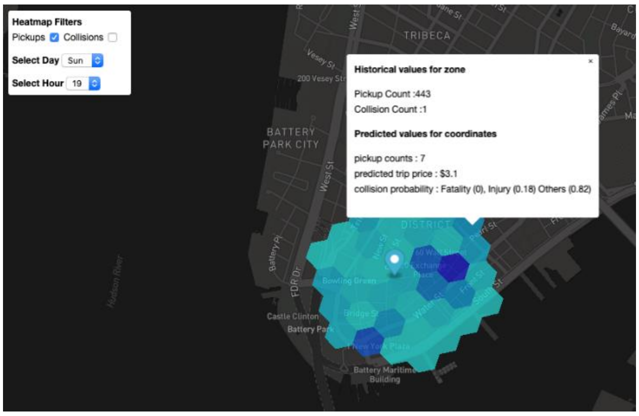
Evaluate Model ✓
Evaluate test set using 'check'/'quantile' loss metric.

Select Columns in Dataset ✓
Extract the original price and quantile predictions

Data Visualization

We used JavaScript, D3, MapBox, and H3 for data visualizations and user interface. The following image depicts the user experience with options to set the current location and selection of heatmaps based on traffic, and collision data either independently or overlaid. Based on these inputs, the map displays the most profitable pickup points in a polygon-boundary. Additional features such as predicting the fare amount and demand for pickup zones and displaying in the tooltip option are also done.

Hex Zone's gradient when User Chooses Weekday and Time for Pick up



Heatmap Filters

Pickups ☑ Collisions ☐

Select Day Sun

Select Hour 19

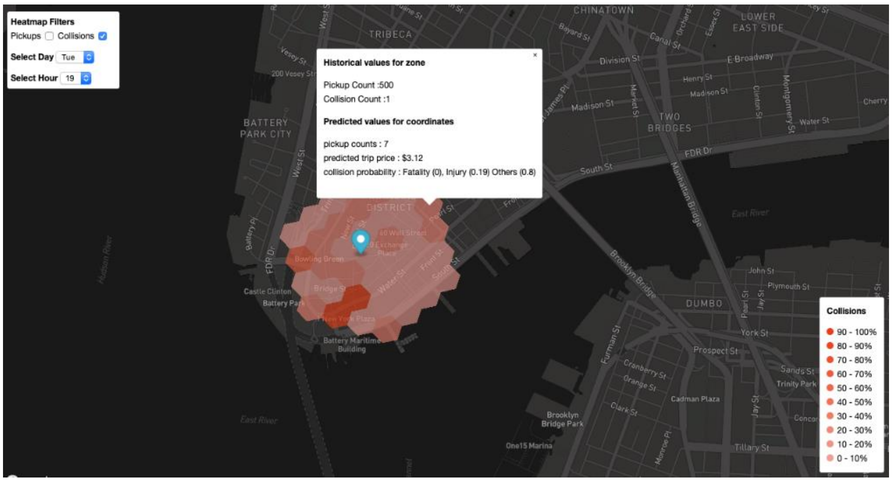
Historical values for zone

Pickup Count :443
Collision Count :1

Predicted values for coordinates

pickup counts : 7
predicted trip price : $3.1
collision probability : Fatality (0), Injury (0.18) Others (0.82)

Hex Zone's gradient when User Chooses Collision for chosen Weekday, Time and Pickup



Heatmap Filters
Pickups □ Collisions ☑
Select Day Tue
Select Hour 19

Historical values for zone

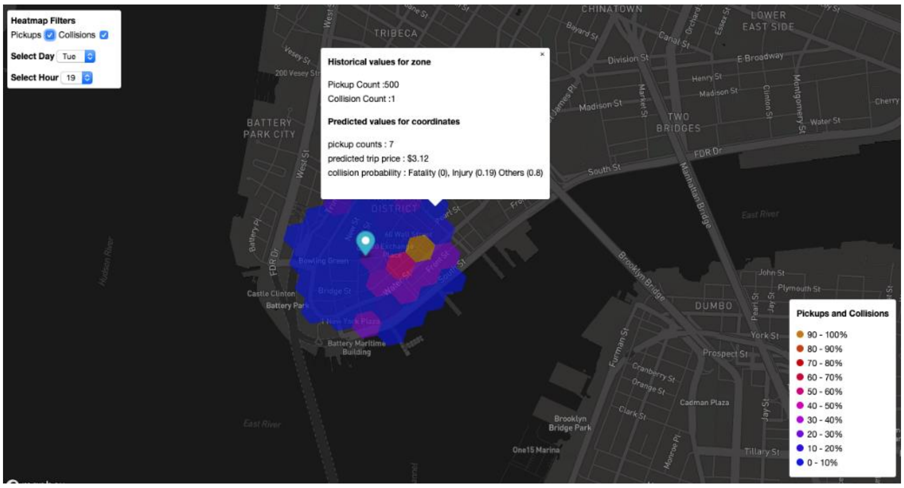
Pickup Count :500
Collision Count :1

Predicted values for coordinates

pickup counts : 7
predicted trip price : $3.12
collision probability : Fatality (0), Injury (0.19) Others (0.8)

Collisions
- ● 90 - 100%
- ● 80 - 90%
- ● 70 - 80%
- ● 60 - 70%
- ● 50 - 60%
- ● 40 - 50%
- ● 30 - 40%
- ● 20 - 30%
- ● 10 - 20%
- ● 0 - 10%

Hex Zone's gradient when User Chooses Collision for chosen Weekday, Time



Heatmap Filters
Pickups ☑ Collisions ☑
Select Day Tue
Select Hour 19

Historical values for zone

Pickup Count :500
Collision Count :1

Predicted values for coordinates

pickup counts : 7
predicted trip price : $3.12
collision probability : Fatality (0), Injury (0.19) Others (0.8)

Pickups and Collisions
- ● 90 - 100%
- ● 80 - 90%
- ● 70 - 80%
- ● 60 - 70%
- ● 50 - 60%
- ● 40 - 50%
- ● 30 - 40%
- ● 20 - 30%
- ● 10 - 20%
- ● 0 - 10%

Experiments and Evaluation

Visualization Experiments:

We first attempted using Tableau (Ref#) for UI mock-ups as a quick and dirty way of conjuring the right direction of our project efforts. As demonstrated in our approach section, for Visualization we considered Licensing, Open Source, Performance, Scalability, Easy integration of cloud-hosted API and a new way of visualizing spatial data. All these factors led us to choose a web interface using D3, JavaScript, H3, and MapBox API, over Tableau. H3 (Hexagonal Hierarchical Spatial Index) is open source by Uber; It is a geospatial indexing system using hexagonal grids for granular analysis. Hexagon has 6 equidistant cells/neighbors, unlike other polygons where the distance varies between the nearest cells. However, triangles and squares have neighbors at different distances.

Data Experiments:

We generated a random sampling of our input data using ‘gshuf’ program. We ran the gshuf to shuffle the data and picked the first 50,000 rows of the shuffled dataset as training data. We limited our training data set to 50,000 rows to limit the training time and we observed that it was enough in generating a reasonable prediction. We also reduced the latitude-longitude to use 3 decimal points only.

For individual pickup counts and collision counts, we use the following approach to display the heat map. From the user’s choice of location (anchor), we find the zone the marker belongs to. We calculate 36 neighboring zones (3 layers). We invoke the API for all 36 zones. We normalize the date for all 36 zones using the formula below.

heatmap[index] = (layer[index] - min) / (max - min)

layer[index] is the count for that zone heatmap[index] is the count used for heatmap display max, min are maximum and minimum counts for 36 zones.

For a consolidated view containing both pickup counts and collisions, we experimented with various weights for each value. We chose the weight of 0.1 for a number of pickup points and a factor of 100 for the number of collisions for the zone in generating the heat map.

We sampled the data for various locations and verified the results of the API against running manual queries in the database. We used test locations to verify the empty results from our APIs. (Note: Prediction modeling contain only New York City location points and inputting any other location points, example Hudson river, might result in incorrect prediction results)

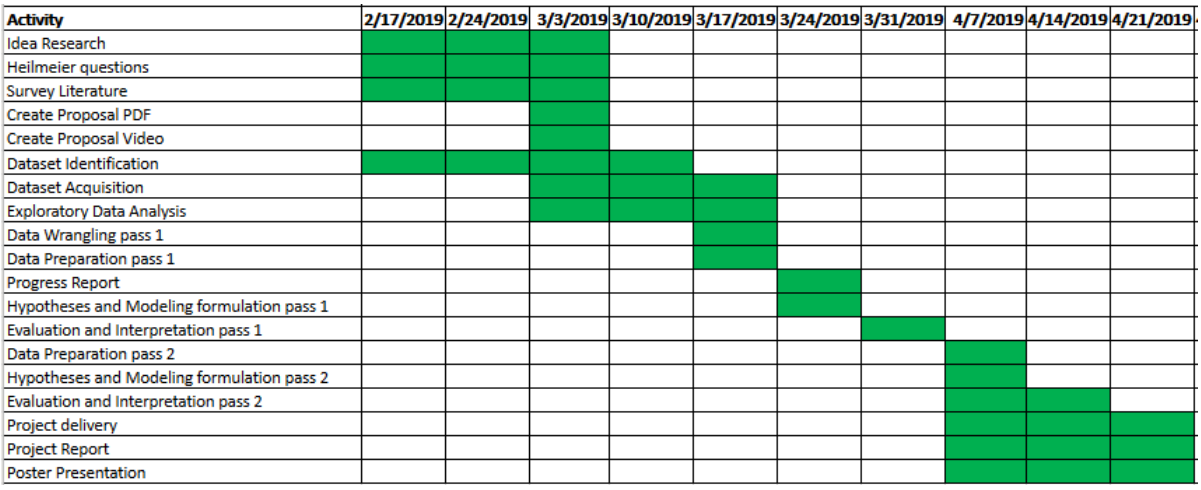
Prediction algorithms:

We experimented with linear regression with cross-validation model for our prediction models. But the coefficient of determination was only ranging from 0.2 - 0.4. We changed the model to use the Multiclass Decision Jungle model and the accuracy of prediction improved to 80%.

CORS Challenge:

We encountered CORS challenge while integrating MapBox (web site originated from AWS domain) with Azure APIs. We resolved the issue by integrating Azure APIs within AWS rather than being integrated within the website.

Plan of activities



Activity	2/17/2019	2/24/2019	3/3/2019	3/10/2019	3/17/2019	3/24/2019	3/31/2019	4/7/2019	4/14/2019	4/21/2019
Idea Research	🟩	🟩	🟩							
Heilmeier questions	🟩	🟩	🟩							
Survey Literature	🟩	🟩	🟩							
Create Proposal PDF			🟩							
Create Proposal Video			🟩							
Dataset Identification	🟩	🟩	🟩	🟩						
Dataset Acquisition			🟩	🟩	🟩					
Exploratory Data Analysis			🟩	🟩	🟩					
Data Wrangling pass 1					🟩					
Data Preparation pass 1					🟩					
Progress Report						🟩				
Hypotheses and Modeling formulation pass 1						🟩				
Evaluation and Interpretation pass 1							🟩			
Data Preparation pass 2								🟩		
Hypotheses and Modeling formulation pass 2								🟩		
Evaluation and Interpretation pass 2								🟩	🟩	
Project delivery								🟩	🟩	🟩
Project Report								🟩	🟩	🟩
Poster Presentation								🟩	🟩	🟩

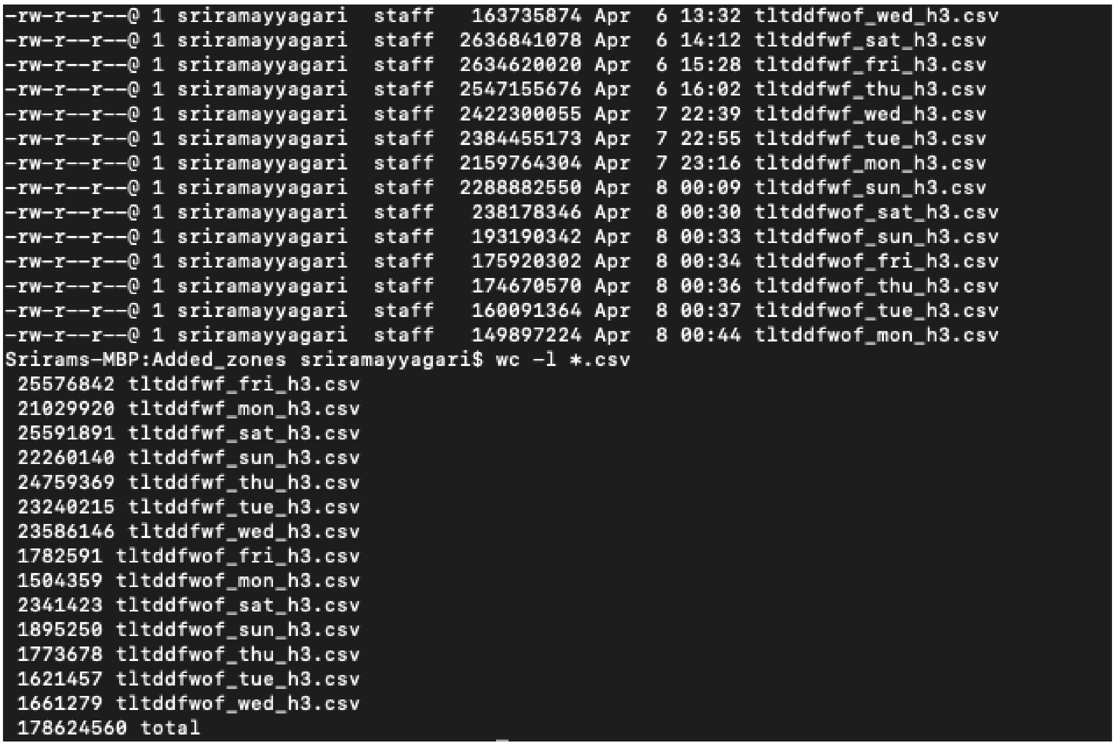
All team members have contributed a similar amount of efforts.

Conclusion

We were able to accomplish most of what we envisioned initially with this project. We had some challenges in terms of getting a better prediction accuracy because some of the data available were a couple of years old and missing data attributes. We also had some steep learning curve for some of the technologies used for this project. Albeit, we had a good opportunity for collaborative work, implement what we learned in the course and deliver a working solution at the end.

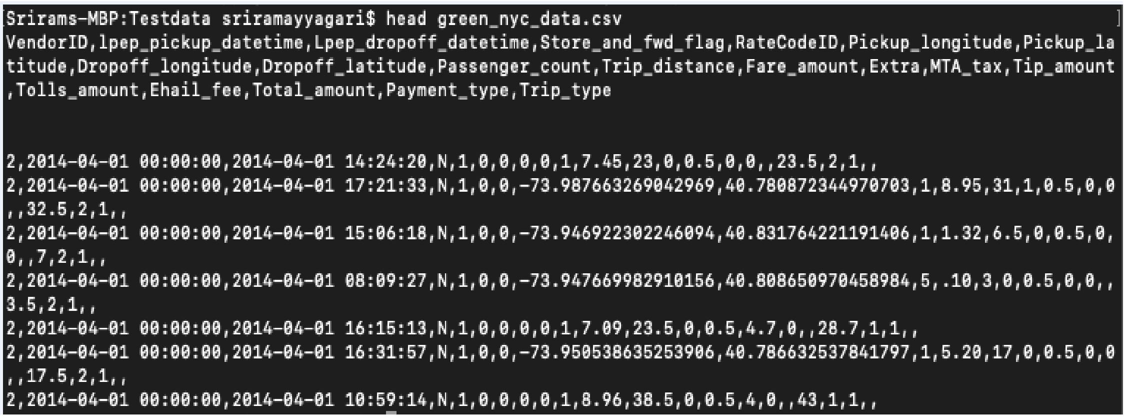
Appendix

- Data sets volume (180 million including collision data, 15 GB)



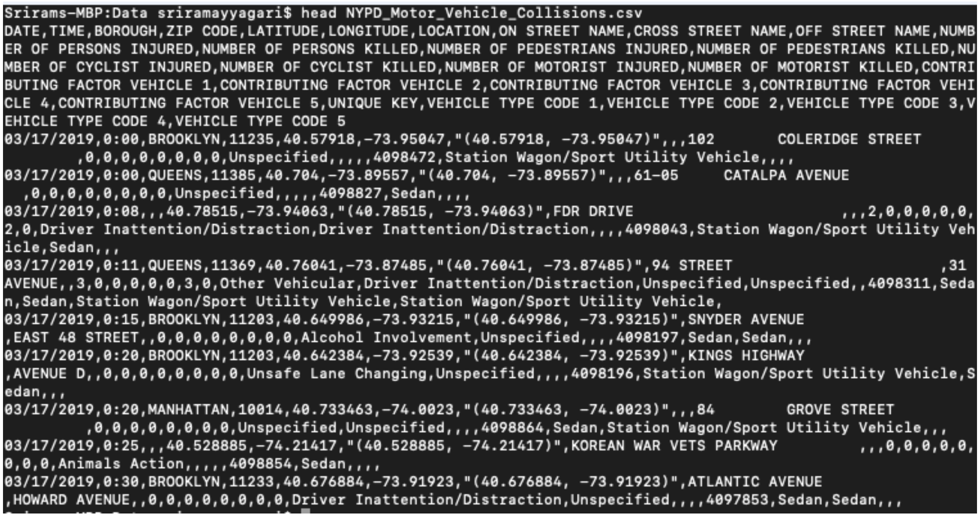
```
-rw-r--r--@ 1 sriramayyagari  staff   163735874 Apr  6 13:32 tltddfwof_wed_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff  2636841078 Apr  6 14:12 tltddfwf_sat_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff  2634620020 Apr  6 15:28 tltddfwf_fri_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff  2547155676 Apr  6 16:02 tltddfwf_thu_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff  2422300055 Apr  7 22:39 tltddfwf_wed_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff  2384455173 Apr  7 22:55 tltddfwf_tue_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff  2159764304 Apr  7 23:16 tltddfwf_mon_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff  2288882550 Apr  8 00:09 tltddfwf_sun_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff   238178346 Apr  8 00:30 tltddfwof_sat_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff   193190342 Apr  8 00:33 tltddfwof_sun_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff   175920302 Apr  8 00:34 tltddfwof_fri_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff   174670570 Apr  8 00:36 tltddfwof_thu_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff   160091364 Apr  8 00:37 tltddfwof_tue_h3.csv
-rw-r--r--@ 1 sriramayyagari  staff   149897224 Apr  8 00:44 tltddfwof_mon_h3.csv
Srirams-MBP:Added_zones sriramayyagari$ wc -l *.csv
 25576842 tltddfwf_fri_h3.csv
 21029920 tltddfwf_mon_h3.csv
 25591891 tltddfwf_sat_h3.csv
 22260140 tltddfwf_sun_h3.csv
 24759369 tltddfwf_thu_h3.csv
 23240215 tltddfwf_tue_h3.csv
 23586146 tltddfwf_wed_h3.csv
 1782591 tltddfwof_fri_h3.csv
 1504359 tltddfwof_mon_h3.csv
 2341423 tltddfwof_sat_h3.csv
 1895250 tltddfwof_sun_h3.csv
 1773678 tltddfwof_thu_h3.csv
 1621457 tltddfwof_tue_h3.csv
 1661279 tltddfwof_wed_h3.csv
 178624560 total
```

- Green Taxi Data



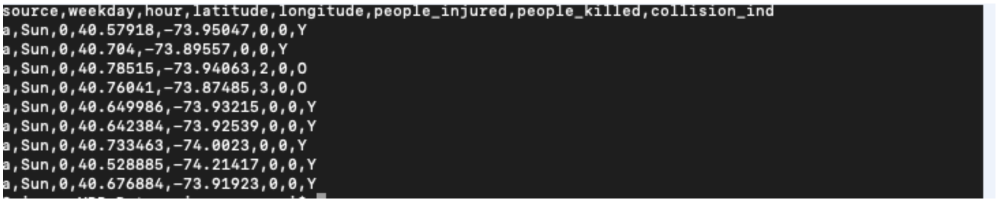
```
Srirams-MBP:Testdata sriramayyagari$ head green_nyc_data.csv
VendorID,lpep_pickup_datetime,Lpep_dropoff_datetime,Store_and_fwd_flag,RateCodeID,Pickup_longitude,Pickup_la
titude,Dropoff_longitude,Dropoff_latitude,Passenger_count,Trip_distance,Fare_amount,Extra,MTA_tax,Tip_amount
,Tolls_amount,Ehail_fee,Total_amount,Payment_type,Trip_type


2,2014-04-01 00:00:00,2014-04-01 14:24:20,N,1,0,0,0,0,1,7.45,23,0,0.5,0,0,,23.5,2,1,,
2,2014-04-01 00:00:00,2014-04-01 17:21:33,N,1,0,0,-73.987663269042969,40.780872344970703,1,8.95,31,1,0.5,0,0
,,32.5,2,1,,
2,2014-04-01 00:00:00,2014-04-01 15:06:18,N,1,0,0,-73.946922302246094,40.831764221191406,1,1.32,6.5,0,0.5,0,
0,,7,2,1,,
2,2014-04-01 00:00:00,2014-04-01 08:09:27,N,1,0,0,-73.947669982910156,40.808650970458984,5,.10,3,0,0.5,0,0,,
3.5,2,1,,
2,2014-04-01 00:00:00,2014-04-01 16:15:13,N,1,0,0,0,0,1,7.09,23.5,0,0.5,4.7,0,,28.7,1,1,,
2,2014-04-01 00:00:00,2014-04-01 16:31:57,N,1,0,0,-73.950538635253906,40.786632537841797,1,5.20,17,0,0.5,0,0
,,17.5,2,1,,
2,2014-04-01 00:00:00,2014-04-01 10:59:14,N,1,0,0,0,0,1,8.96,38.5,0,0.5,4,0,,43,1,1,,
```

- Collision data



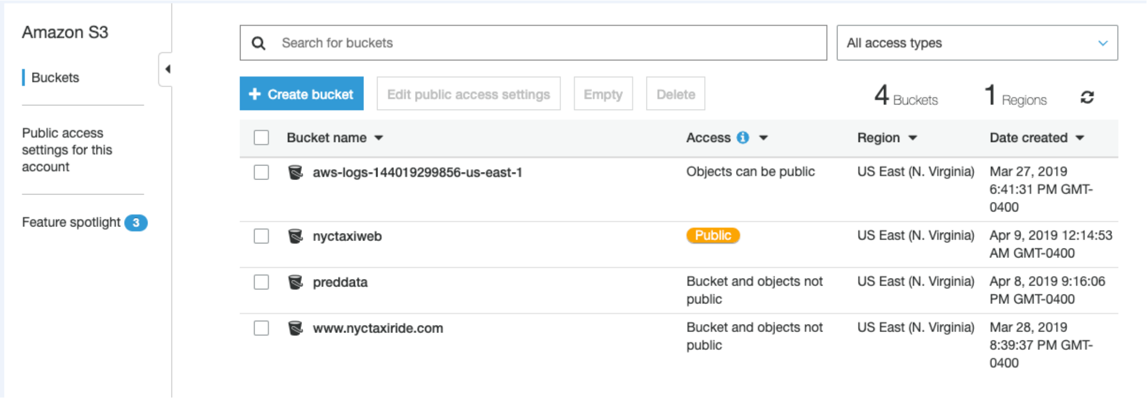
```
Srirams-MBP:Data sriramayyagari$ head NYPD_Motor_Vehicle_Collisions.csv
DATE,TIME,BOROUGH,ZIP CODE,LATITUDE,LONGITUDE,LOCATION,ON STREET NAME,CROSS STREET NAME,OFF STREET NAME,NUMB
ER OF PERSONS INJURED,NUMBER OF PERSONS KILLED,NUMBER OF PEDESTRIANS INJURED,NUMBER OF PEDESTRIANS KILLED,NU
MBER OF CYCLIST INJURED,NUMBER OF CYCLIST KILLED,NUMBER OF MOTORIST INJURED,NUMBER OF MOTORIST KILLED,CONTRI
BUTING FACTOR VEHICLE 1,CONTRIBUTING FACTOR VEHICLE 2,CONTRIBUTING FACTOR VEHICLE 3,CONTRIBUTING FACTOR VEHI
CLE 4,CONTRIBUTING FACTOR VEHICLE 5,UNIQUE KEY,VEHICLE TYPE CODE 1,VEHICLE TYPE CODE 2,VEHICLE TYPE CODE 3,V
EHICLE TYPE CODE 4,VEHICLE TYPE CODE 5
03/17/2019,0:00,BROOKLYN,11235,40.57918,-73.95047,"(40.57918, -73.95047)",,,102       COLERIDGE STREET
        ,0,0,0,0,0,0,0,0,Unspecified,,,,,4098472,Station Wagon/Sport Utility Vehicle,,,,
03/17/2019,0:00,QUEENS,11385,40.704,-73.89557,"(40.704, -73.89557)",,,61-05     CATALPA AVENUE
  ,0,0,0,0,0,0,0,0,Unspecified,,,,,4098827,Sedan,,,,
03/17/2019,0:08,,,40.78515,-73.94063,"(40.78515, -73.94063)",FDR DRIVE                       ,,,2,0,0,0,0,0,
2,0,Driver Inattention/Distraction,Driver Inattention/Distraction,,,,4098043,Station Wagon/Sport Utility Veh
icle,Sedan,,,
03/17/2019,0:11,QUEENS,11369,40.76041,-73.87485,"(40.76041, -73.87485)",94 STREET                       ,31
AVENUE,,3,0,0,0,0,0,3,0,Other Vehicular,Driver Inattention/Distraction,Unspecified,Unspecified,,4098311,Seda
n,Sedan,Station Wagon/Sport Utility Vehicle,Station Wagon/Sport Utility Vehicle,
03/17/2019,0:15,BROOKLYN,11203,40.649986,-73.93215,"(40.649986, -73.93215)",SNYDER AVENUE
,EAST 48 STREET,,0,0,0,0,0,0,0,0,Alcohol Involvement,Unspecified,,,,4098197,Sedan,Sedan,,,
03/17/2019,0:20,BROOKLYN,11203,40.642384,-73.92539,"(40.642384, -73.92539)",KINGS HIGHWAY
,AVENUE D,,0,0,0,0,0,0,0,0,Unsafe Lane Changing,Unspecified,,,,4098196,Station Wagon/Sport Utility Vehicle,S
edan,,,
03/17/2019,0:20,MANHATTAN,10014,40.733463,-74.0023,"(40.733463, -74.0023)",,,84        GROVE STREET
         ,0,0,0,0,0,0,0,0,Unspecified,Unspecified,,,,4098864,Sedan,Station Wagon/Sport Utility Vehicle,,,
03/17/2019,0:25,,,40.528885,-74.21417,"(40.528885, -74.21417)",KOREAN WAR VETS PARKWAY          ,,,0,0,0,0,0,
0,0,0,Animals Action,,,,,4098854,Sedan,,,,
03/17/2019,0:30,BROOKLYN,11233,40.676884,-73.91923,"(40.676884, -73.91923)",ATLANTIC AVENUE
,HOWARD AVENUE,,0,0,0,0,0,0,0,0,Driver Inattention/Distraction,Unspecified,,,,4097853,Sedan,Sedan,,,
```

- Collision data after processing



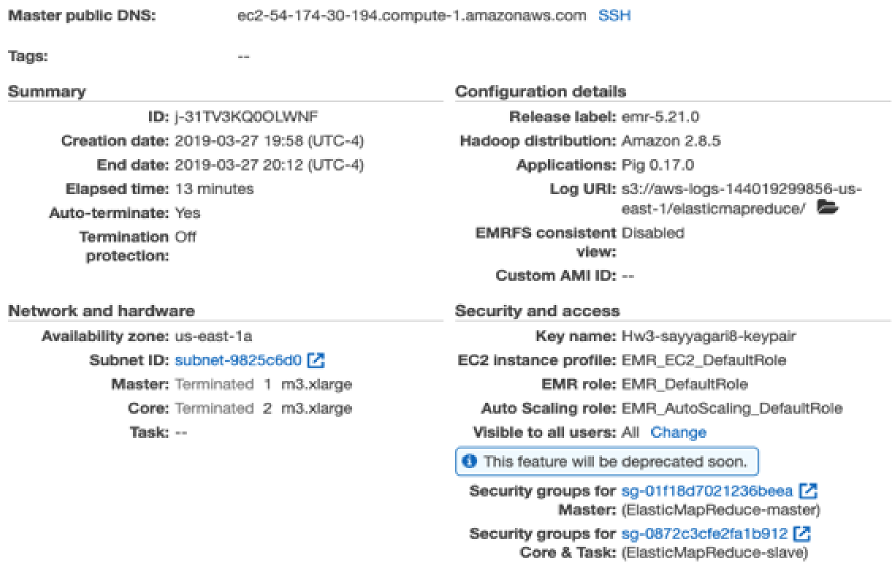
```
source,weekday,hour,latitude,longitude,people_injured,people_killed,collision_ind
a,Sun,0,40.57918,-73.95047,0,0,Y
a,Sun,0,40.704,-73.89557,0,0,Y
a,Sun,0,40.78515,-73.94063,2,0,O
a,Sun,0,40.76041,-73.87485,3,0,O
a,Sun,0,40.649986,-73.93215,0,0,Y
a,Sun,0,40.642384,-73.92539,0,0,Y
a,Sun,0,40.733463,-74.0023,0,0,Y
a,Sun,0,40.528885,-74.21417,0,0,Y
a,Sun,0,40.676884,-73.91923,0,0,Y
```

- AWS S3 buckets depicting a collection of datasets for the project



Bucket name	Access	Region	Date created
aws-logs-144019299856-us-east-1	Objects can be public	US East (N. Virginia)	Mar 27, 2019 6:41:31 PM GMT-0400
nyctaxiweb	Public	US East (N. Virginia)	Apr 9, 2019 12:14:53 AM GMT-0400
preddata	Bucket and objects not public	US East (N. Virginia)	Apr 8, 2019 9:16:06 PM GMT-0400
www.nyctaxiride.com	Bucket and objects not public	US East (N. Virginia)	Mar 28, 2019 8:39:37 PM GMT-0400

- AWS EMR/Hadoop Configuration details



Master public DNS: ec2-54-174-30-194.compute-1.amazonaws.com SSH

Tags: --

Summary

ID: j-31TV3KQ0OLWNF
Creation date: 2019-03-27 19:58 (UTC-4)
End date: 2019-03-27 20:12 (UTC-4)
Elapsed time: 13 minutes
Auto-terminate: Yes
Termination protection: Off

Configuration details

Release label: emr-5.21.0
Hadoop distribution: Amazon 2.8.5
Applications: Pig 0.17.0
Log URI: s3://aws-logs-144019299856-us-east-1/elasticmapreduce/
EMRFS consistent view: Disabled
Custom AMI ID: --

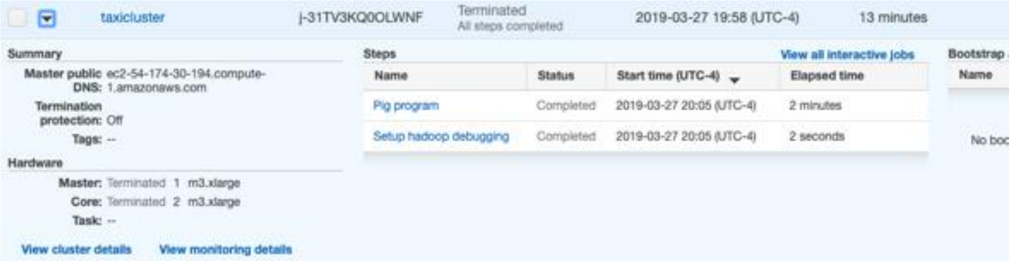
Network and hardware

Availability zone: us-east-1a
Subnet ID: subnet-9825c6d0
Master: Terminated 1 m3.xlarge
Core: Terminated 2 m3.xlarge
Task: --

Security and access

Key name: Hw3-sayyagari8-keypair
EC2 instance profile: EMR_EC2_DefaultRole
EMR role: EMR_DefaultRole
Auto Scaling role: EMR_AutoScaling_DefaultRole
Visible to all users: All Change
This feature will be deprecated soon.
Security groups for Master: sg-01f18d7021236beea (ElasticMapReduce-master)
Security groups for Core & Task: sg-0872c3cfe2fa1b912 (ElasticMapReduce-slave)

- Pig Script Run monitoring



taxicluster j-31TV3KQ0OLWNF Terminated All steps completed 2019-03-27 19:58 (UTC-4) 13 minutes

Summary

Master public DNS: ec2-54-174-30-194.compute-1.amazonaws.com
Termination protection: Off
Tags: --

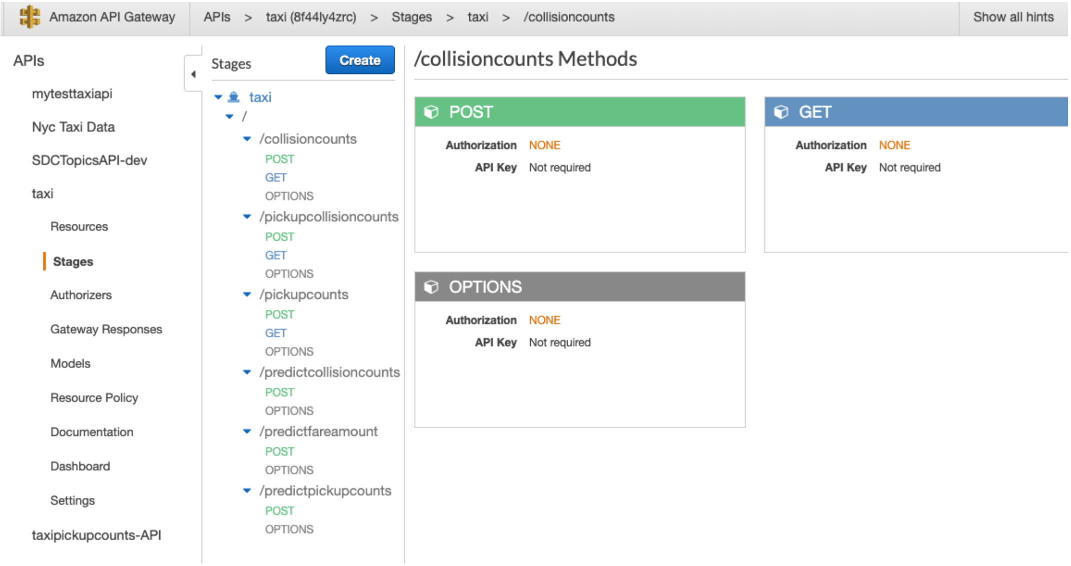
Hardware

Master: Terminated 1 m3.xlarge
Core: Terminated 2 m3.xlarge
Task: --

Steps

Name	Status	Start time (UTC-4)	Elapsed time
Pig program	Completed	2019-03-27 20:05 (UTC-4)	2 minutes
Setup hadoop debugging	Completed	2019-03-27 20:05 (UTC-4)	2 seconds

- AWS API Gateway deployment



taxi
/
/collisioncounts
POST
GET
OPTIONS
/pickupcollisioncounts
POST
GET
OPTIONS
/pickupcounts
POST
GET
OPTIONS
/predictcollisioncounts
POST
OPTIONS
/predictfareamount
POST
OPTIONS
/predictpickupcounts
POST
OPTIONS

/collisioncounts Methods

POST
Authorization NONE
API Key Not required

GET
Authorization NONE
API Key Not required

OPTIONS
Authorization NONE
API Key Not required

References

- NYC TLC Trip Record Data https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- Uber Dataset https://github.com/fivethirtyeight/uber-tlc-foil-response

- Project Survey http://nyctaxiweb.s3-website-us-east-1.amazonaws.com/CSE6242_Group4_Survey.pdf

- Fast Forest Quantile Regression https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/fast-forest-quantile-regression

- Multiclass Decision Jungle. https://docs.microsoft.com/en-us/azure/machine-learning/studio-module-reference/multiclass-decision-jungle

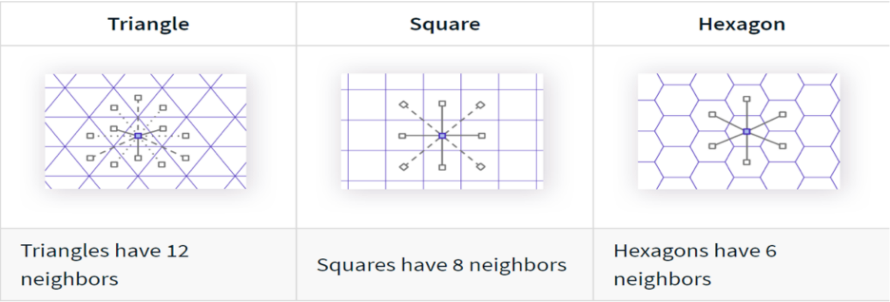
- Data Dictionaries

  - Yellow Taxi Trips https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_yellow.pdf

  - Green Taxi Trips https://www1.nyc.gov/assets/tlc/downloads/pdf/data_dictionary_trip_records_green.pdf

  - Uber Trips Data Dictionary
    1. April-Sep 2014 Dataset
       Date/Time, Lat, Lon, Base
    2. Jan-June 2015
       Dispatching_base_num, Pickup_date, Affiliated_base_num, location

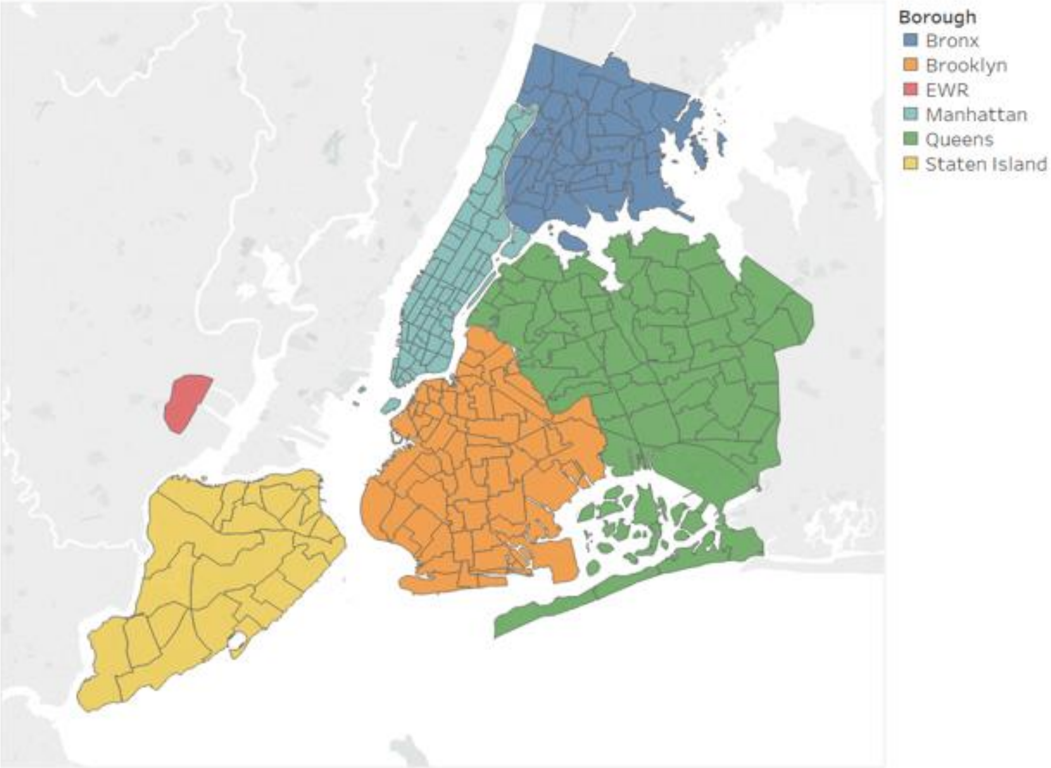
- H3 Hexagon model comparison



Triangle	Square	Hexagon
Triangles have 12 neighbors	Squares have 8 neighbors	Hexagons have 6 neighbors

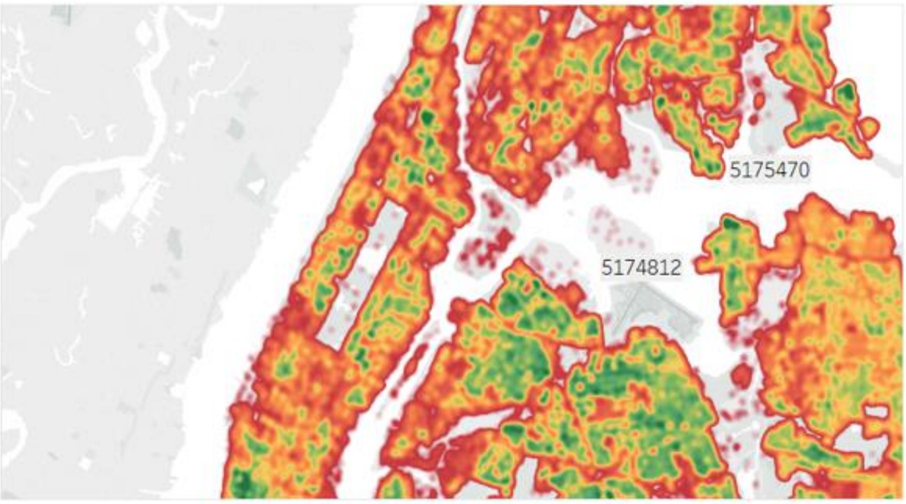
- Tableau Experiment Visualization

Mockup of NYC Taxi Zones



Map based on Longitude (generated) and Latitude (generated). Color shows details about Borough. Details are shown for Zone.

Mockup of Heatmap of Taxi Pickup Points and Hotspots of Uber and other Cab Services



Map based on Longitude (generated) and Latitude (generated). The marks are labeled by Address Id. Details are shown for Zipcode and Full Stree.