Capstone Project: Ride BigApple 🚕 🍎

Pipistrello {analytics: LLC}

by **Luis Ulloa**
Chief Data Scientist & CEO

Problem Statement

- **Premise:** Can yellow cab fares be predicted within New York City's five boroughs based on time of the day, time of the year and certain high passenger areas?
    - Scenario: Dataset of rides with just the fare, coordinates, number of passengers and a time stamp.
    - Aim: Engineer new features and build a model that takes user inputs and predicts the ride’s fare and distance.

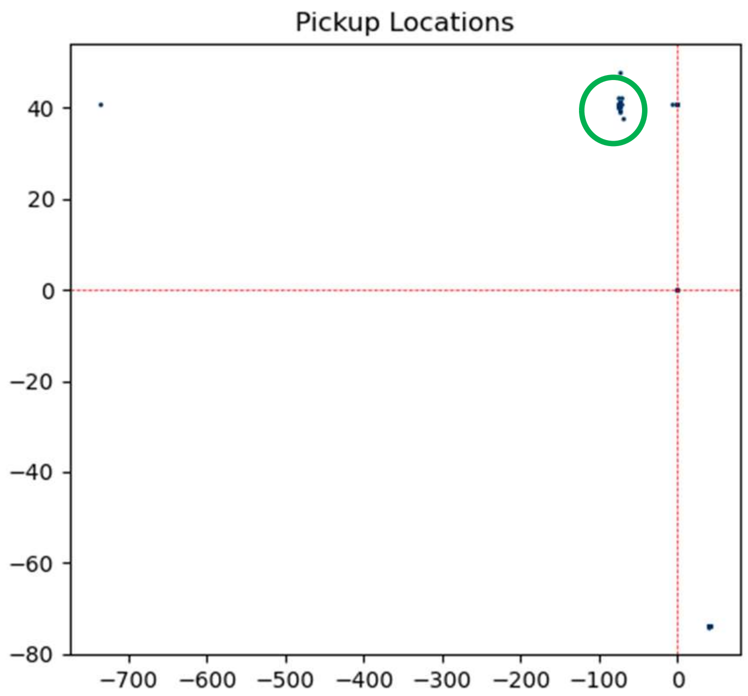
Dataset

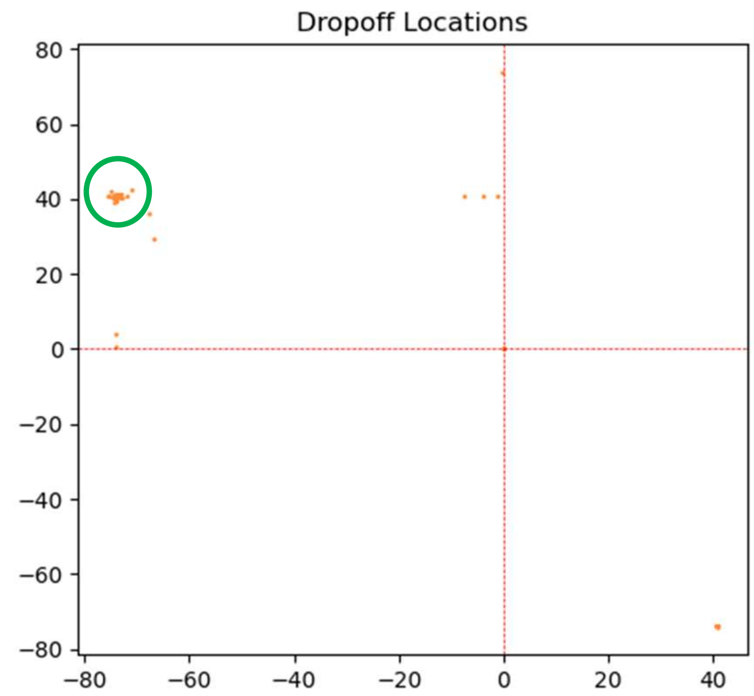
- 55.4 million rows 
- Sampled 60 thousand observations 

	key	fare_amount	pickup_datetime	pickup_longitude	pickup_latitude	dropoff_longitude	dropoff_latitude	passenger_count
0	2010-10-31 03:32:00.000000130	3.7	2010-10-31 03:32:00 UTC	-73.982163	40.762762	-73.987518	40.760543	1
1	2014-11-20 22:50:22.0000002	14.5	2014-11-20 22:50:22 UTC	-73.995560	40.759405	-73.968201	40.804051	1
2	2010-01-26 19:27:55.0000002	11.7	2010-01-26 19:27:55 UTC	-74.001313	40.736943	-73.994137	40.699002	1
3	2010-03-22 19:00:00.000000191	4.5	2010-03-22 19:00:00 UTC	-74.005897	40.770640	-74.008753	40.769735	1
4	2011-02-26 08:57:11.0000002	4.9	2011-02-26 08:57:11 UTC	-73.984917	40.749153	-74.000801	40.757591	1

Exploratory Data Analysis

- **Pickup and Dropoff Coordinates**



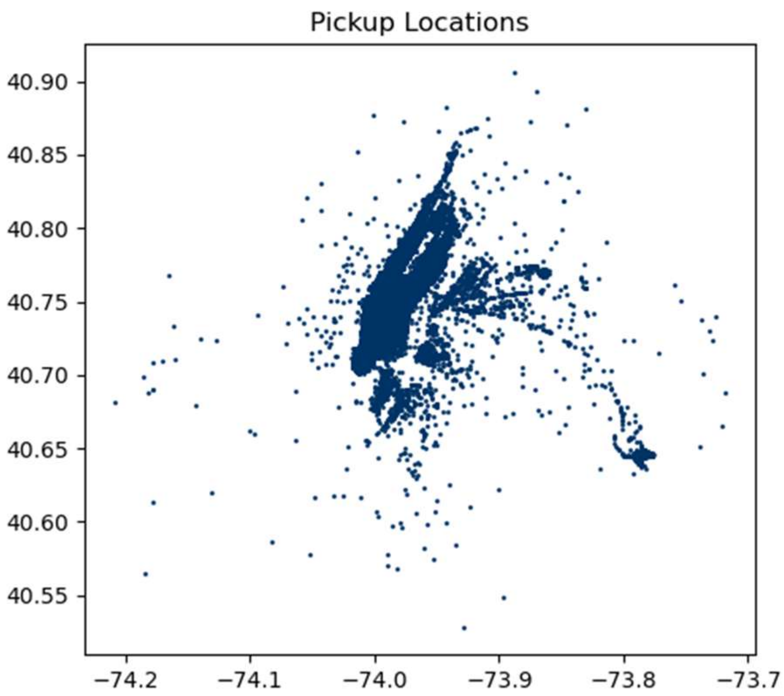


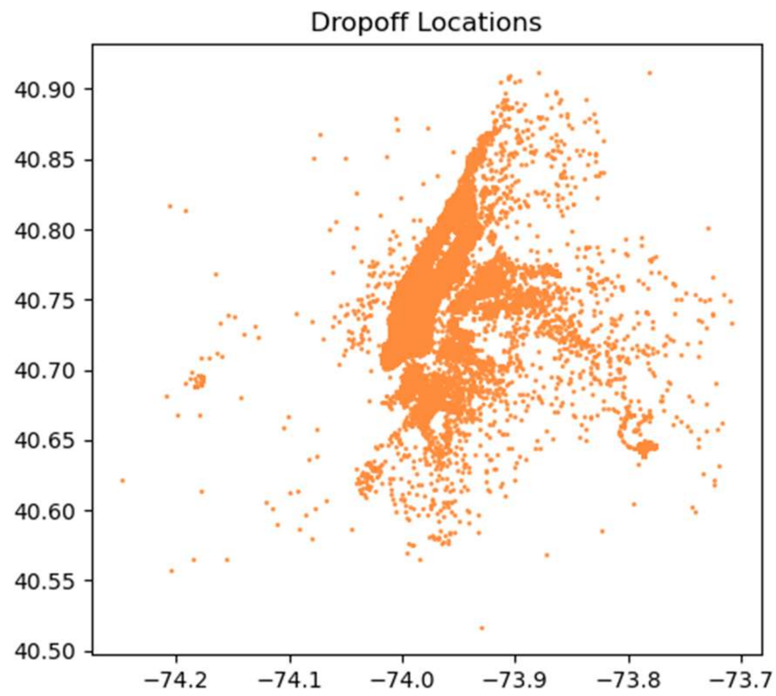
Exploratory Data Analysis

- **Pickup and Dropoff Coordinates**
  - Northernmost point: 40.915 degrees N latitude
  - Southernmost point: 40.496 degrees N latitude
  - Westernmost point: -74.256 degrees W longitude
  - Easternmost point: -73.702 degrees W longitude

Exploratory Data Analysis

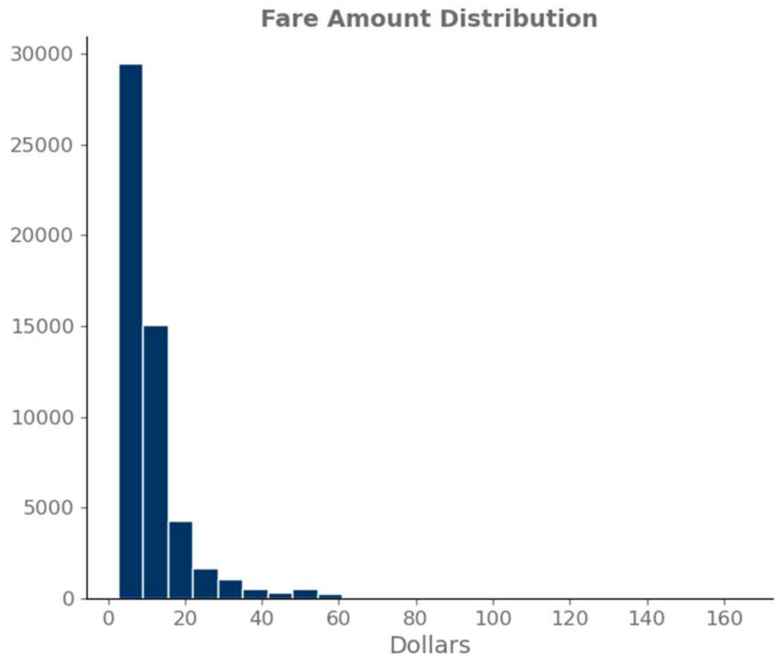
- **Pickup and Dropoff Coordinates**

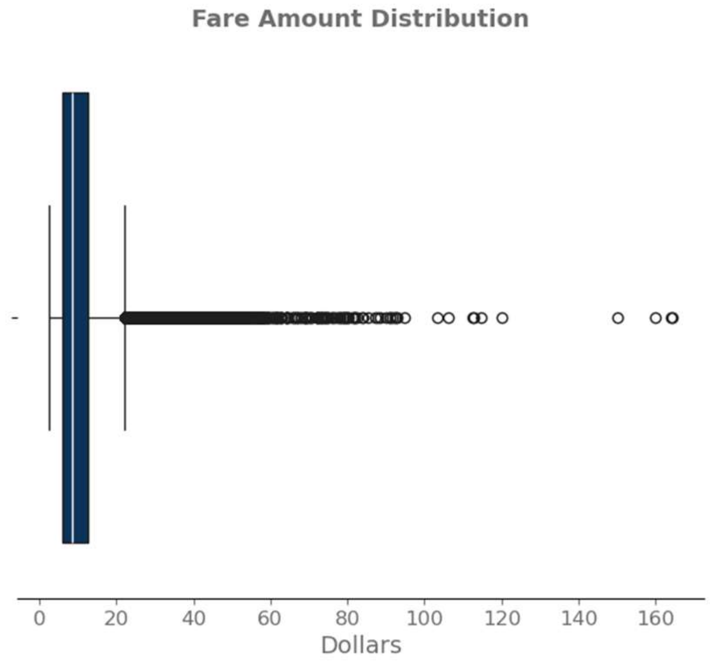




Exploratory Data Analysis

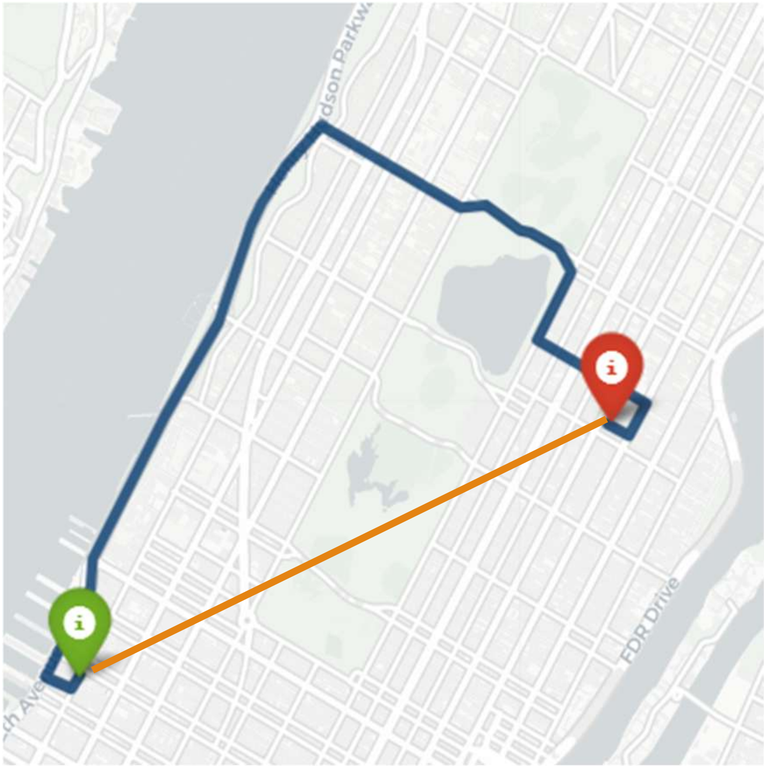
- **Target Variable: Fare Amount**
  - Right-skewed
  - Mean: 11.16 USD





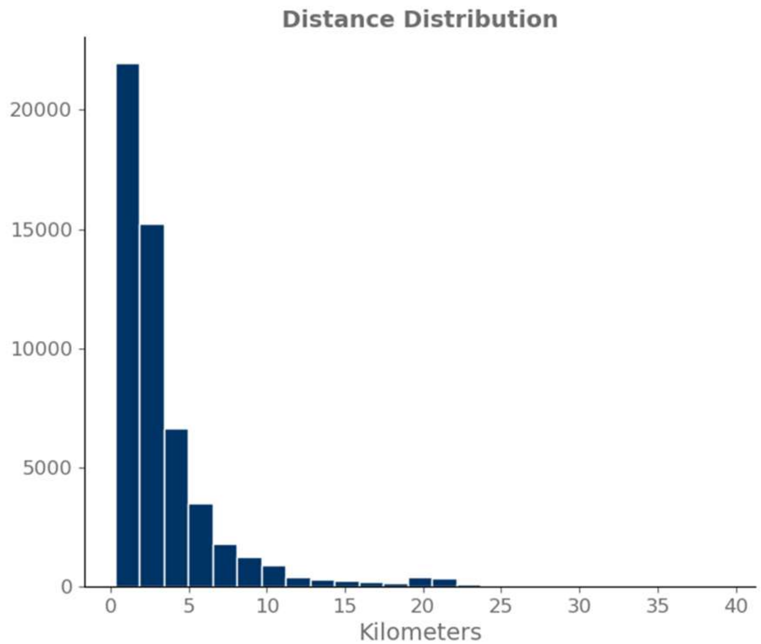
Exploratory Data Analysis

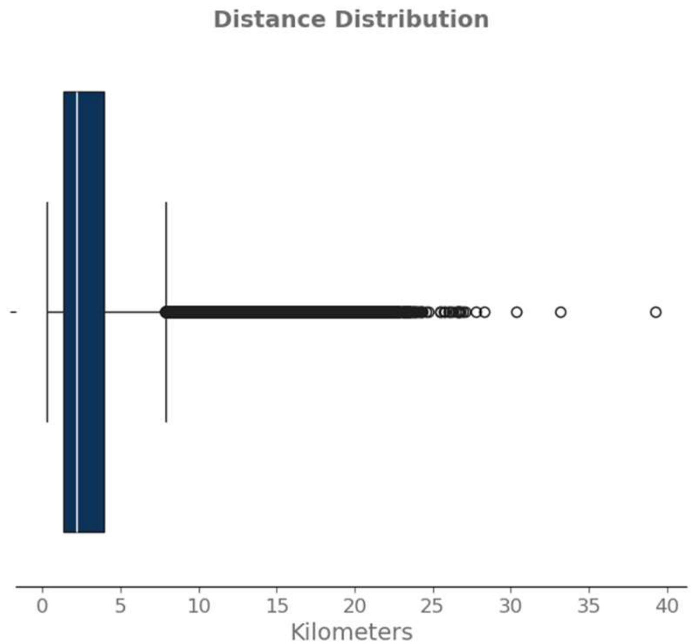
- **Dominant feature: Distance, 0.9 corr.**
  - 1st: Geodesic Distance (from coordinates)
    - Straight orange line
  - 2nd: Estimated Distance (after applying factors)
    - Aims to approximate blue line
    - 1.15 under 10 kms
    - 1.2 otherwise



Exploratory Data Analysis

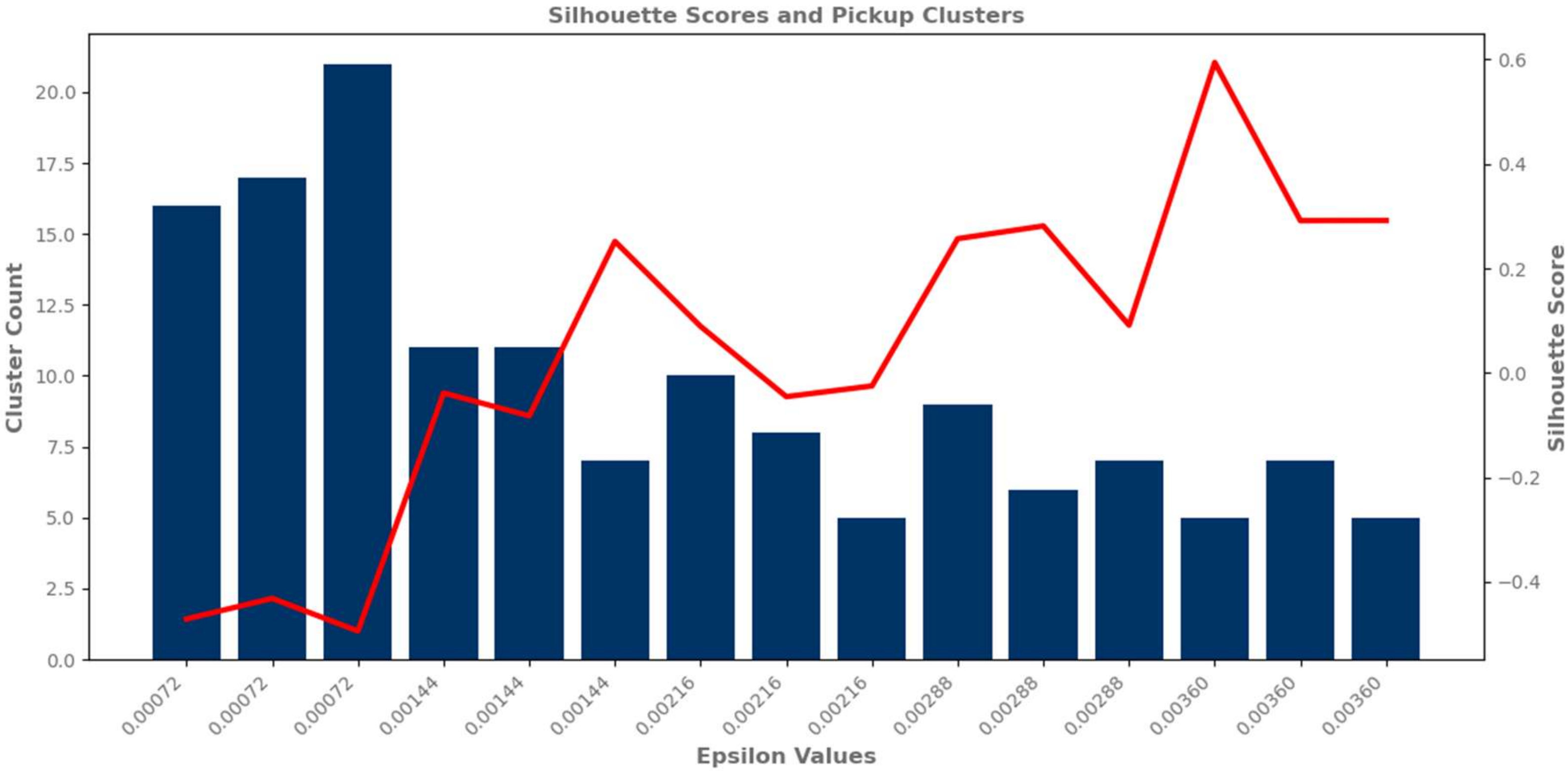
- **Dominant feature: Distance, 0.9 corr.**
  - Also, right-skewed
  - Mean: 3.37 kms





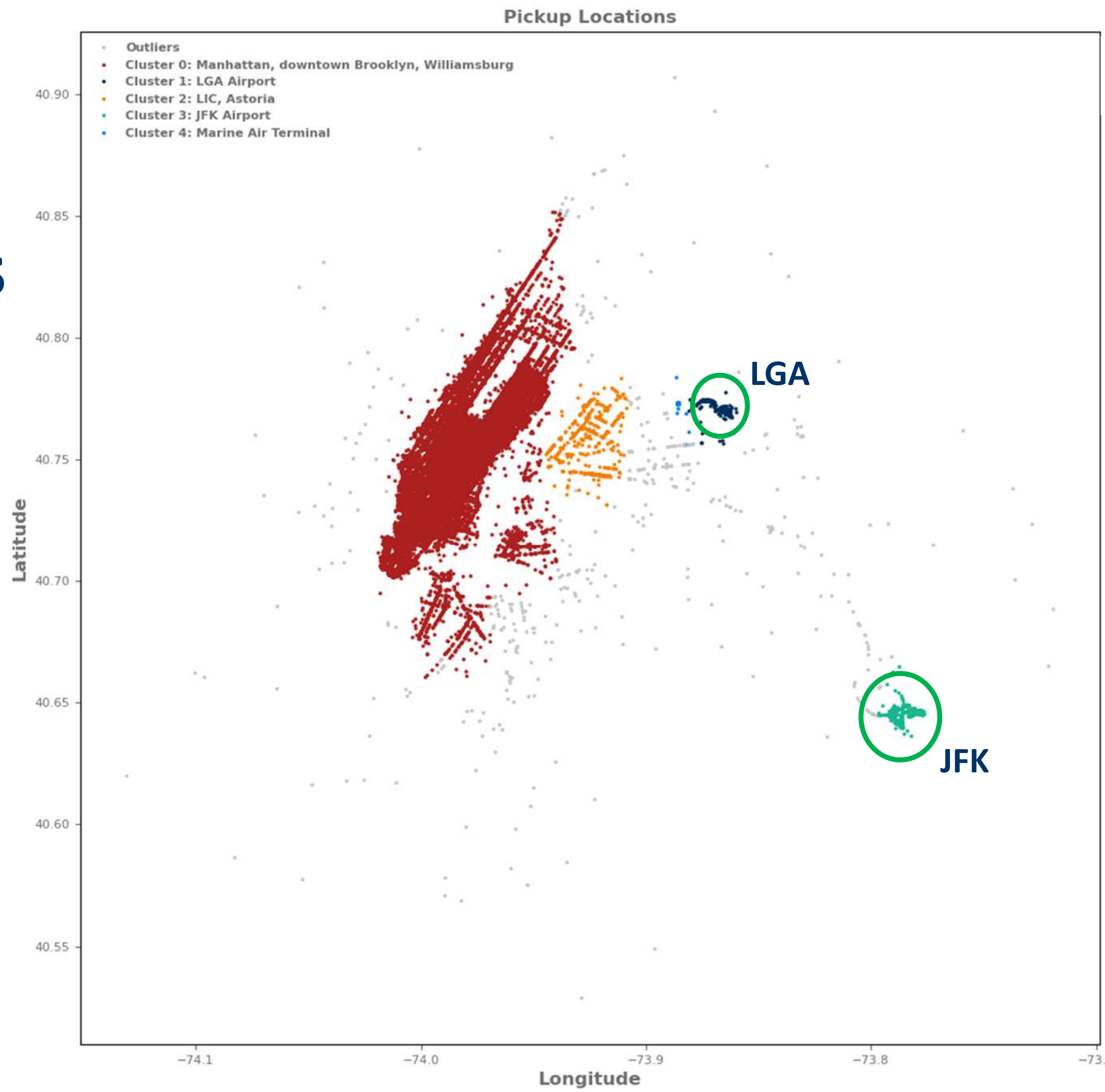
Modeling: Geospatial Clusters

- **Pickup Coordinates**



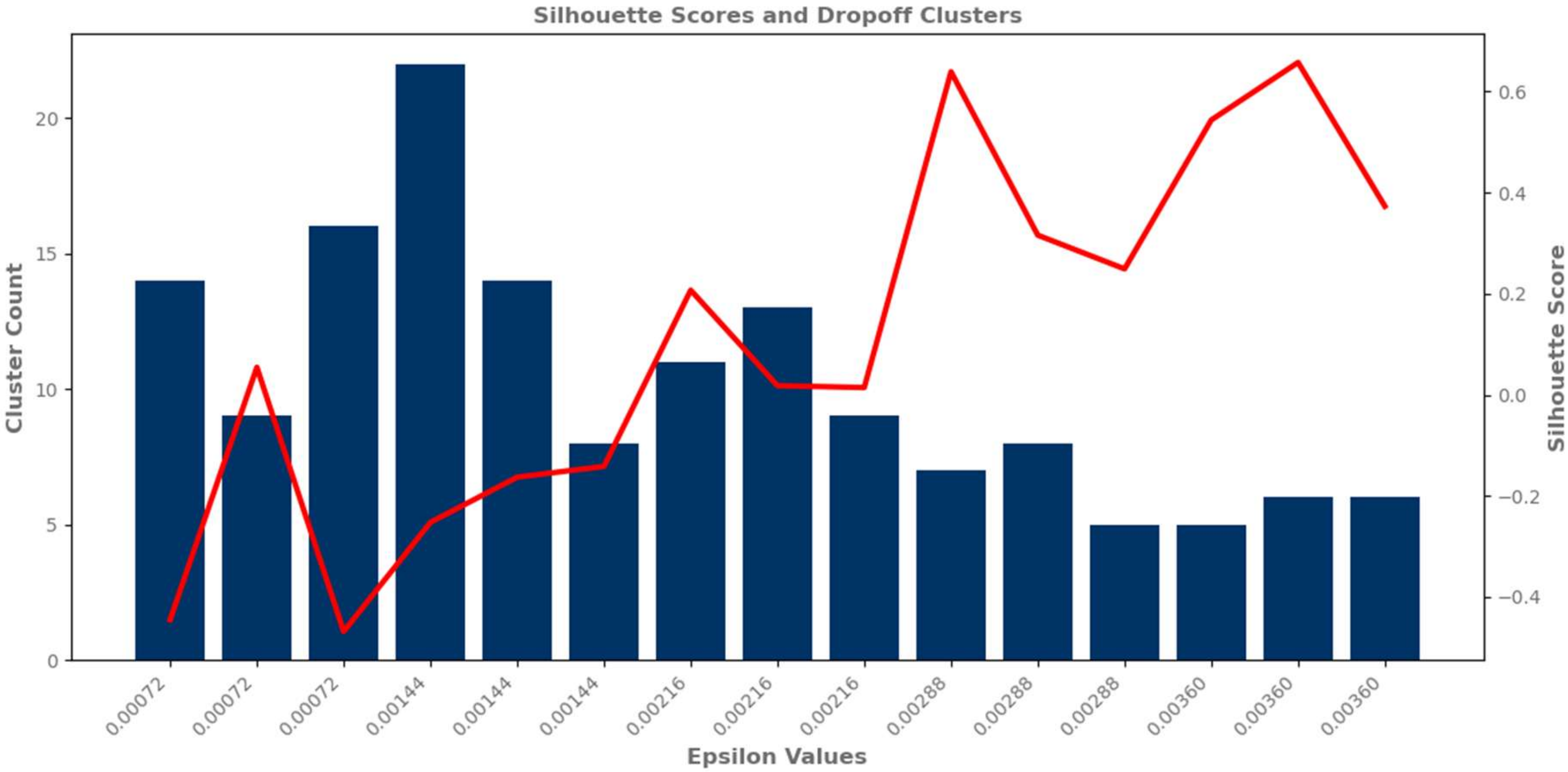
Modeling: Geospatial Clusters

- Pickup Clusters



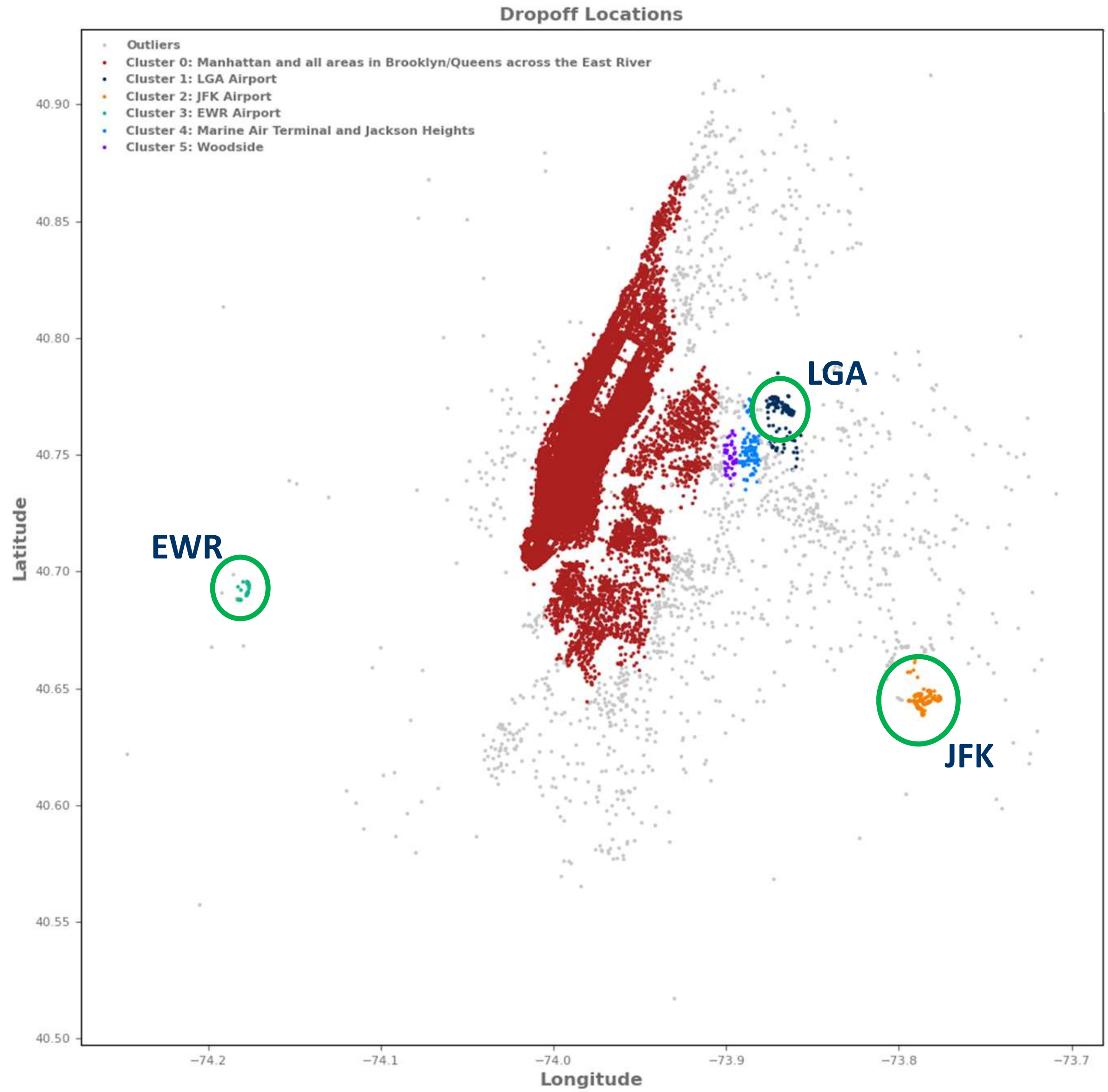
Modeling: Geospatial Clusters

- **Dropoff Coordinates**



Modeling: Geospatial Clusters

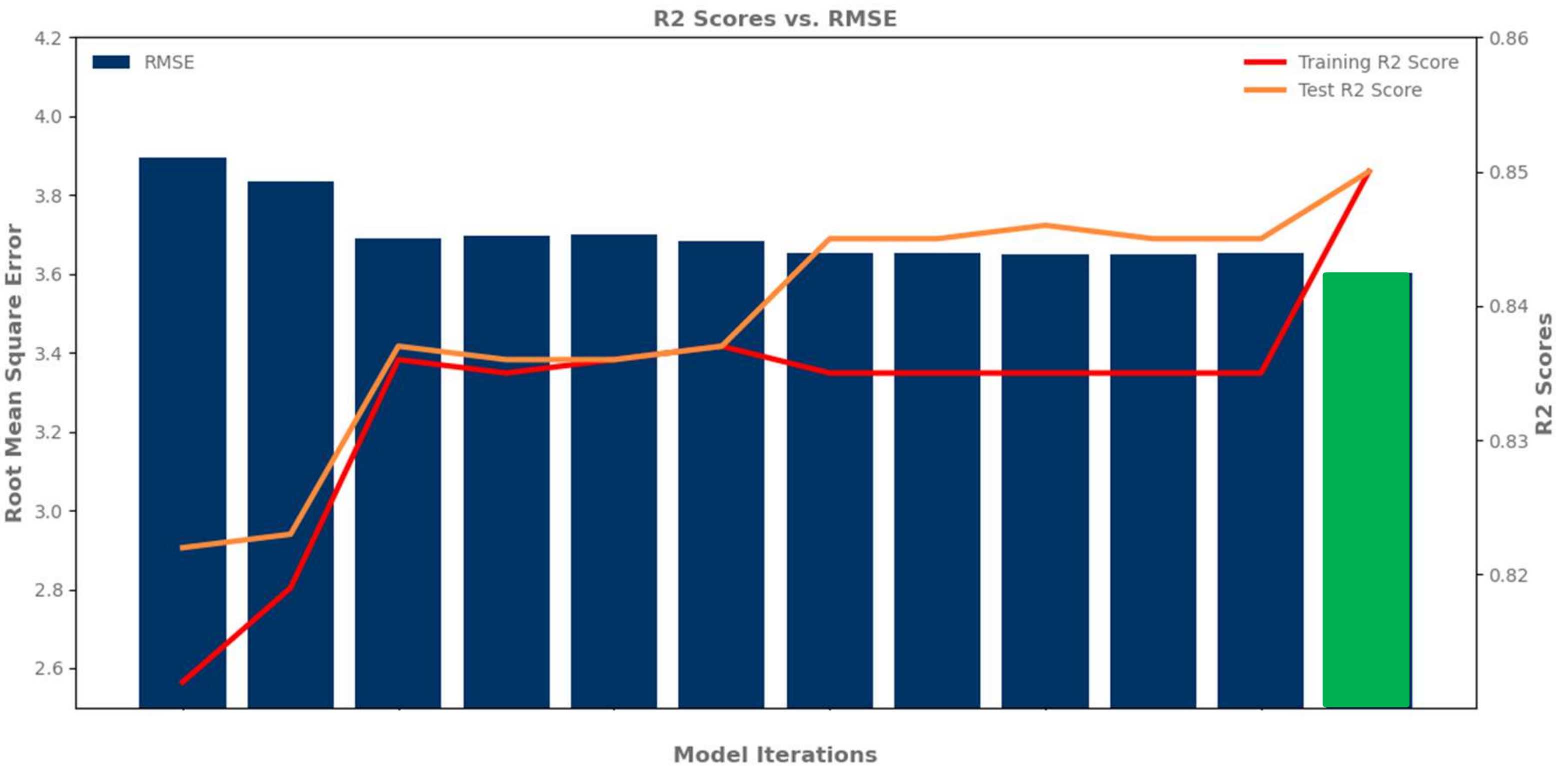
- Dropoff Clusters



Modeling: XGBRegressor

- **Engineered Features**
  - **p_0:** pickups, Manhattan+
  - **p_1:** pickups, LGA
  - **p_3:** pickups, JFK
  - **p_4:** pickups, Marine Air Terminal
  - **d_0:** dropoffs, Manhattan+
  - **d_1:** dropoffs, LGA
  - **d_2:** dropoffs, JFK
  - **d_3:** dropoffs, EWR
- **Engineered Features**
  - **estimated_distance:** from coordinates
  - **distance_hour:** distance/hour interaction
  - **JFK:** rides to/from JFK
  - **LGA:** rides to/from LGA
  - **weekend_rides:** rides on Sat. or Sun.
  - **holiday_rides:** rides in Nov. or Dec.

Modeling: XGBRegressor



MODEL ITERATIONS

- **First nine**: Linear Regression
- **Next two**: Lasso
- **Last**: XGBRegressor
  - RMSE: 3.602
  - R2 Training: 0.85
  - R2 Test: 0.85

Next Steps: Project

- More EDA,
  - Explore identified relationships at deeper level (better granularity)
  - Discover new relationships
  - Tweak engineered features

- Aim for more, smaller, localized clusters, specially in Manhattan

- Upgrade Streamlit App
  - Accept landmark names instead of addresses
  - Yankee Stadium, Columbus Circle, etc.

Next Steps: Myself

- **Bootcamp got me to the tip of the iceberg**

- **Future Focus**
  - Feature Engineering
    - *Python Feature Engineering Cookbook*
      by Soledad Galli
  - Visualizations
    - *Storytelling with Data*
      by Cole Nussbaumer Knaflic

Questions?