PREDICTING TAXI ETA TIMES IN NYC

Final Project for APMA 4990: Introduction to Data Science Industry

Joseph Archer and John Vahedi

Preview

- Data Integrity
- Feature Creation
- Exploratory Analysis
- Model Selection
- Results and model comparison

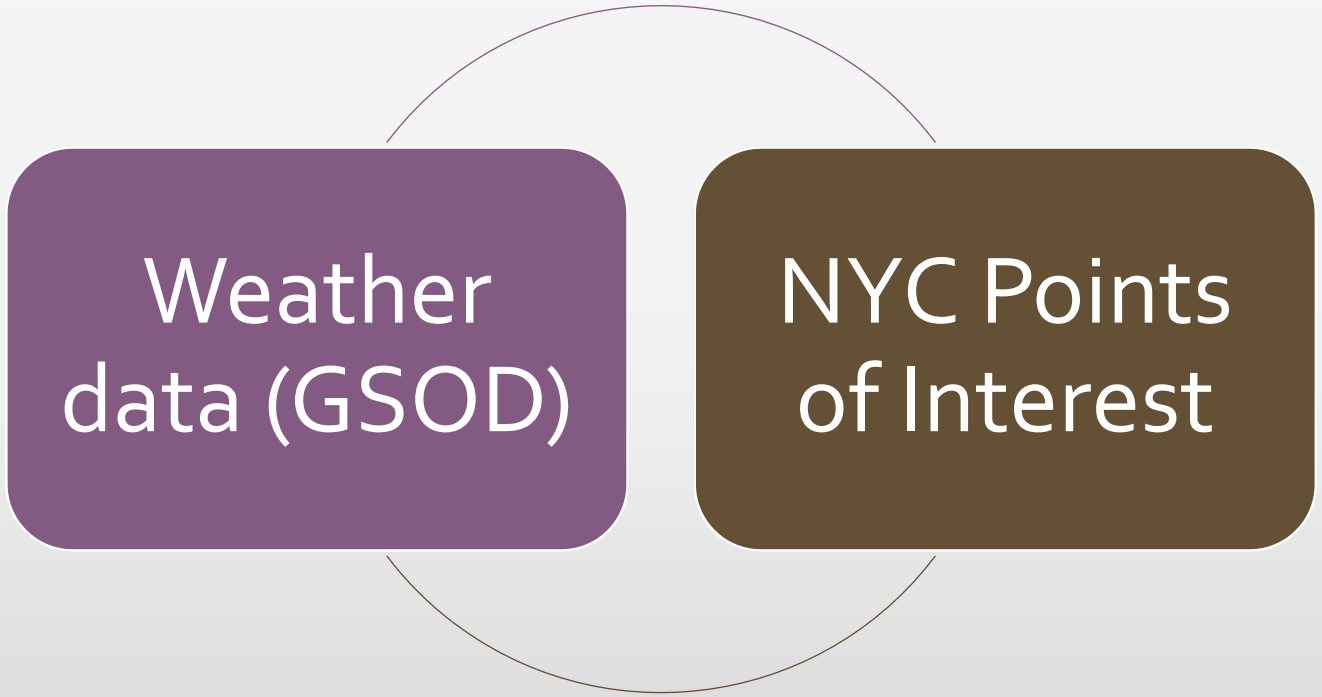
DATA INTEGRITY

Filters

- Eliminate nulls/na
- Latitude and longitude
  - latitude between 40.5 and 41
  - longitude between -74.3 and -73.5
- Speed
  - filtered between 2 mph and 70 mph
  - not used as a feature due to data leakage

FEATURE CREATION

Data Gathering



Geospatial

Points of interest

River latitude and longitude

L1 and L2 norm

Geohash

Time

Time cos and sin

Weekend

Rush hour

Categorical time

Unique

Rushed distance

Higher order latitude and longitude

Google maps distance

Feature Creation

Geospatial

- Points of interest
- River latitude and longitude
- L1 and L2 norm
- Geohash

Feature Creation

Time

- Time cos and sin
- Weekend
- Rush hour
- Categorical time

Feature Creation

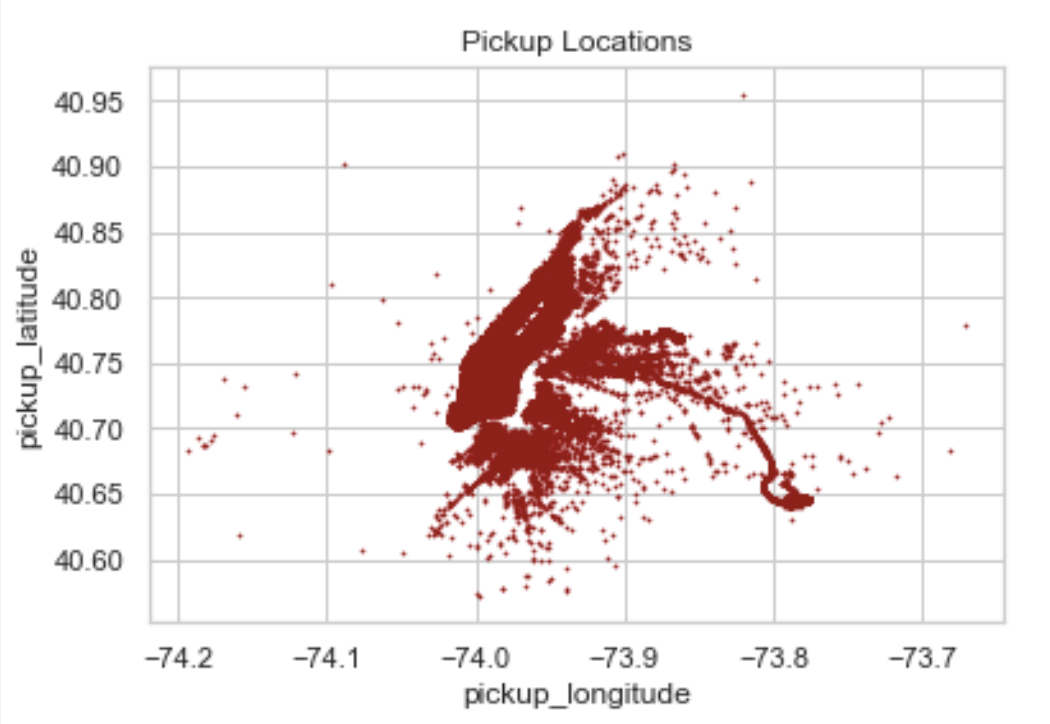
Unique

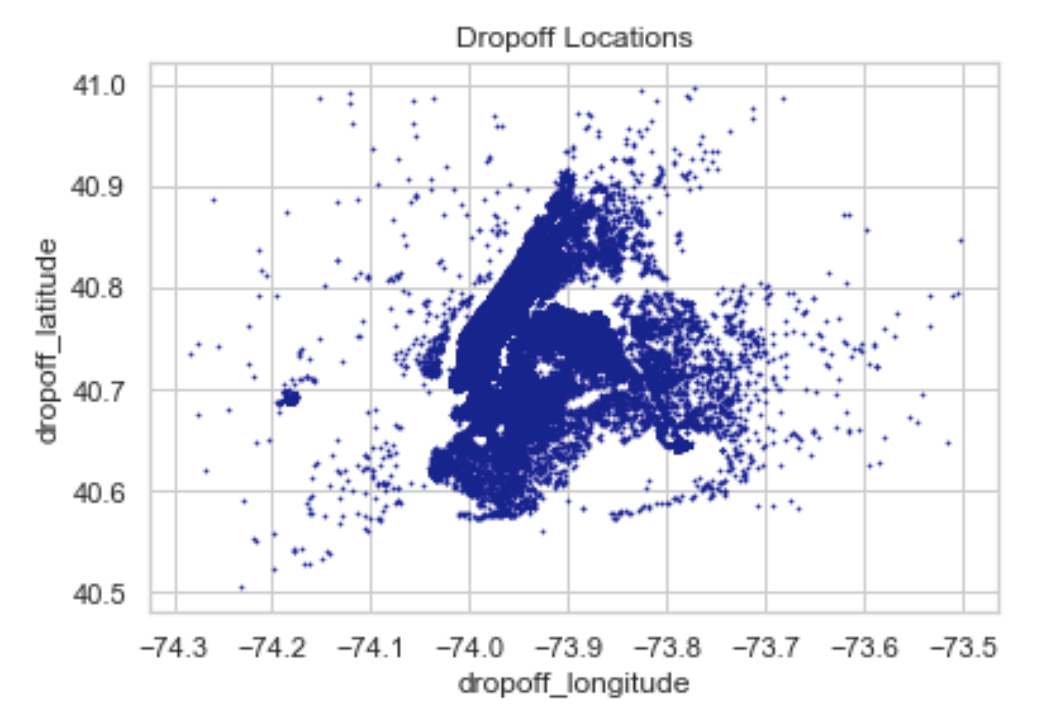
- Rushed distance
- Higher order latitude and longitude
- Google maps distance

Feature Creation

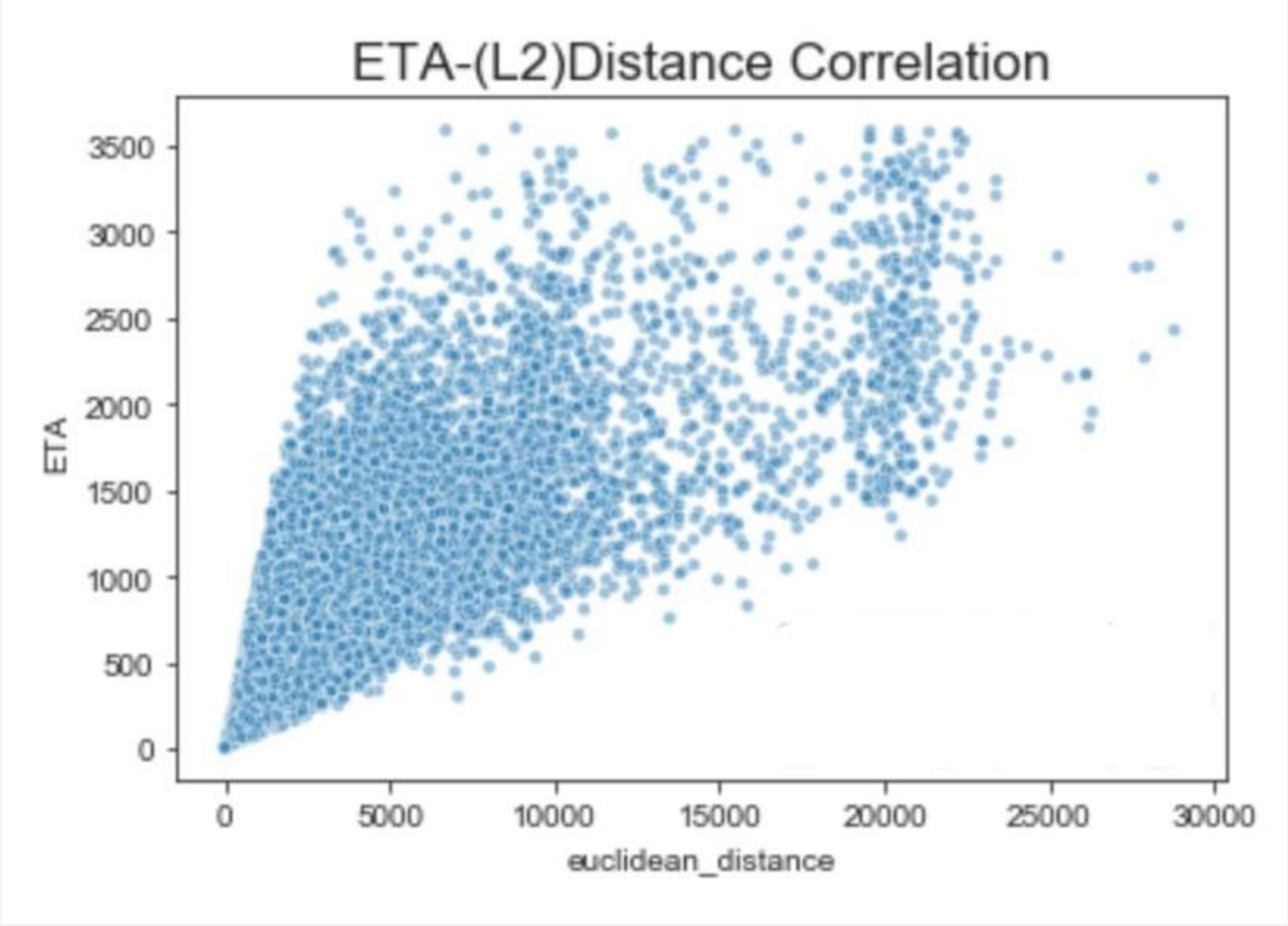
MODEL SELECTION AND COMPARISON

EXPLORATORY ANALYSIS

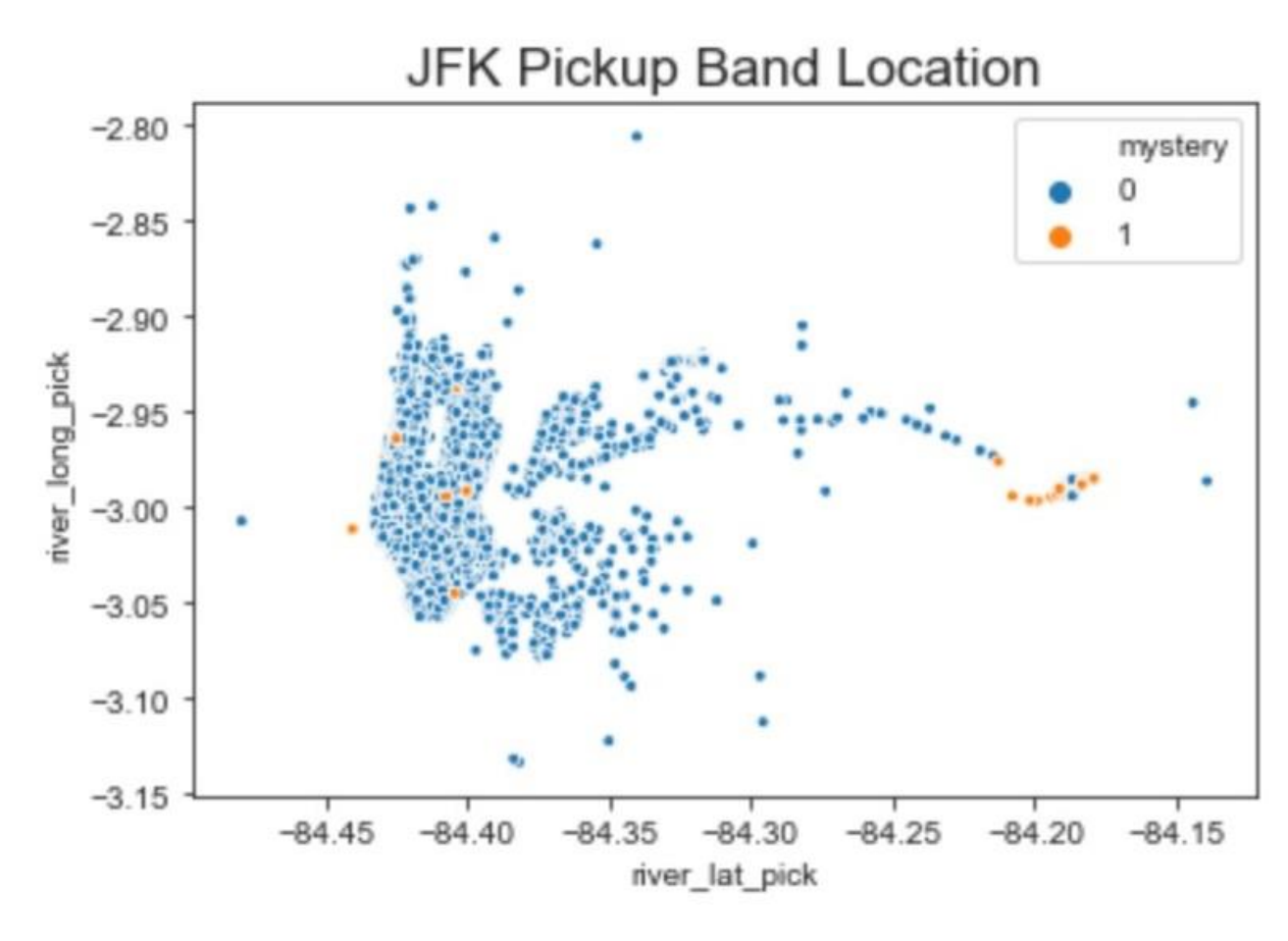


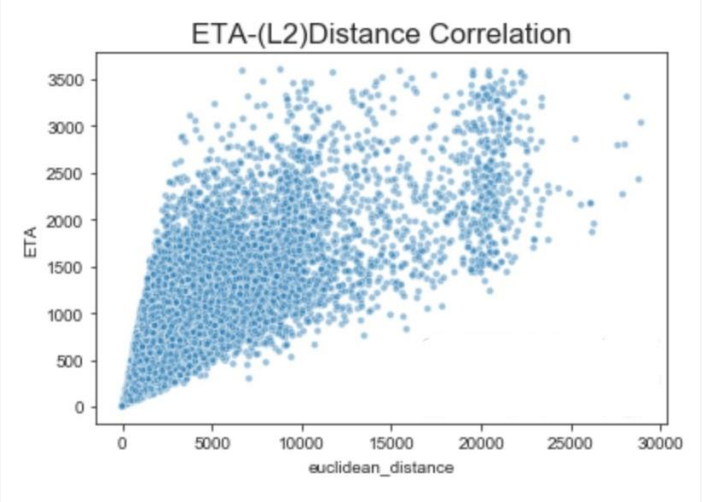


Exploratory Analysis

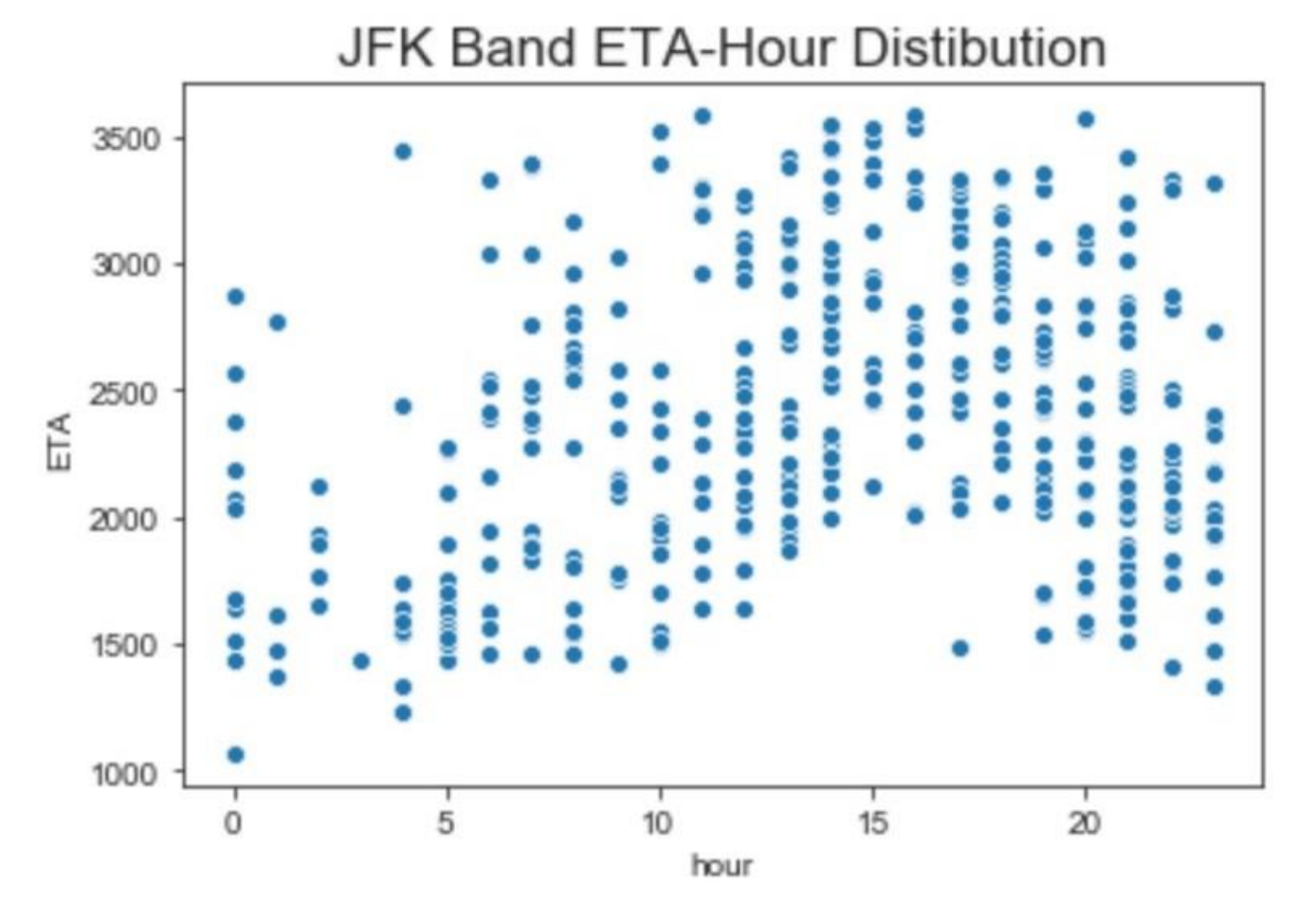


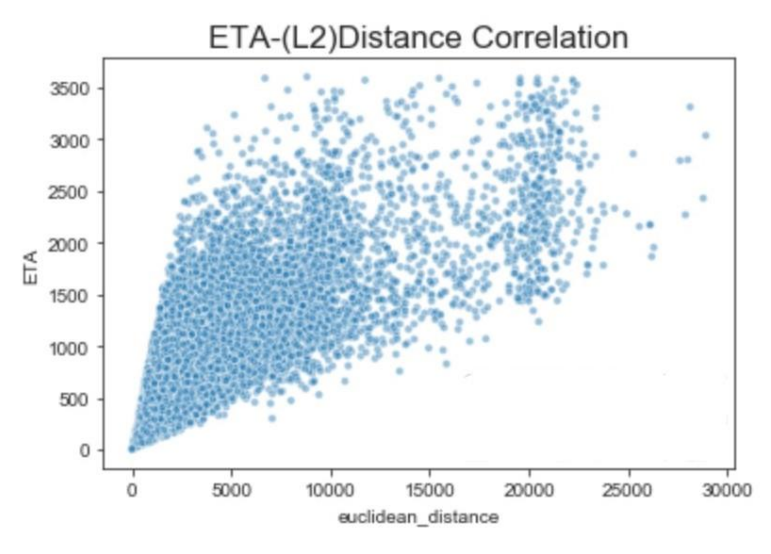
Exploratory Analysis

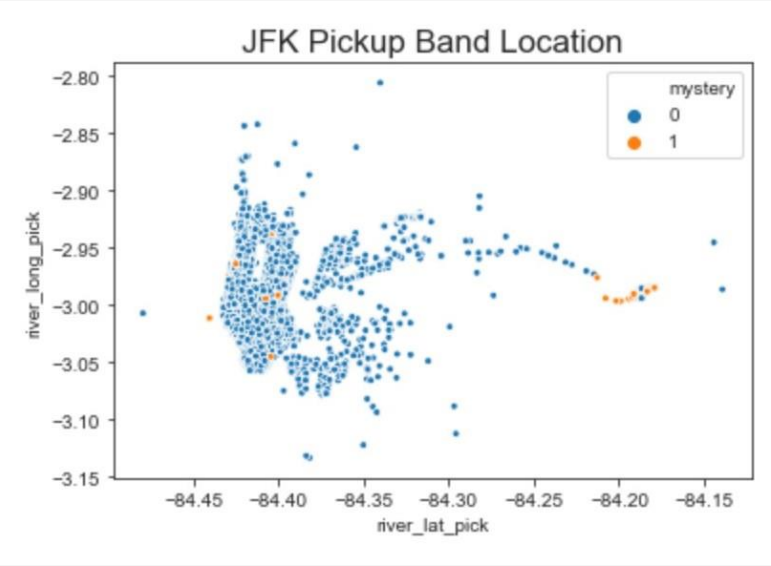




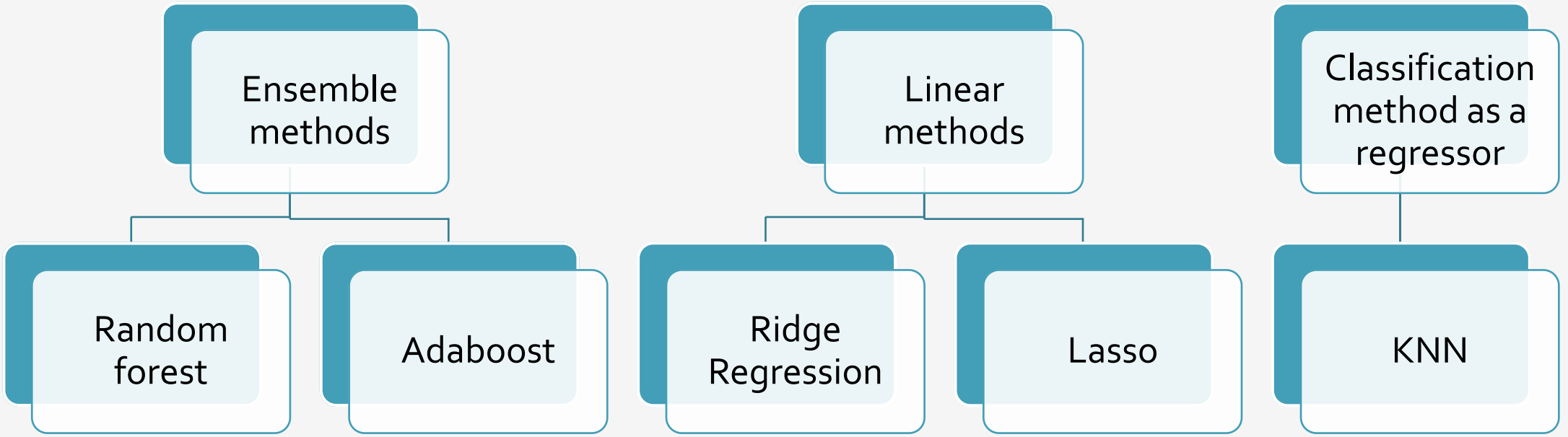
Exploratory Analysis

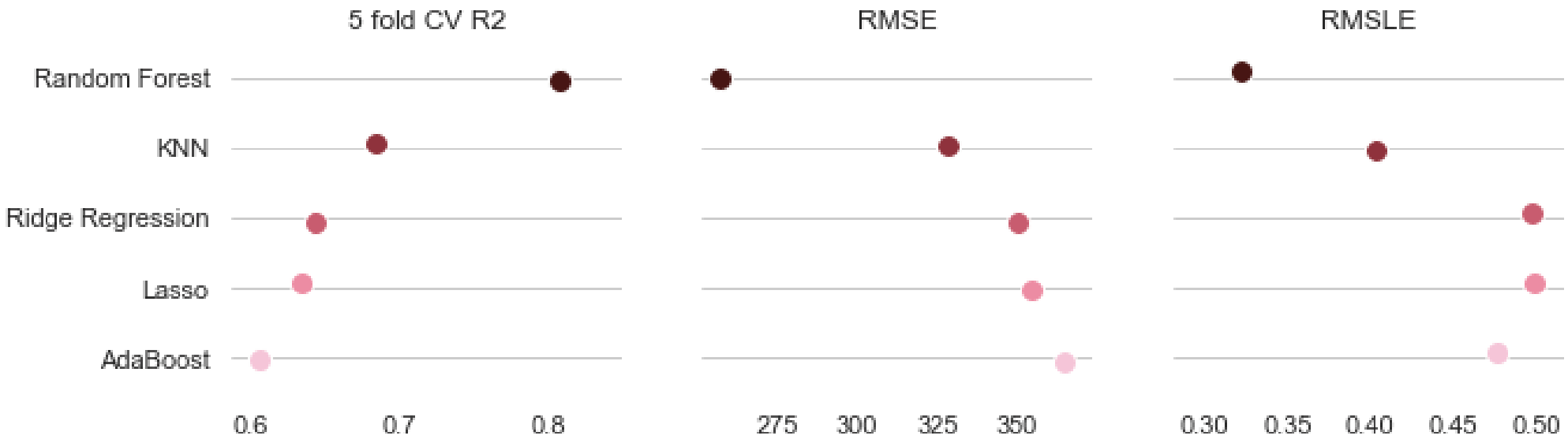






Model Selection





MODEL COMPARISON

5 fold CV R2	RMSE	RMSLE	Name
0.808012	257.449794	0.322079	Random Forest
0.605875	364.648080	0.476695	AdaBoost
0.642688	349.855128	0.496594	Ridge Regression
0.633013	354.223491	0.498373	Lasso
0.684499	328.438417	0.403756	KNN

MODEL COMPARISON

THANK YOU

Questions?