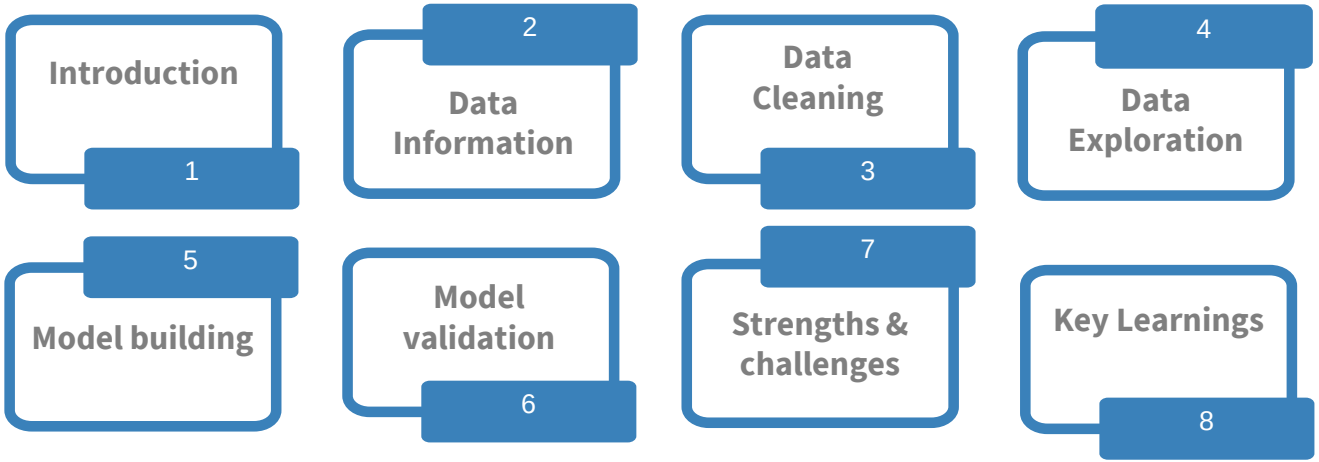
STAT 689 Class Project

To Tip or not to Tip – that's the question!

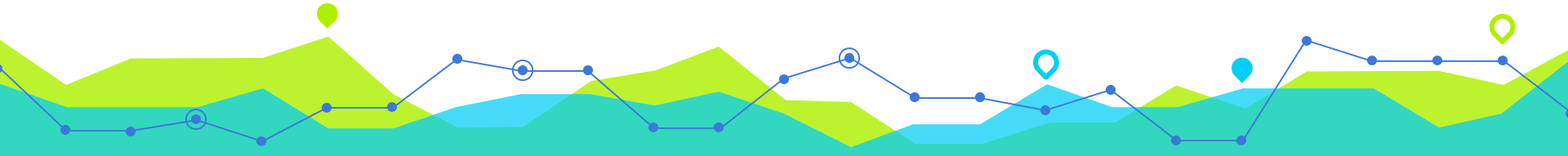
Presented by –
Abhilash Tangadpalliwar & Debapriyo Paul

Our Agenda


Introduction
1
2
Data Information
Data Cleaning
3
4
Data Exploration
5
Model building
Model validation
6
7
Strengths & challenges
Key Learnings
8

Introduction

- The City of Chicago in November of 2016 released a public dataset containing information over 100 million taxi rides since 2013 (https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/data)
- This public dataset does not include any data from the rideshare services like Uber and Lyft, but in 2015, the taxi-owners association of Chicago claimed that Uber and Lyft have caused them a loss of 30-40% in business
- Uber and Lyft started their operations in Chicago in 2011 and 2013 respectively



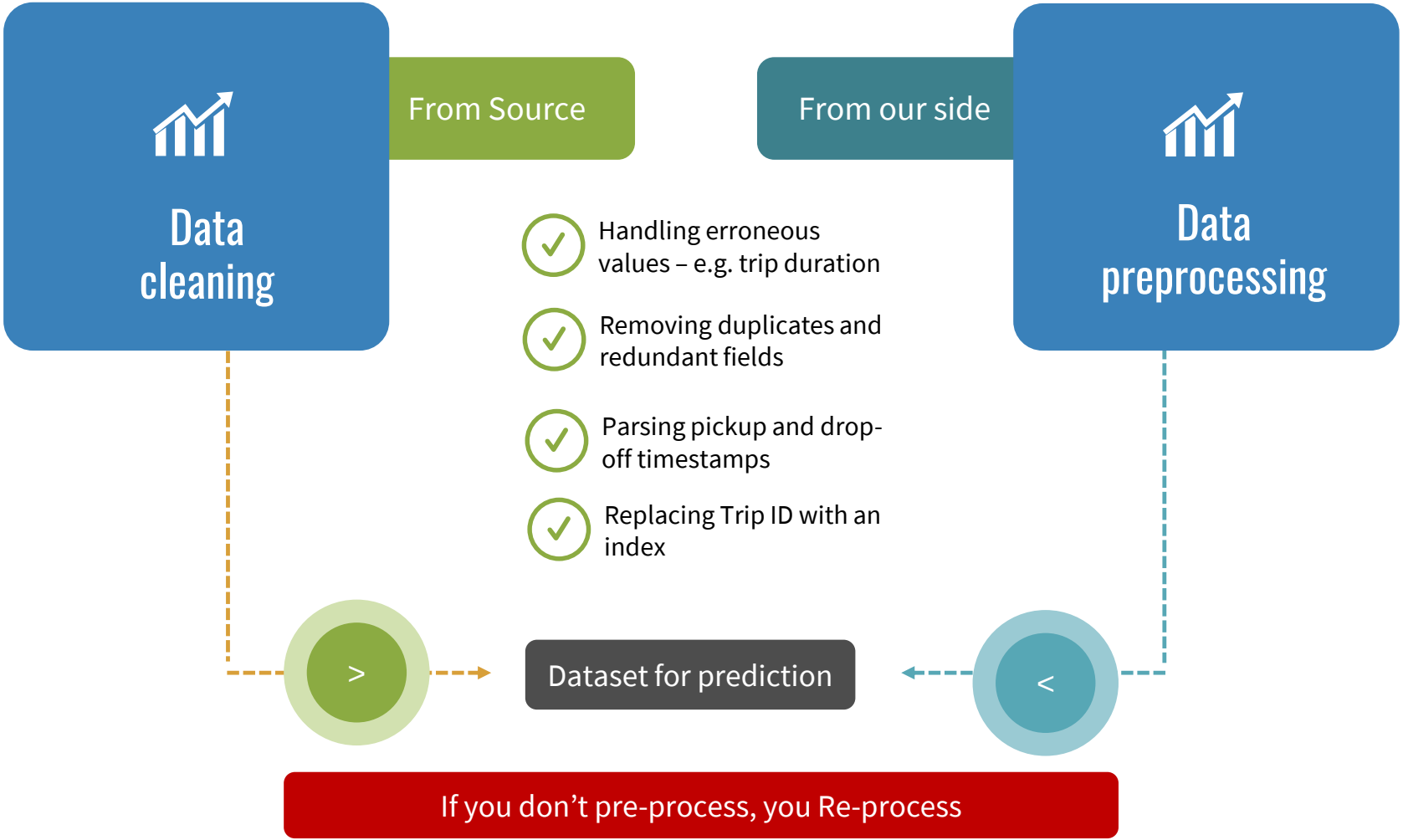
Data Information

Fields

- Taxi ID
- Trip ID
- Trip Start and End Time
- Trip Duration
- Trip Distance
- Fare
- Payment Type
- Taxi Company
- Pickup & Dropoff Location, etc.

Limitations

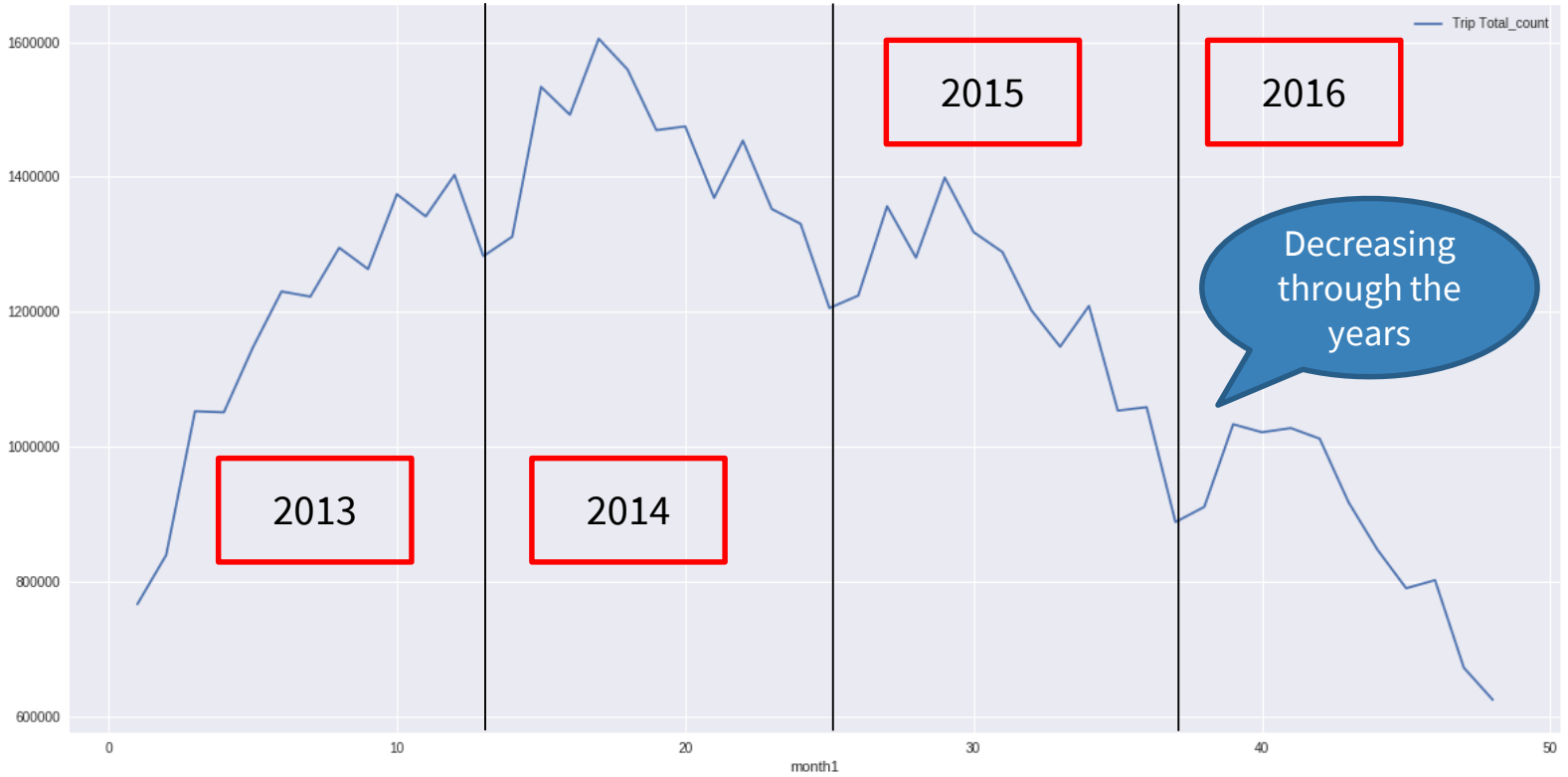
- Trips not reported in real time
- Masking of Taxi ID
- Exact Pickup & Dropoff Location unknown
- Location available on Census Tract and Community area level
- Census Tracts not available for ¼ trips


Data cleaning
From Source
From our side
Data preprocessing
Handling erroneous values – e.g. trip duration
Removing duplicates and redundant fields
Parsing pickup and drop-off timestamps
Replacing Trip ID with an index
>
Dataset for prediction
<
If you don’t pre-process, you Re-process

Data Loading

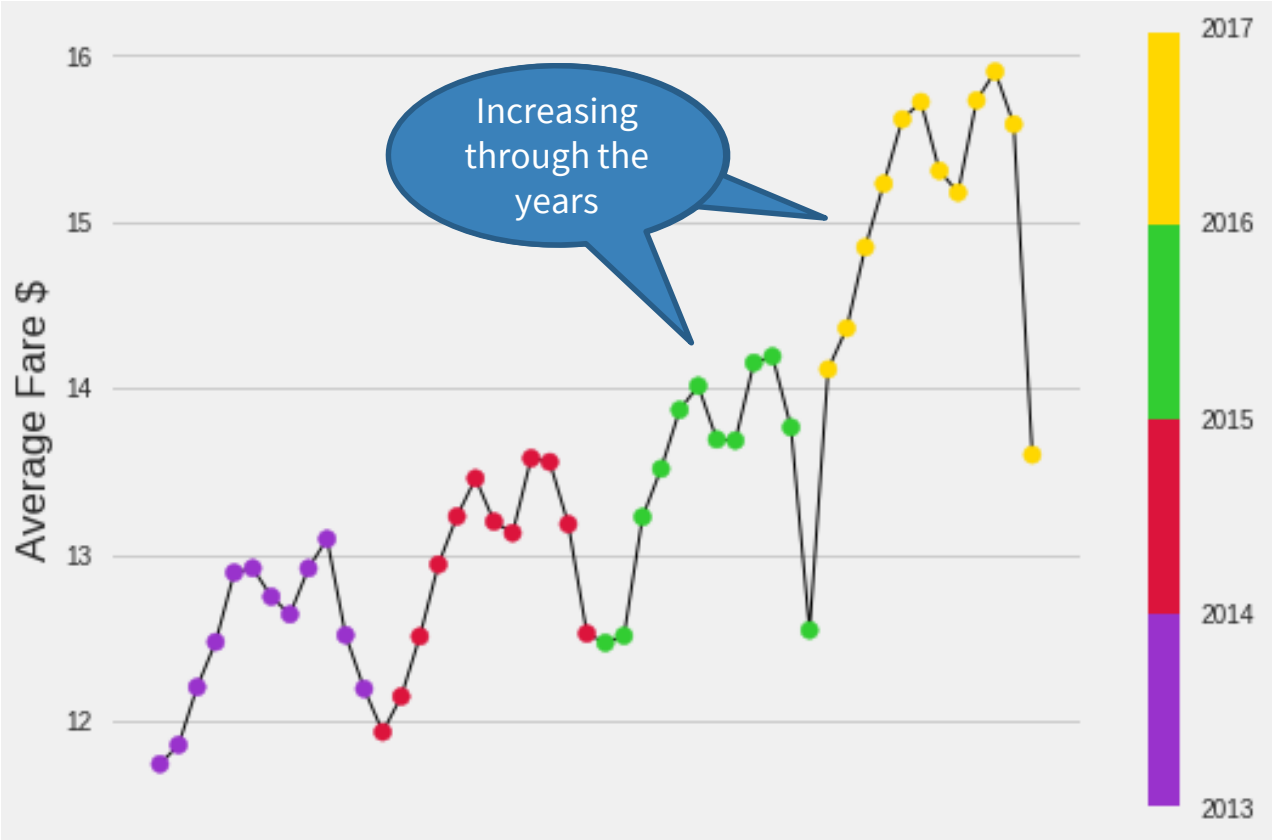
- Data Exploration and Modeling was performed in Google Colaboratory since it uses an accelerated GPU and doesn't require PC's memory for handling ~40 GB data

Data Exploration



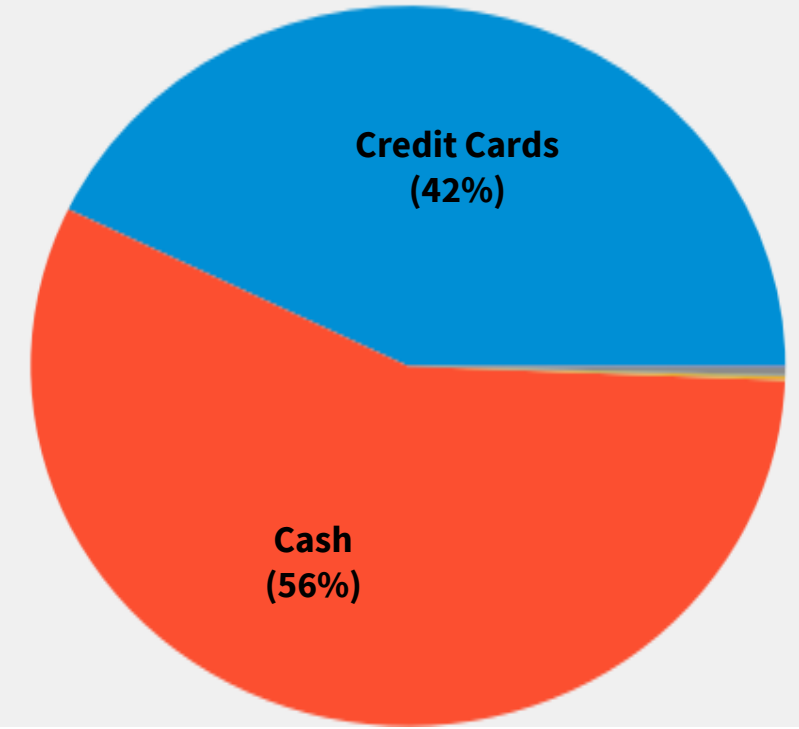
Chicago Taxi Trips in numbers over the years (2013-2016)

Data Exploration



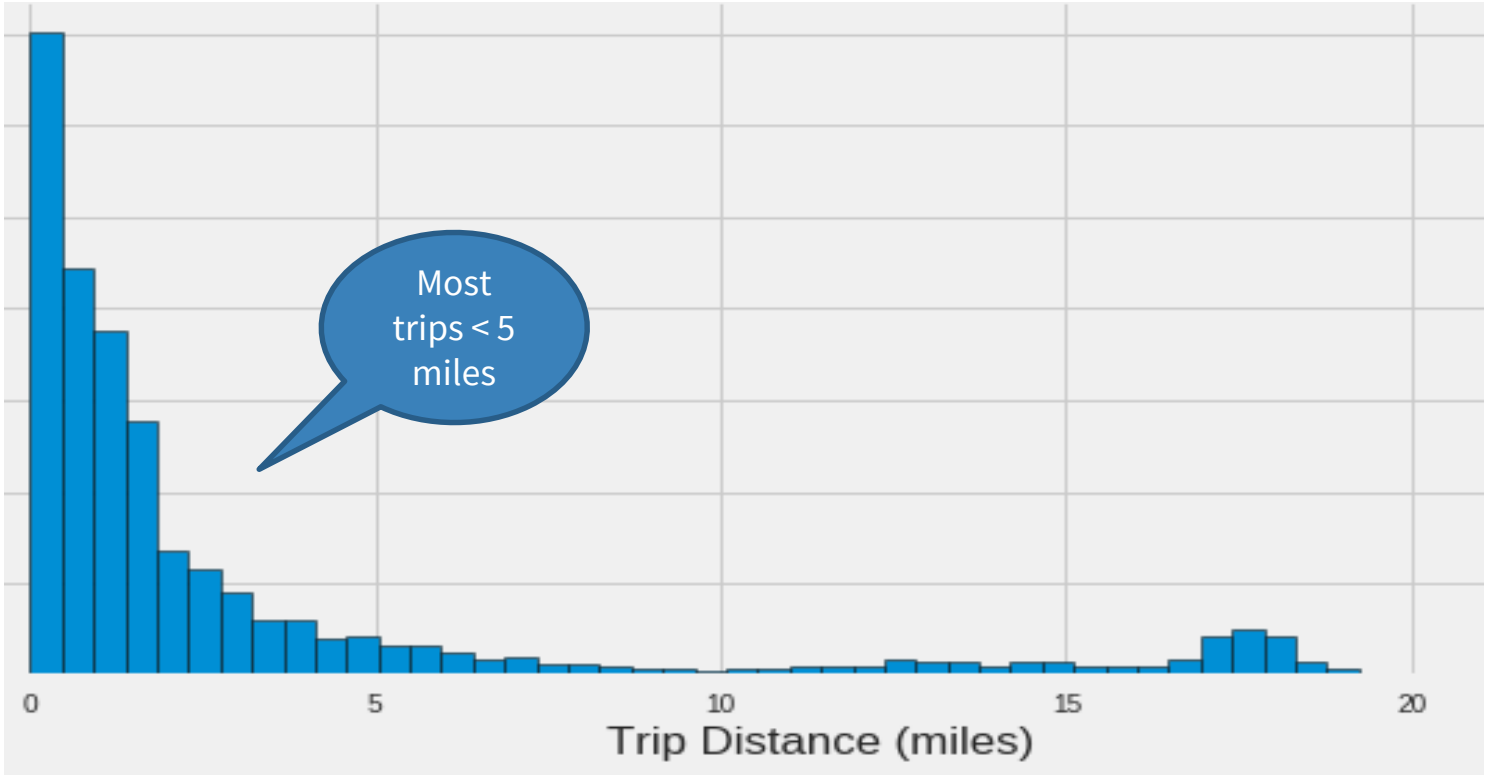
Average Taxi Fares over the years (2013 to 2016)

Data Exploration



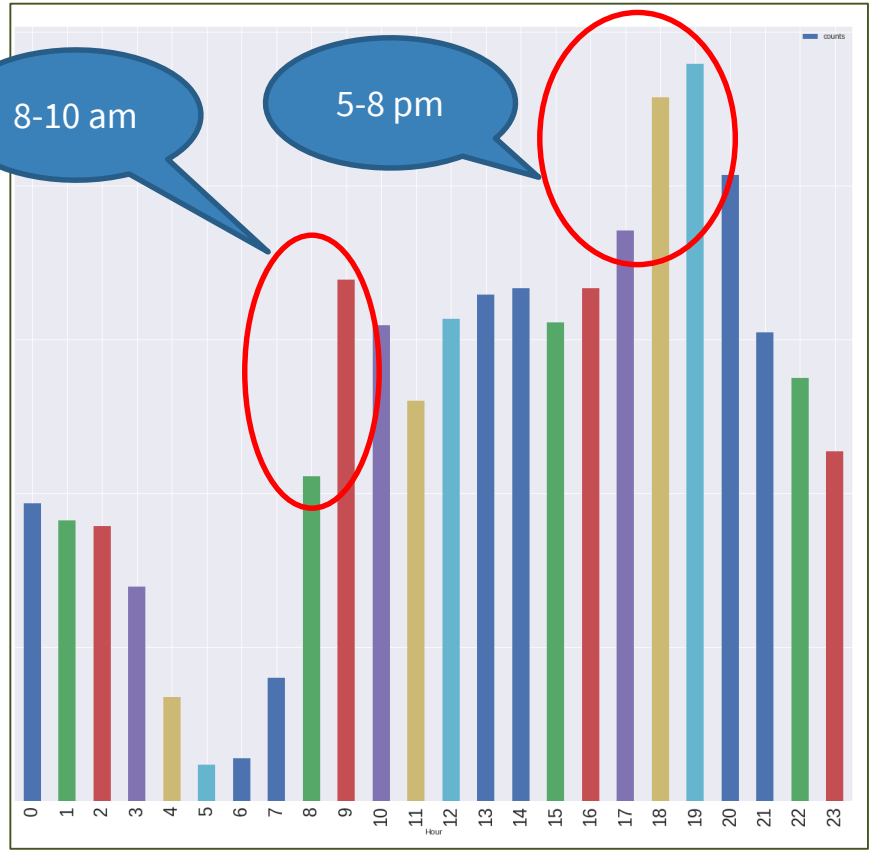
Typical distribution of Payment Types for Taxi fares

Data Exploration

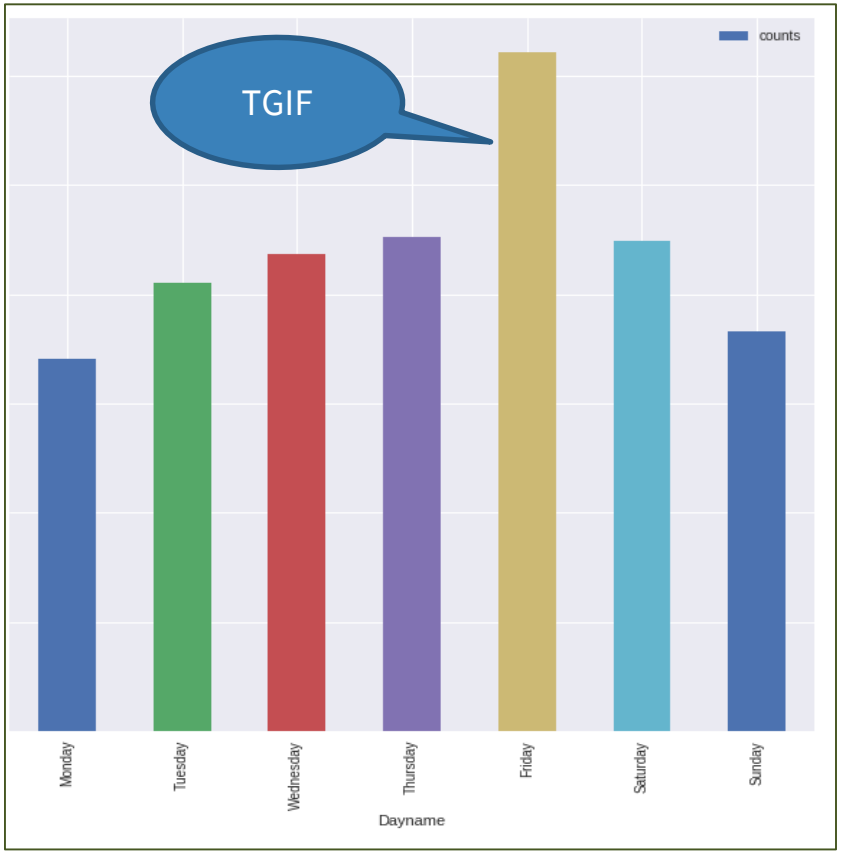


Histogram of no. of trips with Trip Distance

Data Exploration

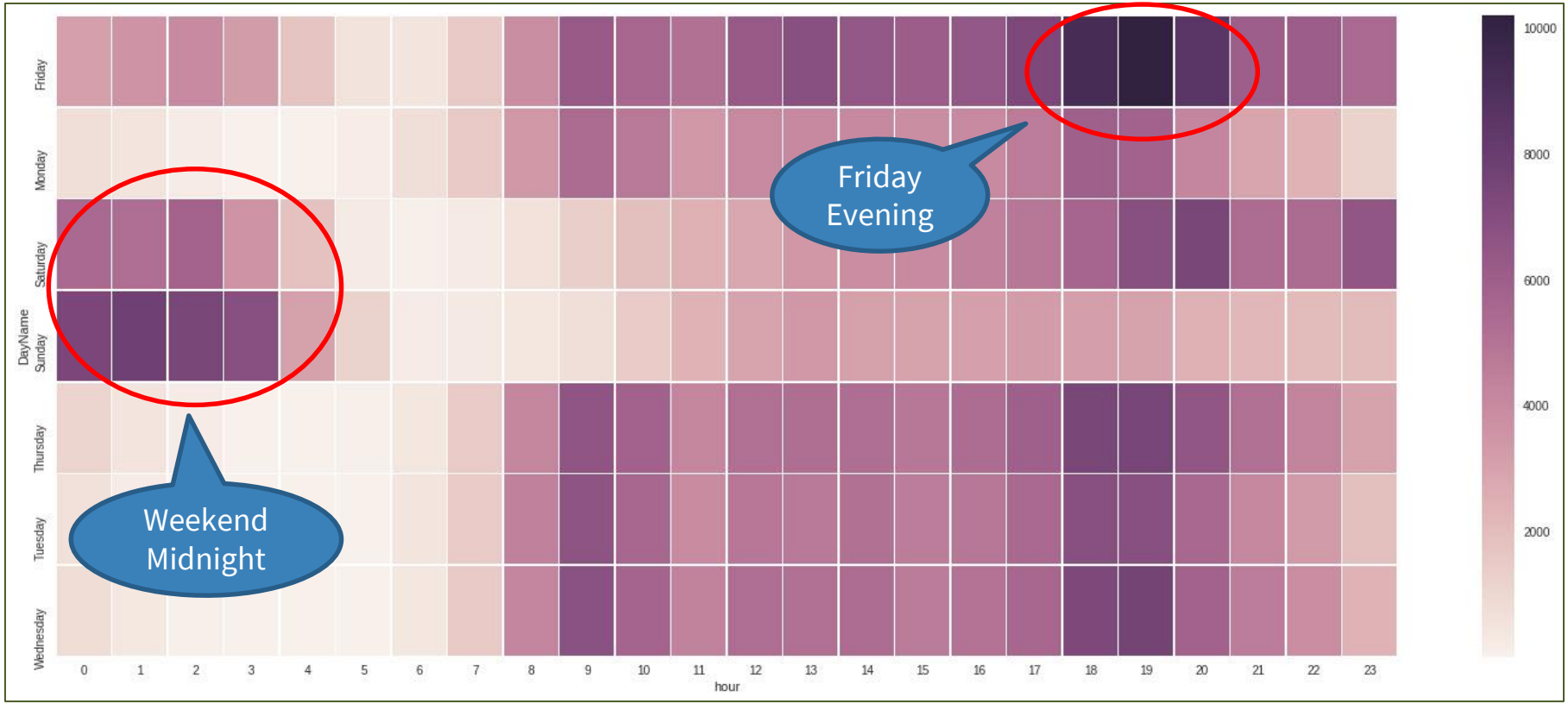


Hour-wise trips on a Typical Day



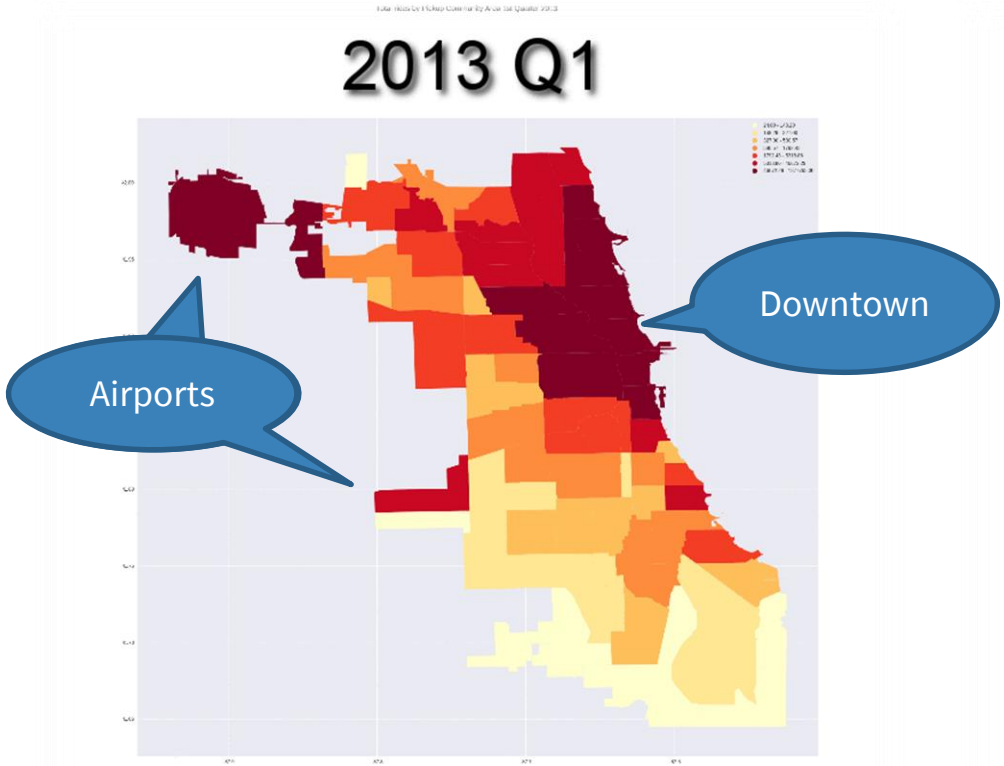
Day-wise trips on a Typical Week

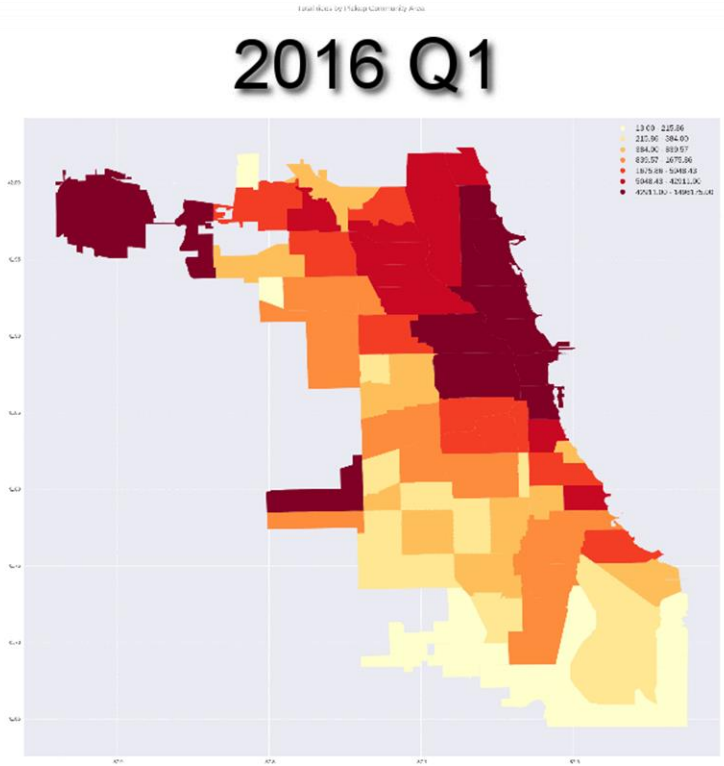
Data Exploration



Heatmap for Day-wise and Hour-wise Trips

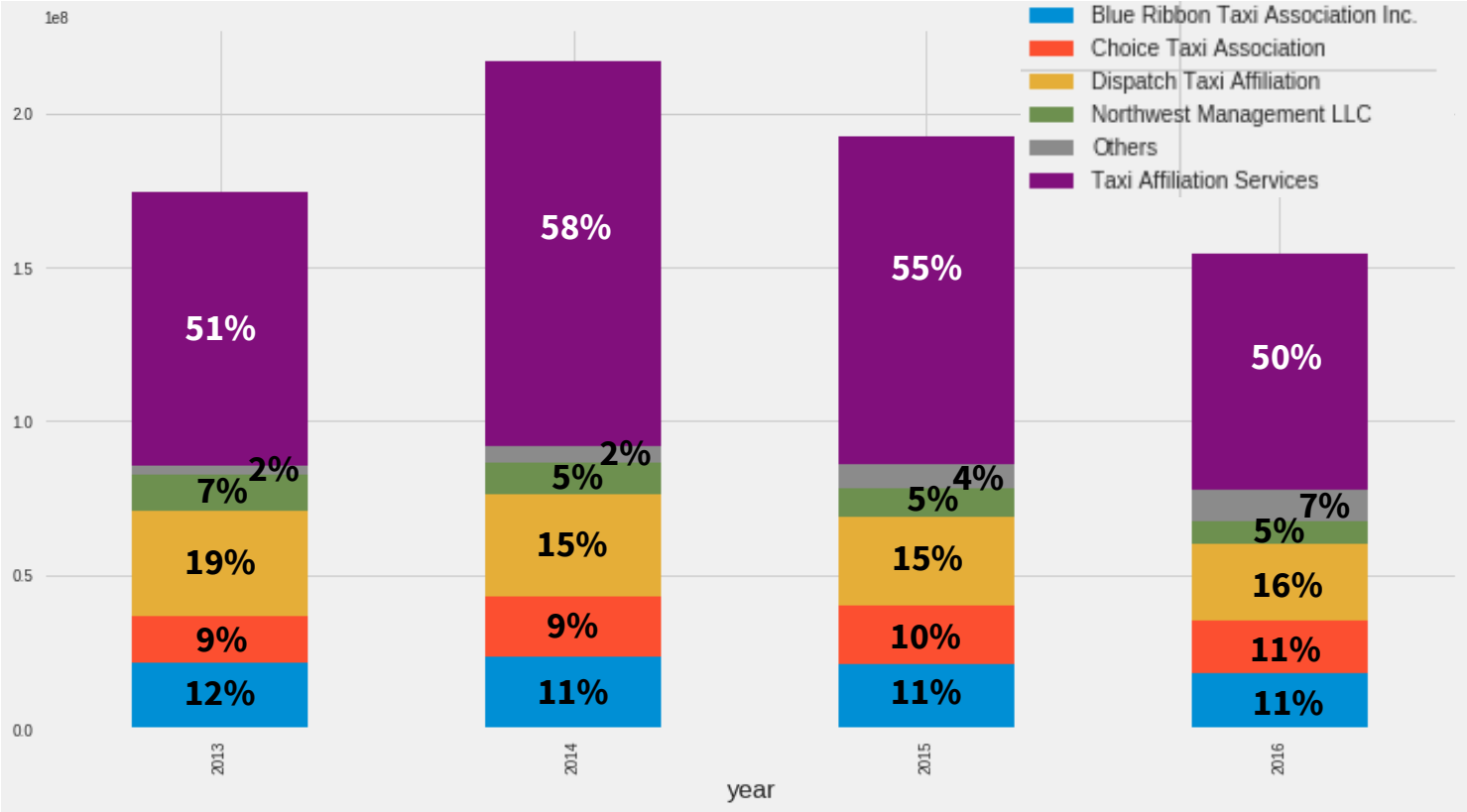
Data Exploration





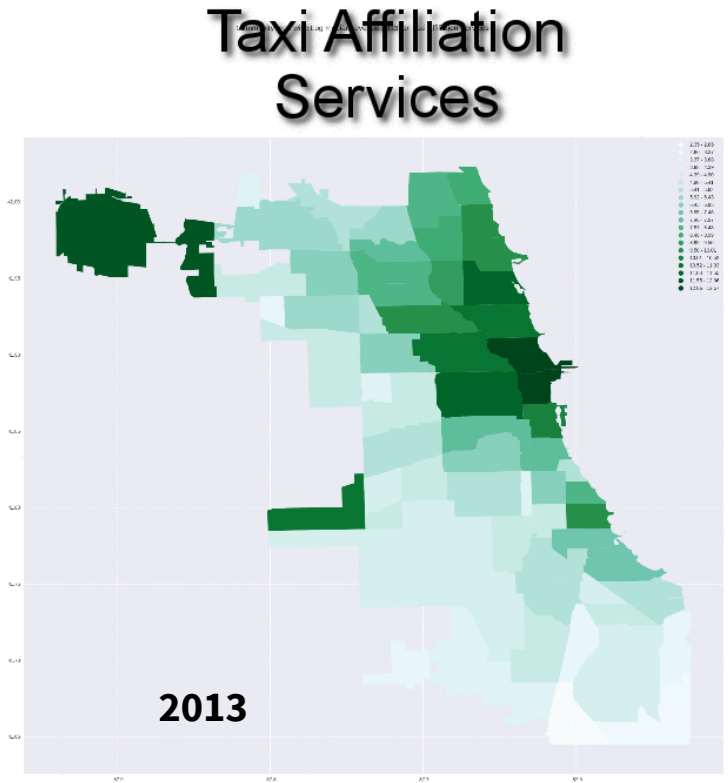
Community-area wise Pickups (2013 v/s 2016)

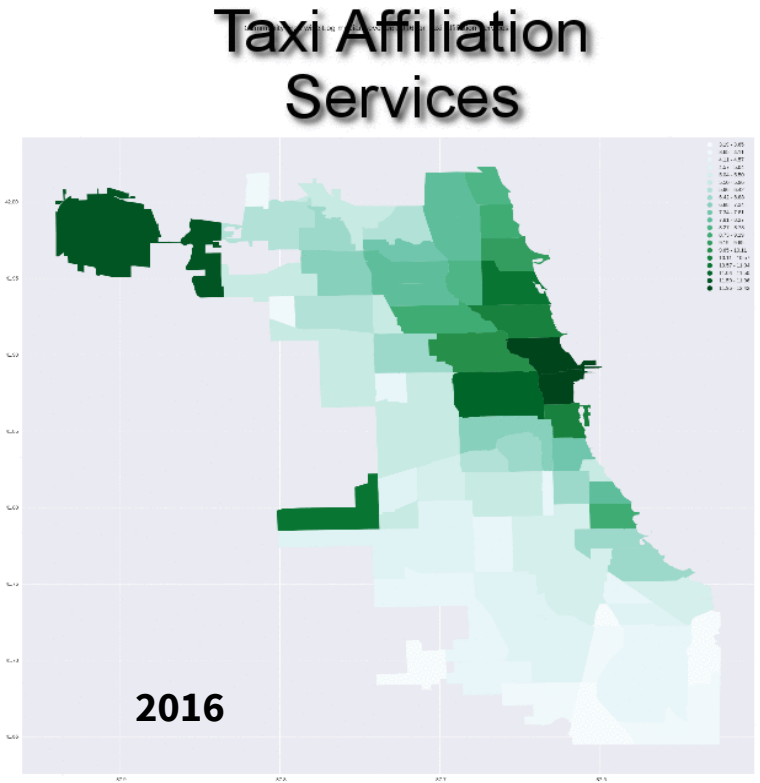
Data Exploration



Market-share of Taxi Companies over the years

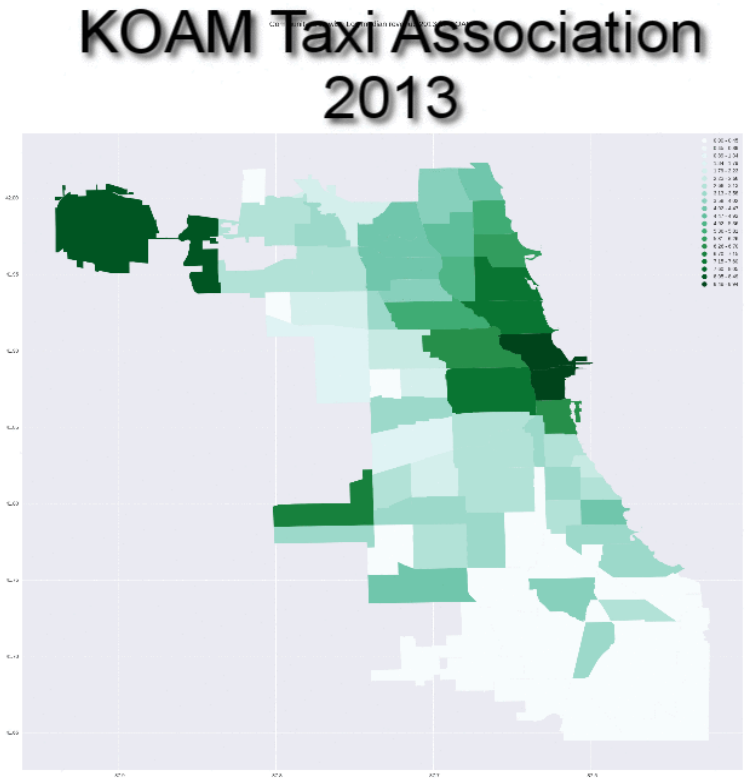
Data Exploration





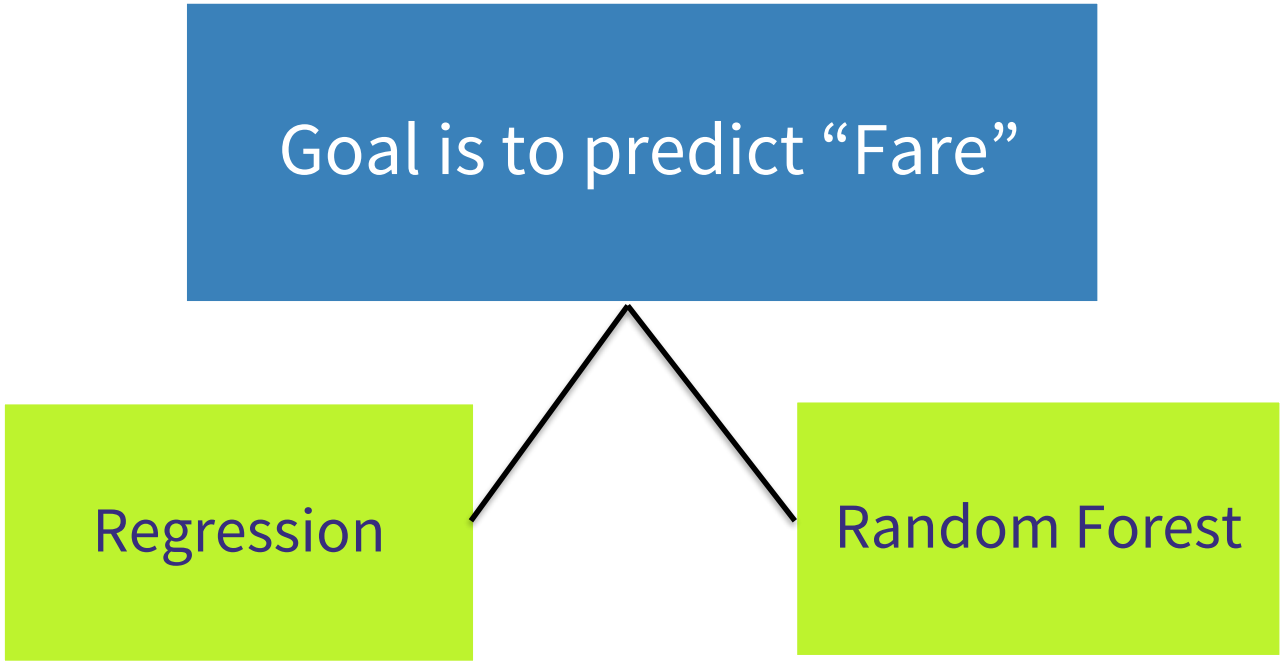
Community area-wise pickups for Top 5 Taxi Companies over the years (2013 v/s 2016)

Data Exploration

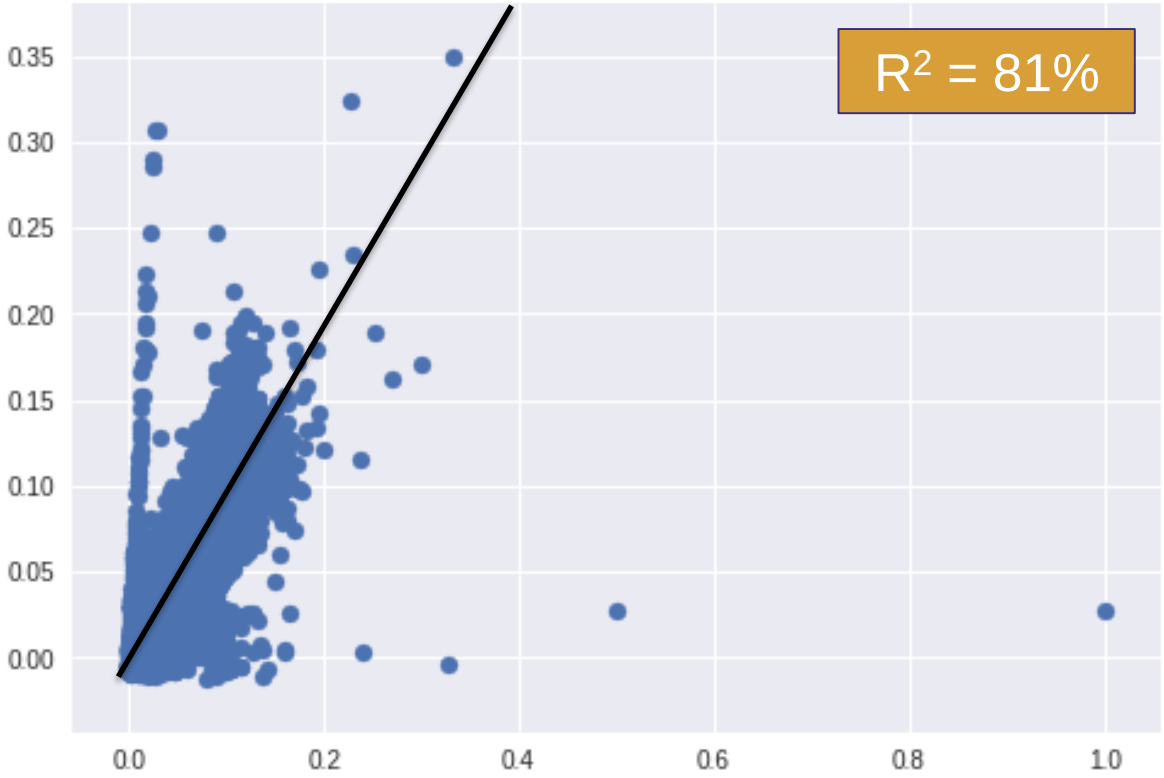


Community area-wise pickups for KOAM Taxi Association (2013 to 2016)

Model Building – Part 1



Regression Results

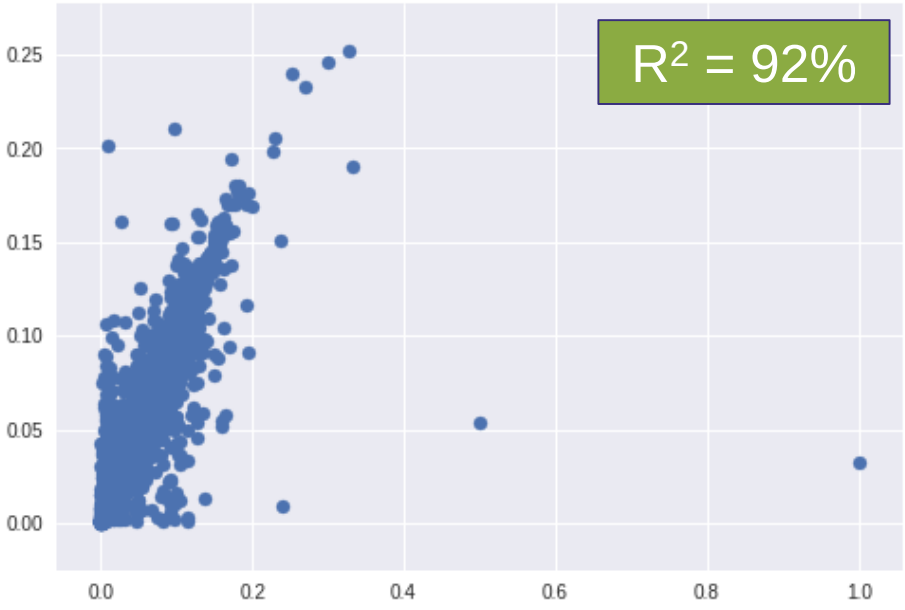


Scatterplot for Predicted V/s Actual Responses (Normalized)

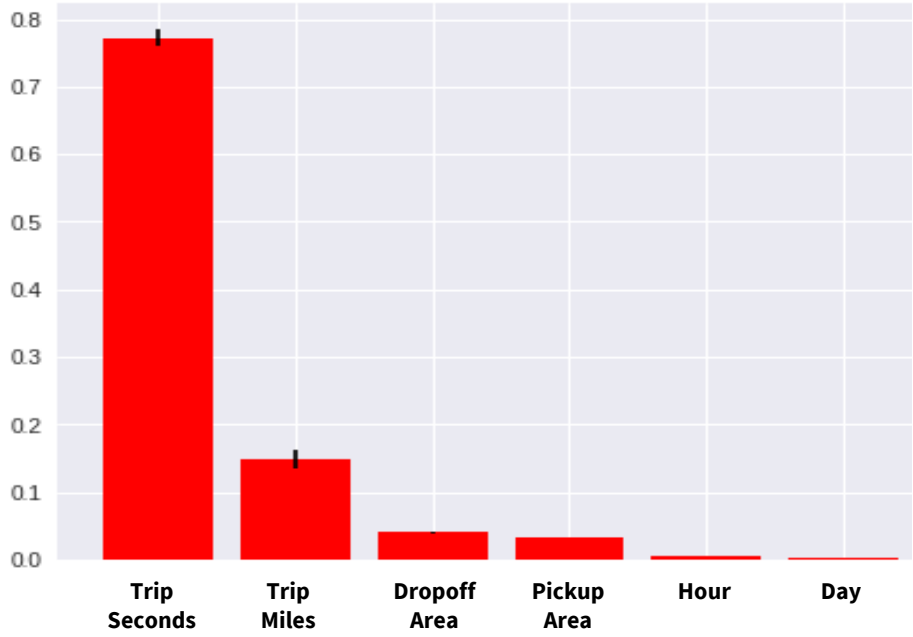
Regression Equation

y *= 0.4678*(trip_seconds) + 0.2904*(trip_miles) + 0.0214* (pickup_community) + 0.0152* (dropff_community) + 0.0015 * (day_name) – 0.0026*(hour)*

Tuned Random Forest Regressor



Scatterplot for Predicted V/s Actual Responses (Normalized)



Feature Importance (Tuned Random Forest)

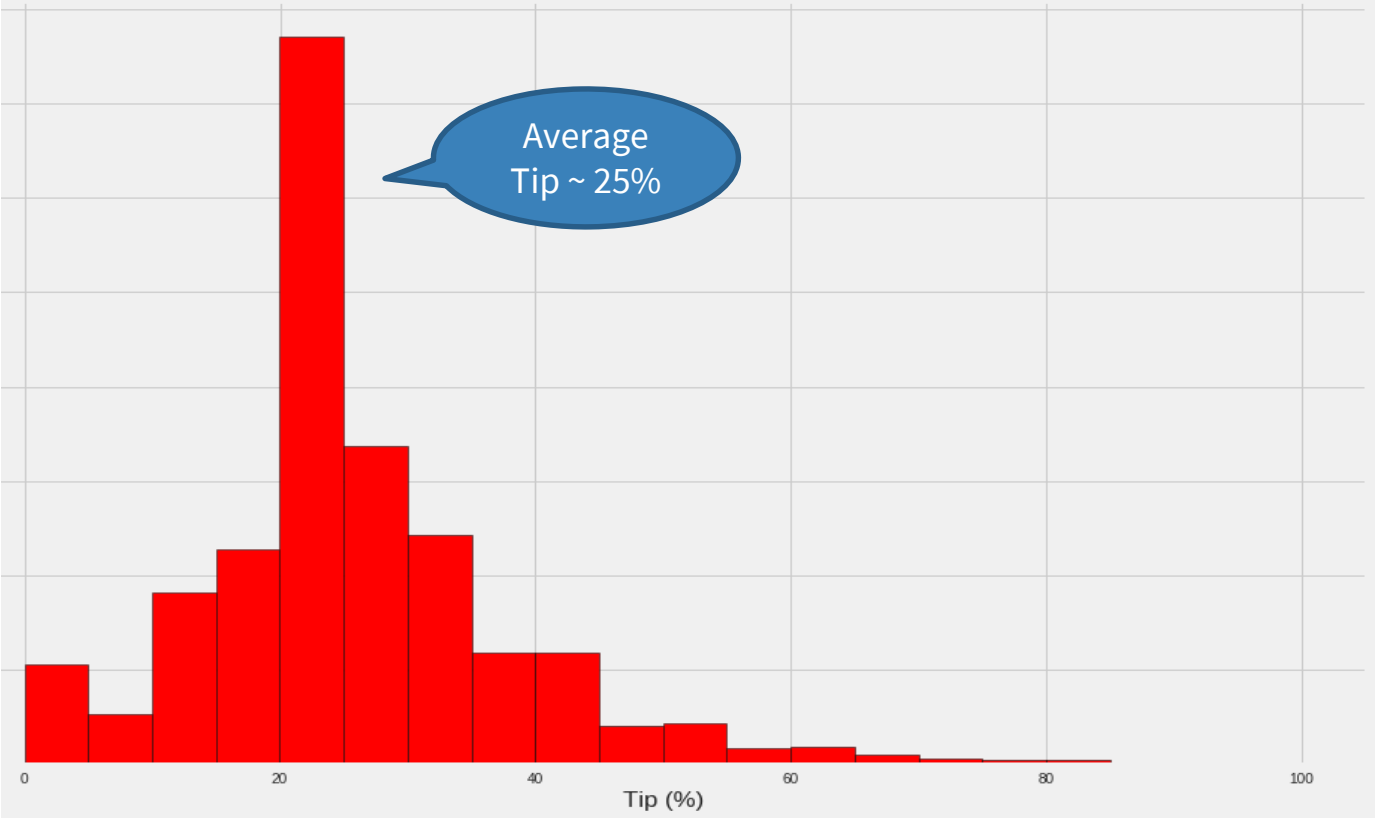
Model Building – Part 2

Attribute of Interest = **Tips**

Tips contribute a major share of taxi drivers' take-home income

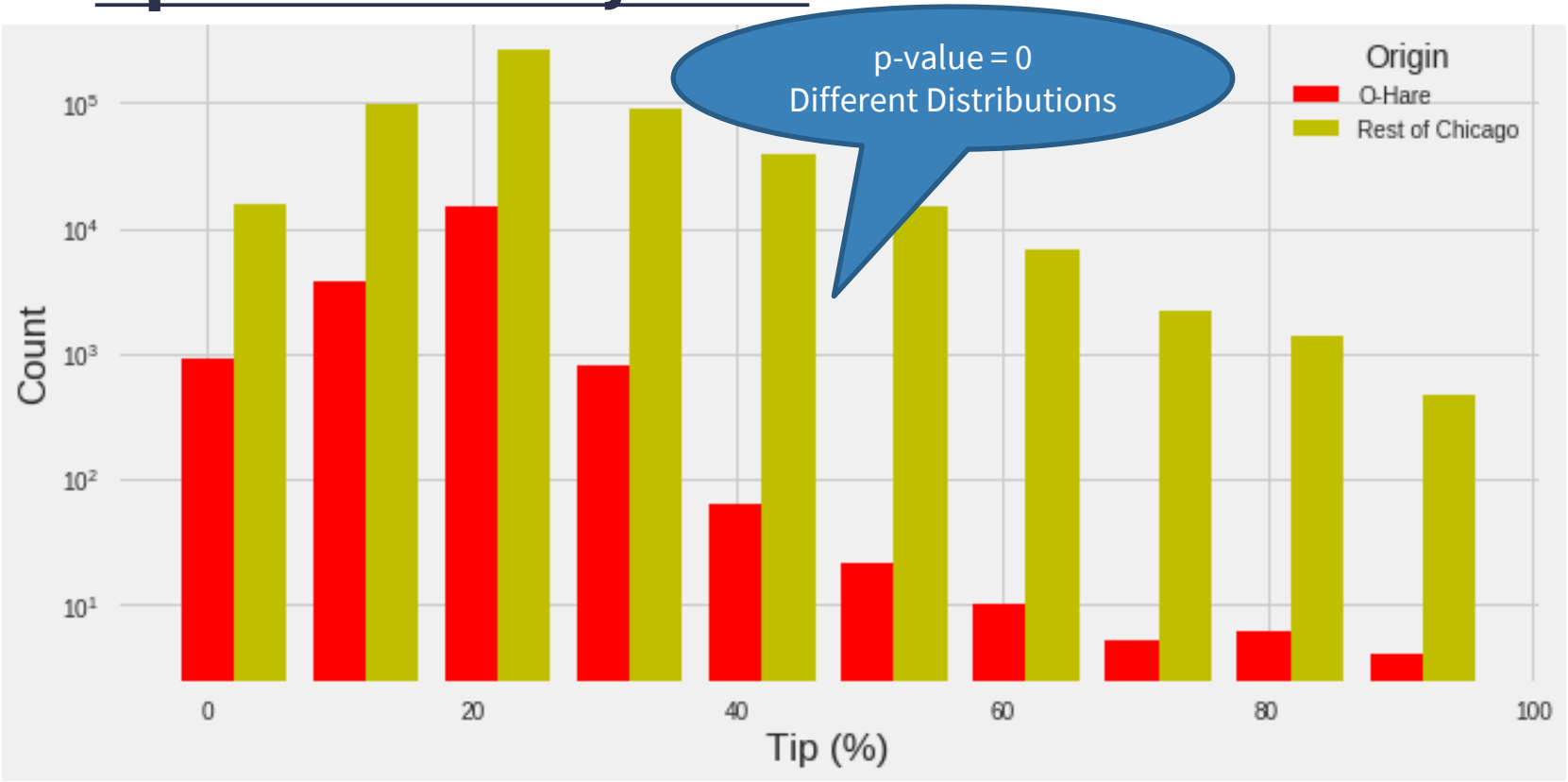
Which factors affect tips?

Tips – How do they look?



Histogram of % Tip value

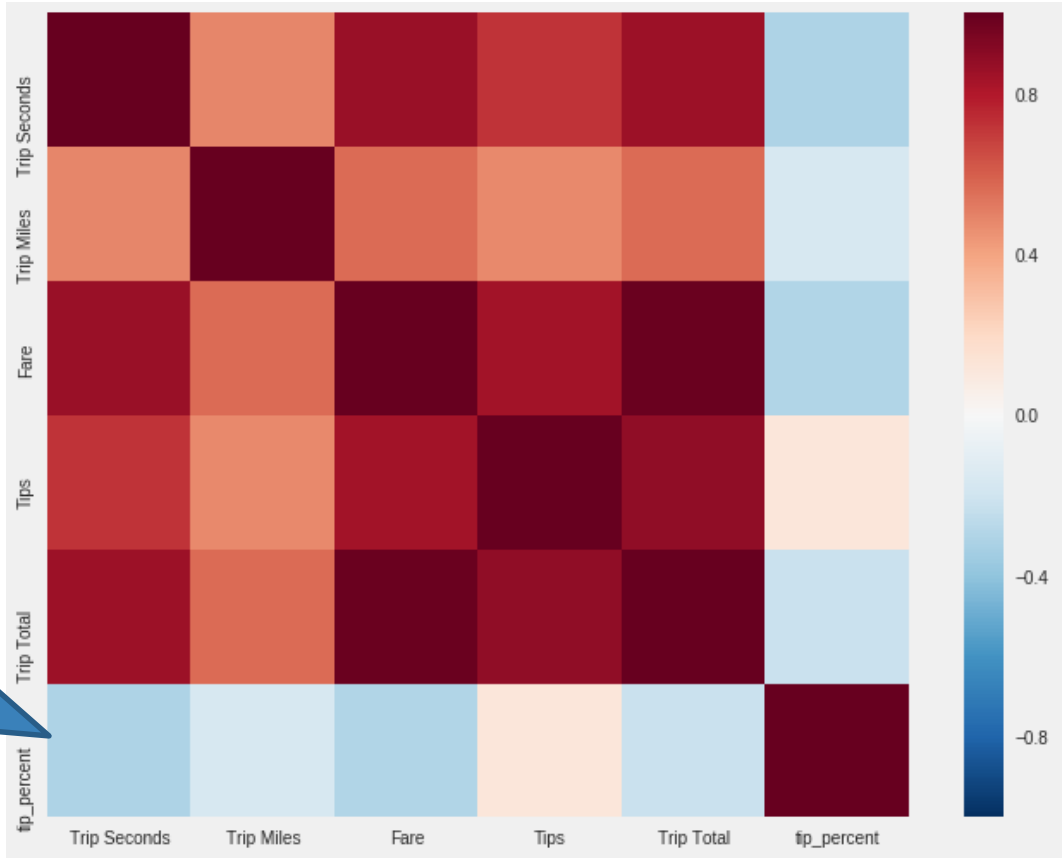
Tips – How do they look?



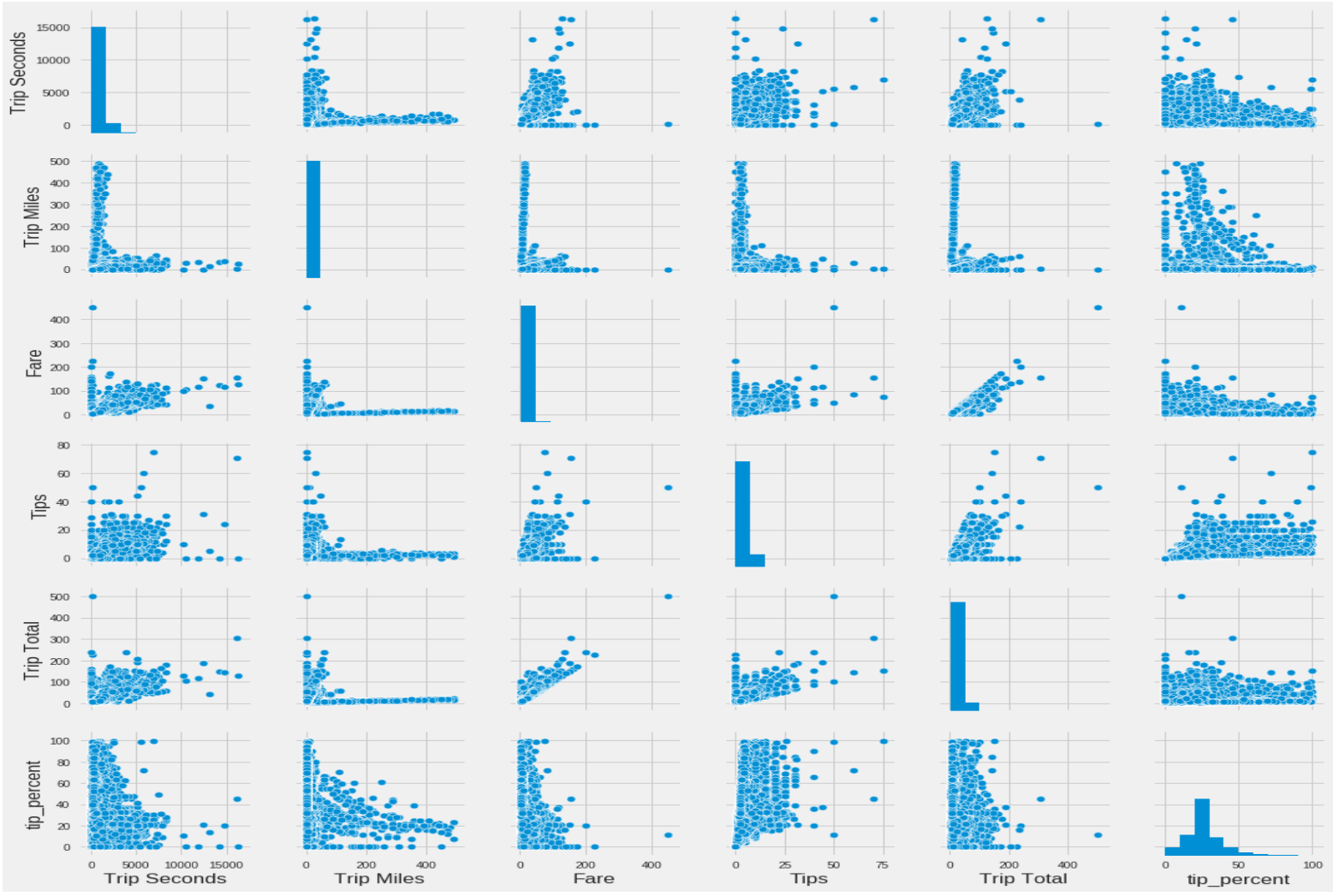
Histogram of % Tip value for O'Hare vs Rest of Chicago Pickups

Tips or Tip % ?

Hypothesis:
Almost no linear relationship with other continuous variables



Relationship heatmap between variables



Pair Plot between variables

Data Imbalancing

4% - 96% imbalance

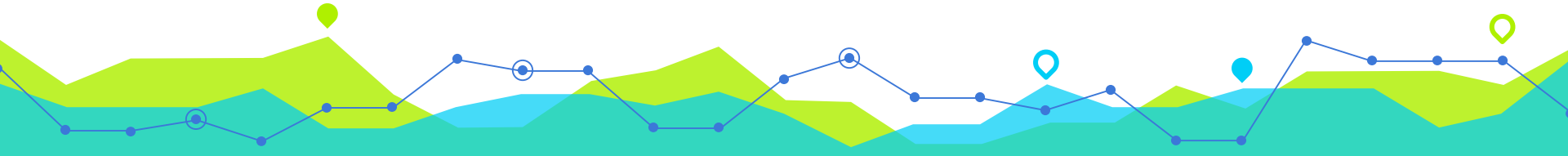
- Only 4% records have 0 tips
- Imbalance poses a challenge in classification

Under sampling

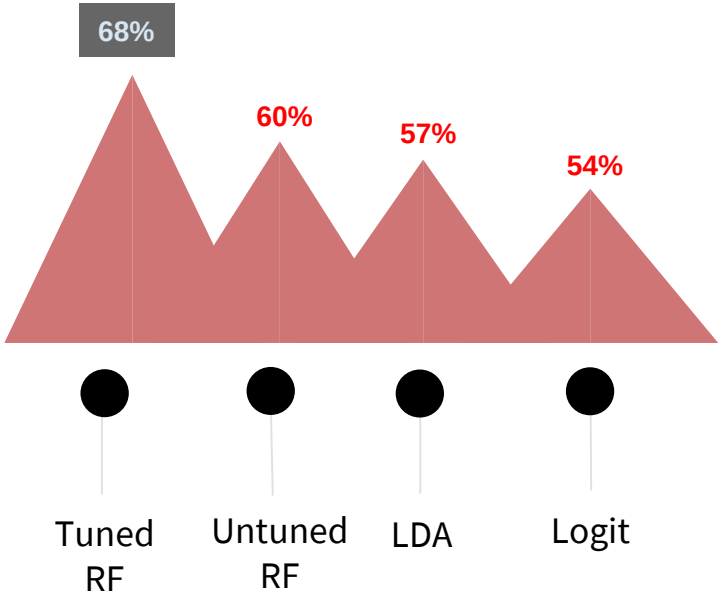
- Separated those 4% records from rest of the data
- Used rest 96% data for sampling

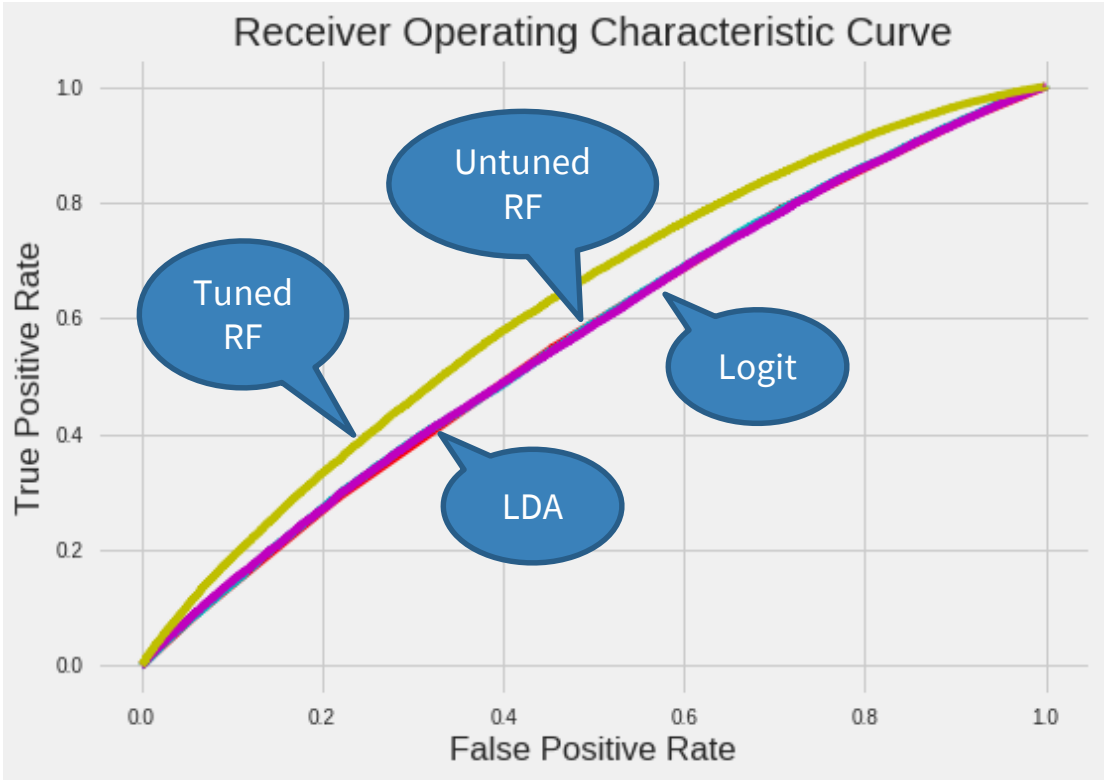
Training Data

- Sampled observations from 96% dataset in a ratio of 3:1 to the 4% dataset

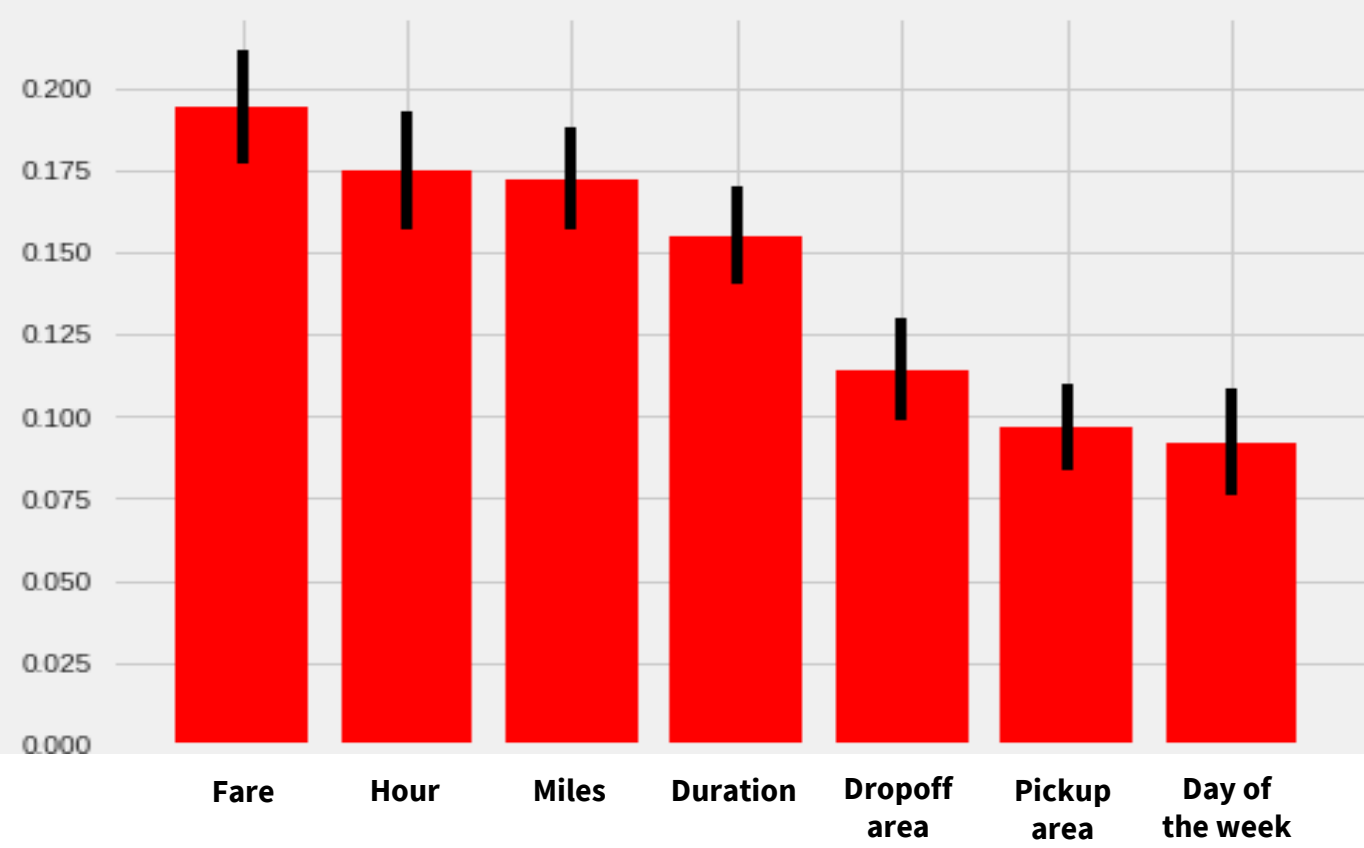


Classification Models and their Accuracy





Feature Importance (Tuned Random Forest)



Strength & Limitations

YES to Big Data

Our methodology helps analyse huge datasets even ~40 GB ones

Collinearity is not a curse

RF model ensures that performance is unhindered by non linear relationships

Highly performant

Easy implementation, validation and testing

Significantly high computation time even with GPU accelerated system

Computation Time

Route Prediction is not possible due lack of relevant information, etc.

Data Inadequacies

What we learnt?

Predictive Modeling is like a sandbox

Requires Creativity
Can (attempt to) predict anything

Perhaps disproving something still carries value

Can change your point of view

Do not underestimate data preprocessing

80/20

Go for it

Taxi Drivers will thank you!

Thank You!

#BTHOFinals

P.S. Don’t forget to tip the cabbie ;)