Look here for the punchline

Predicting Capital BikeShare Usage

Destry Saul

August 19, 2013

General Assembly - Data Science 4

Two systems - same operator - NY much busier

capital bikeshare™

Operated by Alta Bicycle Share.

Equipment by
Public Bike System Company

~200 stations
1800 bikes

~3.2 Million trips in 2011 & 2012 combined

capitalBikeshare.com

citi bike

Operated by NYC Bike Share LLC,
a subsidiary of Alta Bicycle Share.

Equipment by
Public Bike System Company

320+ stations
4000+ bikes

Over 2.1 Million trips since launch (May 2013)

30 - 40k trips per day

citibikeNYC.com

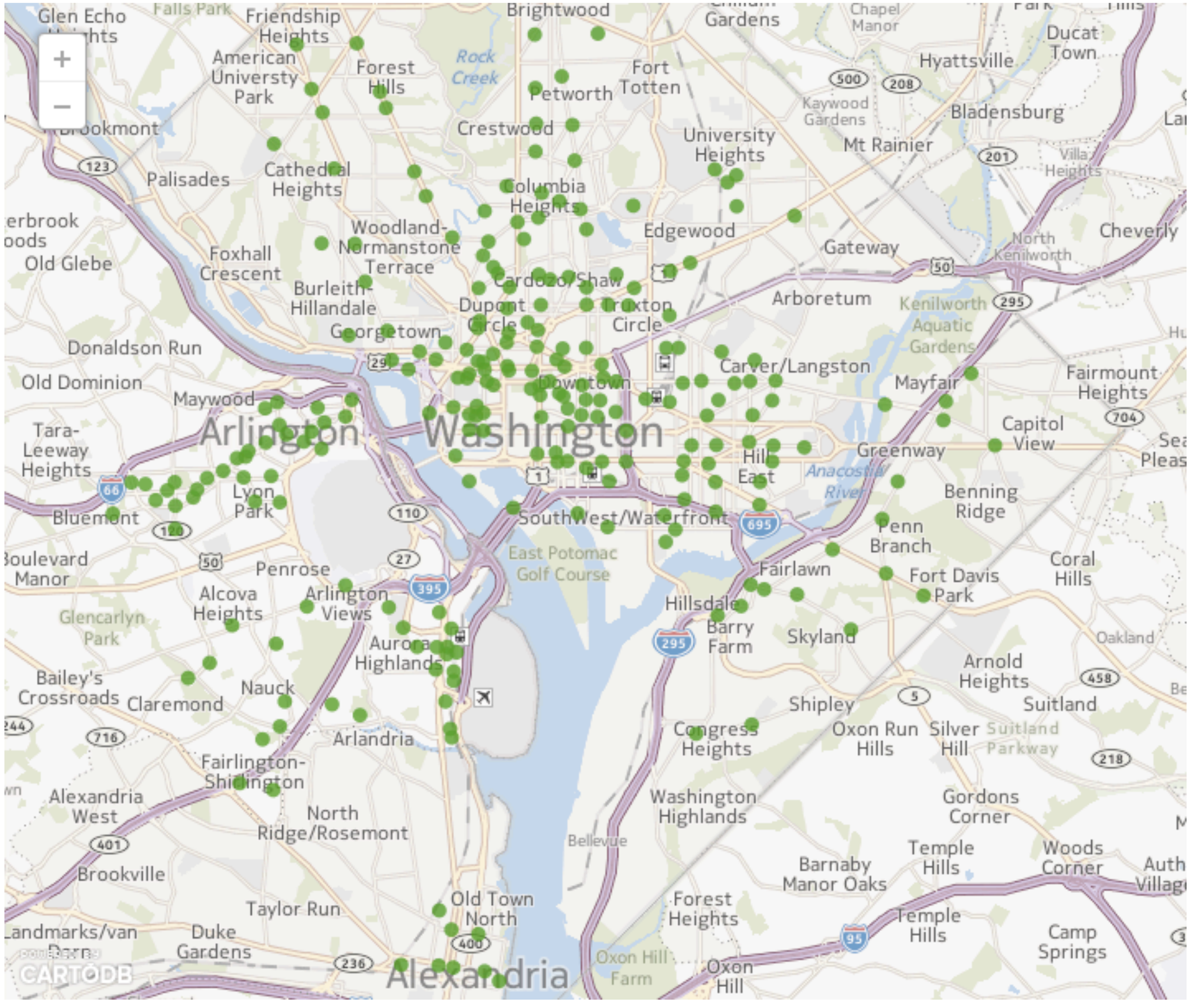
Bikeshare systems require balancing

wnyc.com

Van - 8 bikes
Truck - 20 bikes

wikimedia

Bike Station Locations



Part 1: Predicting the flow of bikes

The Question:

After a given hour, will there be more bikes or less bikes at a certain station, and how many?

The Approach to the Answer:

Gather Bike and Weather Data

Format Data

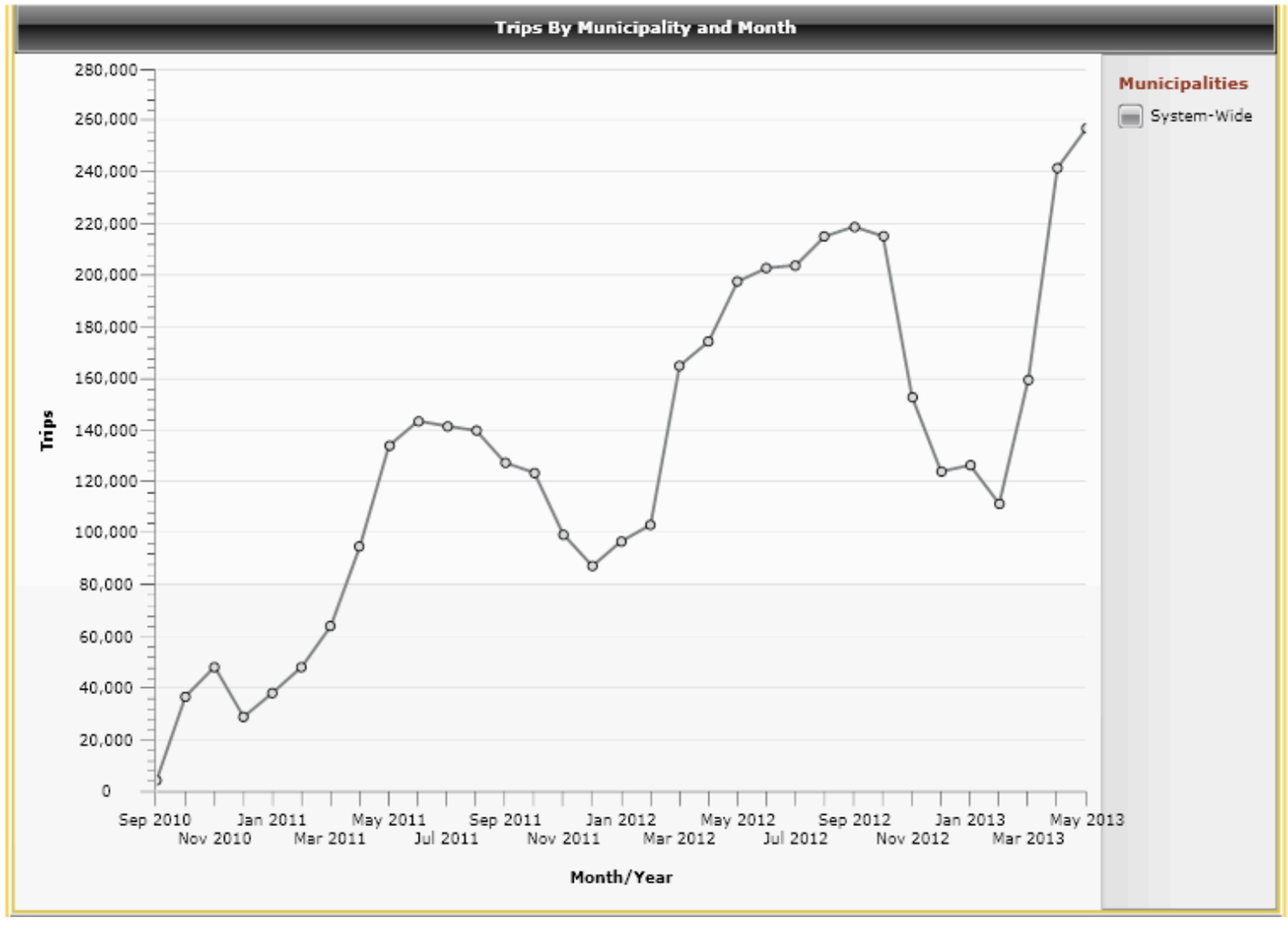
Build Support Vector Regression Model

Capital Bikeshare data is one line = one trip

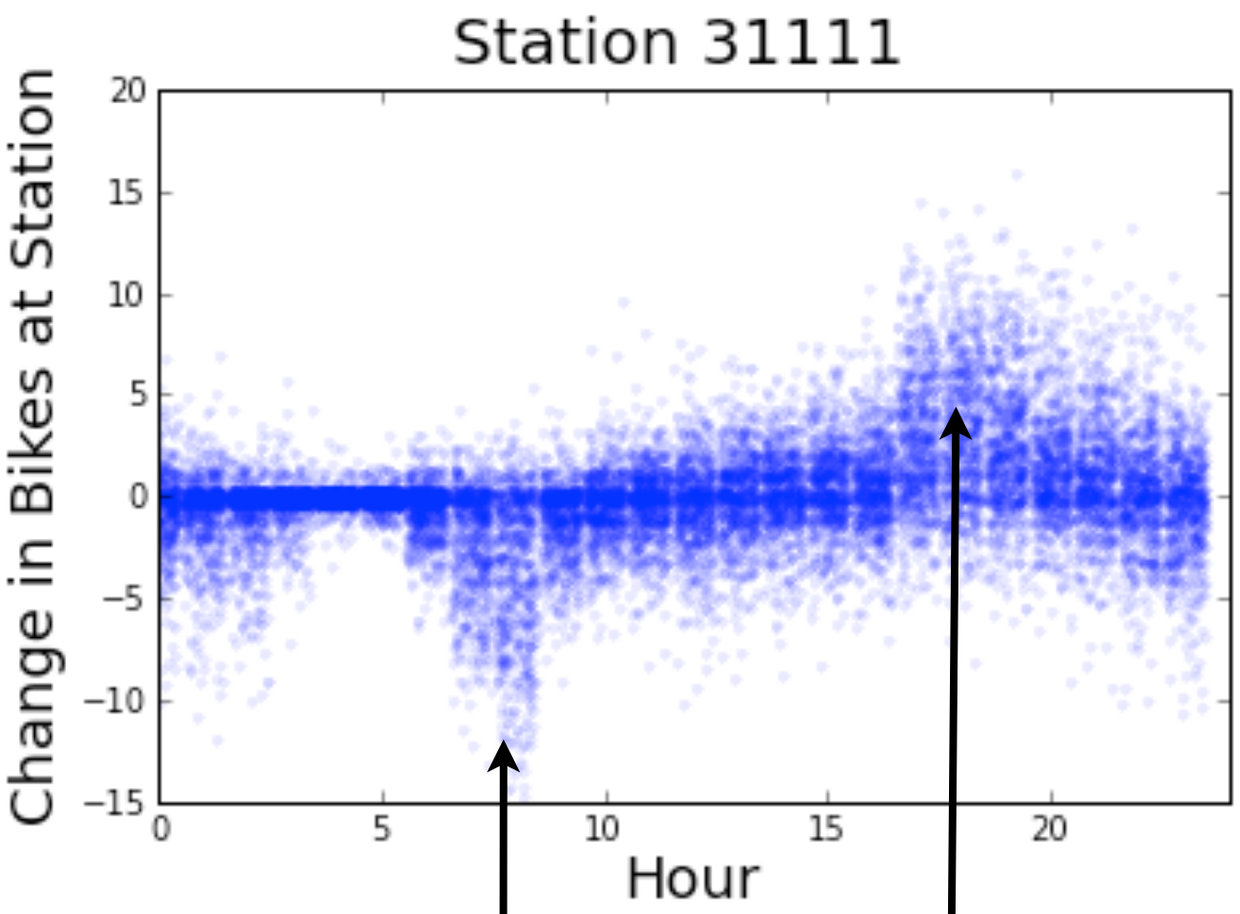
www.capitalBikeshare.com/system-data

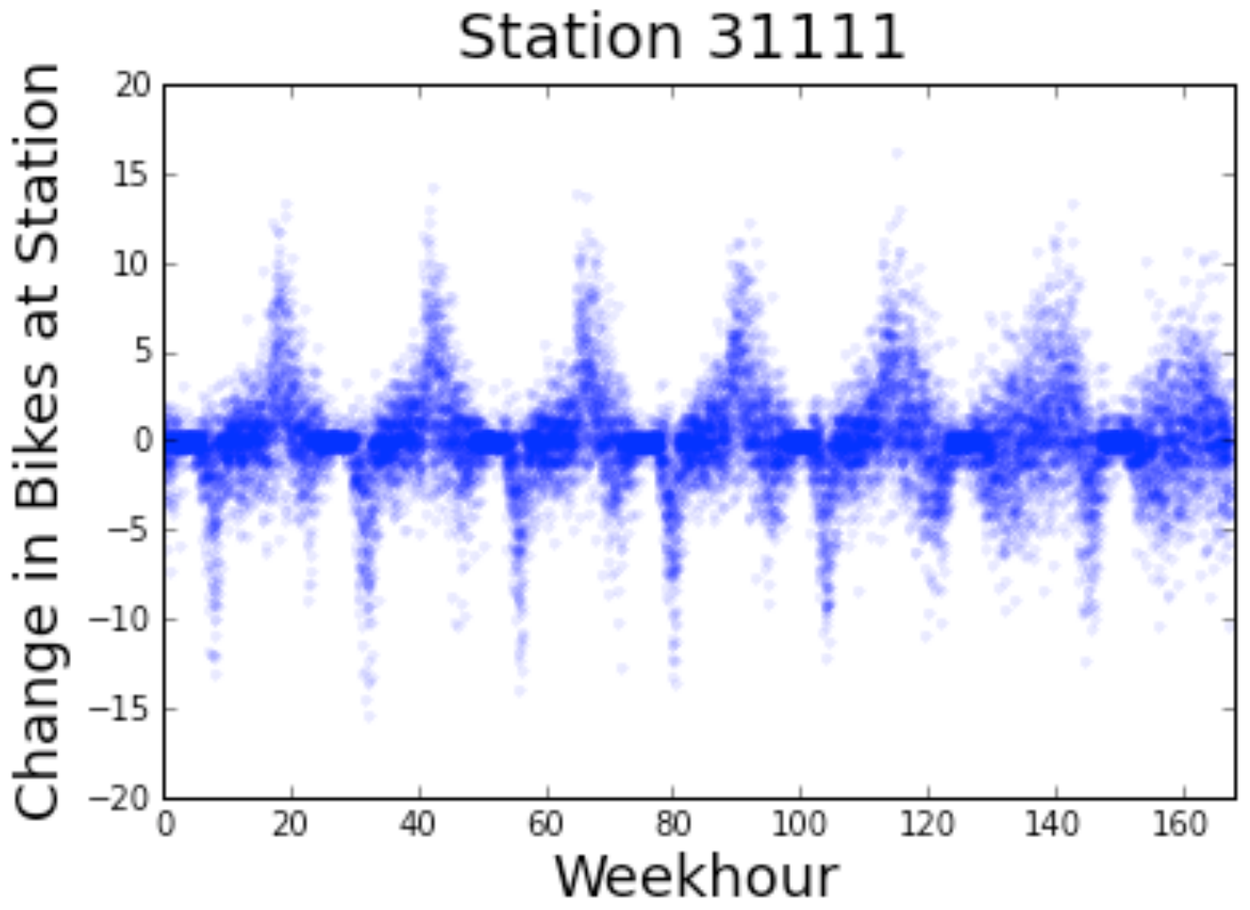
Example rows:

Duration,Start date,End date,Start station,End station,Bike#,Member Type
0h 1min. 50sec.,3/31/2011 23:58,4/1/2011 0:00,14th & Harvard St NW (31105),16th & Harvard St NW (31103),W00749,Registered
0h 16min. 21sec.,3/31/2011 23:52,4/1/2011 0:08,19th & L St NW (31224),7th & Water St SW / SW Waterfront (31609),W01048,Casual



Bike Flow is nonlinear with time





Coming Home?

Leaving for work?

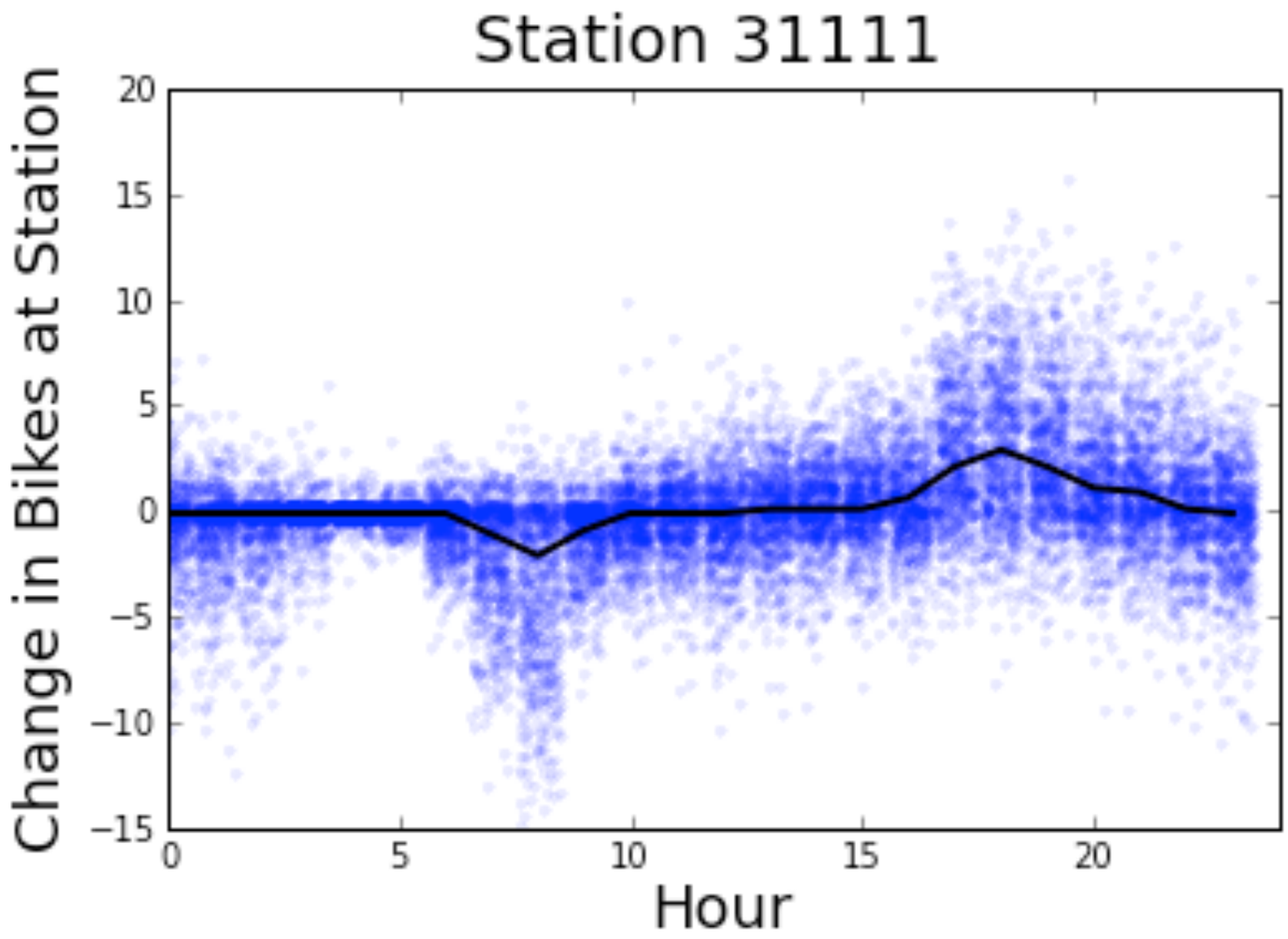
Predicting Usage = Support Vector Regression

> from sklearn.svm import SVR()

-rbf kernel

Model 0 - Hour of the day only
Model 1- Hour of the week (weekhour)
Model 2 - weekhour & temperature
Model 3- Year, weekhour, temperature

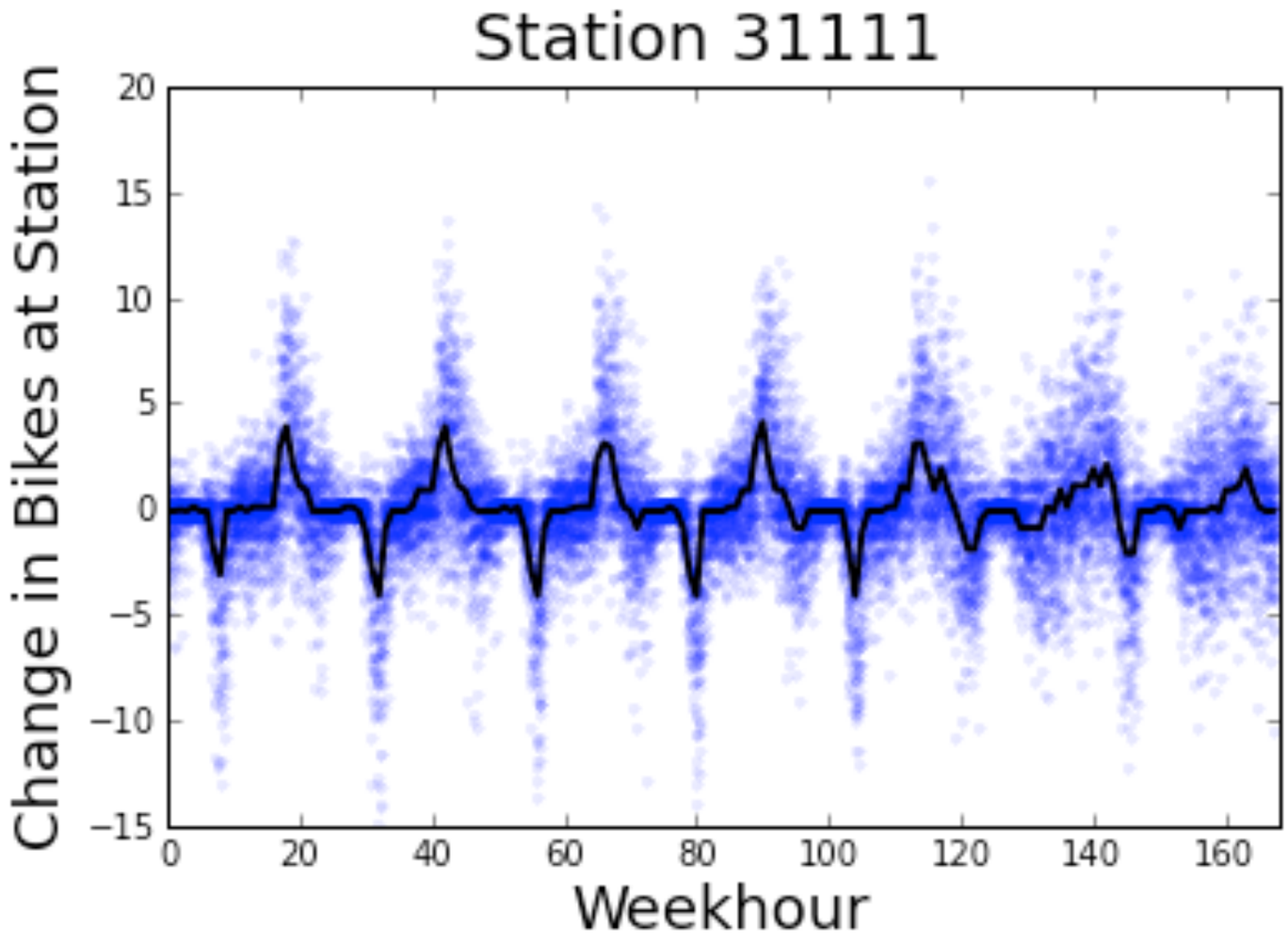
Hour Only = Poor.



Training MSE = 5.76
Mean Cross Validation = 5.82 (5-fold)

No Model MSE = 7.47

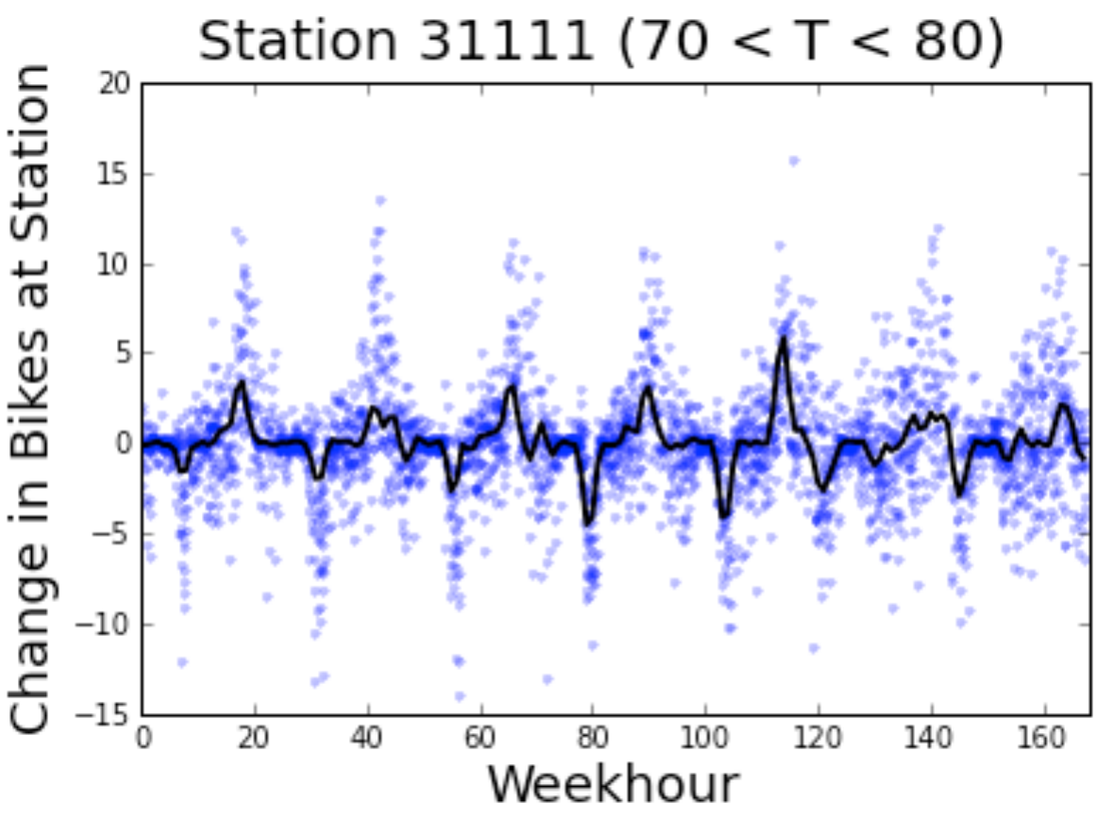
Weekhour Only = Better.

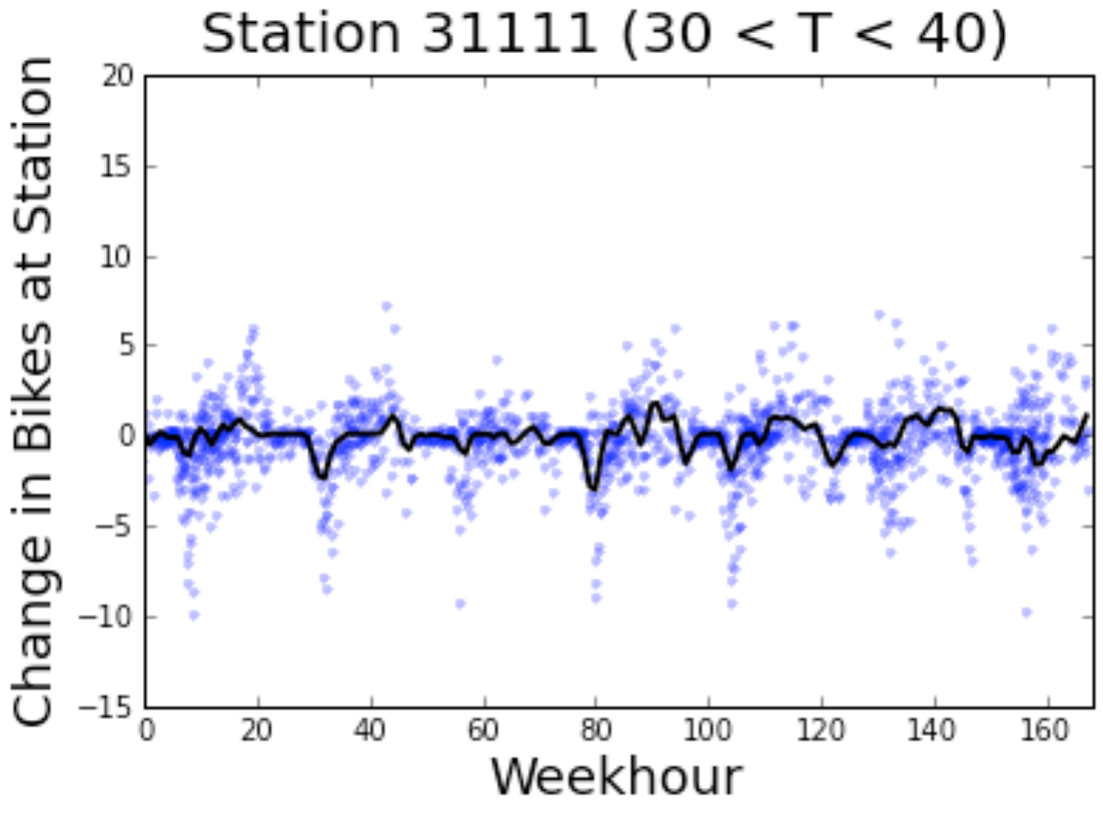


Training MSE = 5.09

Mean Cross Validation = 5.29 (5-fold)

Weekhour + Temp = Not exactly better.

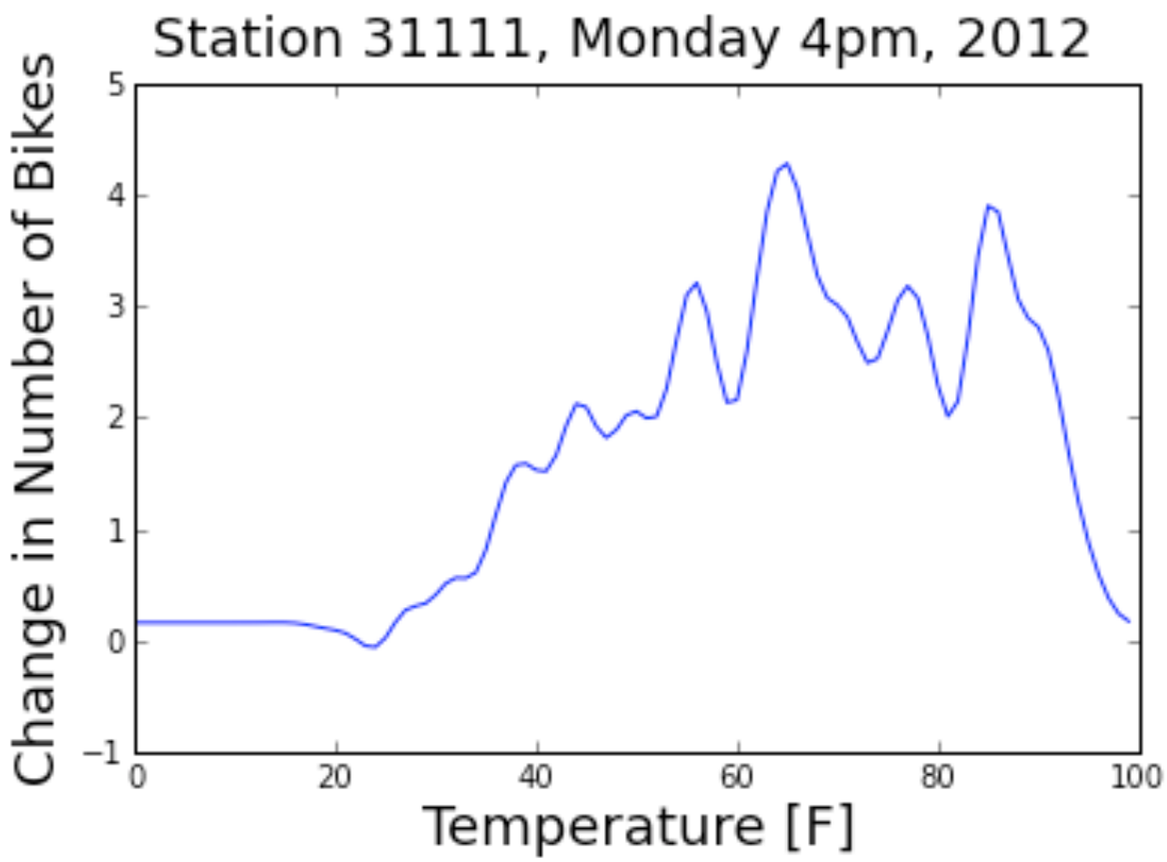




Training MSE = 5.12

Mean Cross Validation = 5.67 (10-fold)

Weekhour + Temp + Yr



Training MSE = 5.07

Mean Cross Validation = 5.67 (10-fold)

Part 2: Cluster Bike Station Models

The Question:

Can we group stations by bike flow?

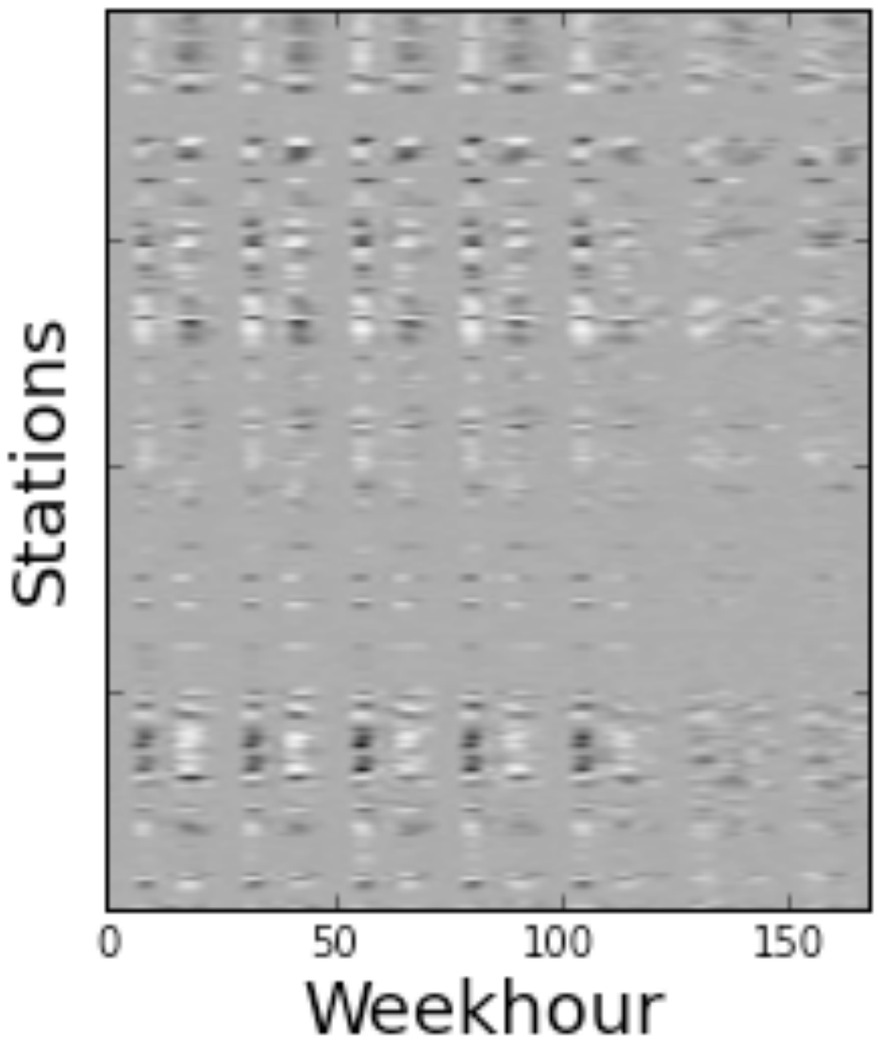
The Approach to the Answer:

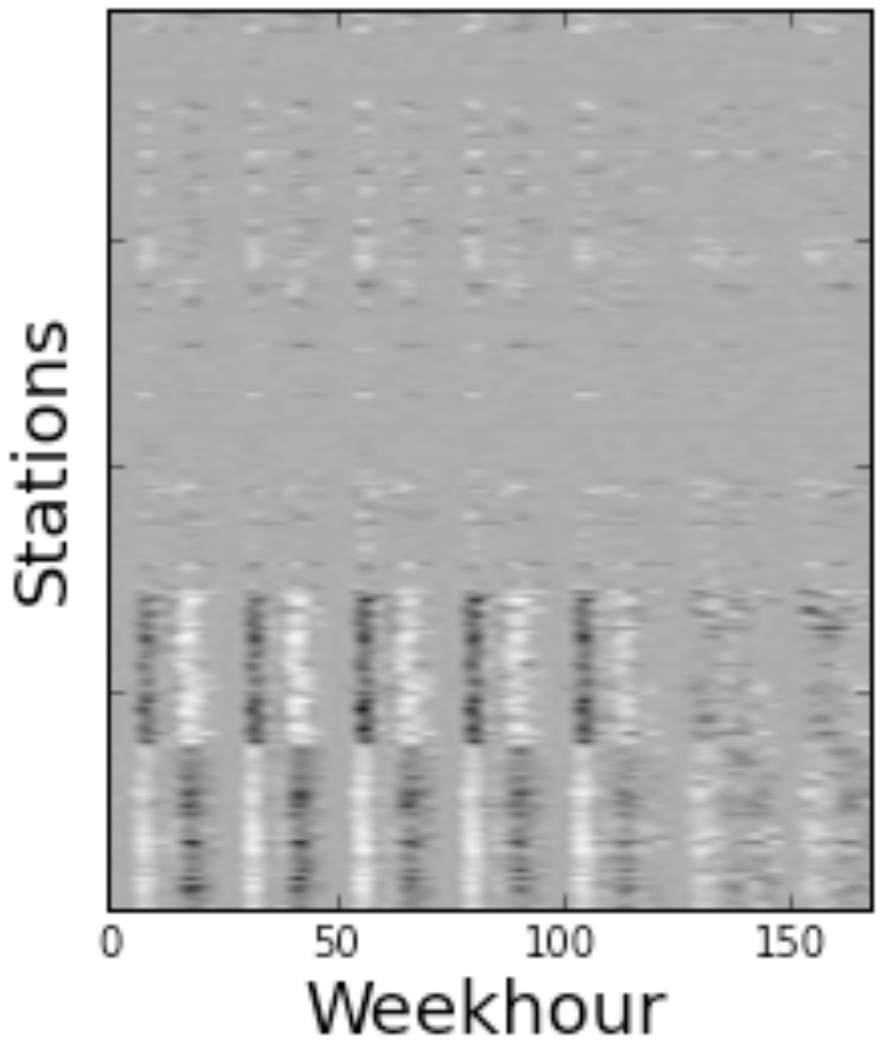
Generate models for each station.

Kmeans

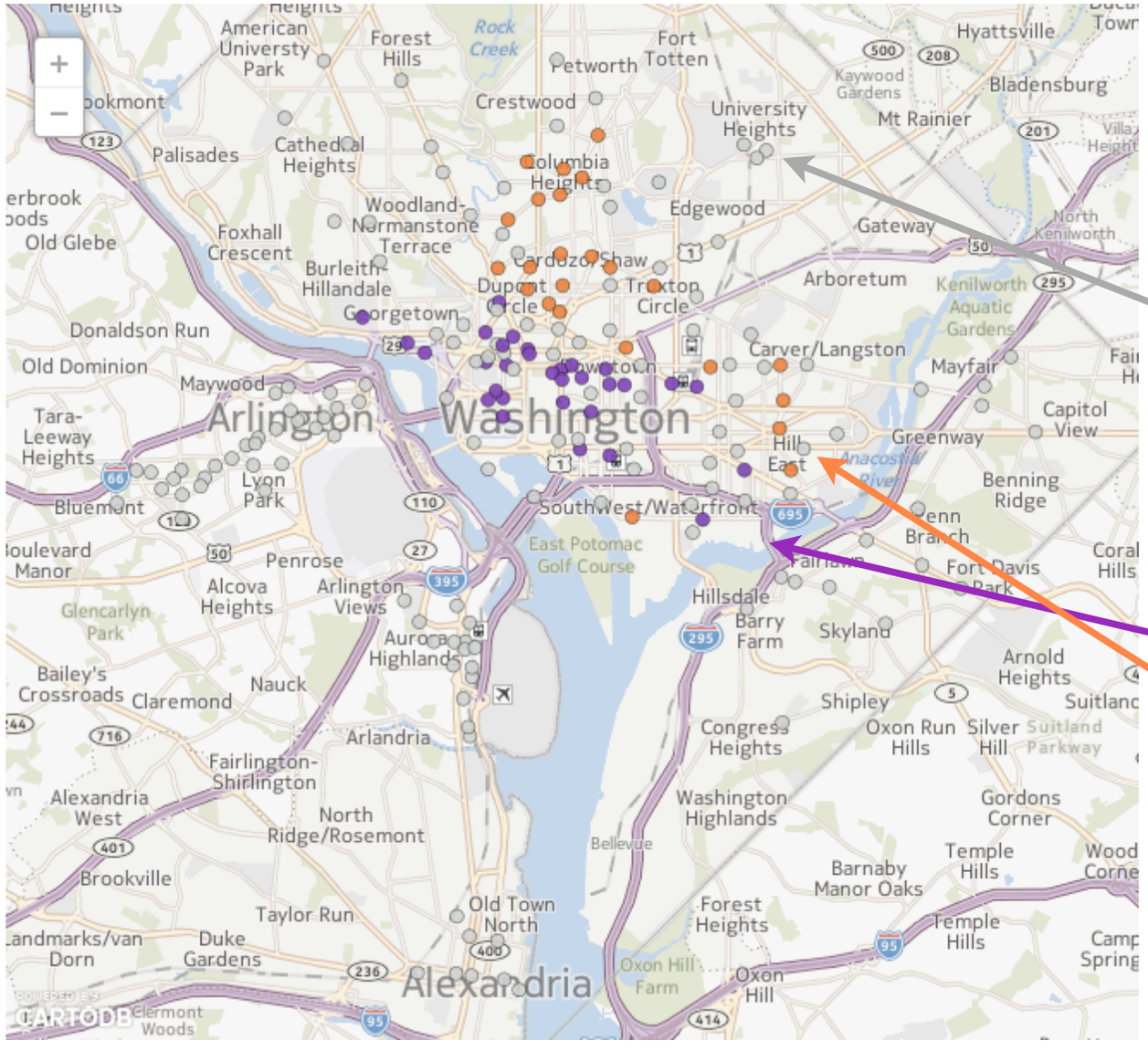
Visualize

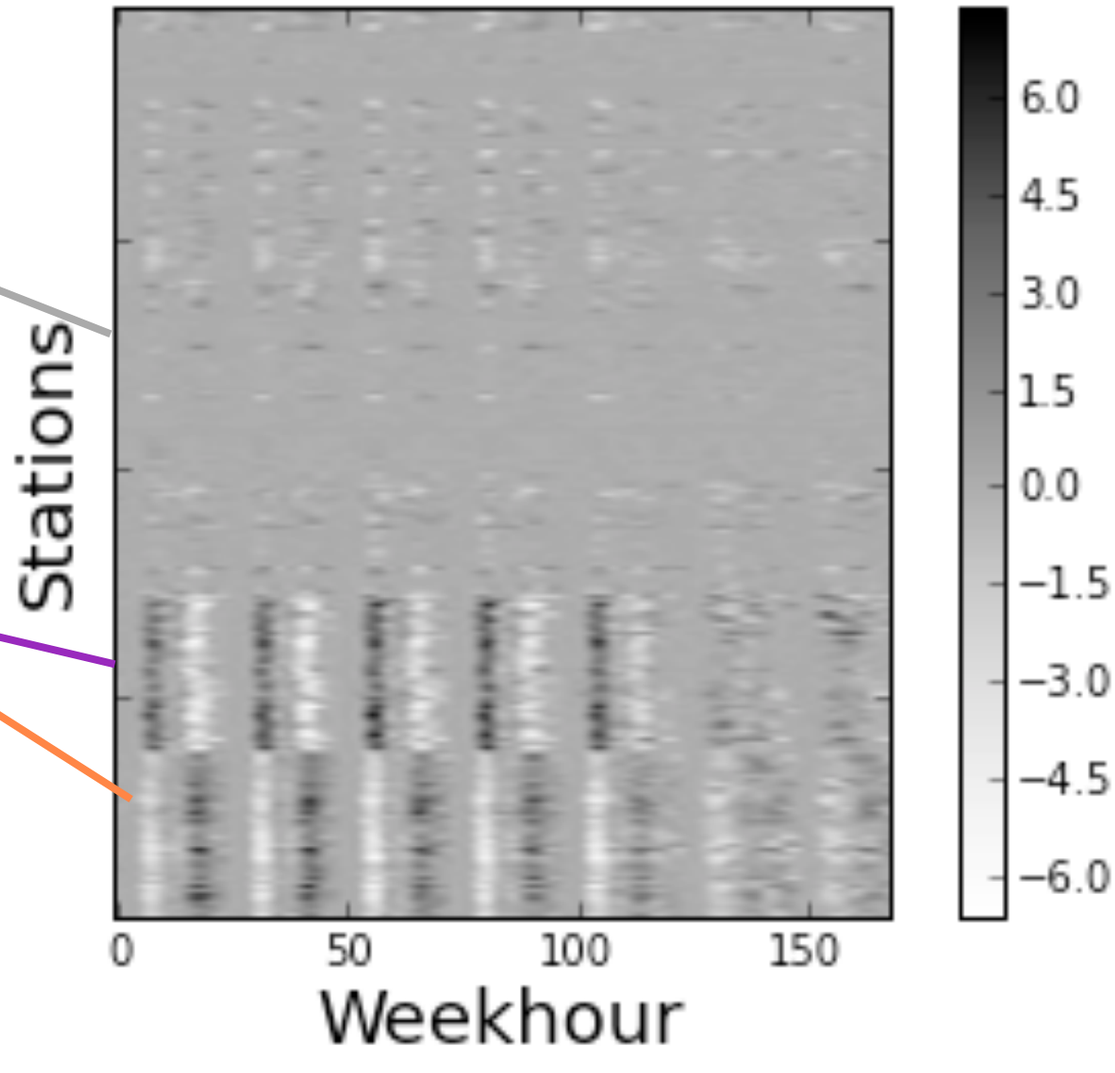
Kmeans works great on station models



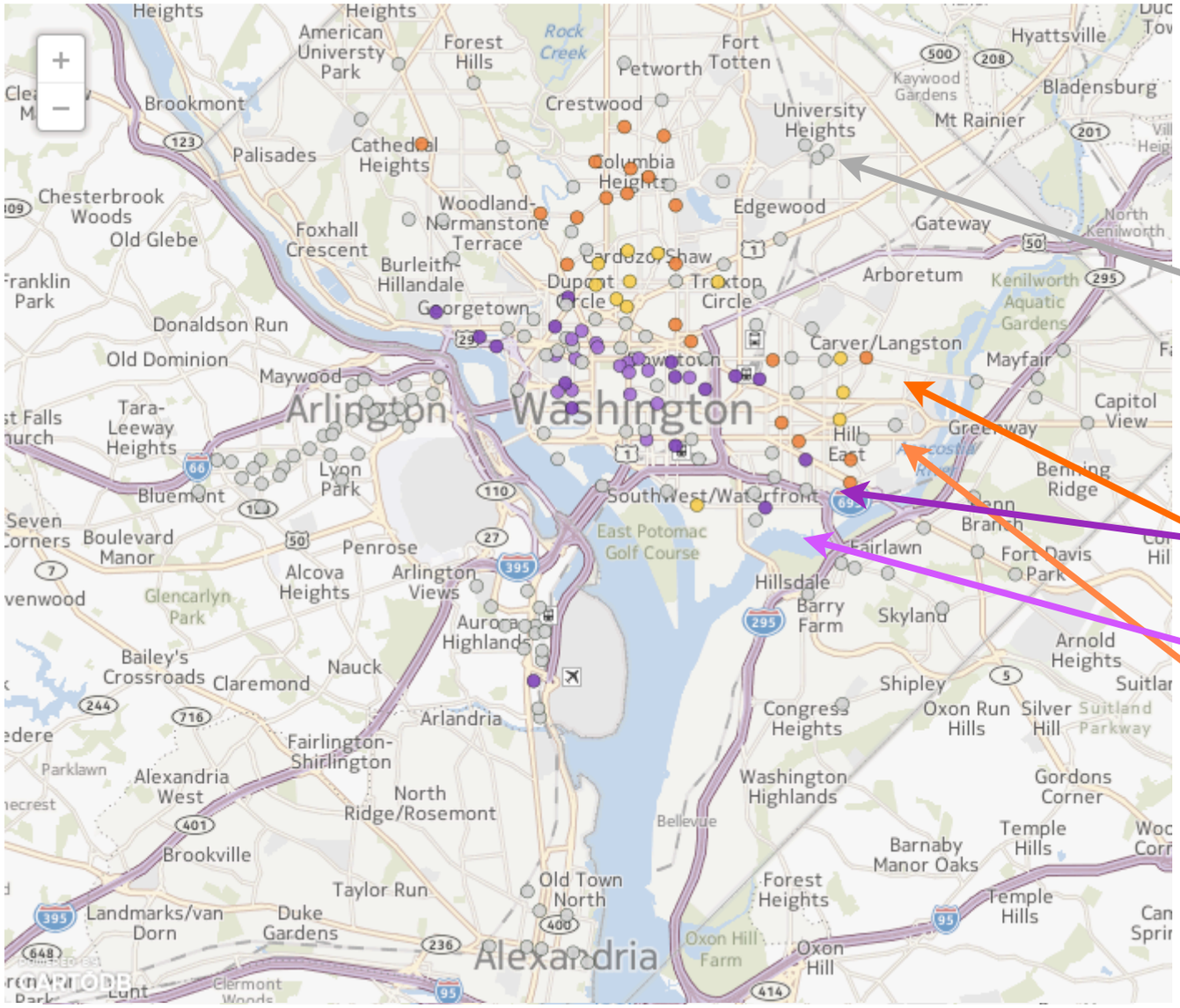


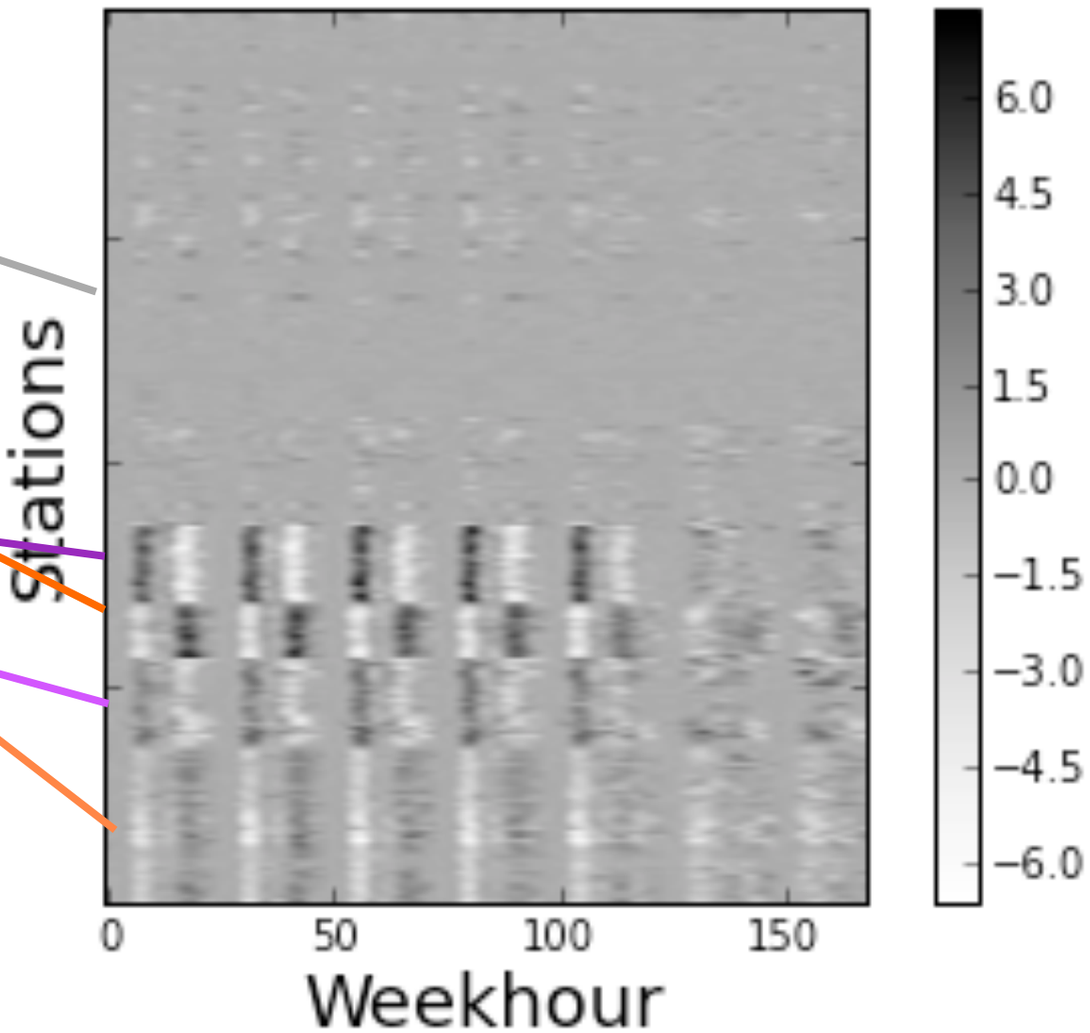
Where bike commuters live and & work





Where bike commuters live and & work





Lots of improvements and new features possible

Model Improvements:

- More weather features - rain, snowdepth
- Merge empty/full station status
- Work/school holidays
- Use station location somehow

Additional Features:

- Detect when a station will be full/empty