# Estimating Citibike Demand to Predict Up-and-Coming Neighborhoods

Dumbo, Brooklyn

# Team Members

Vittorio Bisin
Raghav Singhal
Pranjal Thapar
Gita Ventyana

# Business Understanding

Bike sharing programs have been gaining popularity throughout the world. But a major cost for most systems is the re-allocations of bikes on a daily basis, due to the naturally asymmetric flow of traffic. To that effect we attempt to build a demand function for each station which can predict the daily demand from each station for the entire network.  Furthermore we also intend to investigate the relationship between real estate prices and changes in Citibike station demand - to predict up and coming neighborhoods.

# CitiBike Dataset

Citi Bike Online has complete (41 months) of daily data (~25,000 daily rides). 7.6 GB of data with more than 35 million trips starting from July 2013.

Features include:

- Trip duration
- Start time and date
- End time and date
- Start station name
- End station name
- Bike ID
- User type
- Gender
- Year of Birth

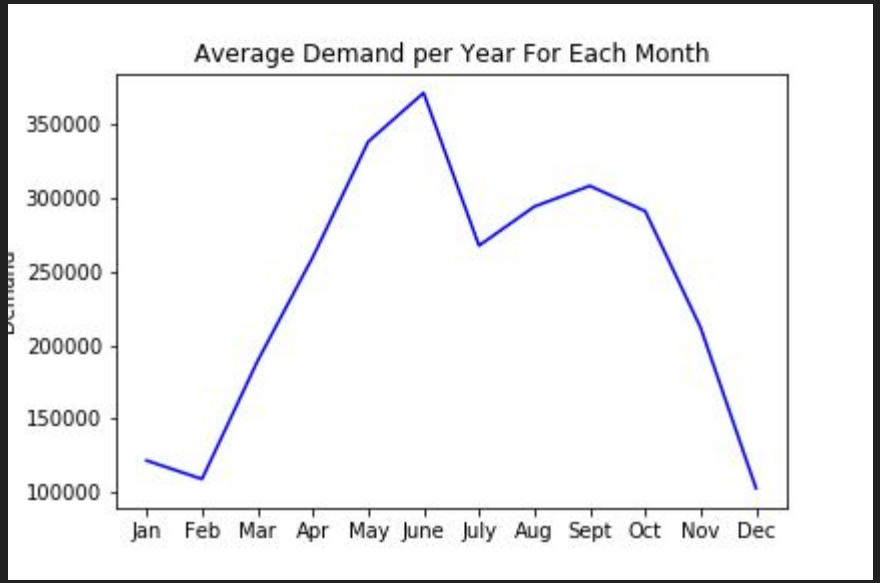
# National Oceanic and Atmospheric Administration (NOAA) Dataset

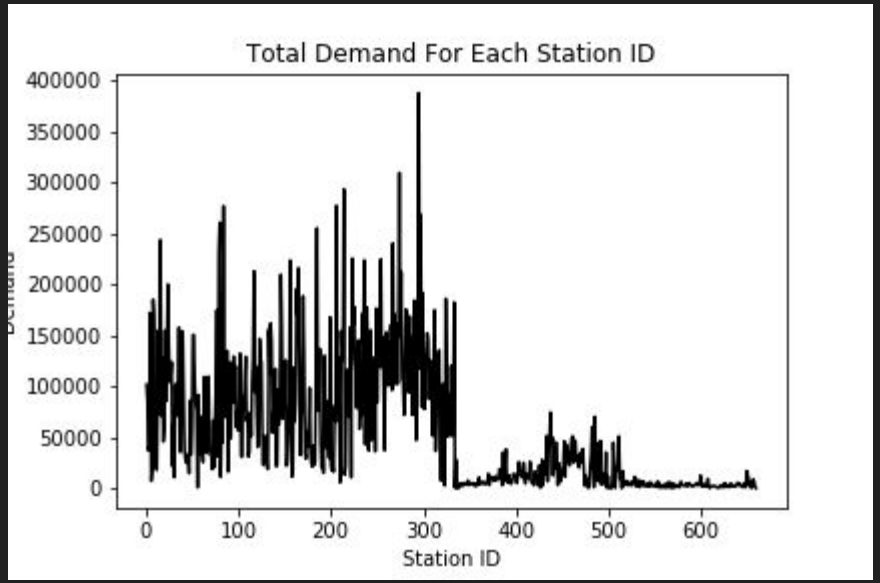
- Measures hourly surface weather data from Central Park
- Rather complicated and unintuitive data set
- Contains 47,335 observations (about 1,279 days)
- Each observation has 1,009 features
- Many of these features are qualitative variables  (e.g. data source flag) or not relevant for us (e.g.  wave measurement)

# Zillow New York Real Estate Data

- Monthly median value (per square ft) of homes in New York
- Categorized by neighborhood
  - 160 neighborhoods
  - Detailed analysis of neighborhoods in Brooklyn
- Dataset contains values from 2007-2017
  - Also categorized by type of home
- Contains separate list price, foreclosures, and estimated rent prices

# Citi Bike Preliminary Data Exploration



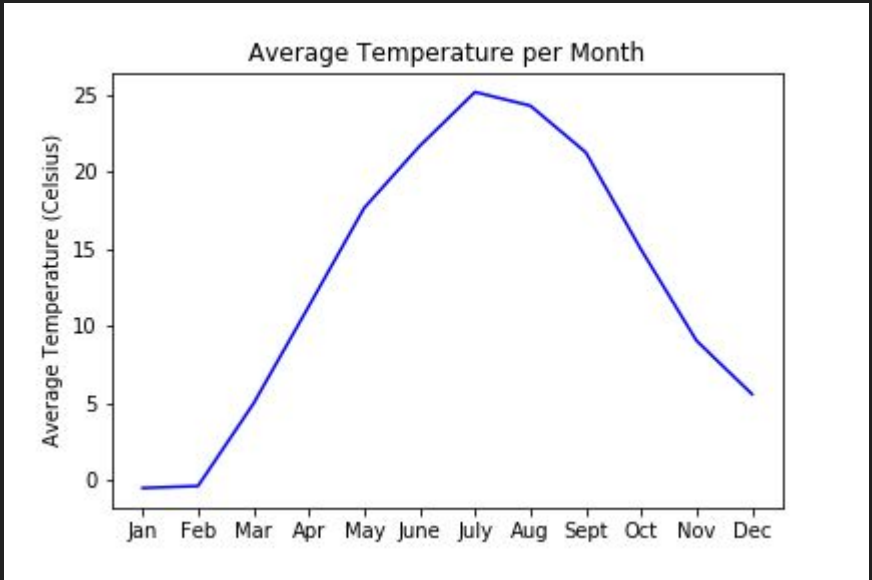
# Citibike Data Preparation

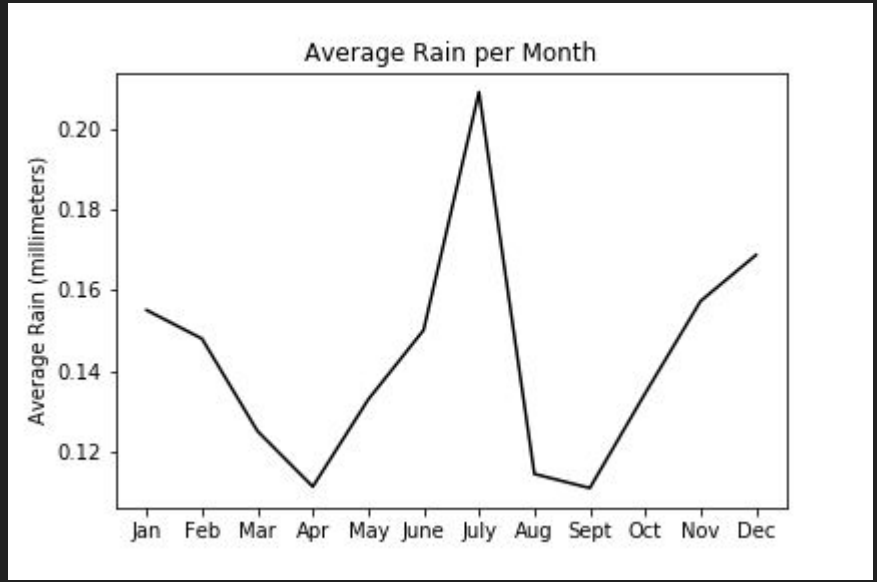
- Already quite clean and well-organized
- Fixed small mistakes (e.g. bike trips lasting over 24 hours)
    - Either error in the system or a stolen bike
    - Timestamp for start and end times had varying formats that needed to be cleaned
- Chose only relevant features
    - E.g. removed user data
- Saved as a dataframe in Pyspark

# NOAA Data Preparation

- Not well-organized dataset - required a great deal of cleaning
  - Incomplete and inconsistent (hopefully not noisy)
  - There were also numerous discrepancies between the dataset and the readme file
- Removed indicator variables not relevant to our analysis
- Used the mean binning method to deal with incomplete entries
  - In some instances the feature had to be removed
- Feature Reduction
  - First used a High Feature Correlation Threshold (FCT)
  - SVD

# Two Features Extracted from NOAA Dataset



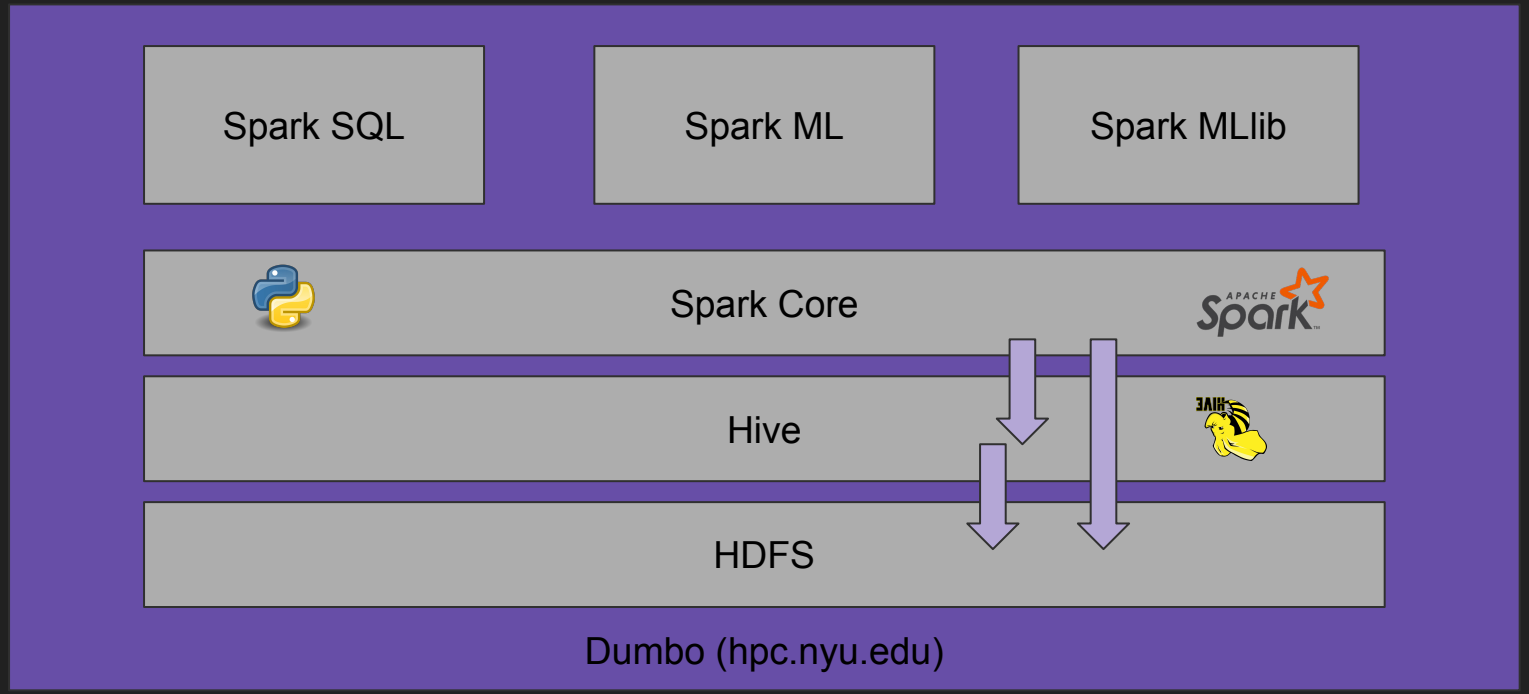
# Features Extracted from Citibike Dataset

The features we used from the Citibike Dataset were:

- Week Number (integer)
- Weekend/Weekday (indicator variable)
- Bike demand between all pairs of stations.
- Mean and median Trip duration from each station for a month

These features were then combined with the weather data to model the demand function and analyze the Citibike Network.

# Architecture



# Modelling: Linear Models

Generalized Linear Models - In Linear Regression the output is assumed to follow a Gaussian Distribution, however for our purposes we have employed a more general family of exponential distributions, the response variable is distributed according to $Y_i \sim f(\cdot|\theta_i, \tau)$

where $f_Y(y|\theta, \tau) = h(y, \tau) \exp\left(\frac{\theta \cdot y - A(\theta)}{d(\tau)}\right)$

# Modelling: Linear Models

The parameters are related to the expected value of the response variable and we select the parameters such that

$$\theta_i = A'^{-1}(g^{-1}(\vec{x_i} \cdot \vec{\beta}))$$

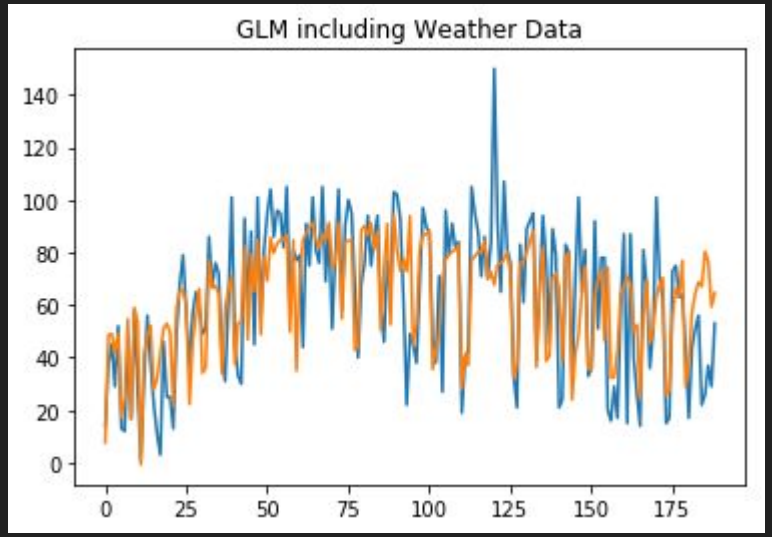
$$\mu_i = A'(\theta_i)$$

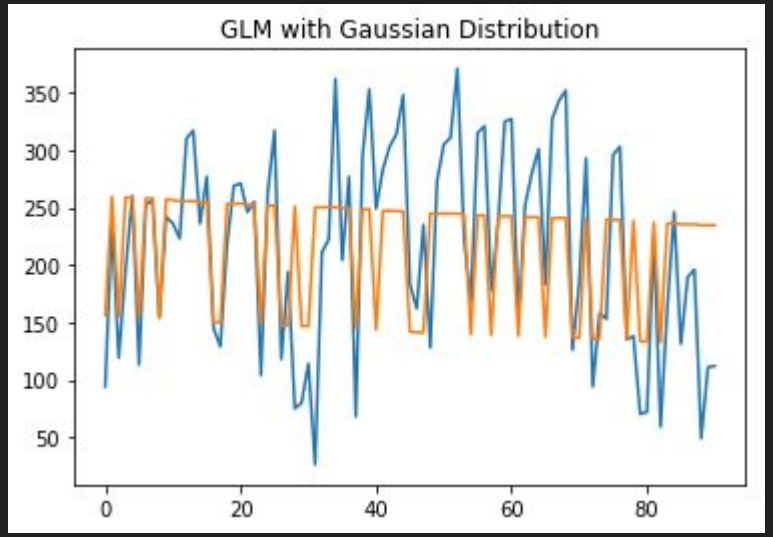
And then we maximize the likelihood function over the regression parameter.

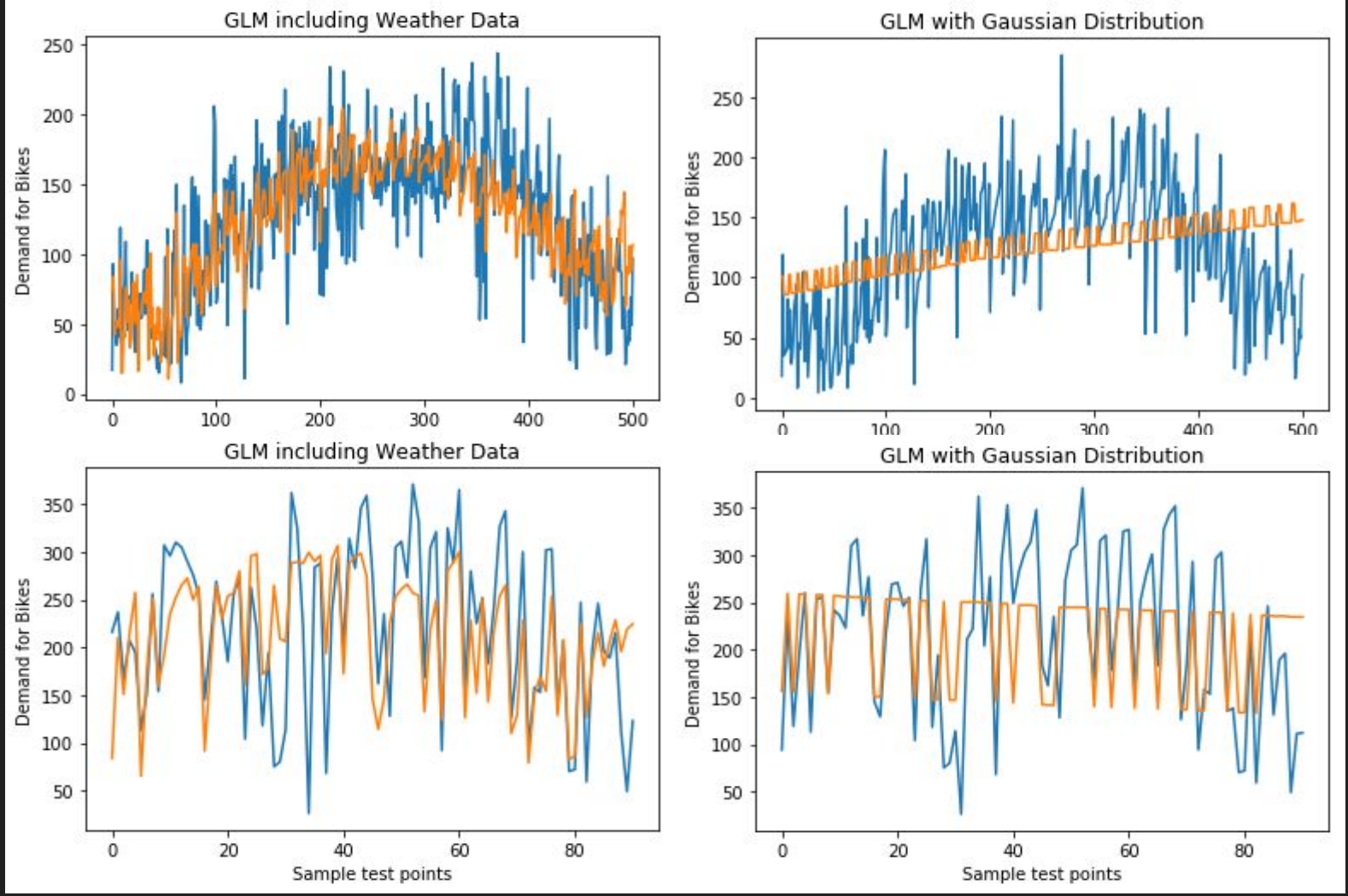
$$\max_{\vec{\beta}} \mathcal{L}(\vec{\theta}|\vec{y}, X) = \prod_{i=1}^{N} h(y_i, \tau) \exp\left(\frac{y_i\theta_i - A(\theta_i)}{d(\tau)}\right)$$

Here, all the functions mentioned above are dependent on the distribution we select, for our purpose where we have to work with count data, we use the Poisson distribution, Gaussian distribution  and a General linear regression.

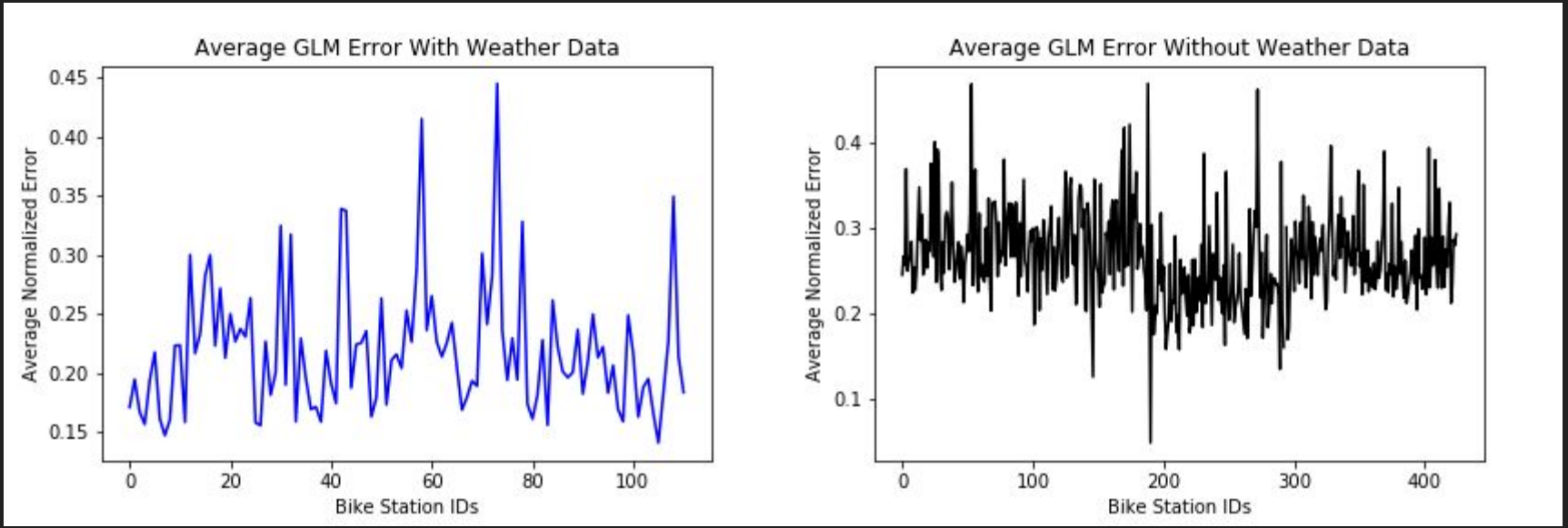
# Model Performance: Bike Station Error Plots




GLM including Weather Data
Demand for Bikes
GLM with Gaussian Distribution
Demand for Bikes
GLM including Weather Data
Demand for Bikes
Sample test points
GLM with Gaussian Distribution
Demand for Bikes
Sample test points

# Full Data Graphs



# Modelling: Clustering

We wanted  to find out group of pairs of stations that experienced similar demand, pairing this with the real estate data could predict trends in the housing market

We used K-means clustering on the feature set to cluster pairs of stations based on the daily demand between them

The feature set includes: Start Station, End Station, weeknumber, weekday, daily demand between pairs of stations.

We used the PySpark MLlib library to build a K- means clustering model that predicts daily demands between each pair of stations

# Model Results: Clustering

Our optimum K=5 gives us 5 clusters that map pairs of stations to 5 categories based on daily demand

```
>>> dfcluster.show()
+-------+------------------+
|Cluster|      Daily Demand|
+-------+------------------+
|      0|25.710392969774283|
|      1| 5.019148586697327|
|      2|3.1435010605346645|
|      3|2.2340122961265534|
|      4|  9.92133366417079|
+-------+------------------+
```

One of the Station Pairs with highest traffic (belonging to Cluster 0) is Yankee Ferry Station probably because it is the only CitiBike Station on Governor's Island.

While one of pairs of stations with lowest traffic were 11th and W52nd, 3rd and E12th and they belonged to Cluster 3.

# Possible Deployment Avenues

Our Linear Models can be deployed to estimate demand for each station every day. This can enable them to distribute bikes more efficiently based on this demand function.

Our Clustering models coupled with weather and the Zillow datasets can be used to predict new housing trends and upcoming neighborhoods.

# Future Work

We plan to:

- Try fitting different regression models available on MLlib
  - Currently Pyspark on HPC crashes because of a lack of memory (EOF error)
- Include our Zillow real estate data set
  - Observe how a change in demand correlates with a change in real estate prices
  - Use Granger causality to determine a possible causal direction

# References

-Networks as Signals, with an application to a Bike-Sharing System.
Ronan Haman et al

-Station Site Optimization in Bike Sharing Systems.
Junming Liu: Qiao Li; Meng Qu; et al

-Inferring bike trip patterns from bike sharing system open data
Longbiao Chen; Jérémie Jakubowicz

- Rebalancing Bike Sharing Systems: A Multi-source Data Smart Optimization
Junming Liu, Leilei Sun, Weiwei Chen, Hui Xiong