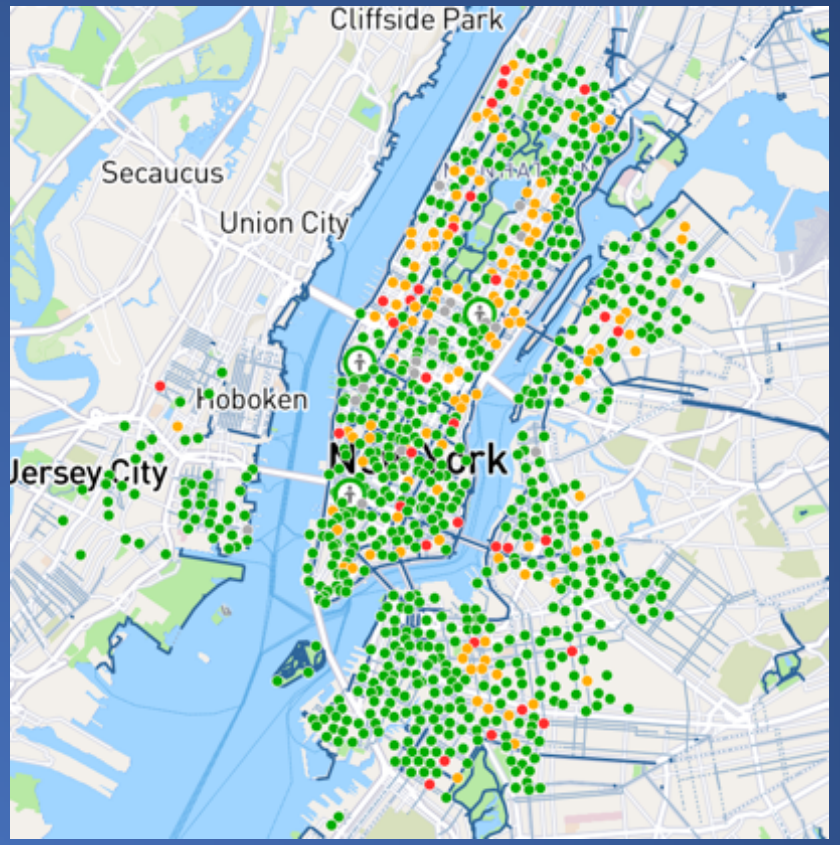
Forecasting Bike Rebalancing

Kristofer
Schobert

Citi Bike

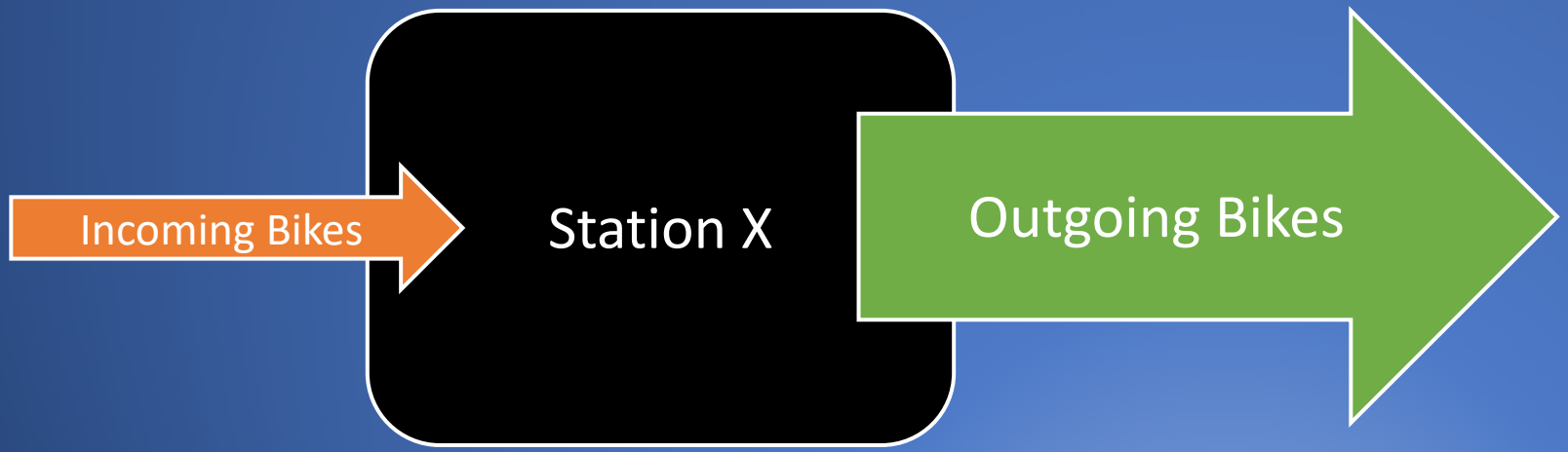
- The largest bike sharing program in the United States
- Roughly 800 stations throughout Manhattan, Brooklyn, Queens and Jersey City



Images taken from Citi Bike website: https://www.citibikenyc.com/

Citi Bike has a problem

- The average flow of bikes into a station is in general not equal to the average flow of bikes out of the station.



Citi Bike’s solution

- The manual rebalancing of bikes

Our Goal

- Predict stations' daily change in bike count
  - Three stations
  - First two weeks of May 2019

Significance

- Anticipate rebalancing needs. How many…
  - workers
  - trucks/trailers
  - stations
  - bikes

Our Methods

- Time Series Forecasting
  - Seasonal ARIMA models
  - Uses Trip data (Training: Mar & Apr 2019)

- Regression Analysis
  - Random Forest Regression and KNN Regression
  - Uses Trip data + JFK Airport weather data

Our Data: Citi Bike Trip Data

Information on each trip taken

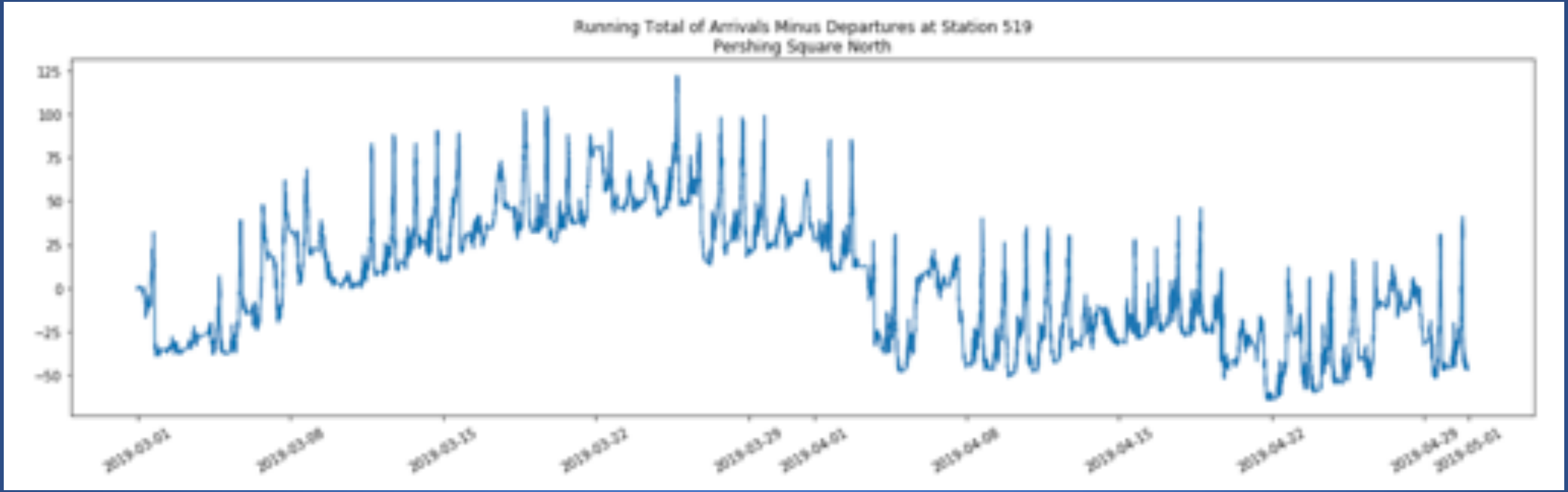
- Starting Station
- Ending Station
- Start Time of Trip
- Stop Time of Trip
- and much more...

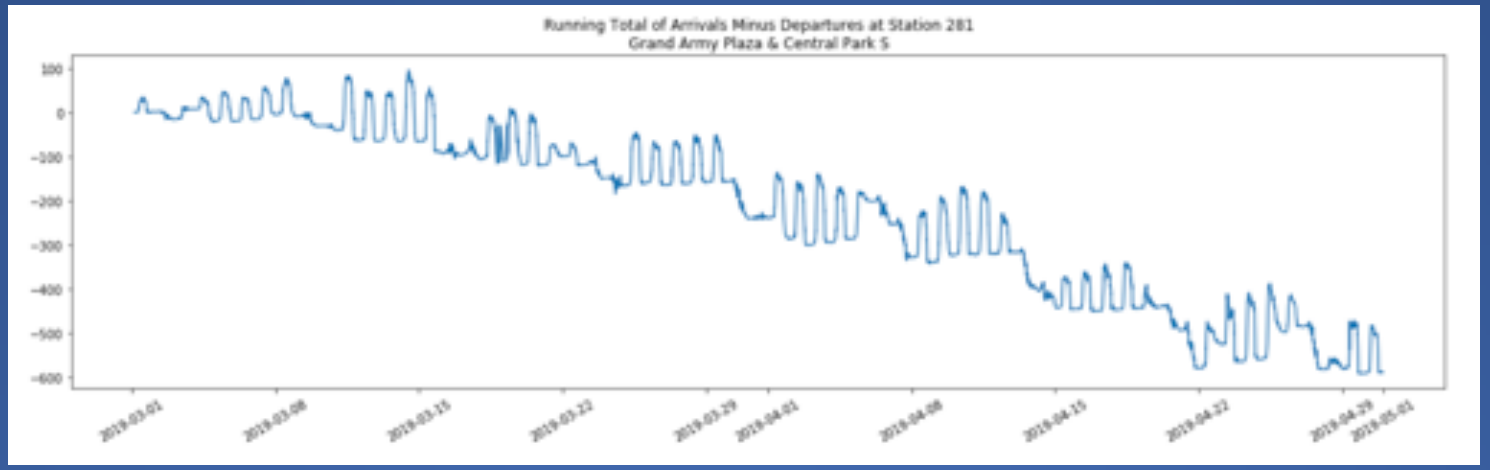
E 17 St & Broadway - “Arrivals minus departures”

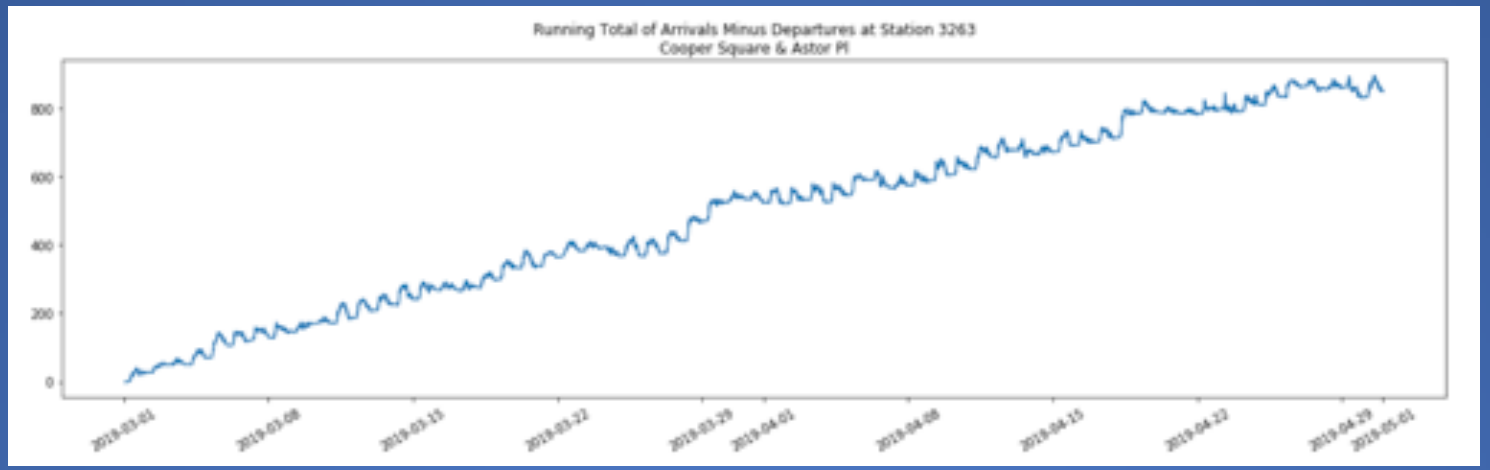
Dataframe containing trip data for E 17 St & Broadway,
the fourth most frequented station.

	starttime	stoptime	start station id	start station name	end station id	end station name	arrive_depart	total_arrive_depart
0	2019-03-01 05:12:41.608	2019-03-01 05:27:07.540	342.0	Columbia St & Rivington St	497.0	E 17 St & Broadway	1	1
1	2019-03-01 06:16:04.964	2019-03-01 06:20:43.263	446.0	W 24 St & 7 Ave	497.0	E 17 St & Broadway	1	2
2	2019-03-01 06:08:36.469	2019-03-01 06:15:52.851	497.0	E 17 St & Broadway	3427.0	Lafayette St & Jersey St	-1	1
3	2019-03-01 06:42:13.258	2019-03-01 06:46:29.974	2003.0	1 Ave & E 18 St	497.0	E 17 St & Broadway	1	2
4	2019-03-01 06:47:49.276	2019-03-01 06:49:56.101	3463.0	E 16 St & Irving Pl	497.0	E 17 St & Broadway	1	3

Most Frequented Station: Pershing Square North

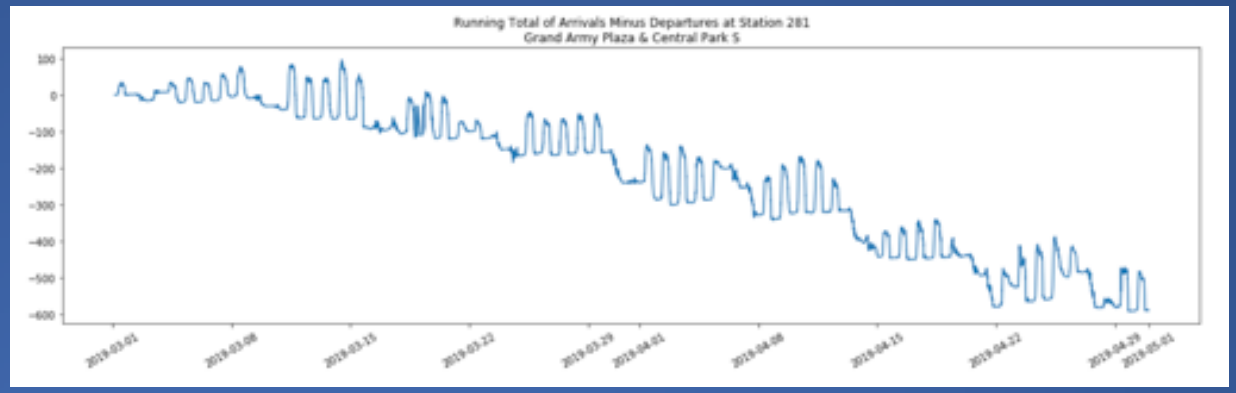



Running Total of Arrivals Minus Departures at Station 281
Grand Army Plaza & Central Park S
100
0
−100
−200
−300
−400
−500
−600
2019-03-01
2019-03-08
2019-03-15
2019-03-22
2019-03-29
2019-04-01
2019-04-08
2019-04-15
2019-04-22
2019-04-29
2019-05-01

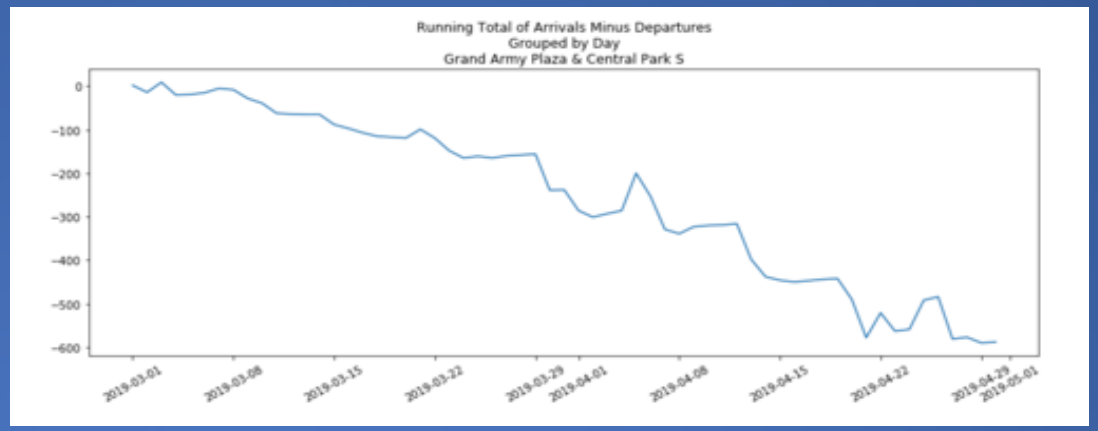

Running Total of Arrivals Minus Departures at Station 3263
Cooper Square & Astor Pl
800
600
400
200
0
2019-03-01
2019-03-08
2019-03-15
2019-03-22
2019-03-29
2019-04-01
2019-04-08
2019-04-15
2019-04-22
2019-04-29
2019-05-01


Running Total of Arrivals Minus Departures at Station 539
Metropolitan Ave & Bedford Ave
400
300
200
100
0
2019-03-01
2019-03-08
2019-03-15
2019-03-22
2019-03-29
2019-04-01
2019-04-08
2019-04-15
2019-04-22
2019-04-29
2019-05-01

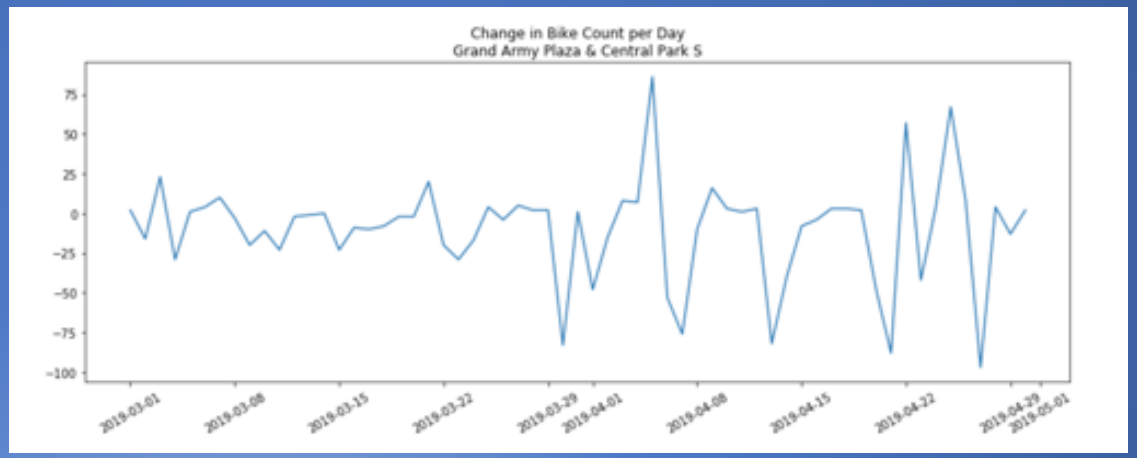
Continuous



Running Total Grouped by Day

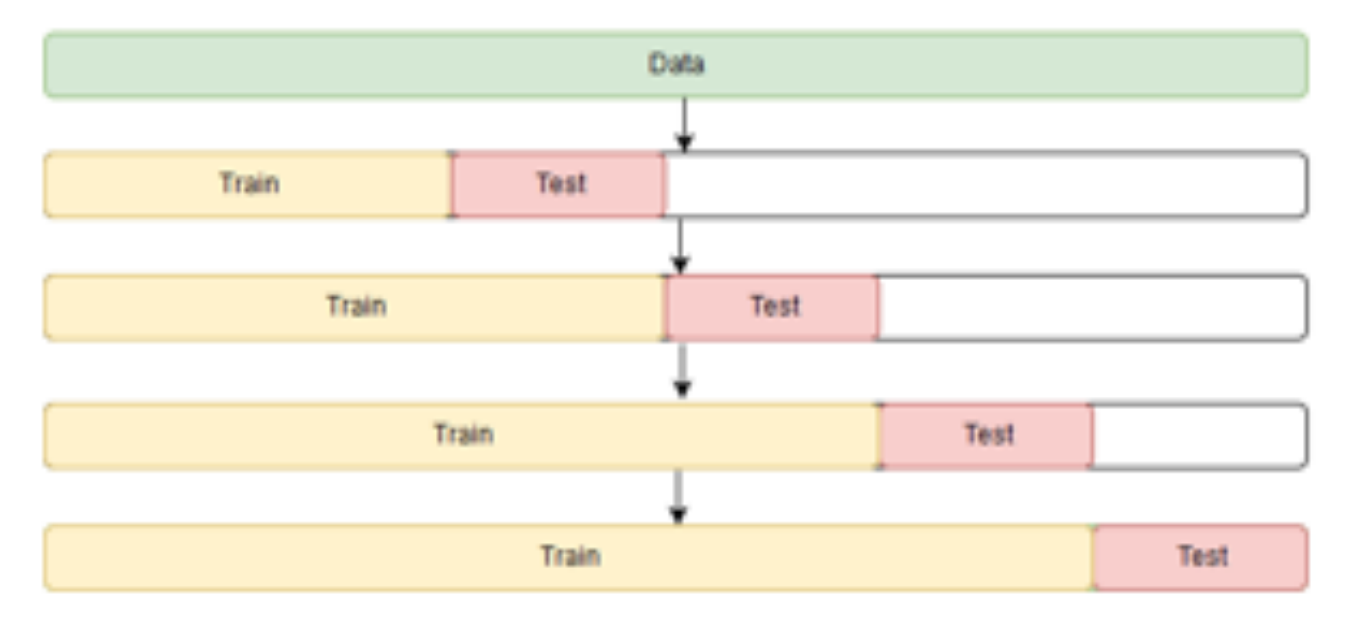


Change in Bike Count versus Day

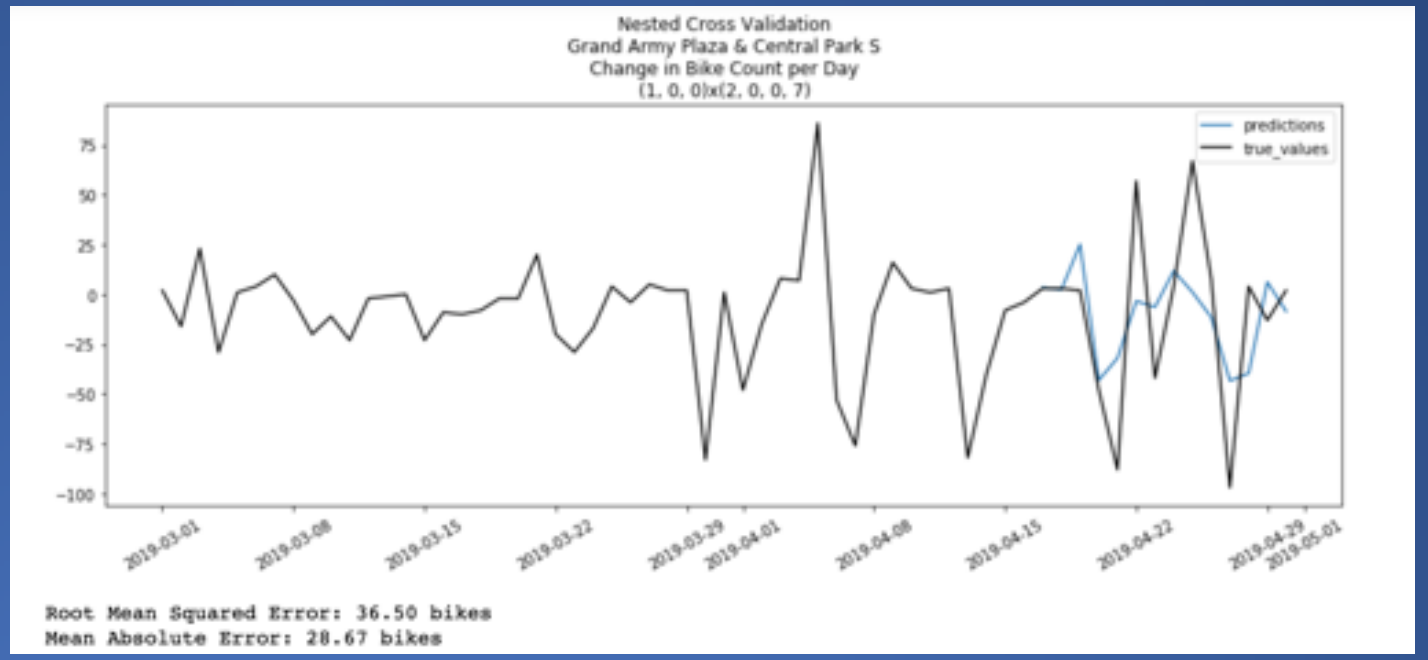


Best SARIMA model Criteria

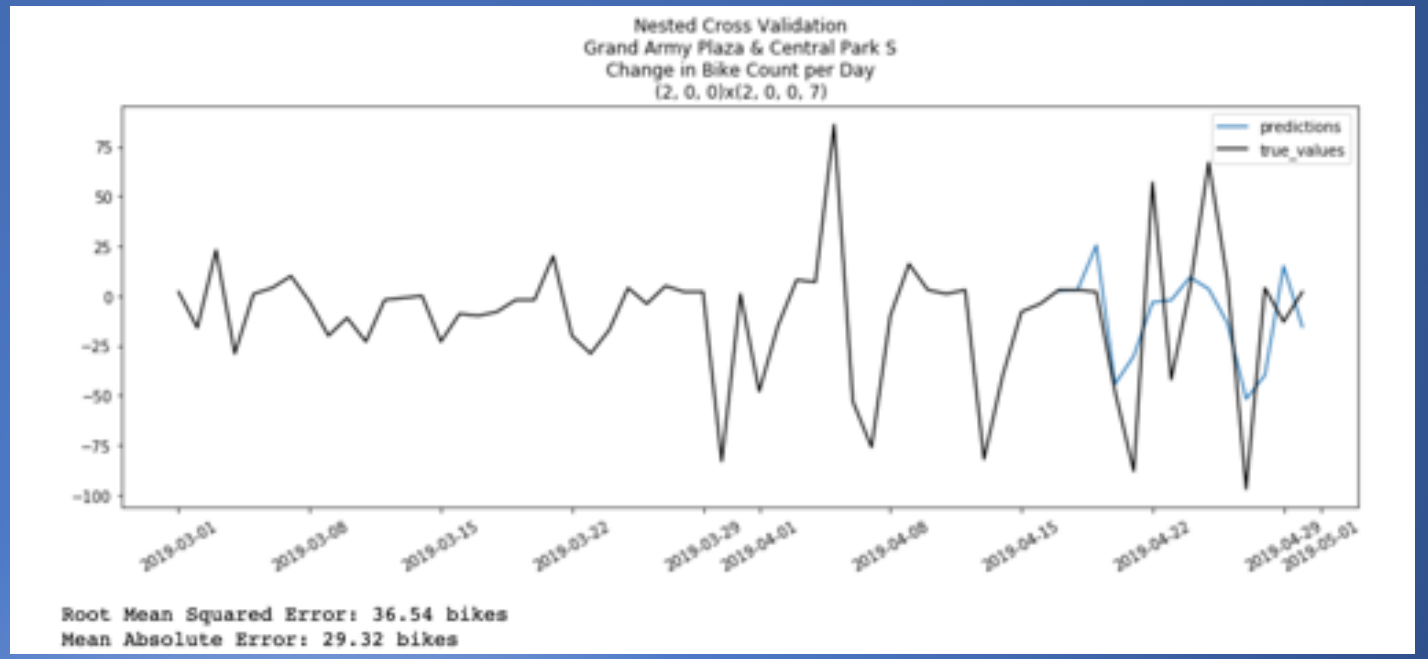
- Three Lowest Akaike Information Criteria (AIC)
  - When modeling training data: Mar & Apr 2019
- Compared via Nested Cross Validation
- RMSE and MAE used as metrics



Lowest AIC:



Second Lowest AIC:



Predict on Testing Data


```
Root Mean Squared Error: 41.93 bikes
Mean Absolute Error: 31.85 bikes
```

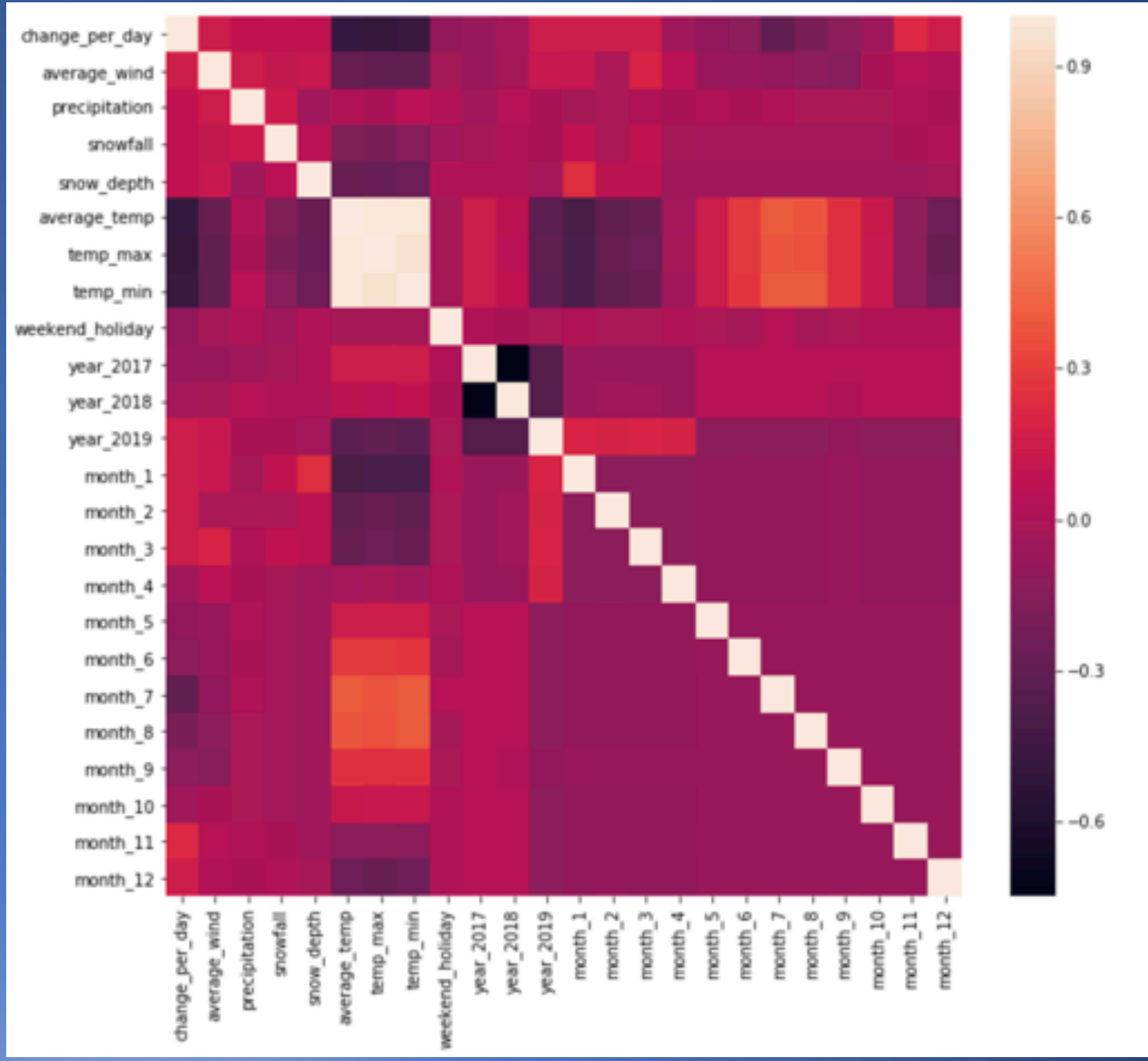
Regression Analysis

- Use weather data
  - Min / Avg / Max Temperature
  - Average Windspeed
  - Precipitation
  - Snowfall / Snow Depth
- Weekday vs. weekend/holiday
- Month

- No stationary requirement
- Use more training data (Jan 2017-Apr 2019)

Correlation Heatmap

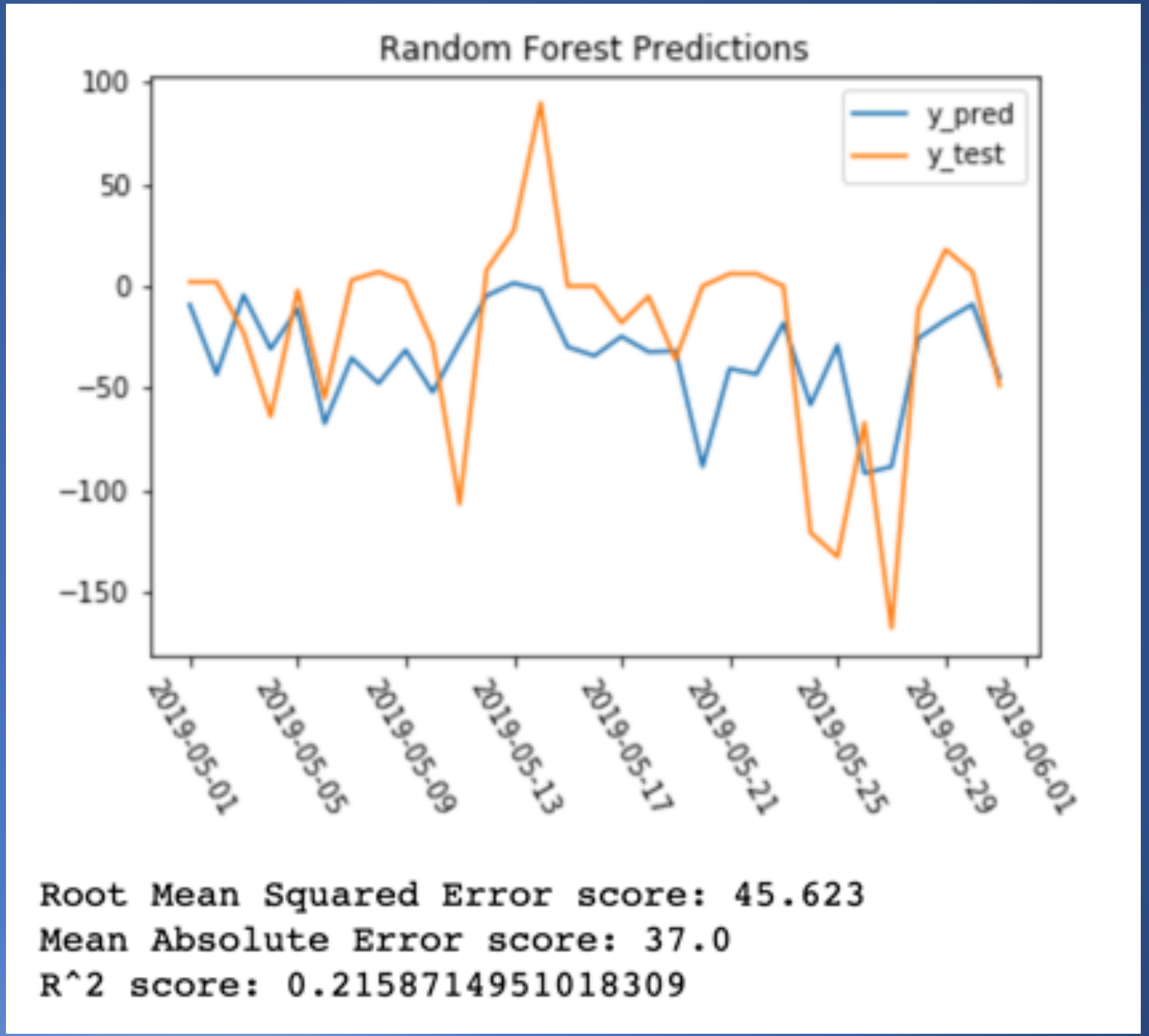
Station 281:
Grand Army Plaza &
Central Park South



Random Forest

Station 281:
Grand Army Plaza &
Central Park South

May 2019 Results

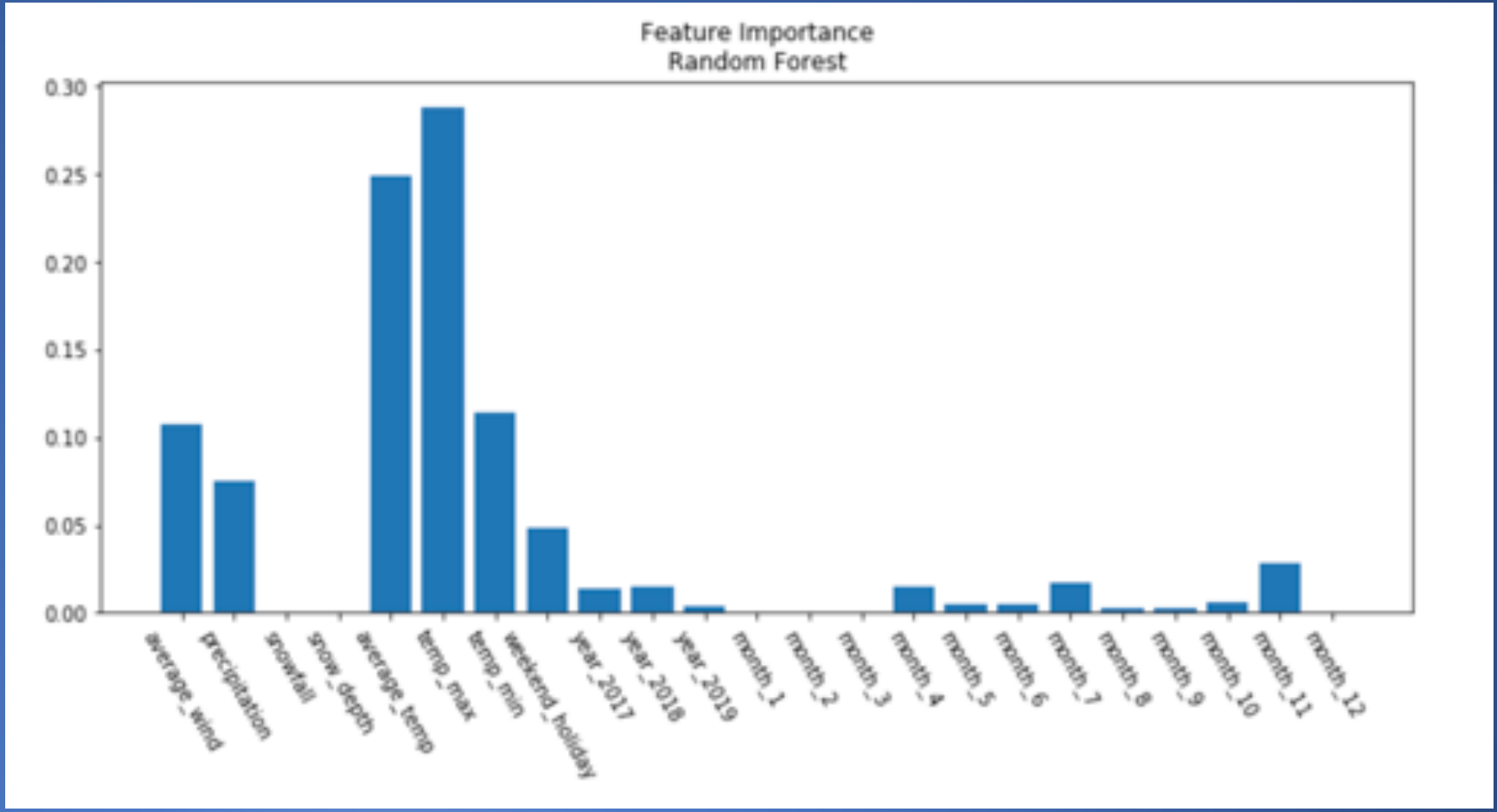


Random Forest

Station 281:
Grand Army Plaza &
Central Park South

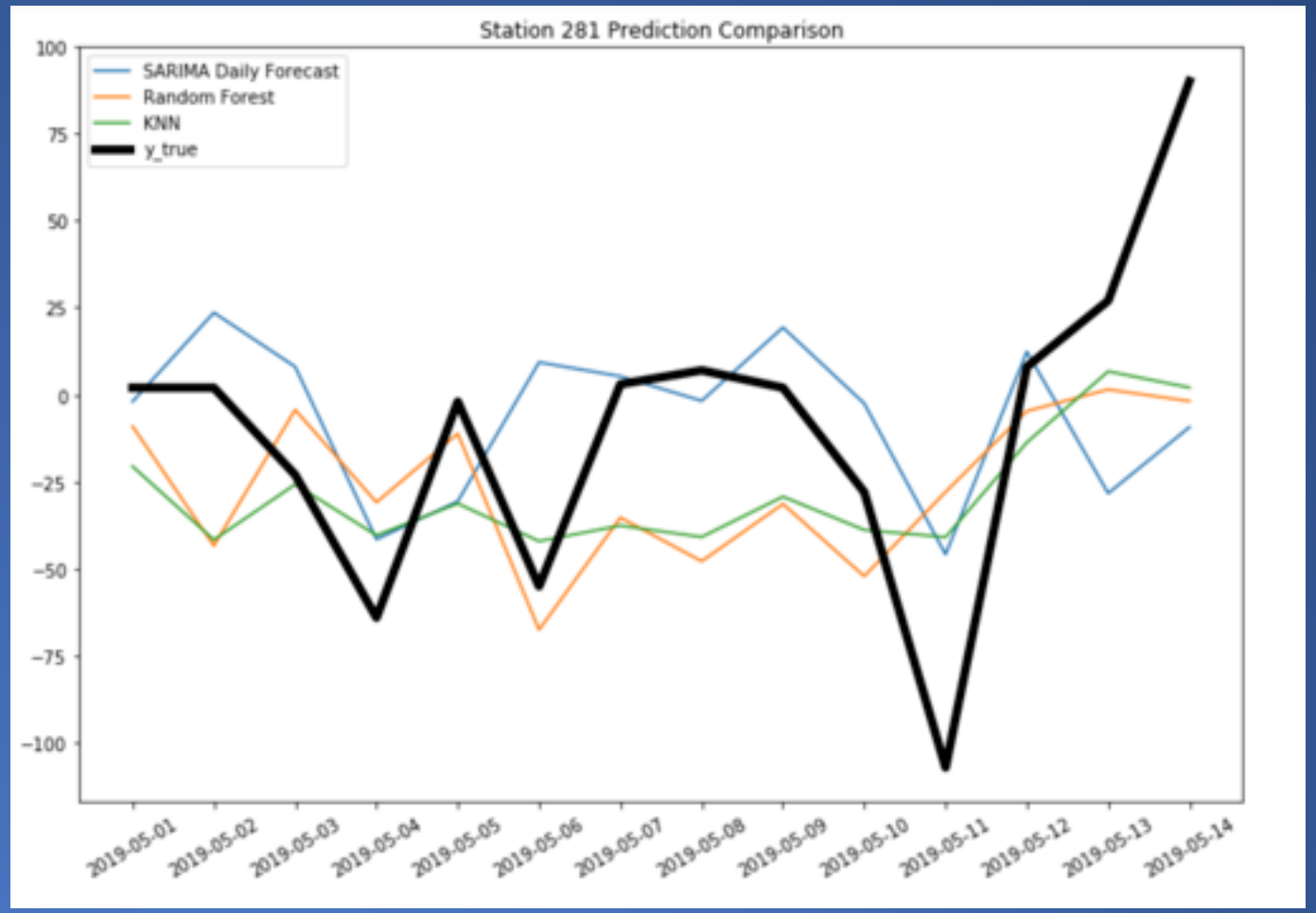
May 2019 Results

```
Root Mean Squared Error score: 45.623
Mean Absolute Error score: 37.0
R^2 score: 0.2158714951018309
```



Model Comparison

Station 281:
Grand Army Plaza &
Central Park South



No clear winner…

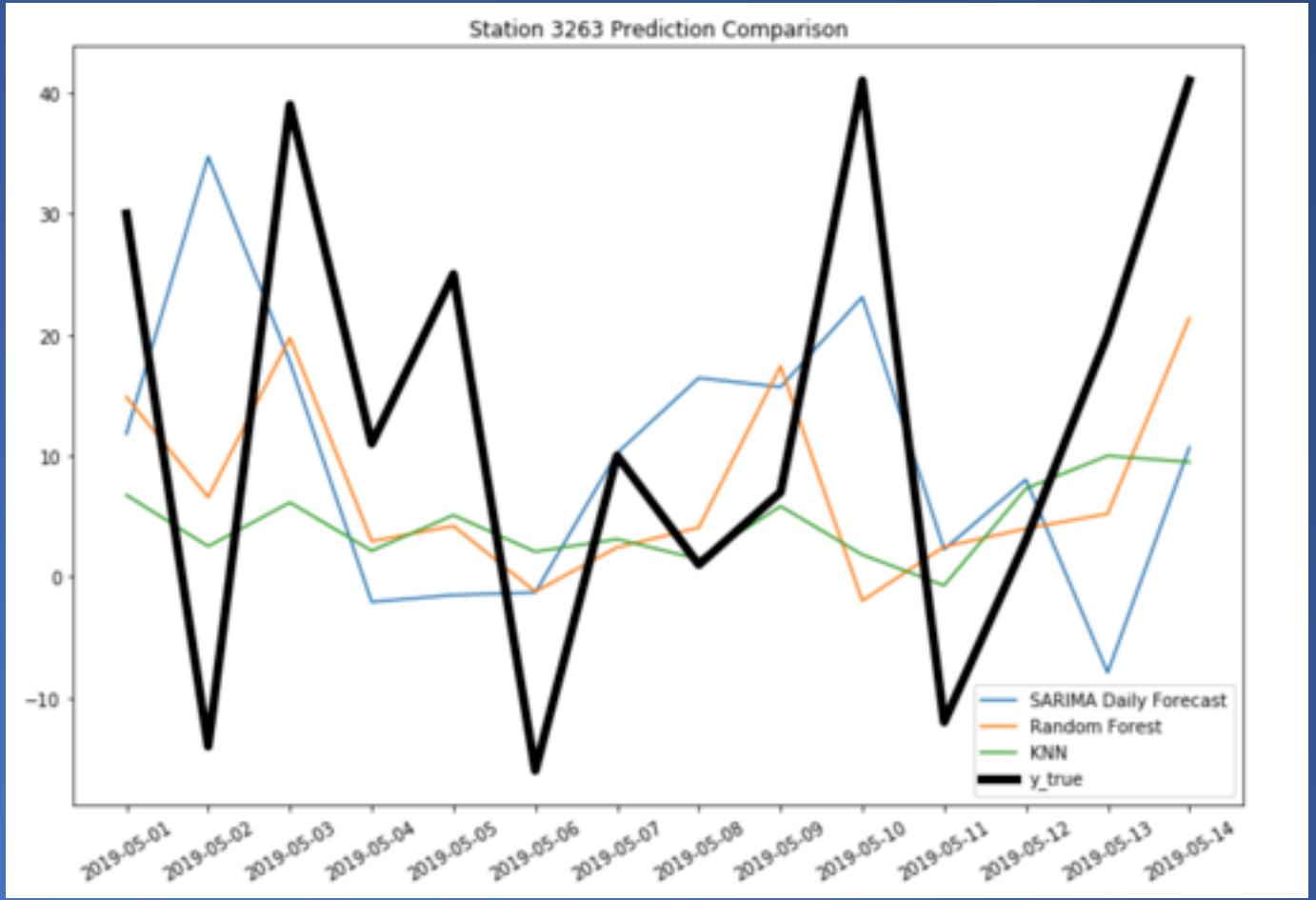
Let’s choose SARIMA

```
Root Mean Squared Error for first two weeks of May 2019
SARIMA Daily Forecast RMSE: 41.927
Random Forest RMSE: 42.699
KNN RMSE: 39.661

Mean Absolute Error for first two weeks of May 2019
SARIMA Daily Forecast MAE: 31.847
Random Forest MAE: 34.964
KNN MAE: 32.952
```

Model Comparison

Station 3263:
Cooper Square & Astor Pl



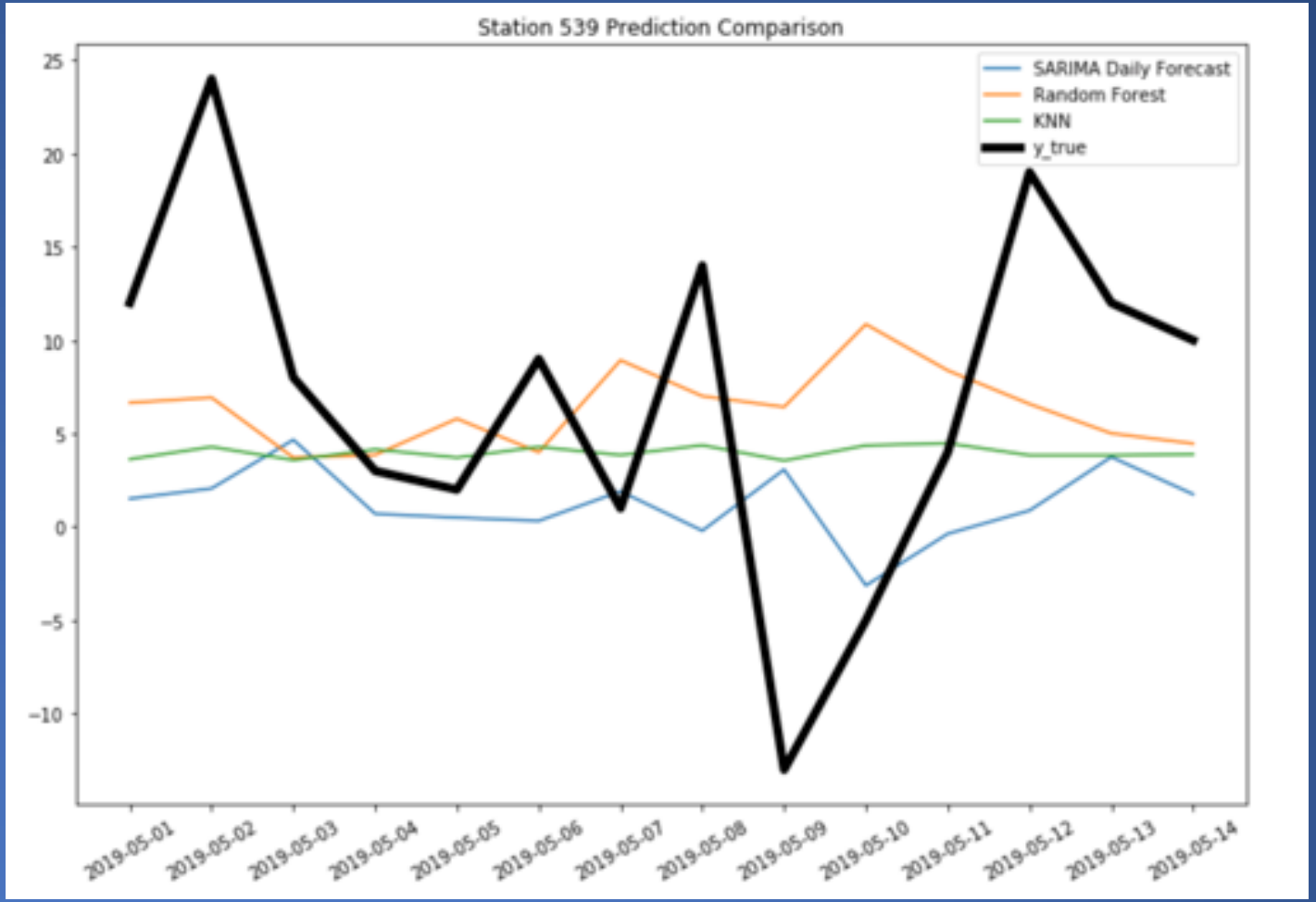
Let's choose Random Forest

```
Root Mean Squared Error for first two weeks of May 2019
SARIMA Daily Forecast RMSE: 22.047
Random Forest RMSE: 18.078
KNN RMSE: 19.851

Mean Absolute Error for first two weeks of May 2019
SARIMA Daily Forecast MAE: 18.725
Random Forest MAE: 15.178
KNN MAE: 16.017
```

Model Comparison

Station 539:
Metropolitan Ave &
Bedford Ave



No clear winner…

Let’s choose KNN

```
Root Mean Squared Error for first two weeks of May 2019
SARIMA Daily Forecast RMSE: 10.799
Random Forest RMSE: 9.8987
KNN RMSE: 9.6402

Mean Absolute Error for first two weeks of May 2019
SARIMA Daily Forecast MAE: 8.5888
Random Forest MAE: 8.2767
KNN MAE: 7.7419
```

Why such poor performance?

- Assumptions of data
- Different time step?
- Long Short-Term Memory recurrent neural networks
- More features for regression?


citi bike