RYDE
NEW YORK CITY TAXI FARE PRICING MODEL RECOMMENDATIONS

DANIEL MITCHELL

CHIEF DATA SCIENTIST

DESCRIPTION OF PROBLEM FORMULATION

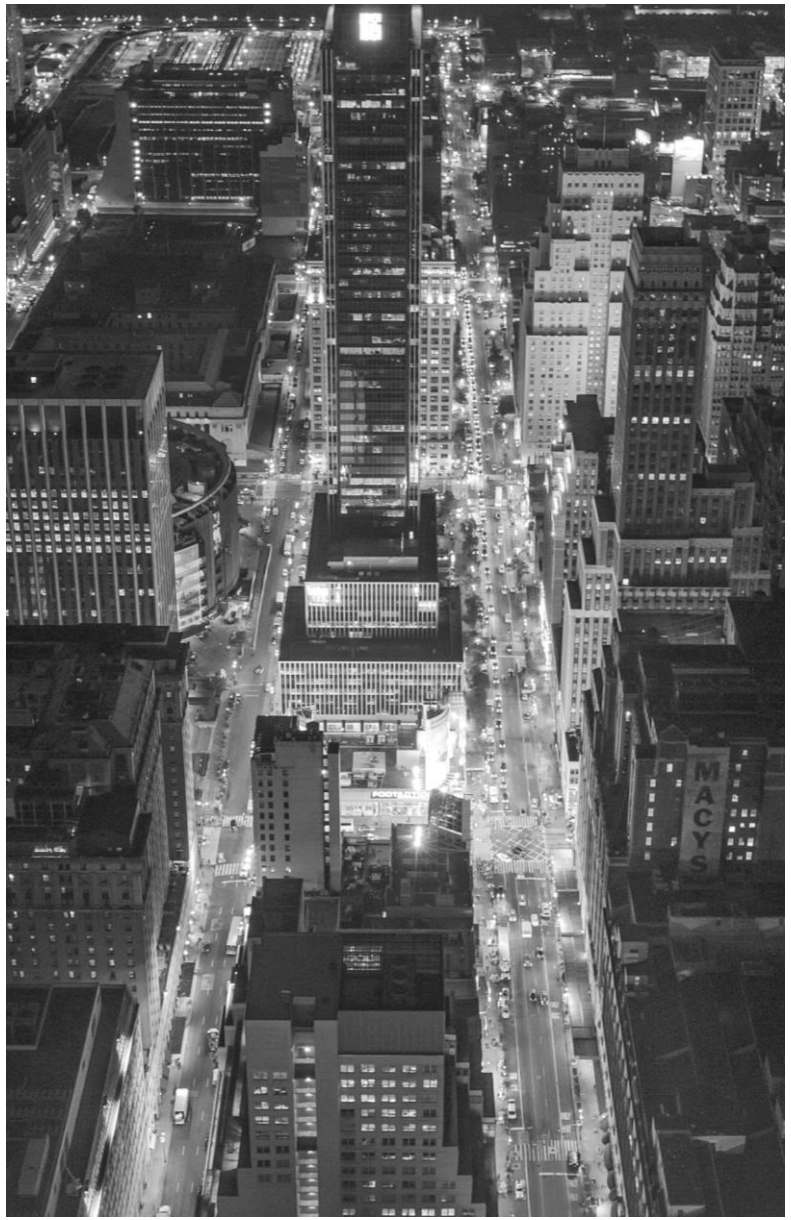
- Ryde relies on an algorithm that estimates reasonable fares for distance
- Ryde does not consider variables (features) such as
  - Travel Time
  - Pickup Locations
  - Dropoff Locations
  - Time of day
  - Day of week
- Ryde is looking for a better prediction model for the next major version update

DATA SETS USED

- Public data from the NYC.gov website
- New York City Taxi and Limousine Commission (TLC)
  - 2018 – Full calendar year of Yellow Taxi Data records
  - 2019 – Full calendar year of Yellow Taxi Data records
  - 2020 – Partial (Jan – Jun) of Yellow Taxi Data records
  - Factbook 2018 - Taxi “Trip Trends”
- Lookup table
  - Mapping (Location ID to Borough, Zone Name)
- 204,051,047 raw records Jan 2018 to Jun 2020
  - 200,448,120 - records after removing non negotiated fares
  - 195,678,879 – records after removal of chares below the minimum fare trips (fare < $3.00)
  - 195,207,603 - records after removing duration of 5 hour+ trips treated as outliers

DATA CLEANING

- The NYC public data for the calendar year 2018 did not include a column for “congestion_surcharge”. Therefore, for the purpose of this pricing model – it was set to 0.0 for this period and for the generated pricing model
- Records for fares < $3.00 (base fare $2.50 + minimum 1/5 mile $0.50 = $3.00)
- Records for negotiated fares (RatecodeID = 5) were removed from the pricing model, as these also will be considered outliers
- Records for taxi rides with a duration more than 300 minutes (5 hours) were removed. After 5 hours, these will be considered outliers and not representative of the average trip duration



MODEL / FEATURE SELECTION / TRAINING / TESTING ERROR

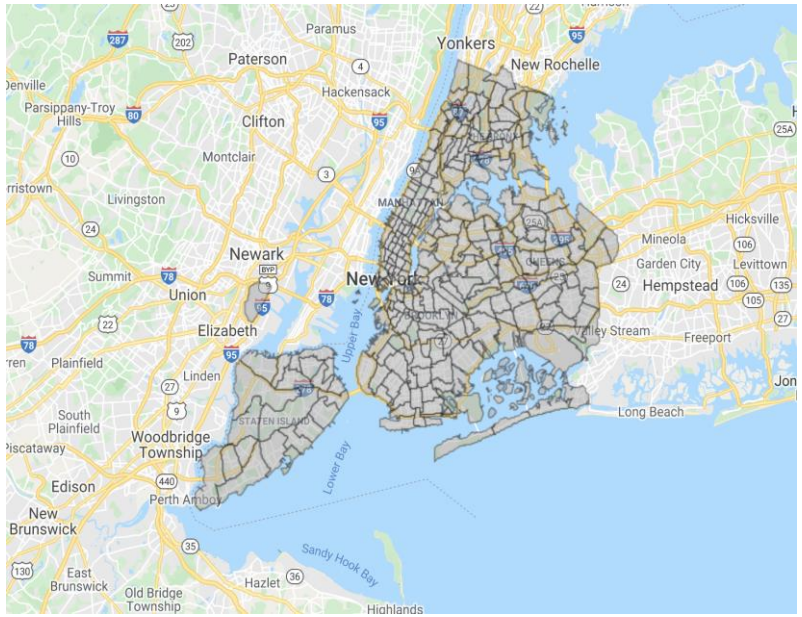
The Approach :

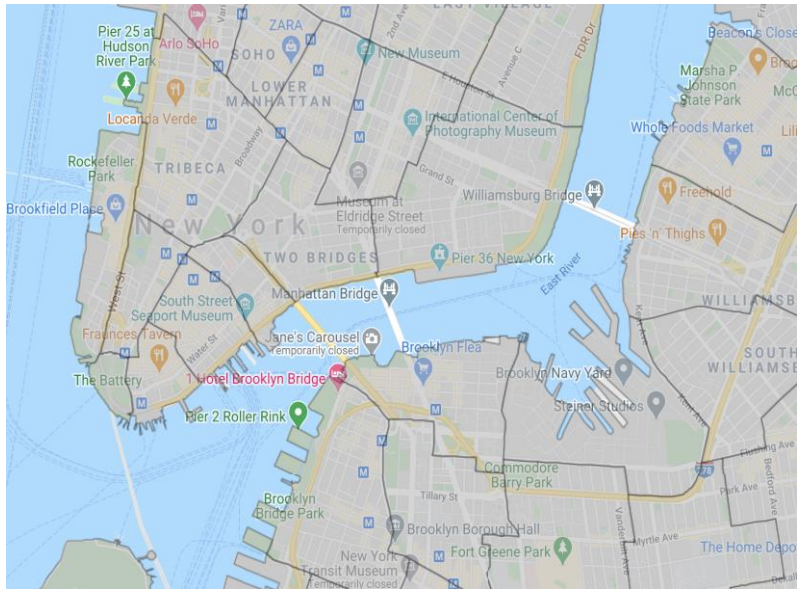
- Collect public data for Yellow Taxi rides from calendar years 2018-present
- Scrub data – remove outliers and invalid records
  - Remove records with null information
  - Remove records that are outliers in time, fare, etc and negotiated fares and long duration hires
- Split the data randomly between a Training (70%) and Testing(30%) data sets
- Create a linear regression model of the Training data
  - Record the correlation values
- Run a check of the Test data against the model to measure its validity (calculate errors)
  - Validate model and results – examine errors of Test data against Trained model
- Combine / join and report on data in combination with other data sets
- Present Findings
- Discuss Potential Opportunities
- Discuss limitations of the findings and opportunities

PLOTS / CHARTS / \
VISUALIZATIONS / SUMMARY / STATS

- Top 10 Location IDs for Fare Revenue

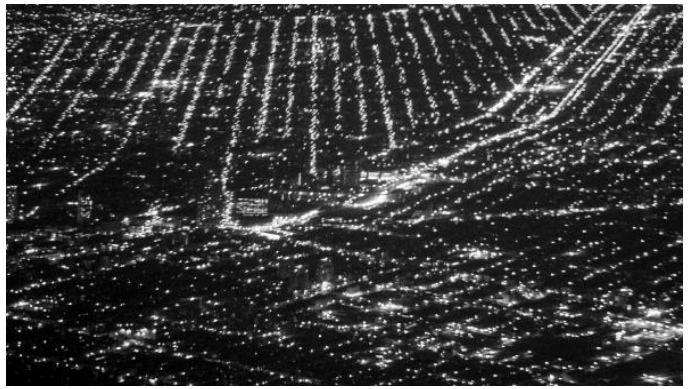
Loc ID	Total Fares	Borough			
132	314715236	Queens	JFK Airport		
138	225789138	Queens	LaGuardia Airport		
161	131767128	Manhattan	Midtown Center		
230	121005706	Manhattan	Times Sq/Theatre District		
162	115840430	Manhattan	Midtown East		
186	114989525	Manhattan	Penn Station/Madison Sq West		
237	114649056	Manhattan	Upper East Side South		
236	106822165	Manhattan	Upper East Side North		
170	100323370	Manhattan	Murray Hill		
48	98613281	Manhattan	Clinton East		





THE PRICING MODEL – BASED ON THE LAST 2.5 YEARS OF DATA * (OUTLIERS REMOVED)

- **Fare Price = - 0.0598737 +**
  - **+ (0.0243) x Passenger Count**
  - **+ (0.0002) x Trip Distance**
  - **+ (0.0141) x Rate Code ID**
  - **+ (-0.0001) x Pick Up Location Id**
  - **+ (-0.0001) x Drop off Location ID**
  - **+ (0.0075) x Payment Type**
  - **+ (0.0) x Taxi Ride duration in minutes** (* needs further investigation)



PLOTS / CHARTS / VISUALIZATIONS / SUMMARY / STATS

- Results of the Linear Regression model – Starting with 195 Million rows of taxi ride data (after scrubbing)
  - for the Training data
    - 136,642,976 Training rows
    - RMSE = 0.41716 - MSE = 0.17402 - MAE = 0.29682
  - For the Test data
    - 58,564,627 Testing rows
    - RMSE = 0.45149 - MSE = 0.20385 - MAE = 0.33386
- The Training and Test data errors are reasonably closely correlated which gives a high confidence in the pricing model
- Visual checks of the 100 rows of random data confirm a close correlation of the model to actual data

features	total_amount	prediction
(14,[0,2,3,4,5,6,...	8.5	8.349898042649947
(14,[0,2,3,4,5,6,...	4.8	5.018197898538492
(14,[0,2,3,4,5,6,...	6.8	7.017423469852911
(14,[0,2,3,4,5,6,...	5.8	6.016687494926326
(14,[0,2,3,4,5,6,...	5.8	6.015525574992489

IMPROVING REVENUE OPPORTUNITIES

- Space for possible promotions that competition does not have
  - Competative discount for multi-passenger trips with extended duration (e.g 3x person driving 45 mins to the airport)
  - Competitive discounts for wheelchair accessible vehicles or young riders requiring special adult excort/supervision – as this
  - Airport trips based on the day of the week – as reported in the NYC 2018 Fact book this statistic is variable on the day of the week and could benefit from competitive rates on strategic days
  - Shared rides and/or requests for special vehicles
  - Drivers able to speak or communicate in multiple languages to accommodate needs/desires of foreign passengers

- The current NYC pricing model shows little or no correlation with the total fare
  - Pick Up Location ID - should be studied at a more granular level which is likely to create further model variables
  - Drop Off Location ID – should be studied at a more granular level which is likely to create further model variables
  - Manhattan and the 2 main NYC Airports are the top 2 pickup and dropoff locations

LIMITATIONS

- Limitations in this report worth noting:
  - The taxi travel duration modelled coefficient came to 0.000 which appears suspicious and in need of additional investigation
- This report does not address modelled variables for the inclusion of
  - Time of Day
  - Day of week
- Based on Intuition and my expériences travelling and visiting NYC
- Based on available web data and reports
  - Fares based on further refinement and modelling of Drop Off and Pick Up location might likely yeild further pricing models that could work as a competative advantage. This is based both on intuition and other studies available on the web such as:
  - 2018 Fact Book – NYC Taxi and Limousine Commision - https://www1.nyc.gov/assets/tlc/downloads/pdf/2018_tlc_factbook.pdf
  - Published academic work - https://chih-ling-hsu.github.io/2018/05/14/NYC


THANK YOU