Chicago Taxi Allocation

Omar ALShaye, Raghav Atal, Rush Samal

Overview

Our Client(s)

- Taxi Service Company
- Future Clients:
    - Chicago Transit Authority
    - Divvy

Problem Statement

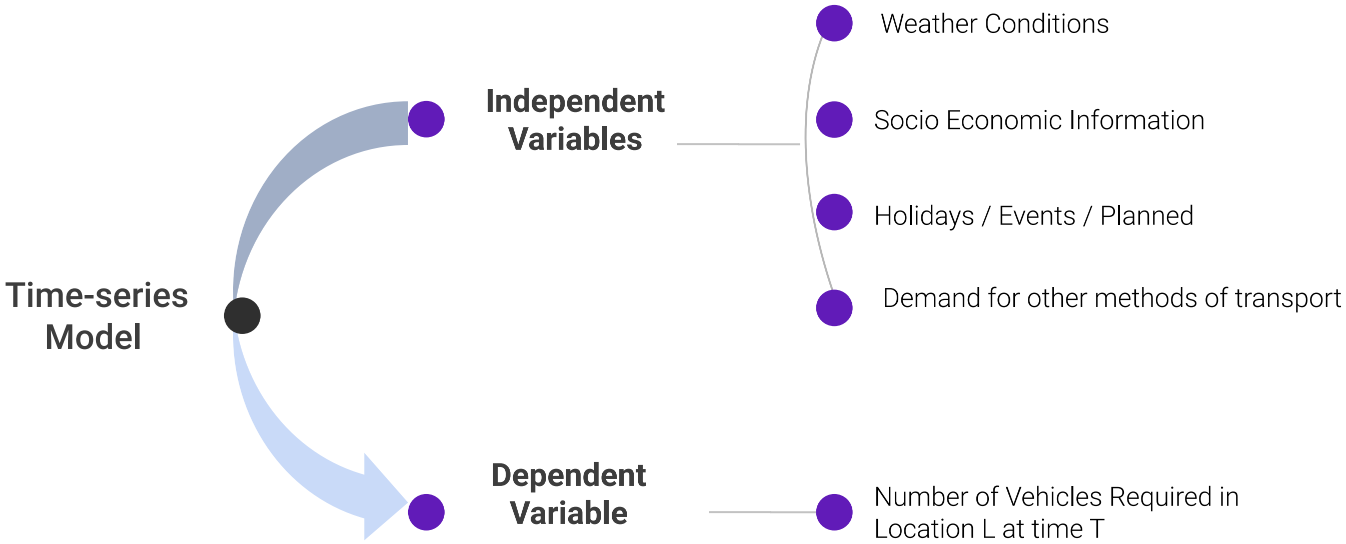
- Taxis face a technological disadvantage when it comes to understanding demand locations of the customers
- Transportation Resources within the city of Chicago are not optimally allocated for different demand patterns

Goals

- Design a comprehensive Big Data/Dashboard solution that allows for better transportation resource allocation and scheduling

Demo

Predictive Modeling Goals



Time-series Model

Independent Variables

- Weather Conditions
- Socio Economic Information
- Holidays / Events / Planned
- Demand for other methods of transport

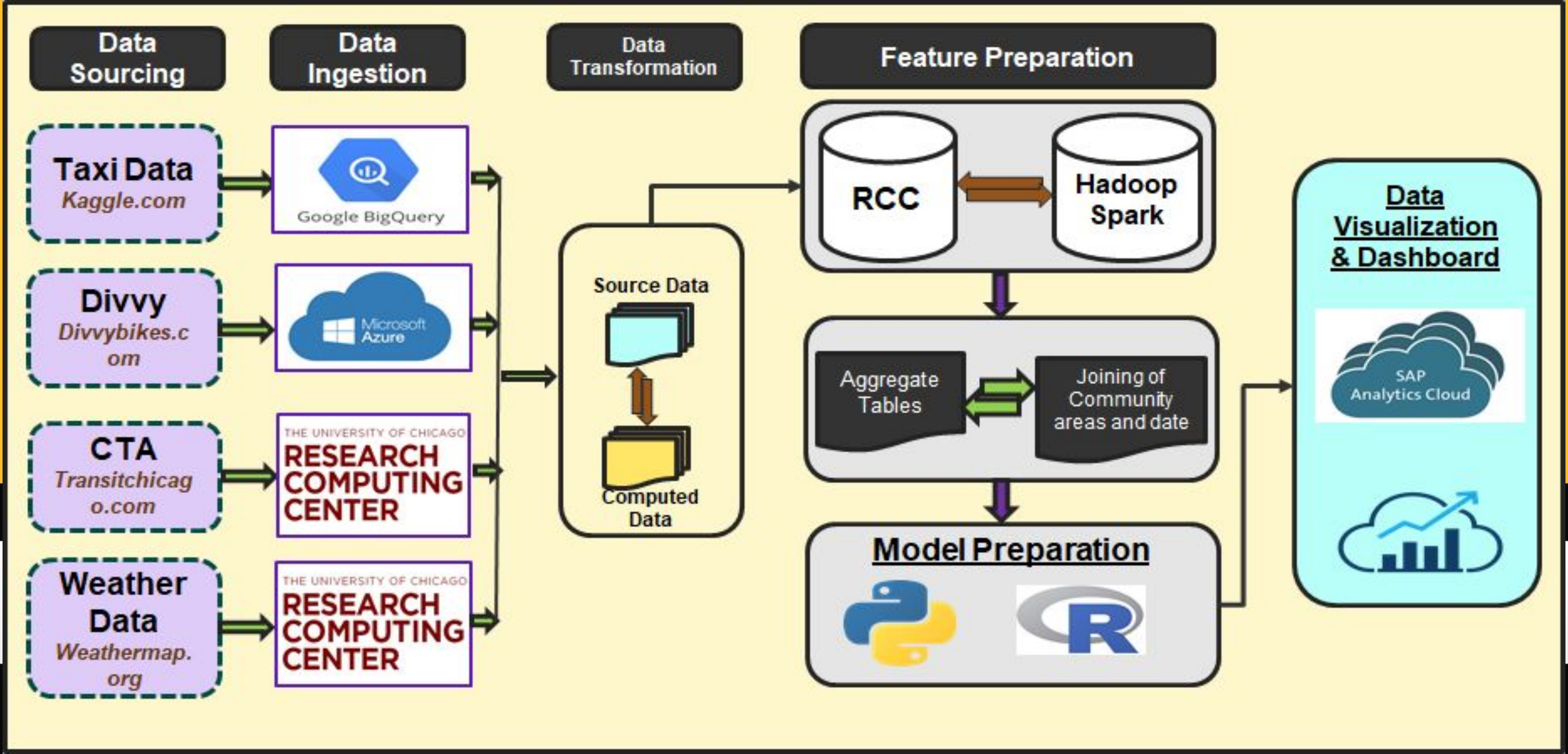
Dependent Variable

- Number of Vehicles Required in Location L at time T

Data Sources

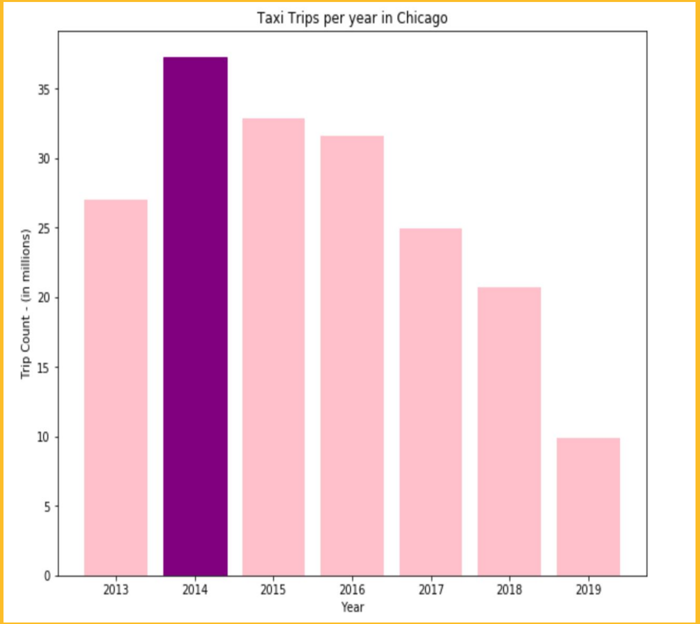
	Taxi Data	Divvy	CTA L Rides	Weather
Size	32 GB	~ 2 GB	~ 33.681 MB	~ 17,987 KB
UoA	Ride Timestamp / Pick-up Lat/Long	Ride Timestamp / Pick-up Station	Daily Ridership/ Station Name	Hourly / ZIP Weather
Timeline	2013 - Present	2013 Q3- 2018 Q4	2001-2018	2014-2018
Resource	https://www.kaggle.com/chicago/chicago-taxi-trips-bq	https://www.divvybikes.com/system-data	https://www.transitchicago.com/data/	https://openweathermap.org/api

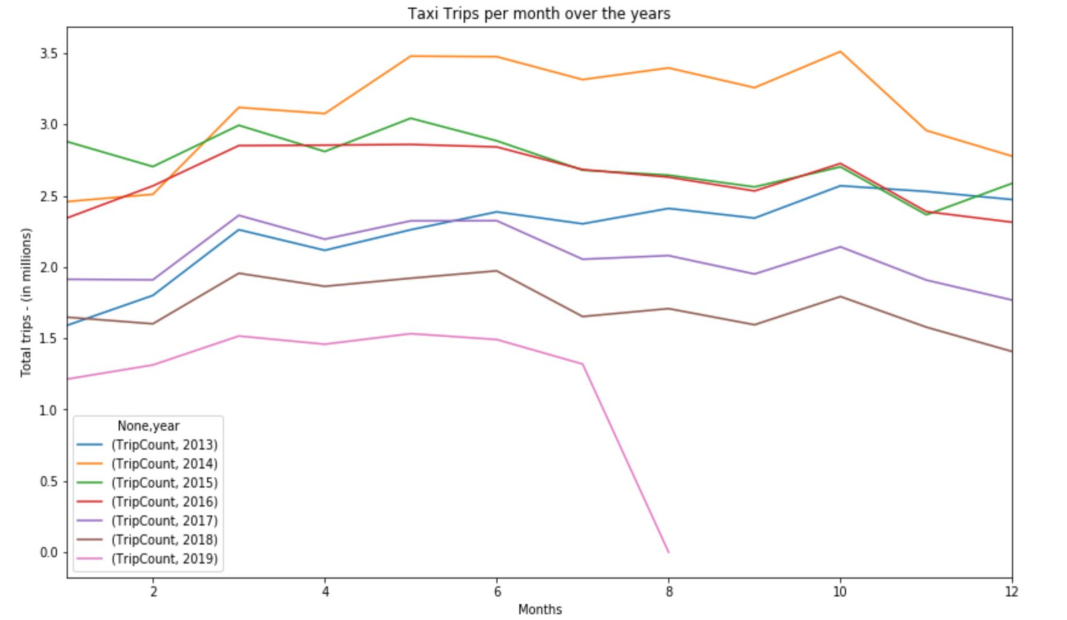
Big Data Architecture



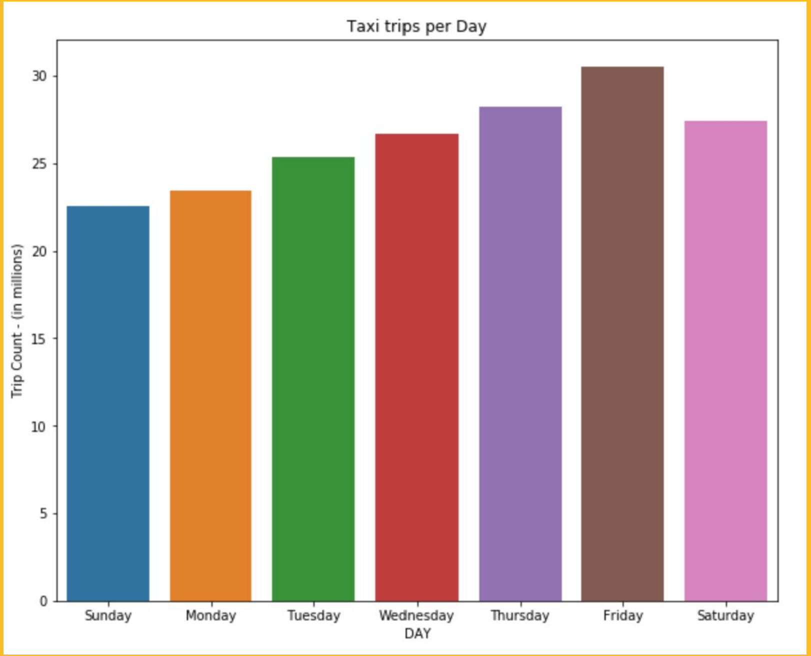
Exploratory Data Analysis

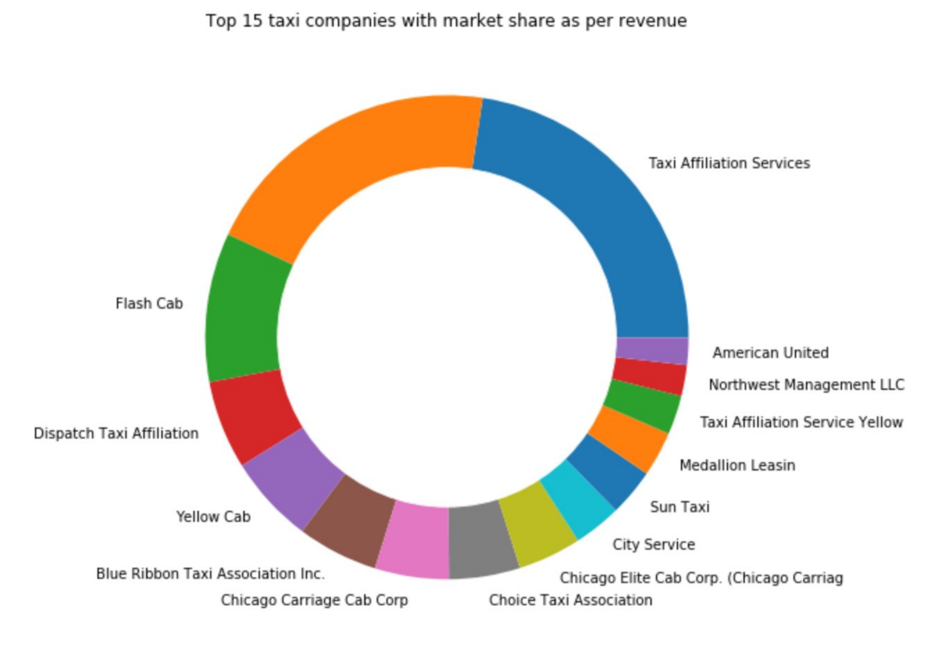
Exploratory Analysis - 1



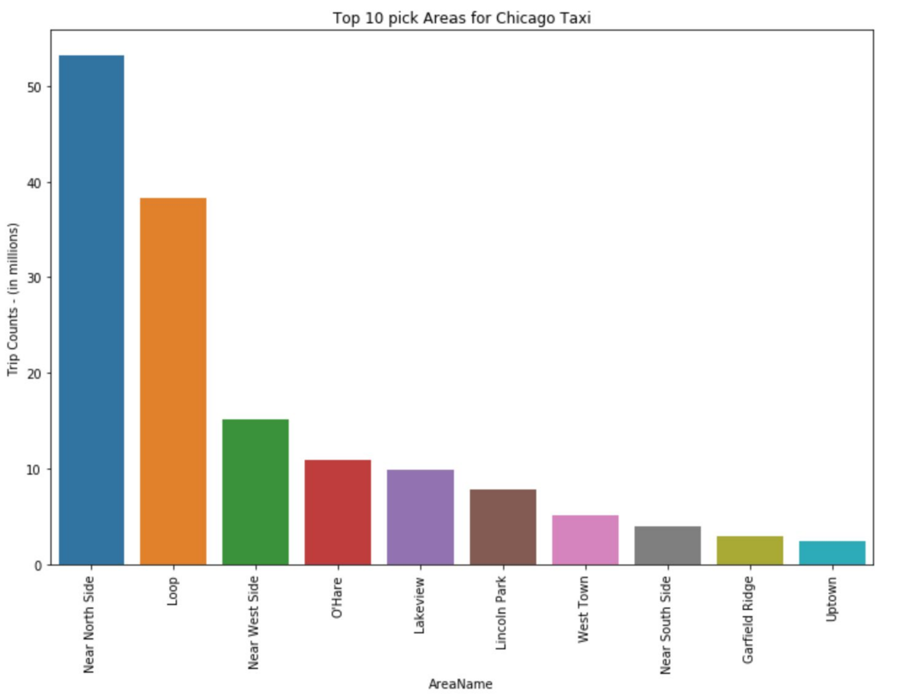


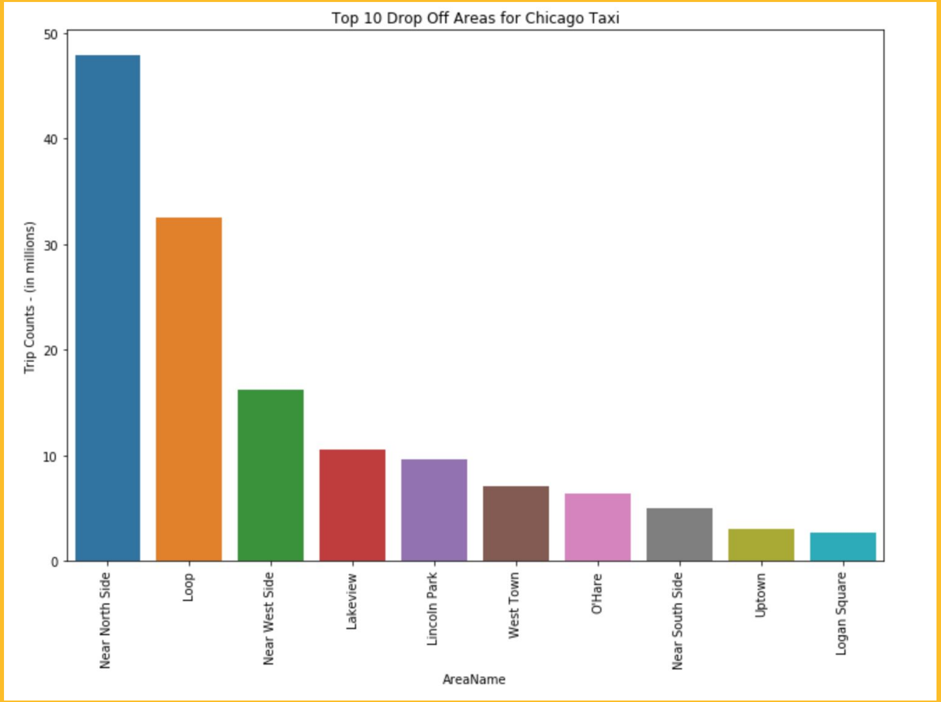
Exploratory Analysis - 2





Exploratory Analysis - 3





Data Preparation

Finding a Common Unit of Analysis

Common UoA

- **Demand:** Count of Vehicle
- **Time:** Daily
- **Location:** Community Area
- **Weather:** Average Temperature

Actuals/Training

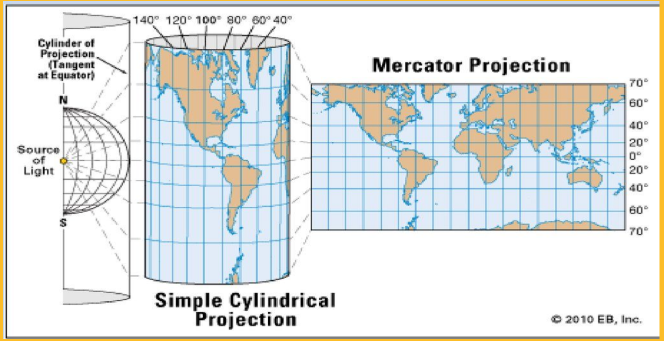
Timeline: 2014-2016

Predictions

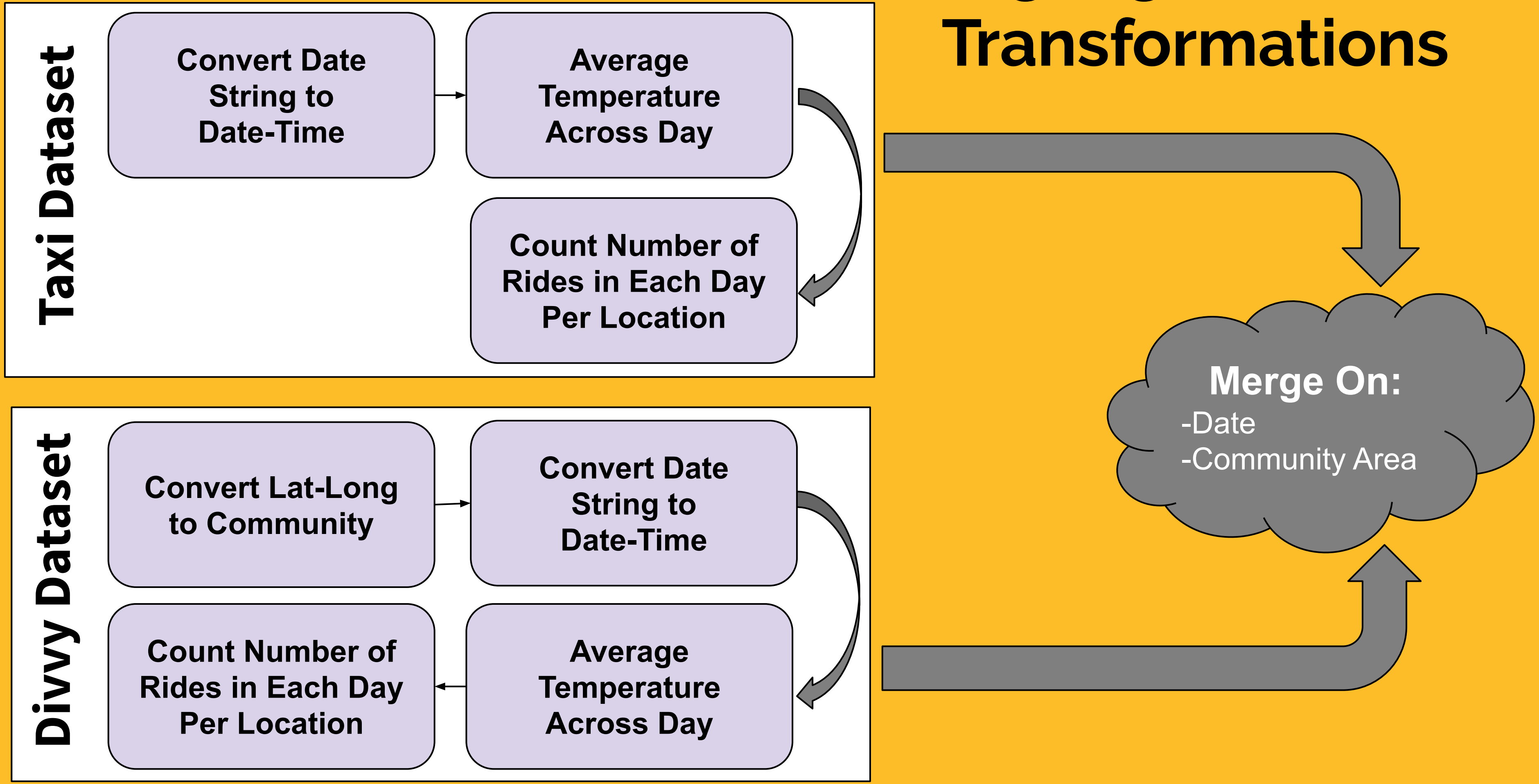
Timeline: 2017

Transforming Coordinates to Community Area

- User defined function to find community area based on coordinate
- Dataset of community area polygons
- Spark function performed to identify community area from coordinate, types of calculations:
  - Mercator Projection
  - Ray Casting Algorithm
  - Boolean Check
- Big Data parallelization allowed complex function to be applied to the large dataset



Highlighted Transformations



Taxi Dataset

Convert Date String to Date-Time

Average Temperature Across Day

Count Number of Rides in Each Day Per Location

Divvy Dataset

Convert Lat-Long to Community

Convert Date String to Date-Time

Average Temperature Across Day

Count Number of Rides in Each Day Per Location

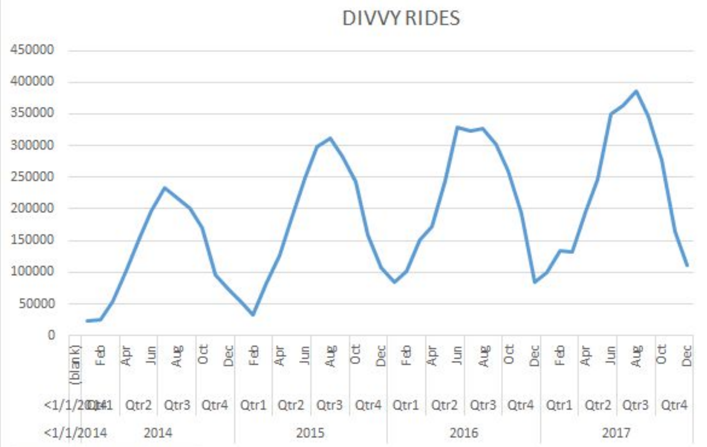
Merge On:

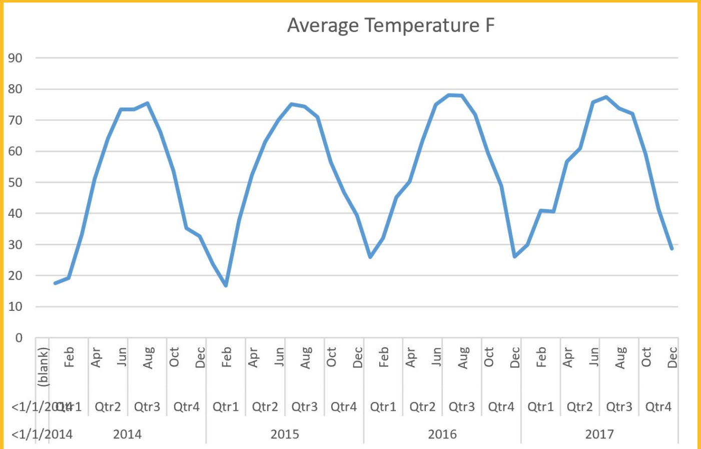
-Date

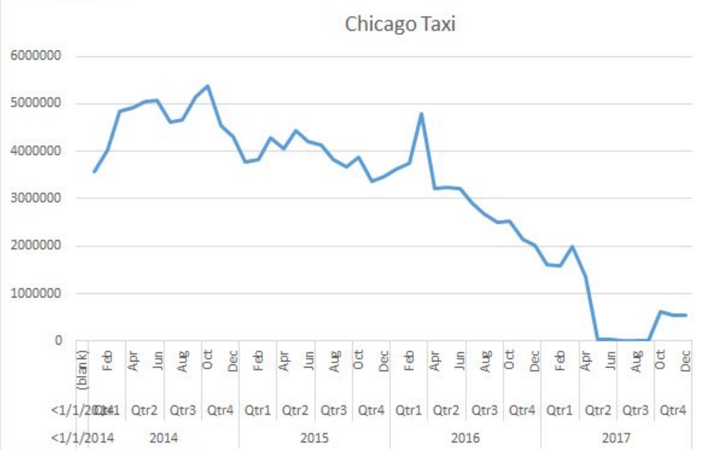
-Community Area

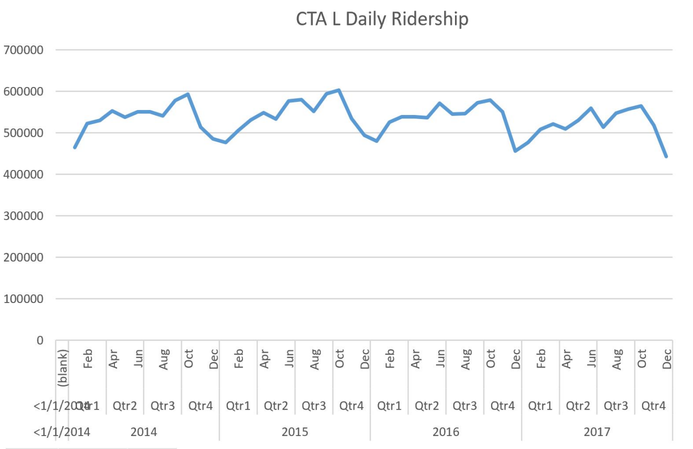
Time-Series Analysis

Chicago Transportation Timeseries



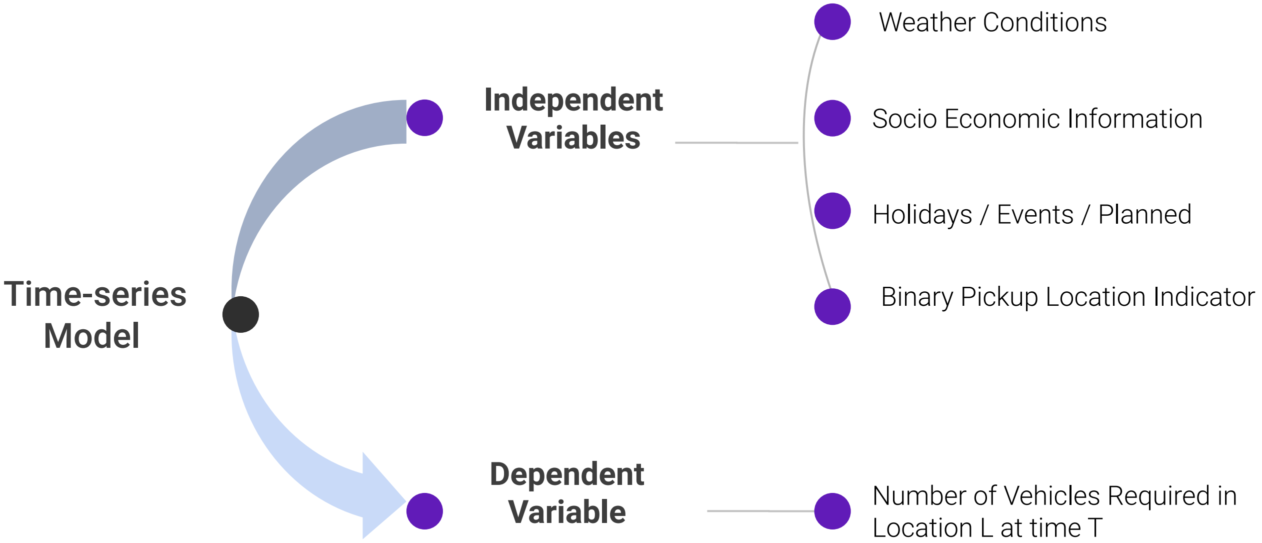






Predictive Modeling

Predictive Model



Time-series Model

Independent Variables

- Weather Conditions
- Socio Economic Information
- Holidays / Events / Planned
- Binary Pickup Location Indicator

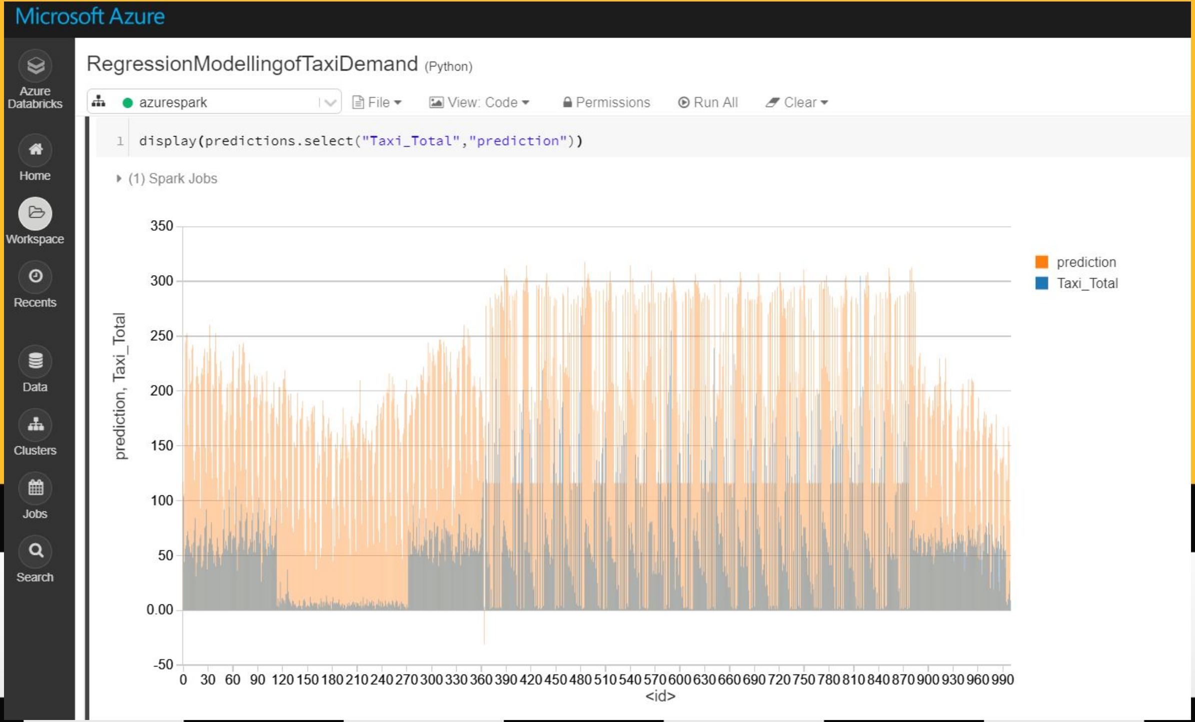
Dependent Variable

- Number of Vehicles Required in Location L at time T

Linear Regression Predictions Vs Actual

RMSE: 2745.040

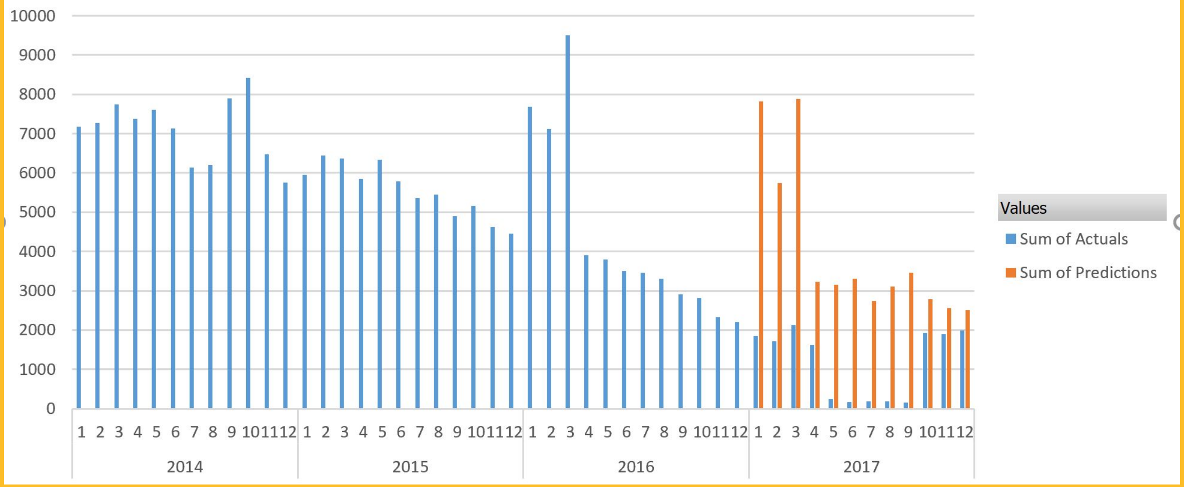
MAE: 1033.385

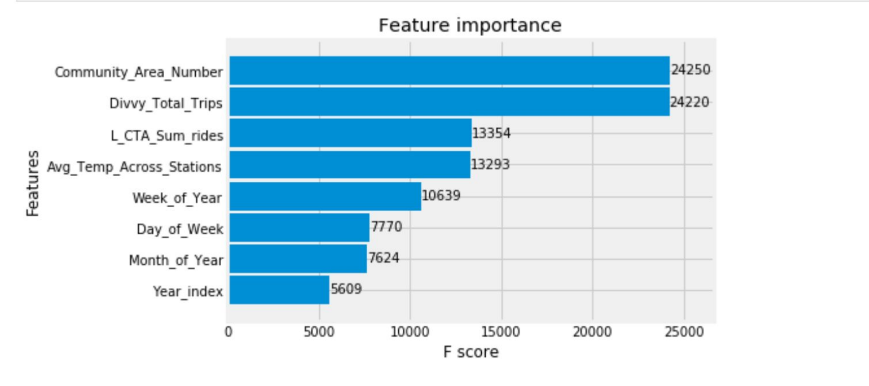


```
display(predictions.select("Taxi_Total","prediction"))
```

XGBoost Predictions Vs Actual

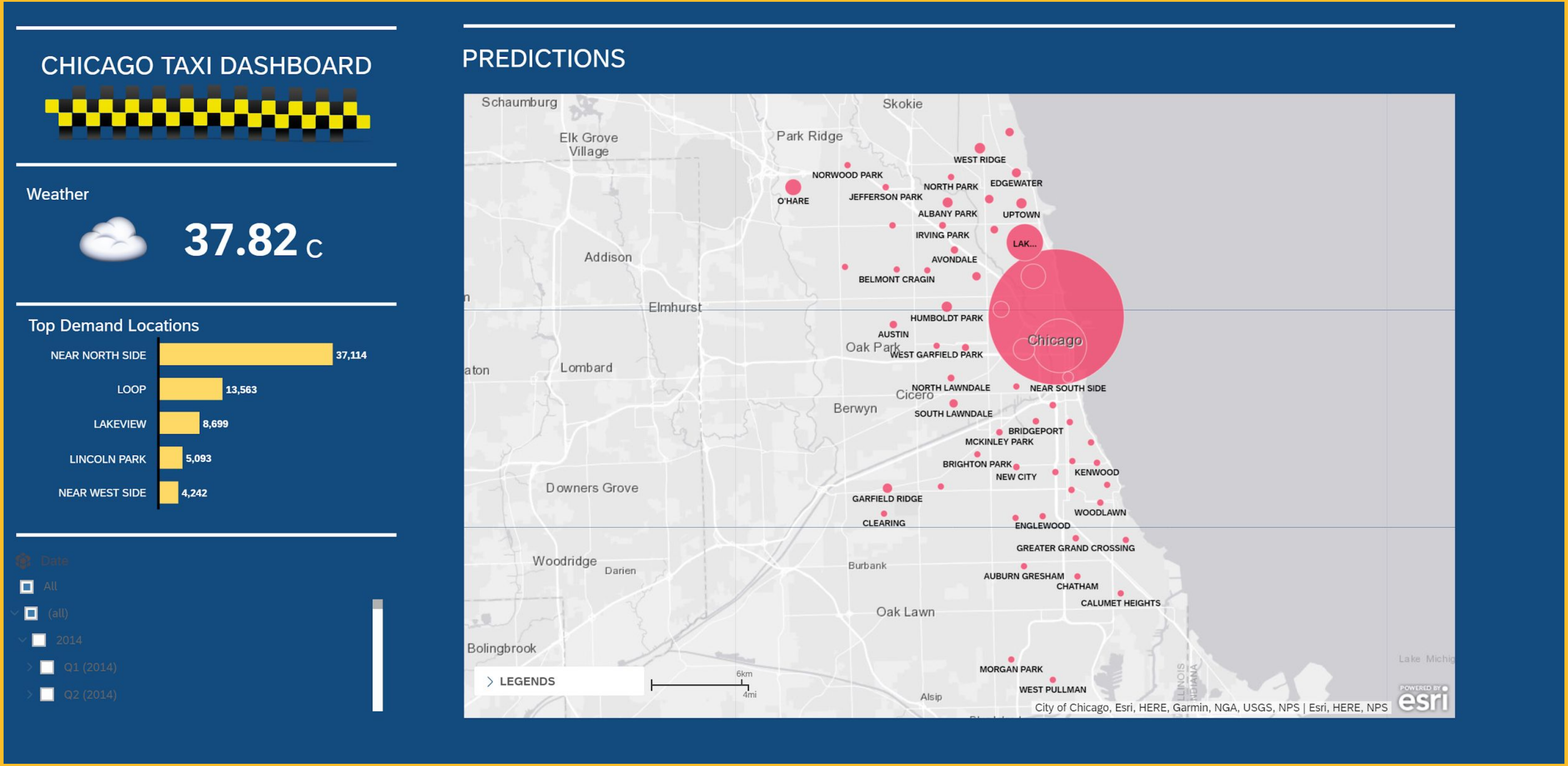
- XGBoost Regressor
  - Trained: 2014-2016
  - Test/Holdout: 2017
- Grid Search
  - N Estimators: 350
- Negative Predictions converted to 0





Data Visualization

Dashboard Display



CHICAGO TAXI DASHBOARD

Weather

37.82 C

Top Demand Locations

NEAR NORTH SIDE	37,114
LOOP	13,563
LAKEVIEW	8,699
LINCOLN PARK	5,093
NEAR WEST SIDE	4,242

Date

□ All

□ (all)

□ 2014

□ Q1 (2014)

□ Q2 (2014)

PREDICTIONS

Future Work

- Merge more sources of data to accurately capture demand and have a holistic view of the entire spectrum.
- Predict independent variables in addition
- Creating pipelines using applications like “Airflow”.
- More studies with respect to the pricing factor to have a comparison between Taxi business vs Network transportation companies like Uber and Lyft.

Conclusion

- The newly created Dashboard should definitely help the Taxi services to optimize their opportunities in Chicago.
- Exploratory analysis performed in this study highlights the year where there is dip in the taxi trips.
- Exploratory analysis also highlights the top pick-up and drops areas which can help the taxi owners to replicate the best practices in the less demand areas.
- Big Data capabilities were essential to creating an actionable dashboard

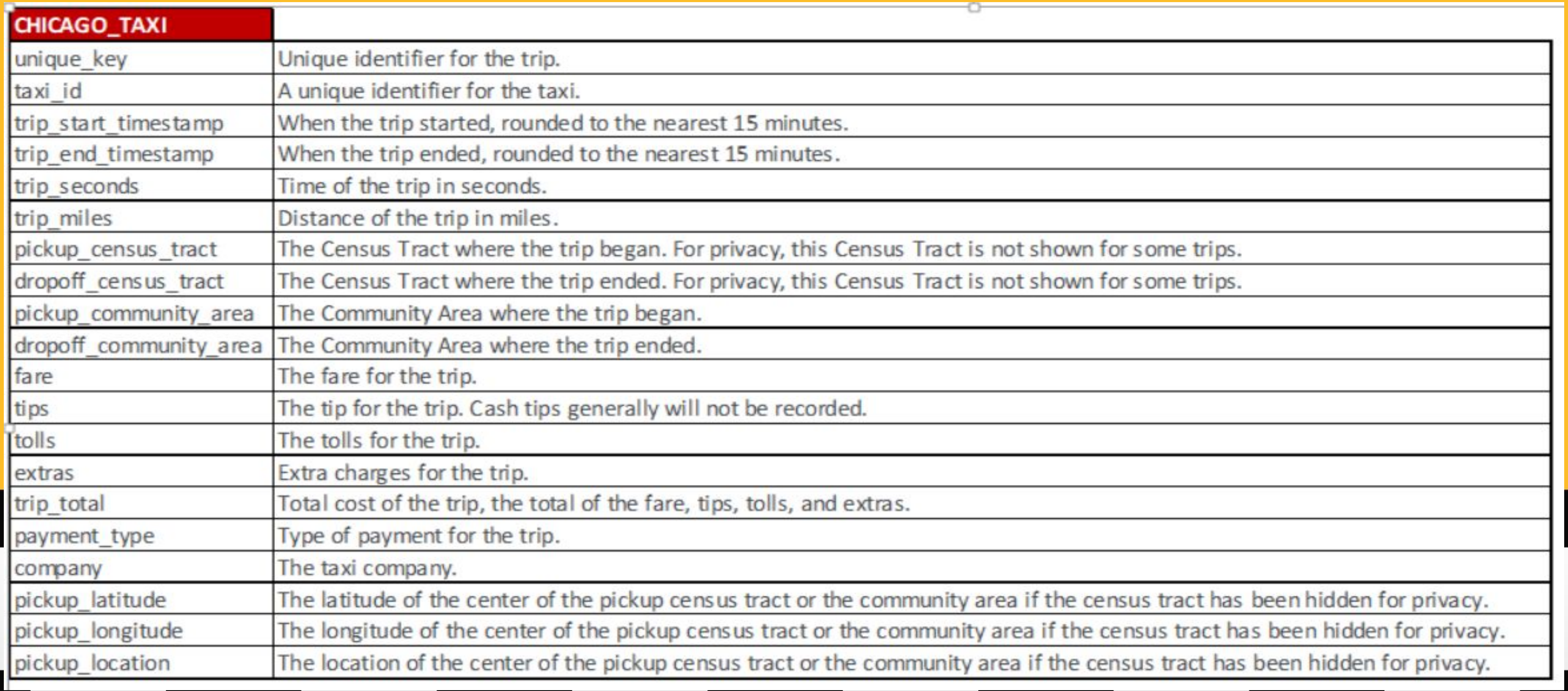
Thank you

Resources

- https://github.com/jkgiesler/parse-chicago-neighborhoods

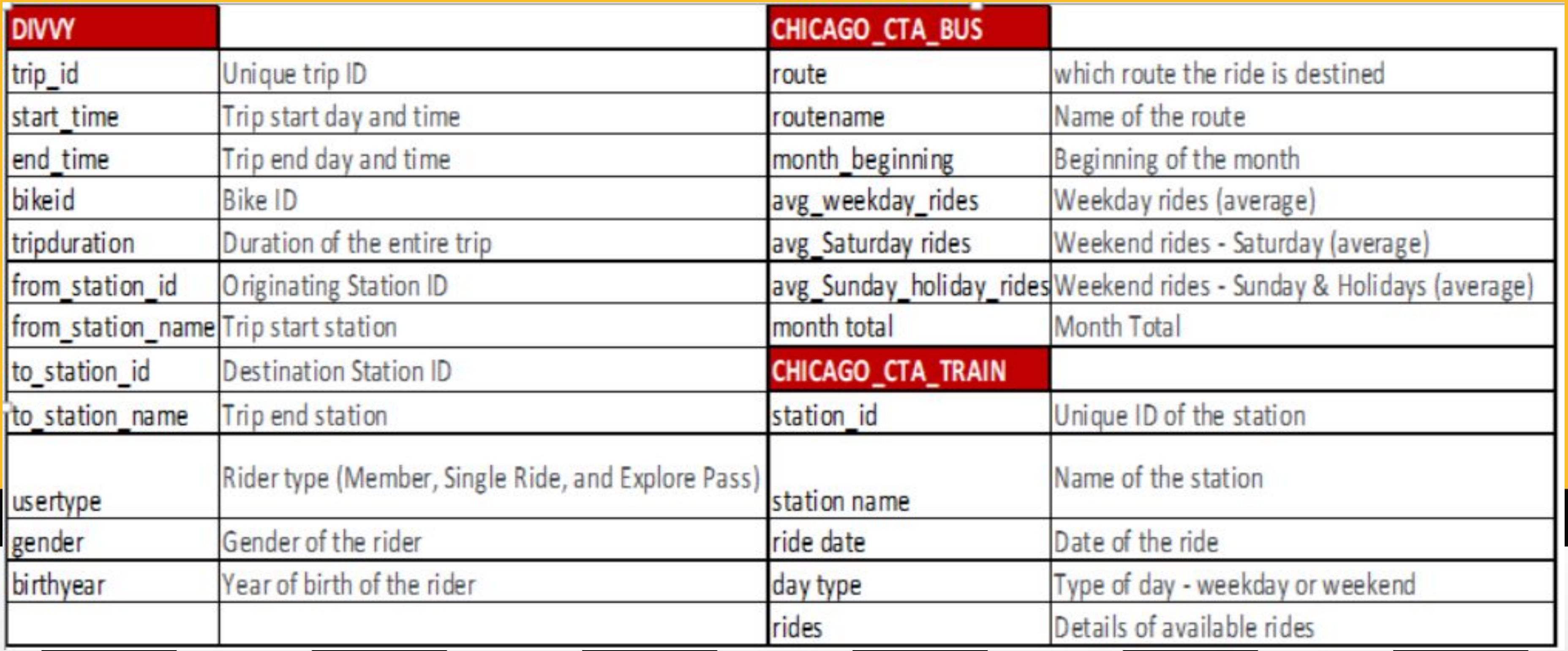
Appendices

Appendix I: Data Definition



CHICAGO_TAXI	
unique_key	Unique identifier for the trip.
taxi_id	A unique identifier for the taxi.
trip_start_timestamp	When the trip started, rounded to the nearest 15 minutes.
trip_end_timestamp	When the trip ended, rounded to the nearest 15 minutes.
trip_seconds	Time of the trip in seconds.
trip_miles	Distance of the trip in miles.
pickup_census_tract	The Census Tract where the trip began. For privacy, this Census Tract is not shown for some trips.
dropoff_census_tract	The Census Tract where the trip ended. For privacy, this Census Tract is not shown for some trips.
pickup_community_area	The Community Area where the trip began.
dropoff_community_area	The Community Area where the trip ended.
fare	The fare for the trip.
tips	The tip for the trip. Cash tips generally will not be recorded.
tolls	The tolls for the trip.
extras	Extra charges for the trip.
trip_total	Total cost of the trip, the total of the fare, tips, tolls, and extras.
payment_type	Type of payment for the trip.
company	The taxi company.
pickup_latitude	The latitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.
pickup_longitude	The longitude of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.
pickup_location	The location of the center of the pickup census tract or the community area if the census tract has been hidden for privacy.

Appendix II: Data Definition



DIVVY		CHICAGO_CTA_BUS	
trip_id	Unique trip ID	route	which route the ride is destined
start_time	Trip start day and time	routename	Name of the route
end_time	Trip end day and time	month_beginning	Beginning of the month
bikeid	Bike ID	avg_weekday_rides	Weekday rides (average)
tripduration	Duration of the entire trip	avg_Saturday rides	Weekend rides - Saturday (average)
from_station_id	Originating Station ID	avg_Sunday_holiday_rides	Weekend rides - Sunday & Holidays (average)
from_station_name	Trip start station	month total	Month Total
to_station_id	Destination Station ID	CHICAGO_CTA_TRAIN	
to_station_name	Trip end station	station_id	Unique ID of the station
usertype	Rider type (Member, Single Ride, and Explore Pass)	station name	Name of the station
gender	Gender of the rider	ride date	Date of the ride
birthyear	Year of birth of the rider	day type	Type of day - weekday or weekend
		rides	Details of available rides