# A Study on New York City Taxi Trips

Team #1: Yunhui Zhao, Chen Chen, Chengxiaoyuan Wang, Yuhao Xie, Sai Geng

05/03/22

# CONTENTS

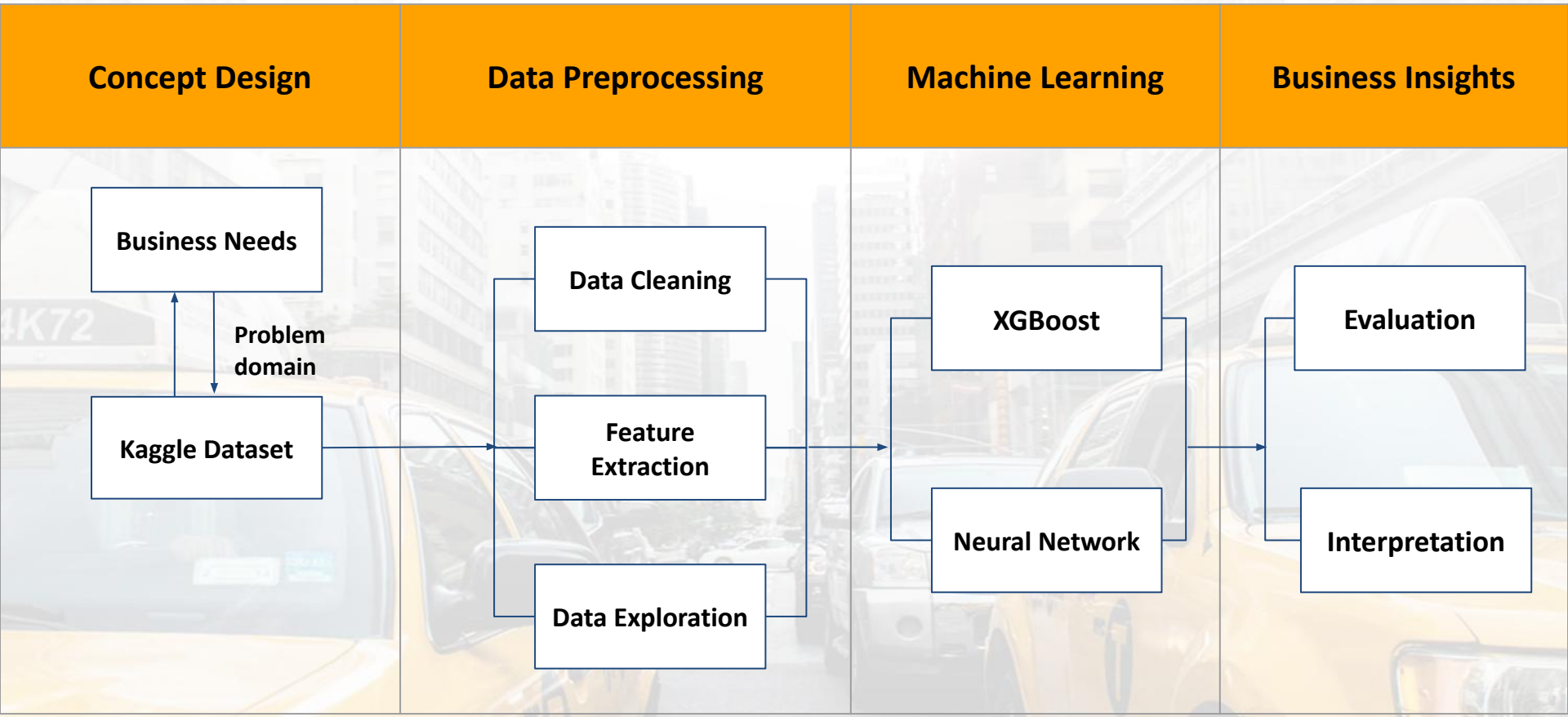
# Problem Statement

- Exploring whether certain pick-up areas in NYC may result in longer trip durations.

- Discovering correlation between the number of passengers and trip duration.

- Deriving actionable insights for taxi companies and Uber/Lyft to allocate their drivers more efficiently and maximize profits.

# System Design - Flow Chart



# Dataset

**Data source**: Kaggle

The data was originally published by the NYC Taxi and Limousine Commission

**Data format**: csv.

**Data Size**: 1,458,644 trip records for training

625,134 trip records for testing

**Data Description**:

- ID – A unique identifier for each trip
- Vendor ID – A code indicating the provider associated with the trip record
- Passenger count - number of passengers
- Pick-up/Drop-off datetime
- Pick-up/Drop-off longitude
- Pick-up/Drop-off latitude
- Trip duration – duration of trip in seconds

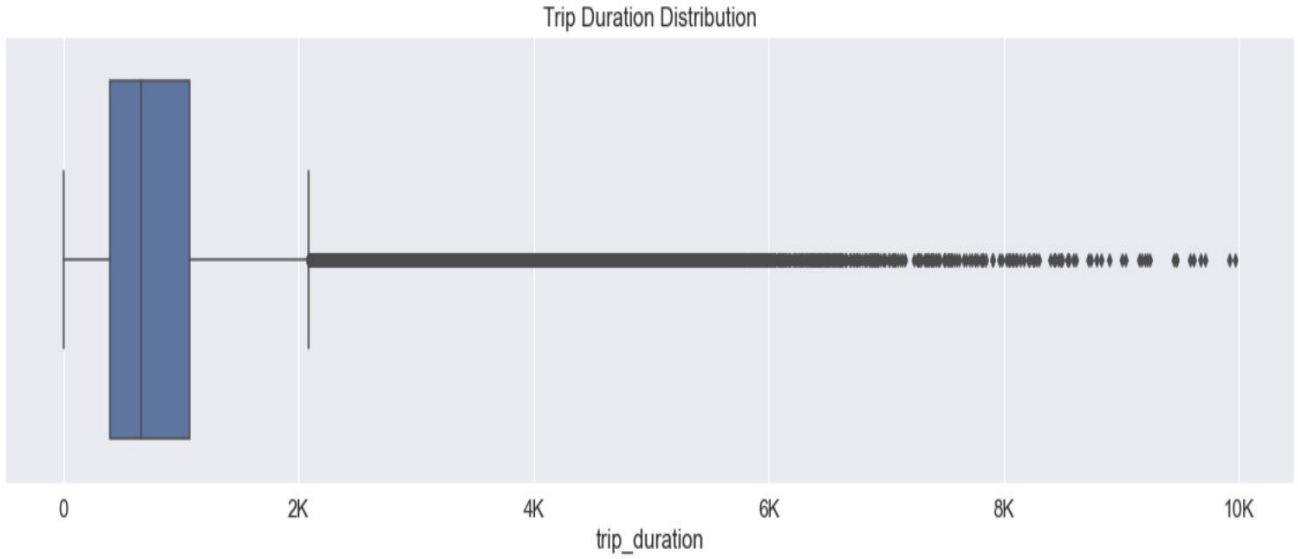
# Data Preprocessing

## Remove Outliers

Removed all the data has trip duration greater than 10000 seconds(2.7hrs)

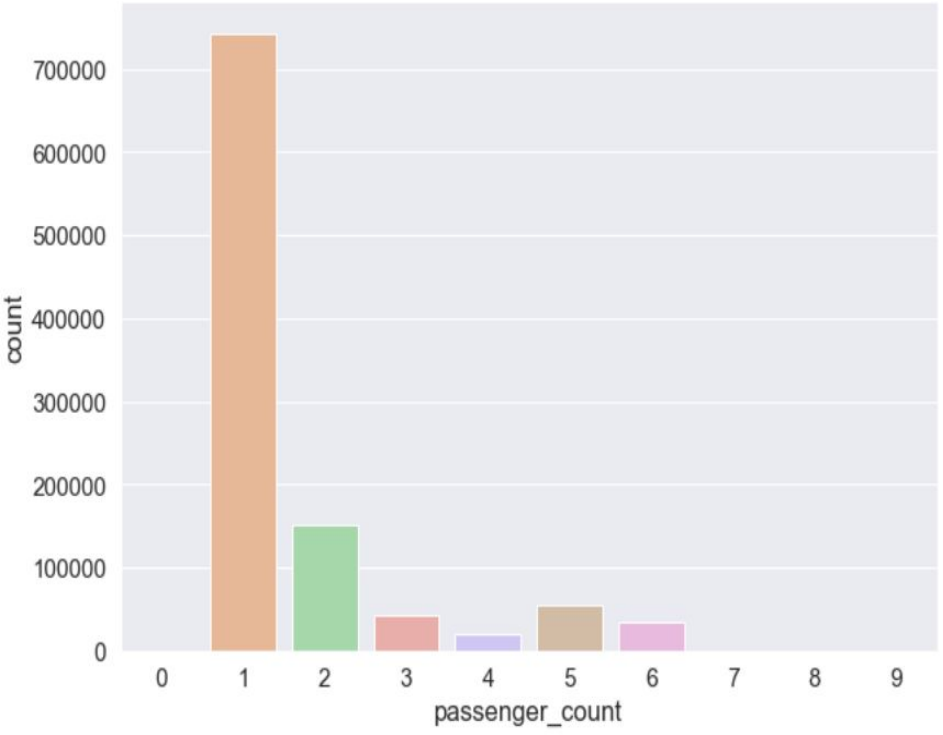
```
#Data cleaning-deal with outliers(trip_duration)-boxplot
#10000seconds = 2.7h
df.shape
print(nytaxi.shape)
df2 = df[(df.trip_duration <10000)]
print(df2.shape)
plt.figure(figsize=(20,5))
sns.boxplot(df2['trip_duration'])
plt.title('Trip Duration Distribution')
ax = plt.gca()
ax.xaxis.set_major_formatter(tick.FuncFormatter(reformat_large_tick_values))
```

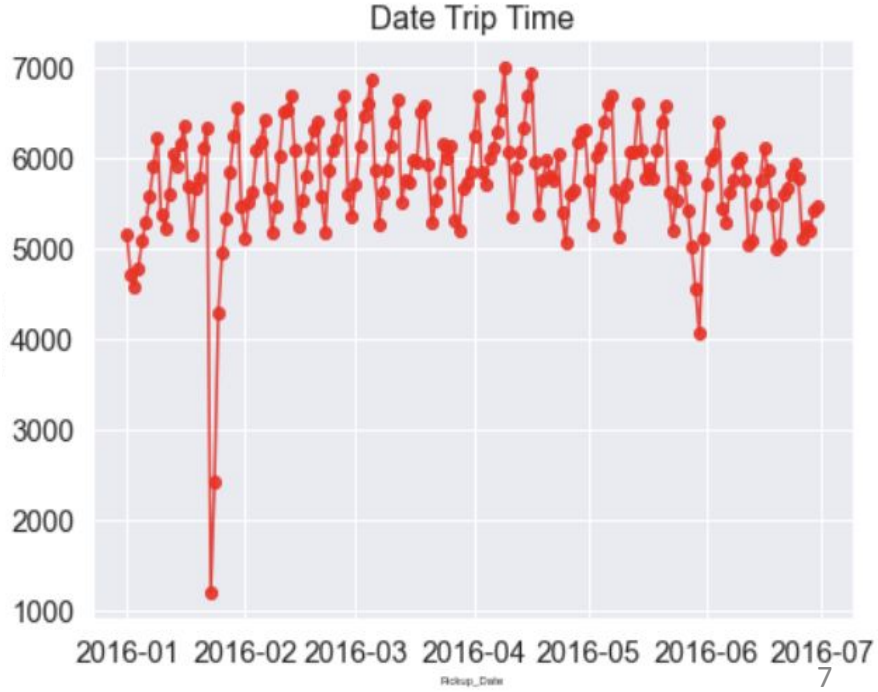
```
(1048575, 11)
(1047048, 11)
```



# Data Exploration

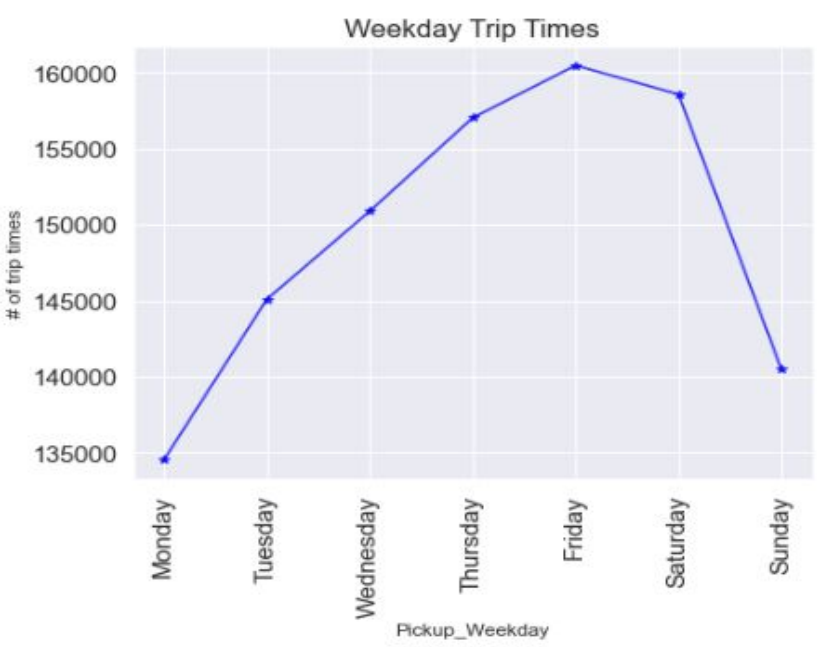
- Most trips have only one passenger.
- 4 - 6 passengers trip implies that cab must be a larger vehicle.
- There was an obvious drop of trips from Jan. 2016 to Feb. 2016 due to inclement weather (snowstorm) in NYC.

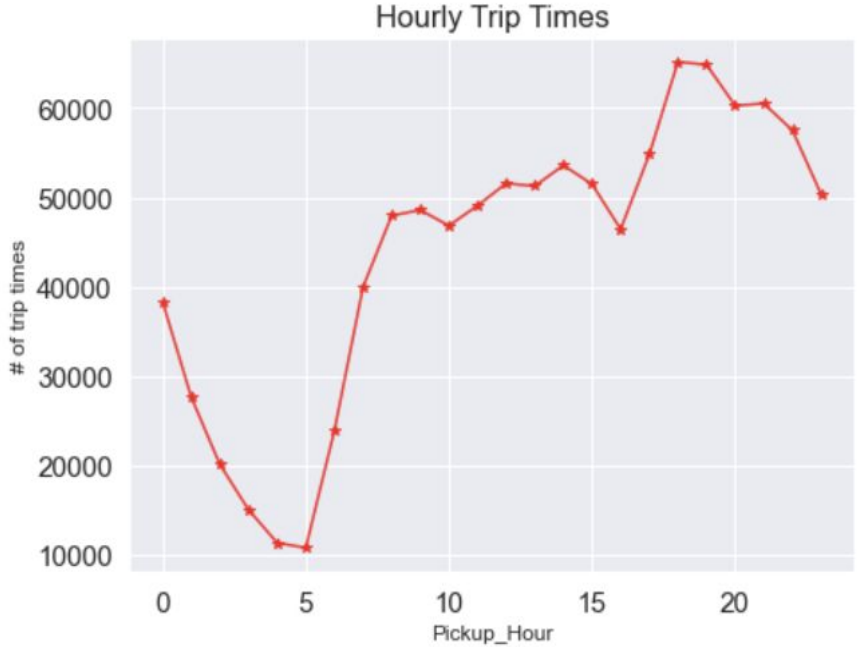


# Data Exploration

- Taxi drivers in NYC experience the highest demand on Fridays.
- Taxi demand peaks around 5 pm, which is the start of rush hours.



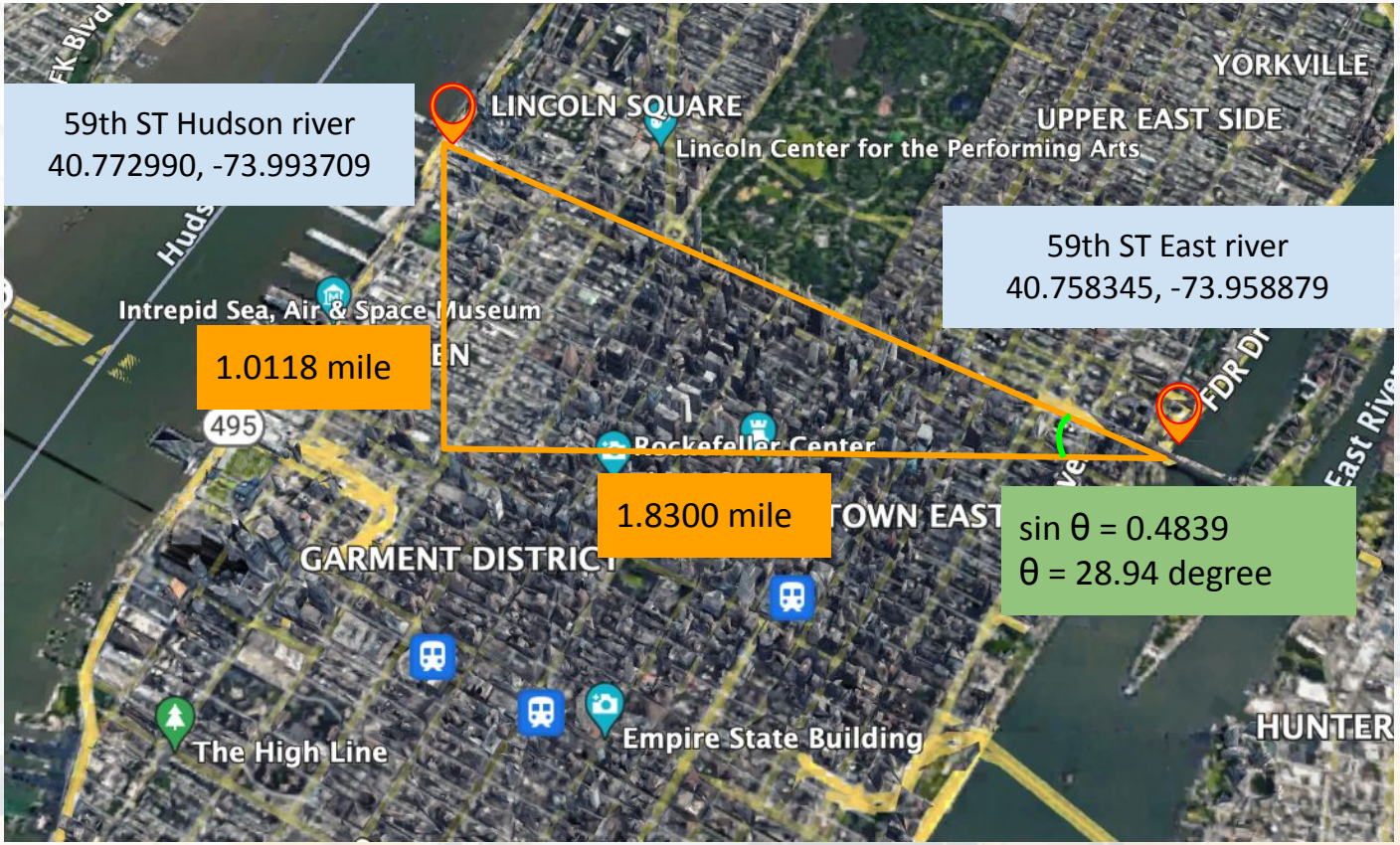
# Manhattan Distance



Manhattan Coordinates We Need

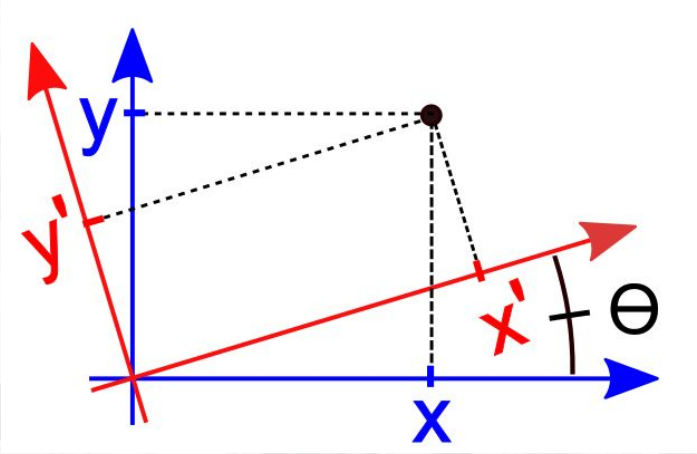
Coordinates We Have

# Latitude and Longitude System




1 Latitude & Longitude ⇨ 2 Difference in Miles ⇨ 3 Angle

# Manhattan Coordinate System




Rotation of Axes

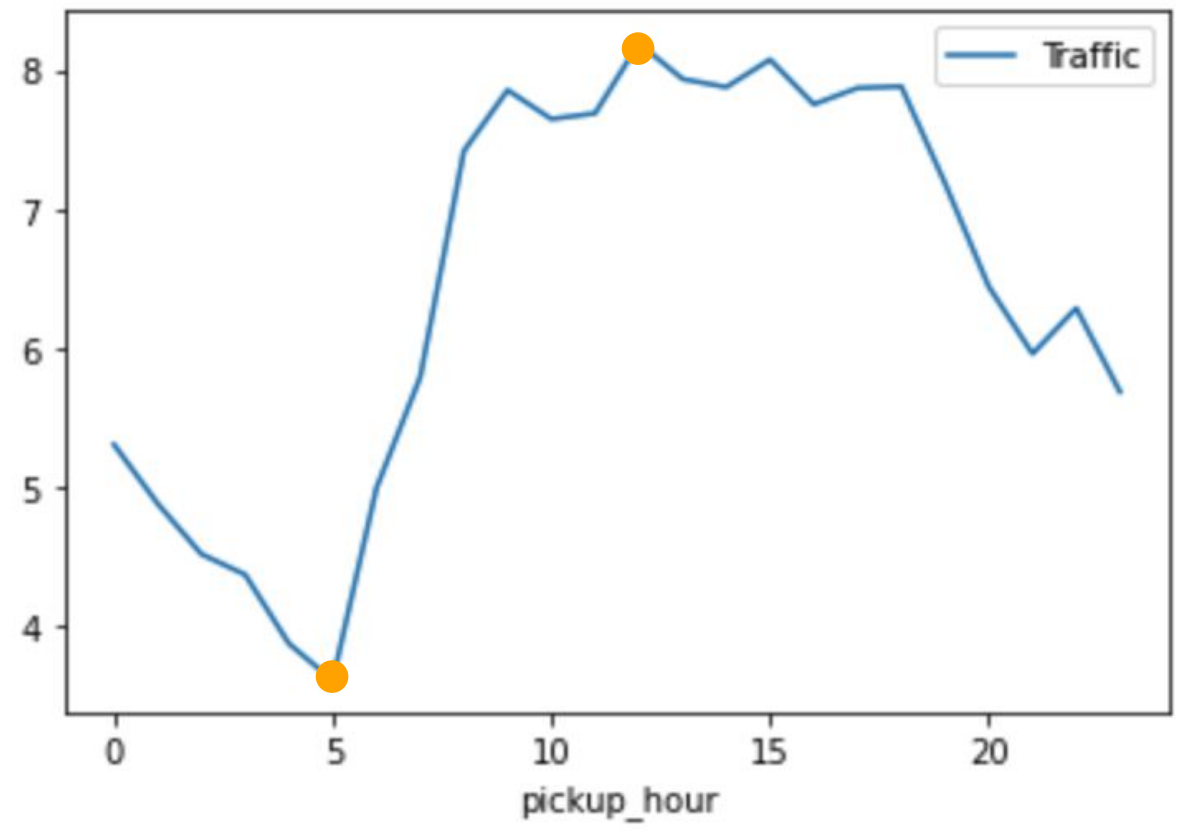
$x = \hat{x} \cos\theta - y\hat{} \sin\theta$

$y = \hat{x} \sin\theta + y\hat{} \cos\theta$

# Manhattan Neighborhood

| | Average Distance (miles) |
|---|---|
| Hell's Kitchen | 2.372860 |
| Time Square | 2.344308 |
| Chelsea | 2.171071 |
| Ktown | 2.143245 |
| Midtown East | 2.140285 |
| Murray Hill | 2.087279 |
| Lincoln Center | 2.026051 |
| Lenox Hill | 1.853281 |

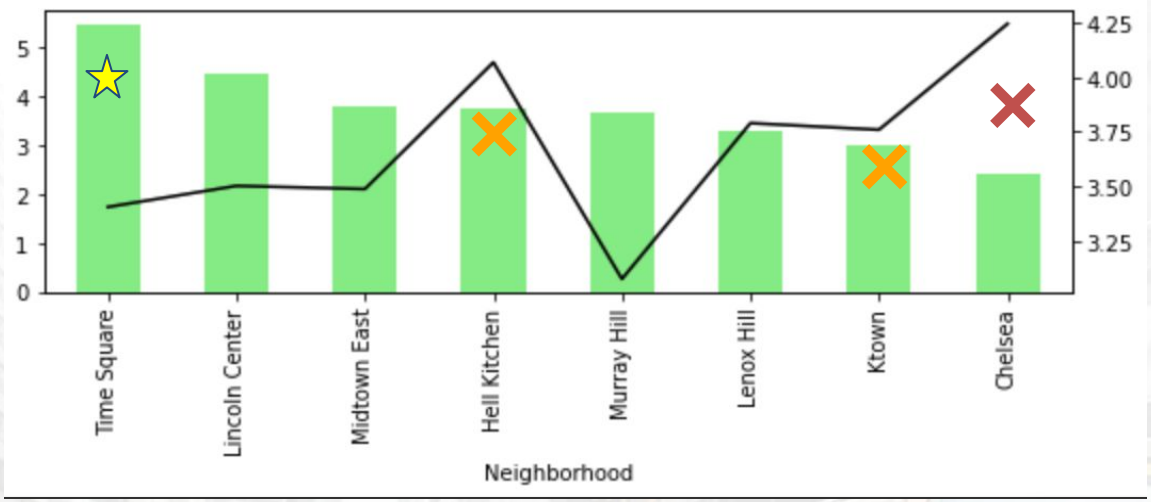
# Traffic




Traffic = Trip duration / Distance
(minutes per mile)

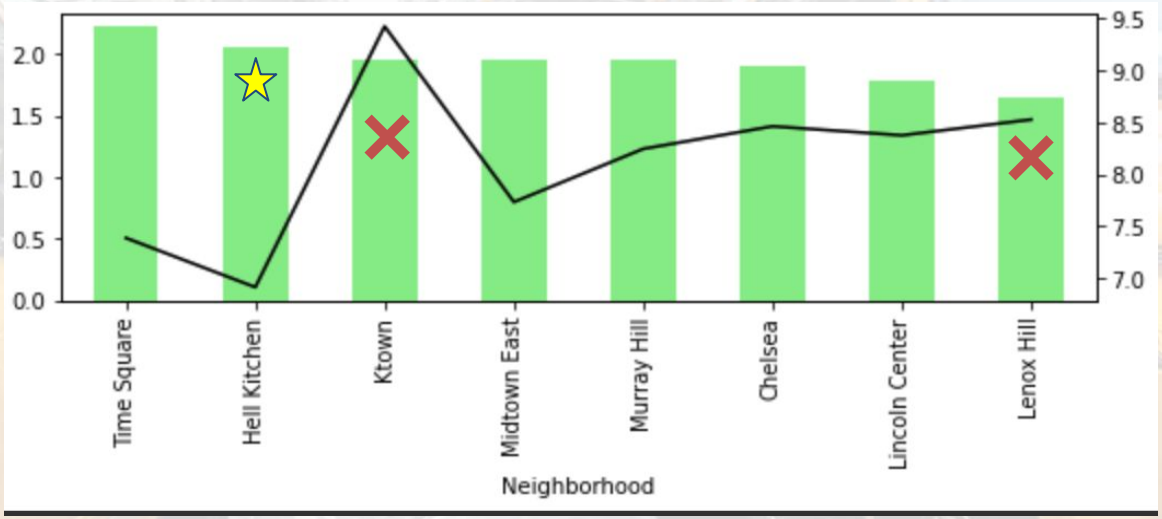
# Where Should Drivers Go @ 5am & 12pm

5am




Bar: Avg. Trip Distance Line: Avg. Traffic

12pm

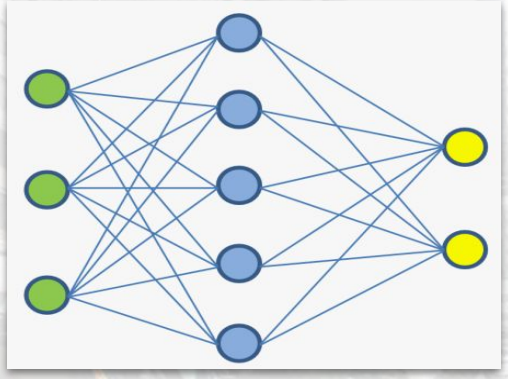


# Model Implementation

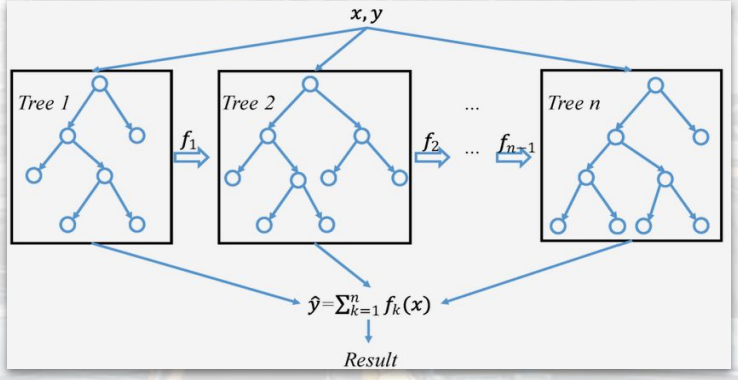
- ***Data Partition:***



- ***Model Selection:*** **Neural Network**



**XGBoost**




- ***Grid Searching/Parameter Tuning: 5-fold cross validation for each model***

```
[ ]  # make a dictionary of hyperparameter values to search.
     param_grid = {'hidden_layer_sizes': [(50,50,50), (50,100,50), (100,1)],
                   'activation': ['relu','tanh','logistic'],
                   'alpha': [0.0001, 0.05],
                   'learning_rate': ['constant','adaptive'],
                   'solver': ['adam']}

[ ]  # Cross validation on Neural Network.
     gridsearch = GridSearchCV(MLPRegressor(max_iter=2000, random_state=1).fit(X,y),
                               param_grid=param_grid,
                               scoring = ["r2", "neg_root_mean_squared_error"],
                               refit = "r2", cv=5, n_jobs=-1, verbose=4)
     gridsearch.fit(X,y)
     clf_NN = gridsearch.best_estimator_
```

```
[ ]  # make a dictionary of hyperparameter values to search.
     search_space = {
         "n_estimators" : [300, 400, 500],
         "max_depth" : [4, 6, 8],
         "gamma" : [0.01, 0.1],
         "learning_rate" : [0.01, 0.1]
     }

[ ]  #XGBoost Run 5-fold cross validation.
     GS = GridSearchCV(estimator = xgb,
                       param_grid = search_space,
                       scoring = ["r2", "neg_root_mean_squared_error"],
                       refit = "r2",
                       cv = 5,
                       verbose = 4)
     GS.fit(xtrain, ytrain)
```

# Model Evaluation

| Model | R-Squared | Efficiency |
|---|---|---|
| **Neural Network** | **0.0043** | **Low** |
| **XGBoost** | **0.7981** | **High** |

*We used the entire dataset (more than 1.4M rows) for XGBoost and 100,000 rows for Neural Network due to the extremely high training time.

# Conclusions

- Nearly 90% of the taxi trips in NYC are serving 1 or 2 passengers.
- Fridays and Saturdays are the busiest days for taxi drivers.
- Trips starting at Hell Kitchen and Time Square tend to have the longest trip duration.
- Picking up passengers at Times Square at 5 am and Hell's Kitchen at noon are the most profitable strategies for NYC taxi drivers.

# THANK YOU