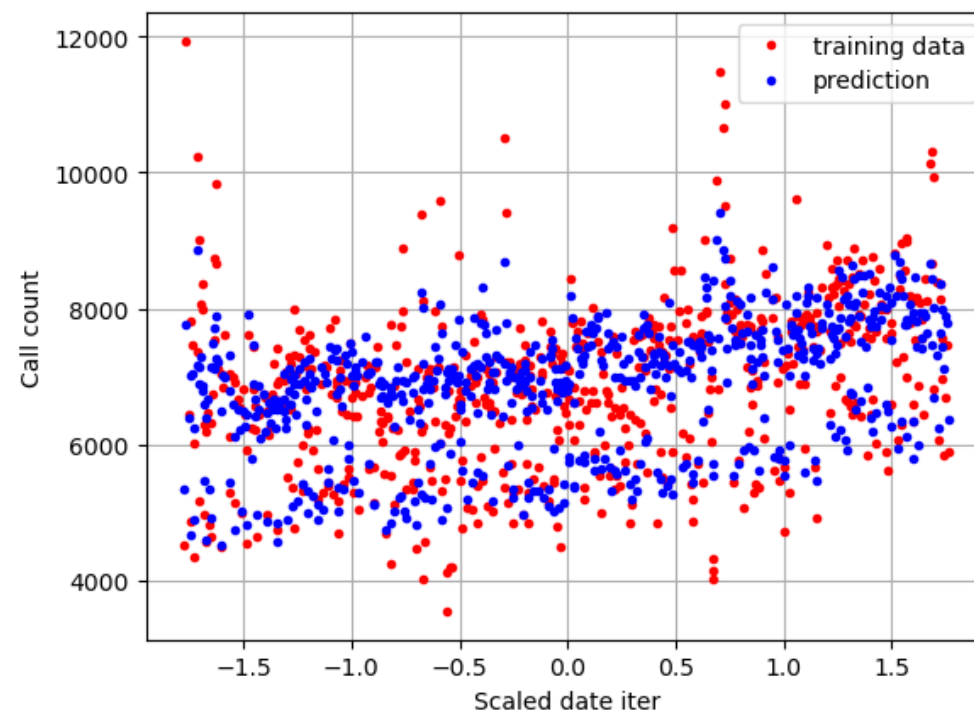
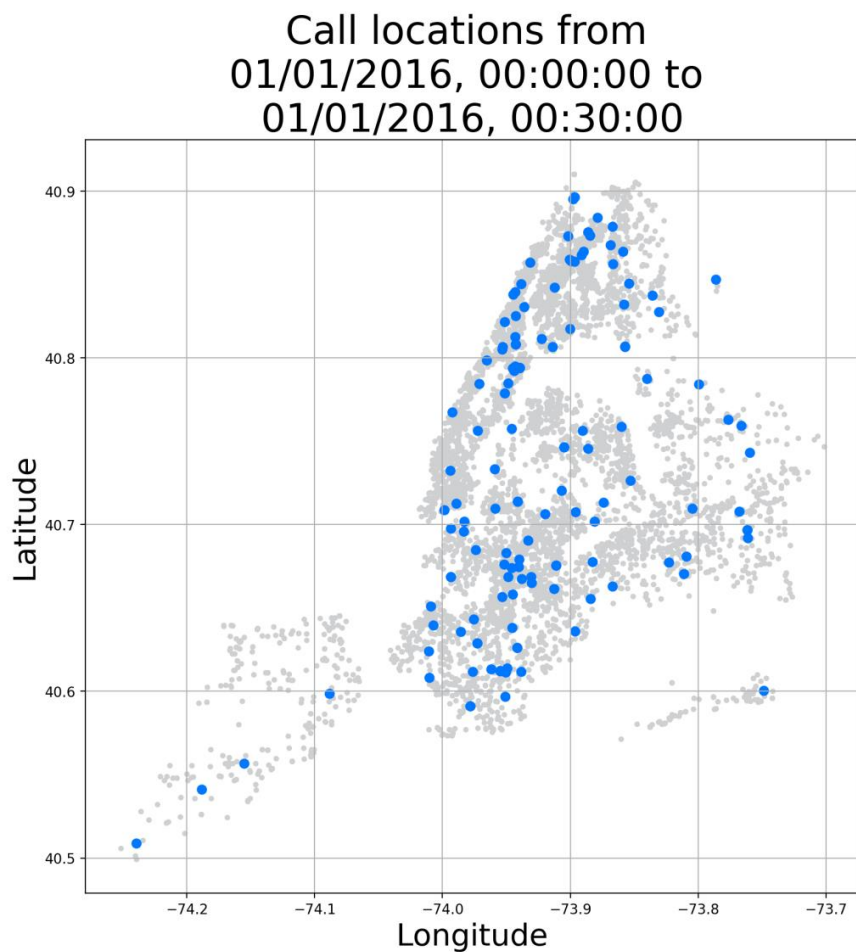


Data Science Case Study

311 Service Requests & Weather Data
























Eric Schwen
June 29th, 2023

311 Service Requests – Data Exploration

What is a 311 request/call?

- “A Service Request is your request for the City to provide you with assistance, perform an inspection, or address a problem”.
- Non-emergency city service requests
- Dataset has 3 years of data (2016, 2017, 2018)
- 7,631,721 total calls; 42 columns
- ~7000 calls per day
- Public dataset available

From nyc.gov:

NYC 311		
 Benefits & Support	 Business & Consumers	 Courts & Law
 Culture & Recreation	 Education	 Employment
 Environment	 Government & Elections	 Health
 Housing & Buildings	 Noise	 Pets, Pests & Wildlife
 Public Safety	 Records	 Sidewalks, Streets, Highways
 Taxes	 Transportation	 Trash & Recycling
 A to Z	 About 311	 Get The NYC311 App
Report Problems ▶	Look Up Service Requests ▶	Make Payments ▶

Eventual Goal: Predict daily inbound 311 calls for the next 7 days

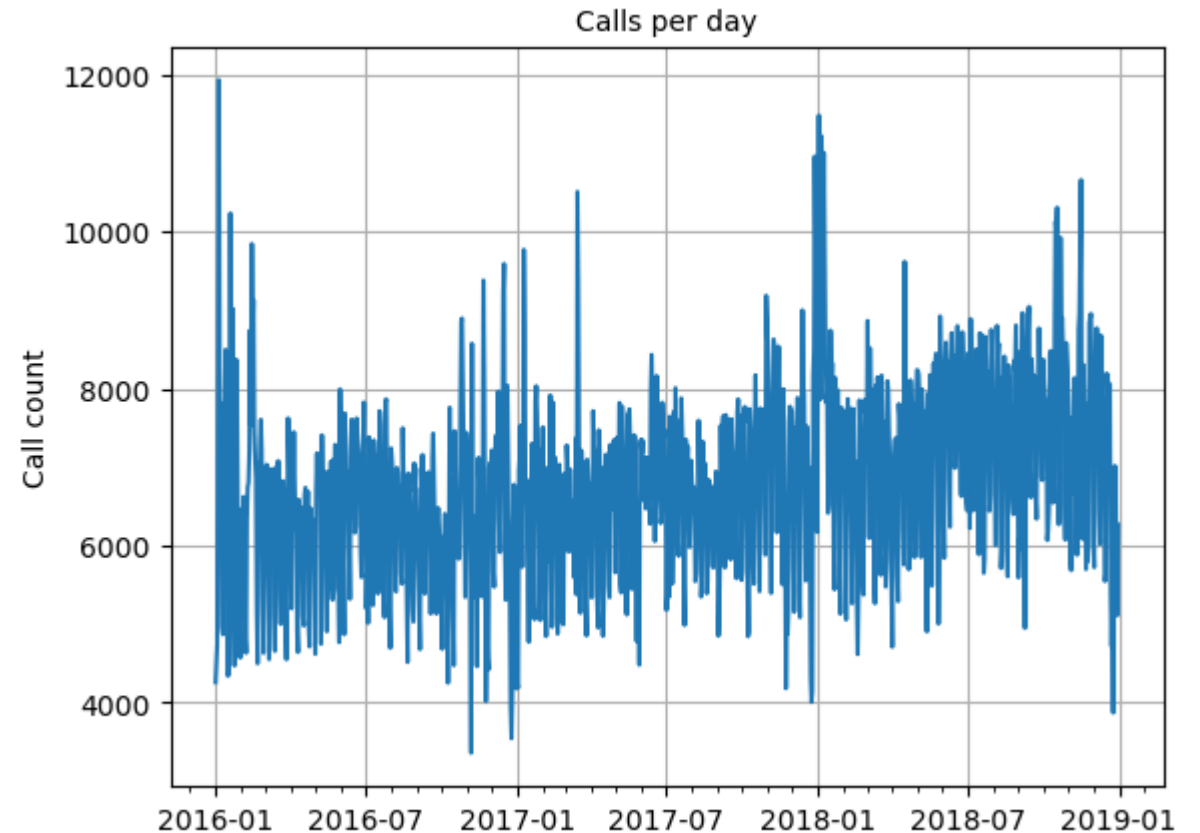
What data do we have? — Single Request Row

• Unique Key	38237851
• Created Date	2018-01-17 14:39:00
• Closed Date	01/24/2018 12:00:00 AM
• Agency	DSNY
• Agency Name	Department of Sanitation
• Complaint Type	Request Large Bulky Item Collection
• Descriptor	Request Large Bulky Item Collection
• Location Type	Sidewalk
• Incident Zip	11222
• Incident Address	95 GREENPOINT AVENUE
• Street Name	GREENPOINT AVENUE
• Cross Street	1FRANKLIN STREET
• Cross Street	2MANHATTAN AVENUE
• Intersection Street 1	NaN
• Intersection Street 2	NaN
• Address Type	ADDRESS
• City	BROOKLYN
• Landmark	NaN
• Facility Type	NaN
• Status	Closed
• Due Date	NaN
• Resolution Description	NaN

• Resolution Action Updated Date	01/24/2018 12:00:00 AM
• Community Board	01 BROOKLYN
• BBL	3025580072
• Borough	BROOKLYN
• X Coordinate (State Plane)	996371.0
• Y Coordinate (State Plane)	205243.0
• Open Data Channel Type	PHONE
• Park Facility Name	Unspecified
• Park Borough	BROOKLYN
• Vehicle Type	NaN
• Taxi Company Borough	NaN
• Taxi Pick Up Location	NaN
• Bridge Highway Name	NaN
• Bridge Highway Direction	NaN
• Road Ramp	NaN
• Bridge Highway Segment	NaN
• Latitude	40.730013
• Longitude	-73.956267
• Location	(40.73001299919553, -73.95626650502489)
• Created Year	2018.0

Daily 311 call totals: Grouping by Created Date

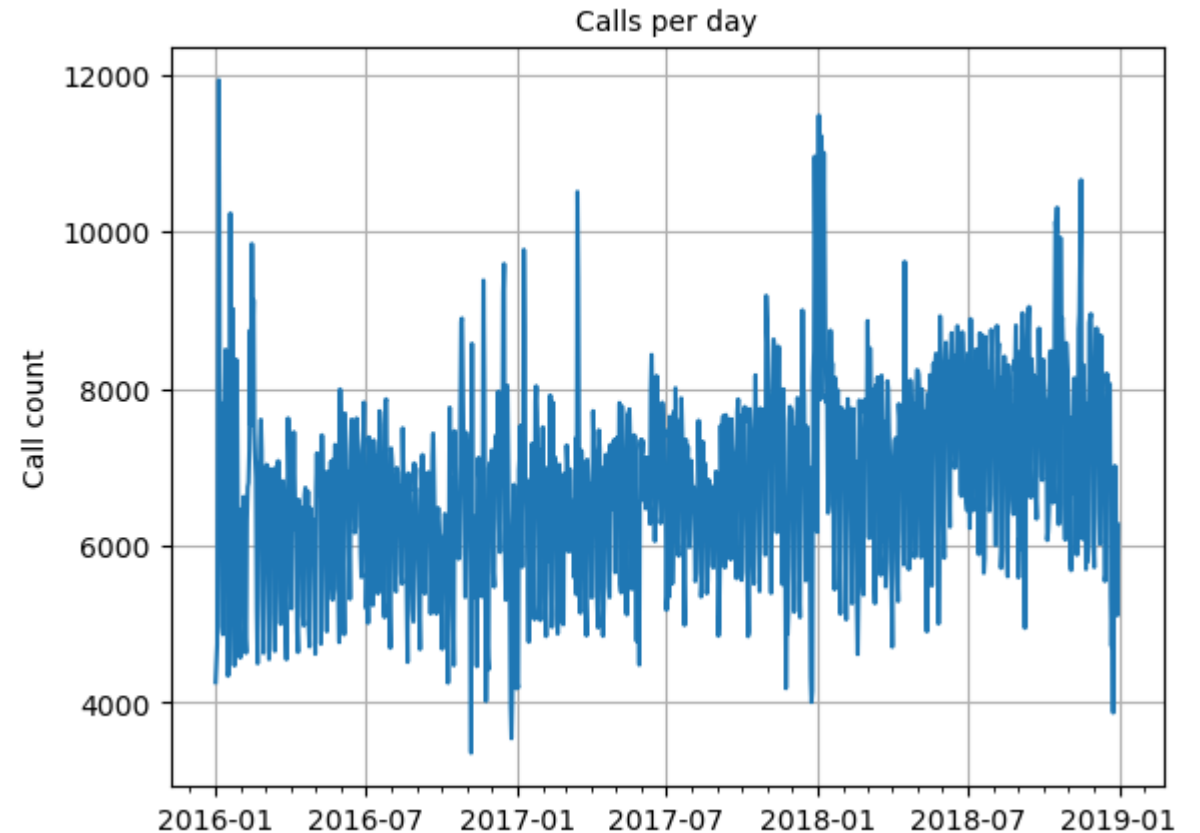
- Import data into pandas DataFrame
- Create an array of all dates that the data covers
- Loop through dates and add call counts
 - Slice DataFrame to locate specific date
 - Record number of calls (length of sliced DataFrame) for each day
- Plot results: Daily 311 calls vs. Date



Daily 311 calls over time: First Impressions

1. Noise/variation over days
 - Also possible weekday vs. weekend difference?
2. Occasional spikes/sharp peaks (ex: after New Years in 2016 and 2018)
 - Bad weather? Other events?
3. General increasing trend over time
4. Seasonal Variation – slowly oscillates?

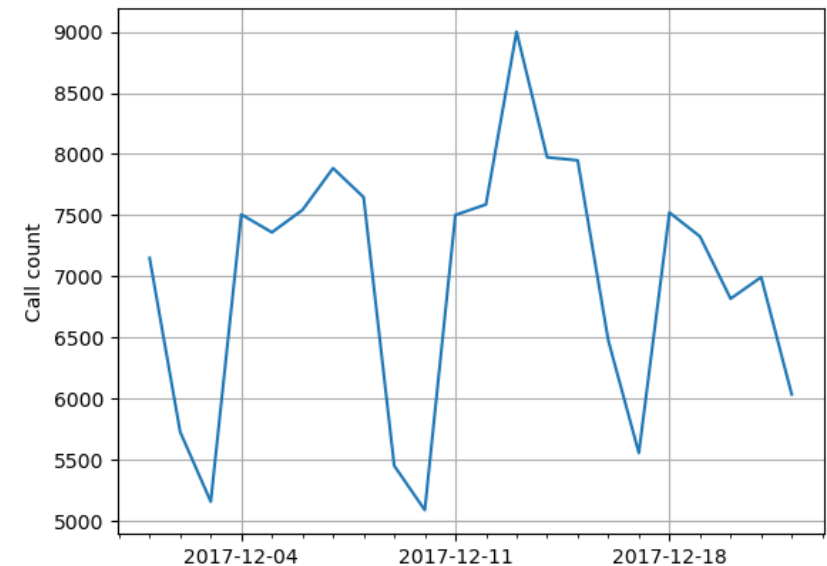
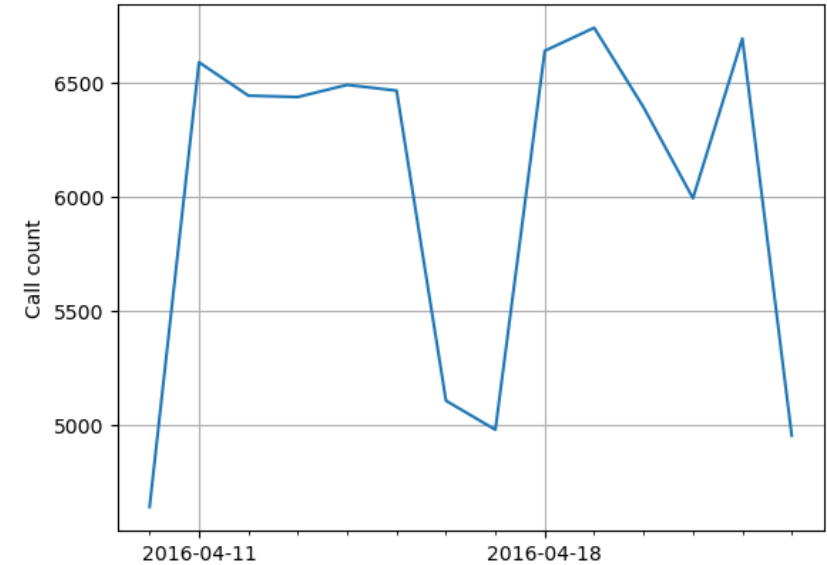
These observations guide my further explorations.



Weekday variation in daily calls?

- Plot daily calls vs time over a 2 weeks in 2016
 - Vertical lines are Mondays
 - 311 calls drop of noticeably on weekends (Saturdays and Sundays).
- Same trend is seen for 3 weeks plotted in December 2017

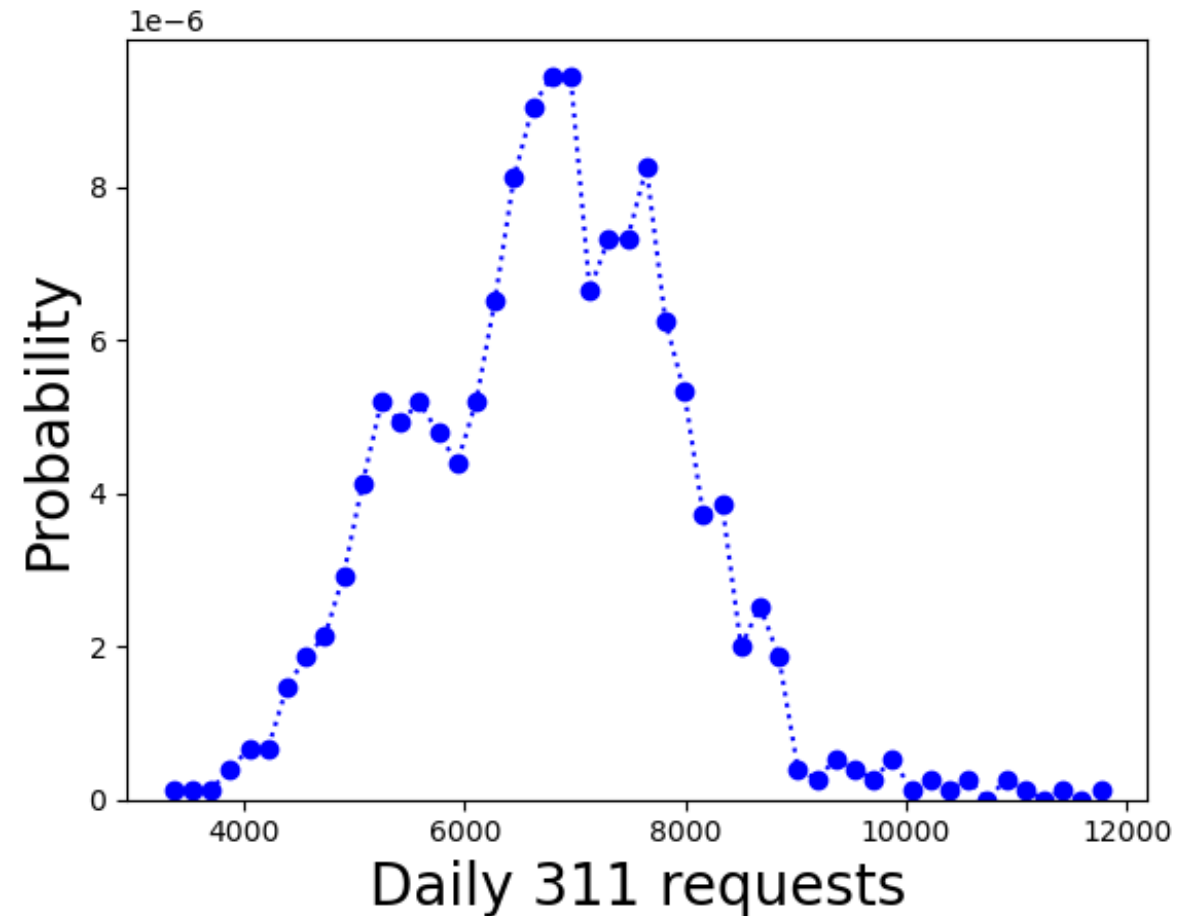
Conclusion: Daily call count prediction should include day of week



Daily calls: Probability distribution

Probability distribution of number of daily 311 calls.

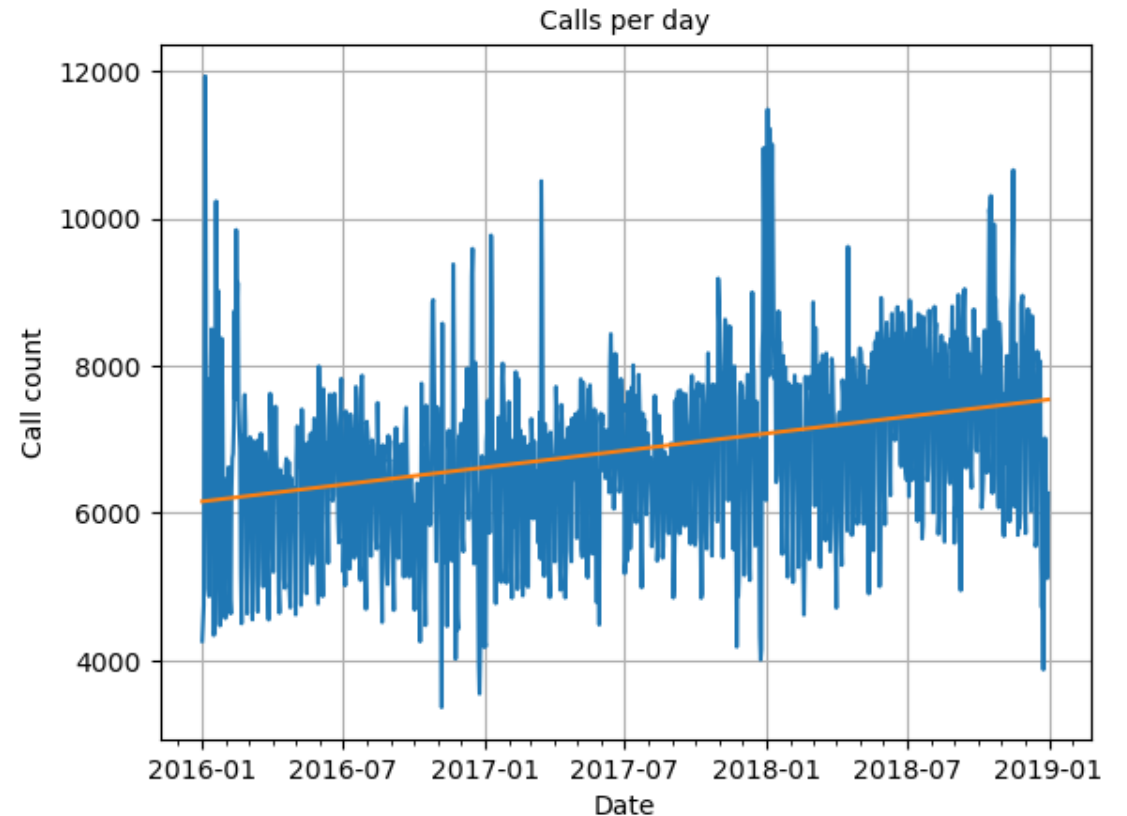
- Shape looks mostly like a normal distribution (though is a bit noisy with this binning)
- Obvious long tail at large call numbers
 - Extreme events/extreme weather?



Daily calls: increasing trend over time?

- Service requests appear to increase over time
- Check if that is true using a simple linear regression.
 - `sklearn.linear_model`
 - `LinearRegression()`
 - Create `date_iter` variable to track time (number of days passed since 2016-01-01)
- Linear regression results:
 - R^2 : 0.115
 - intercept: 6383.96
 - slope: 0.02712 (increase in call count per day)

Conclusion: 311 requests increase over time in this dataset. I should include this variation in any modeling.



Next variables of interest?

- **Complaint type**
 - 371 unique complaint types
 - Could be interesting to look at top few, but likely won't capture whole dataset well
- **Agency Name**
 - 1373 unique agency names
 - Large number of agency names difficult to work with
 - Many unique agency names are potentially redundant, but grouping would be tedious
- **Agency**
 - 30 Unique agencies
 - Look at top agencies to see how calls vary over time!

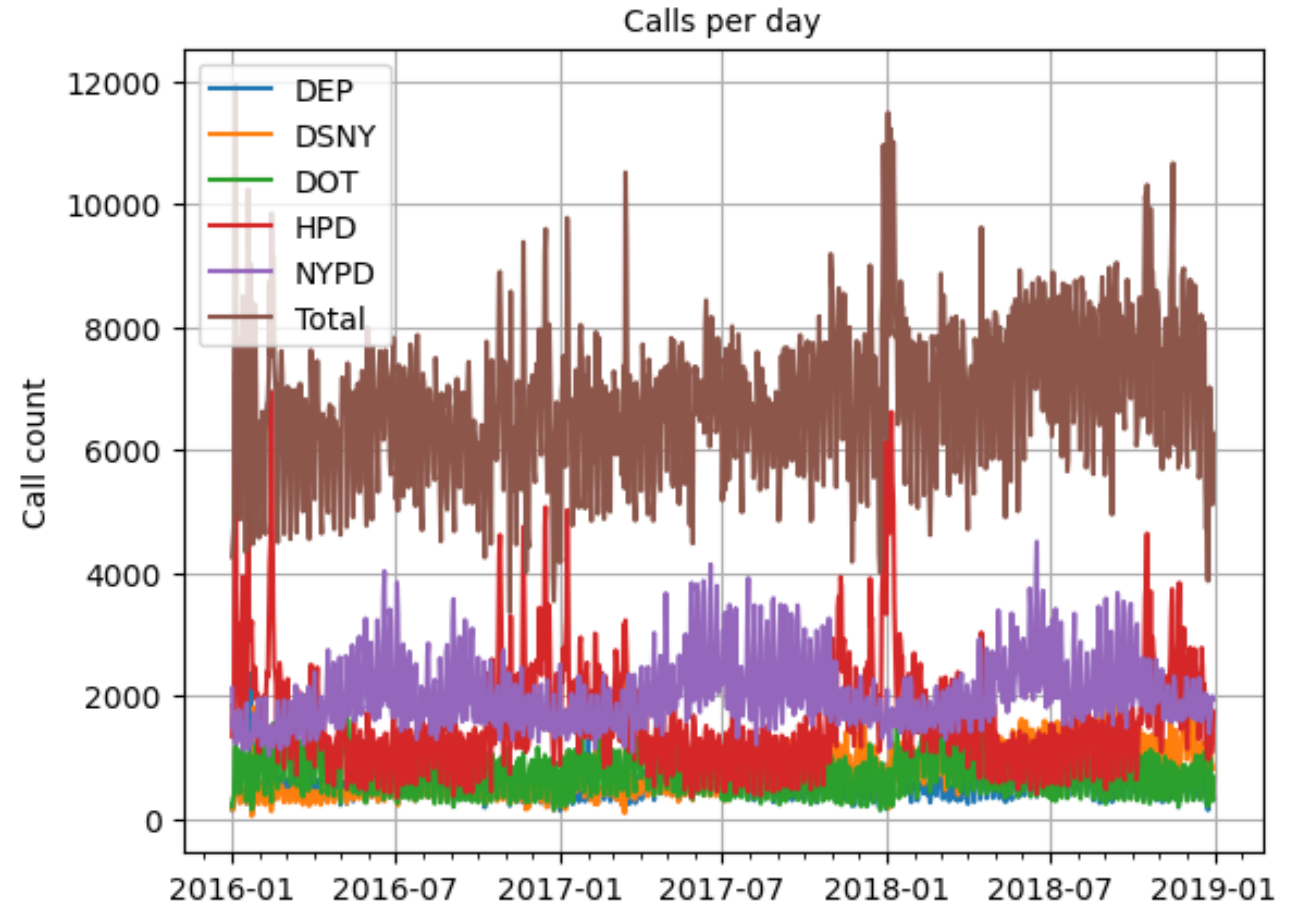
Agencies – What can Agency tell us about 311 calls?

- Main questions:
 - Which agencies get the most calls?
 - How do those totals change over time?

- Slice DataFrame by Agency to get call totals

<u>Agency</u>	<u>Complaint total</u>
NYPD	2167548.0
HPD	1778185.0
DOT	877335.0
DSNY	831025.0
DEP	585582.0
DOB	399518.0
DPR	316820.0
DOHMH	201343.0
DOF	149812.0

- Slice DataFrame by both Agency and Date to get daily totals over time (plotted on the right).

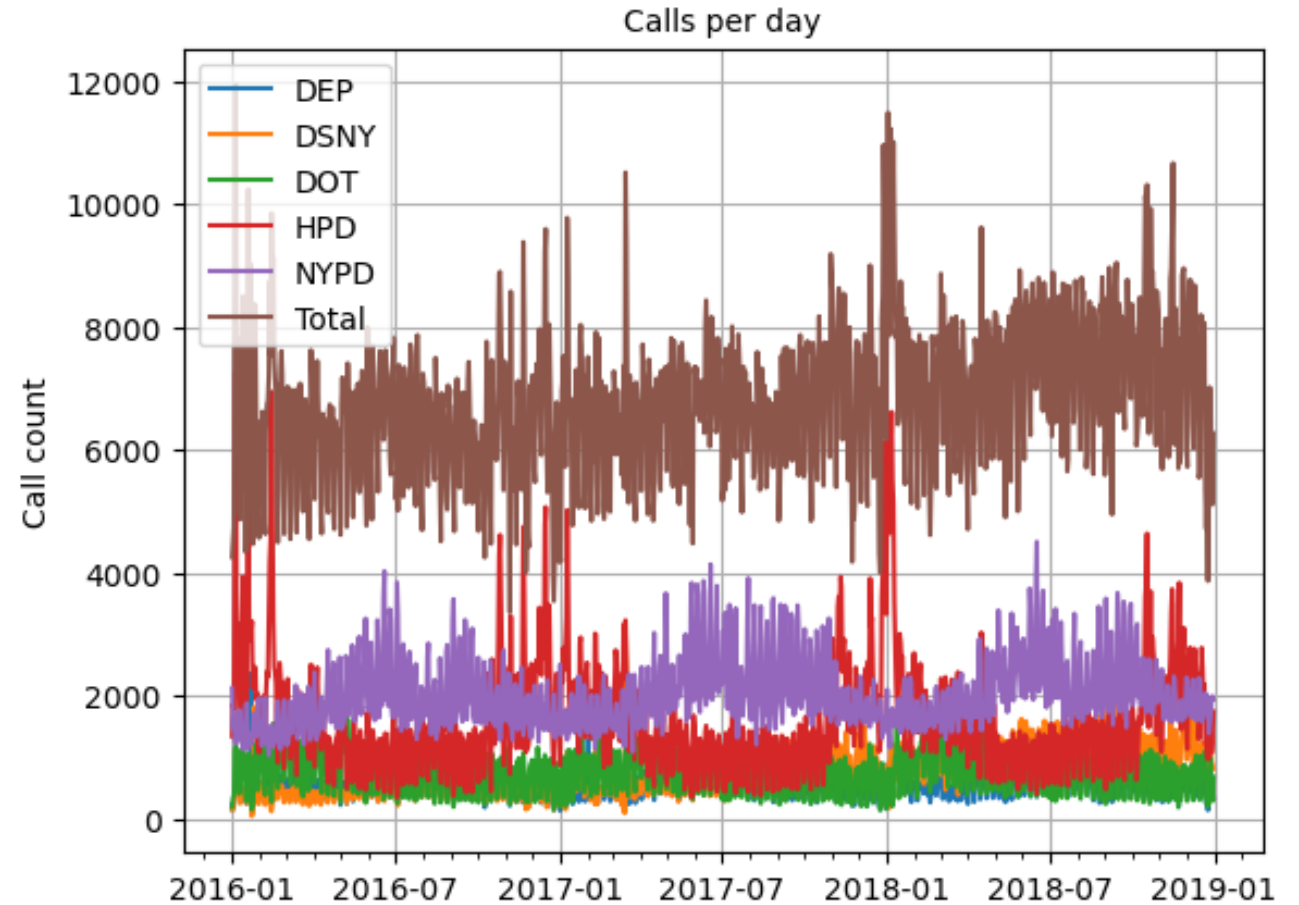


How do Agency totals vary over time?

Main Observations

- NYPD
 - Seasonal variation. Higher in summer.
- HPD (Housing Preservation and Development)
 - Generally higher in winter.
 - Spikes concurrent with spikes in total calls. Extreme weather events?

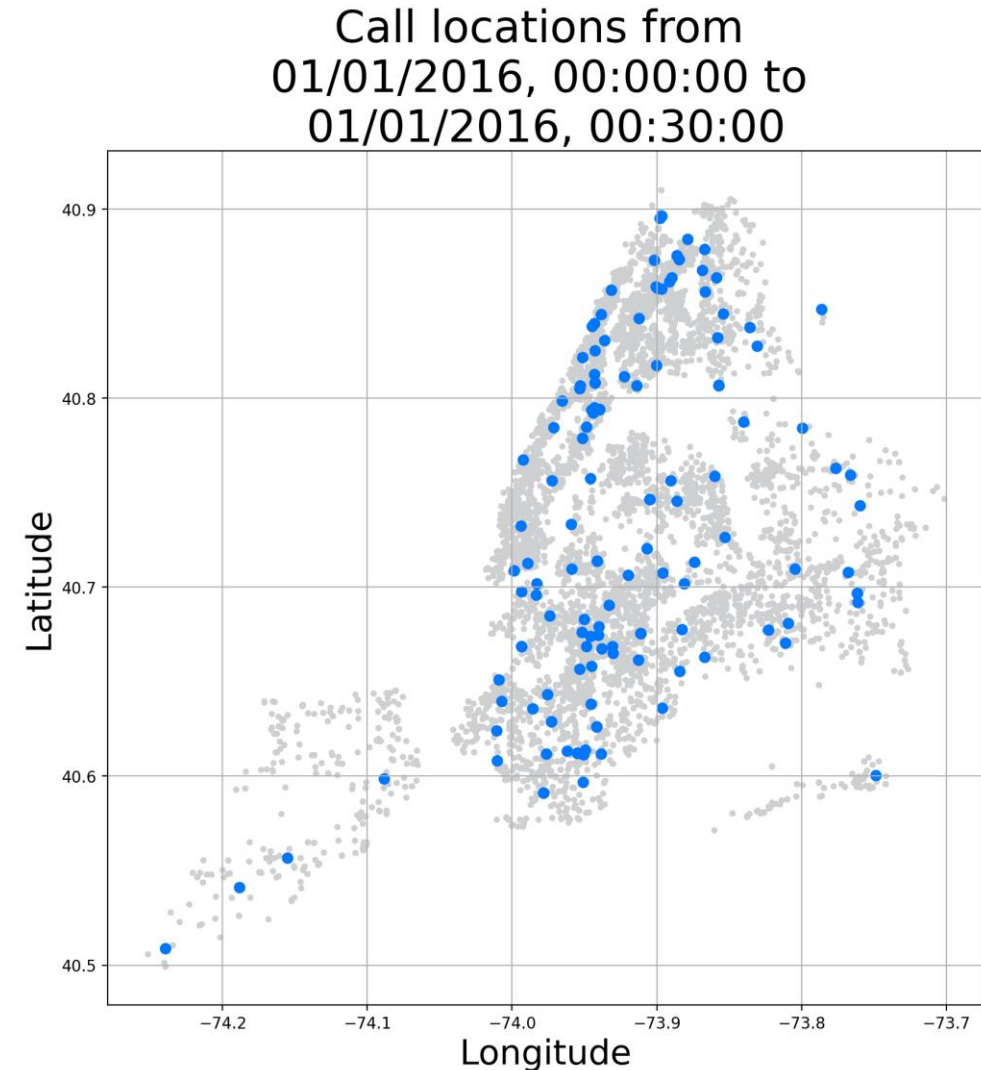
Conclusion: Daily Agency call totals are interesting. I could correlate or model them over time as they relate to weather (Though I haven't had the time to do so).



Location: interesting variable?

- Dataset has rather precise positional and time data
- Could be interesting to look further at Borough totals per day, etc. I haven't had time to investigate thoroughly.

I have a reasonable understanding of 311 data.
Move on to weather!



Weather Data – Data Exploration

- Public data from NOAA weather stations in New York State
- First row of dataset shown on right
- 160755 rows, 21 columns
- Dates range from Jan. 2010 to Nov. 2018
- Some columns have many NaN entries
- Some columns (Rain and SnowIce) are binary

Eventual Goal: Predict daily inbound 311 calls for the next 7 days. Weather may help predict 311 calls.

USAF	726228
WBAN	94740
StationName	ADIRONDACK REGIONAL ARPT
State	NY
Latitude	44.385
Longitude	-74.207
MeanTemp	27.6
MinTemp	24.8
MaxTemp	30.9
DewPoint	25.0
Precipitation	0.07
WindSpeed	1.3
MaxSustainedWind	6.0
Gust	NaN
Rain	0
SnowDepth	NaN
SnowIce	1
Year	2010
Month	1
Day	1

What columns are potentially useful or interesting?

- ❖ MaxTemp (as well as MinTemp, MeanTemp)
 - Will certainly vary over time. Plausibly affect 311 calls when very hot or cold
- ❖ Dew Point
 - Related to relative humidity and heat index
- ❖ Precipitation
 - Lots of rain or snow likely cause problems for the city
- ❖ WindSpeed (and MaxSustainedWind, Gust)
 - High winds maybe down power lines, etc.
- ❖ Latitude/Longitude (or StationName)
 - Will want to restrict analysis to NYC to compare with 311 calls in the city
- ❖ Year/Month/Day

USAF	726228
WBAN	94740
StationName	ADIRONDACK REGIONAL ARPT
State	NY
Latitude	44.385
Longitude	-74.207
MeanTemp	27.6
MinTemp	24.8
MaxTemp	30.9
DewPoint	25.0
Precipitation	0.07
WindSpeed	1.3
MaxSustainedWind	6.0
Gust	NaN
Rain	0
SnowDepth	NaN
SnowIce	1
Year	2010
Month	1
Day	1

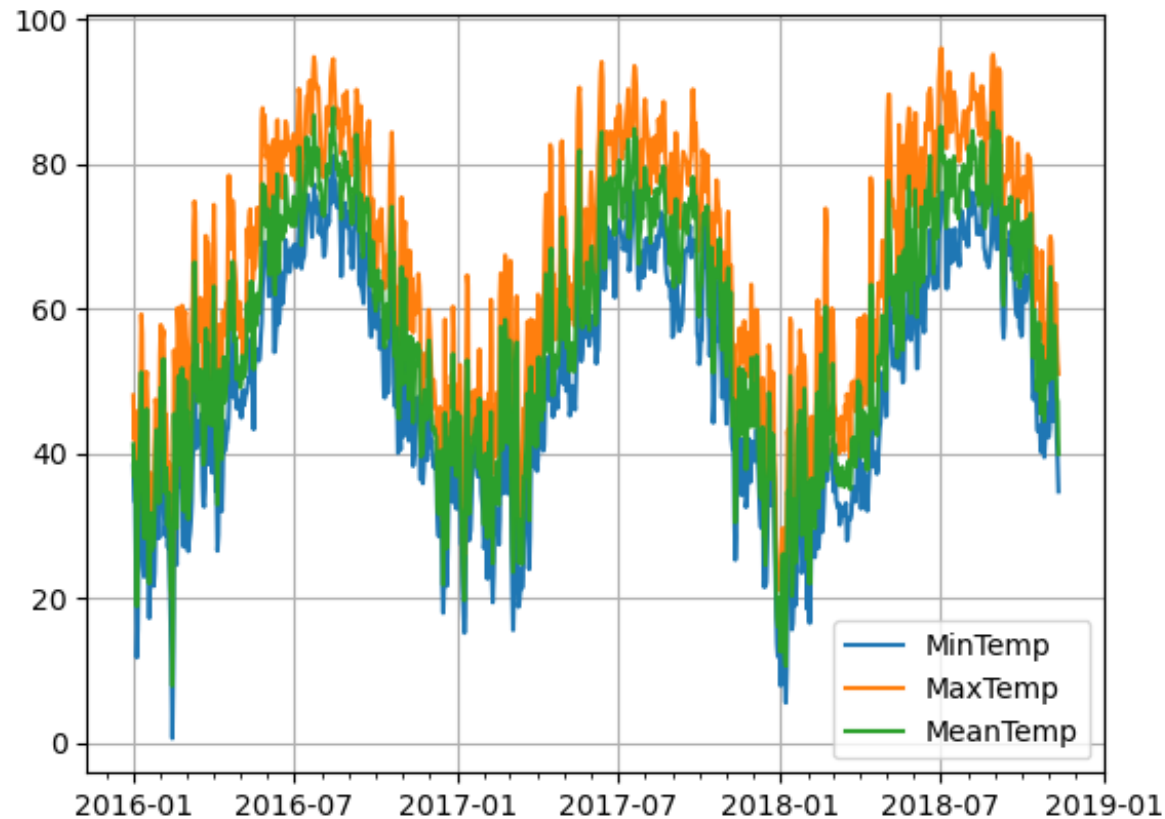
Extract daily NYC weather from full dataset

- Import weather data as DataFrame
- Slice DataFrame to select only stations in NYC
 - Use min. and max. Latitude and Longitude from 311 data
- Remaining weather data is very similar across stations in NYC
 - Max temperatures for a sample day are: 63.0, 66.0, 63.0, 64.9
 - Can reasonably take averages of the data to get weather for the day
- Loop through all dates
 - Average data across all NYC stations to get weather for that day

I now have daily weather data for NYC, which I can analyze and compare with 311 data

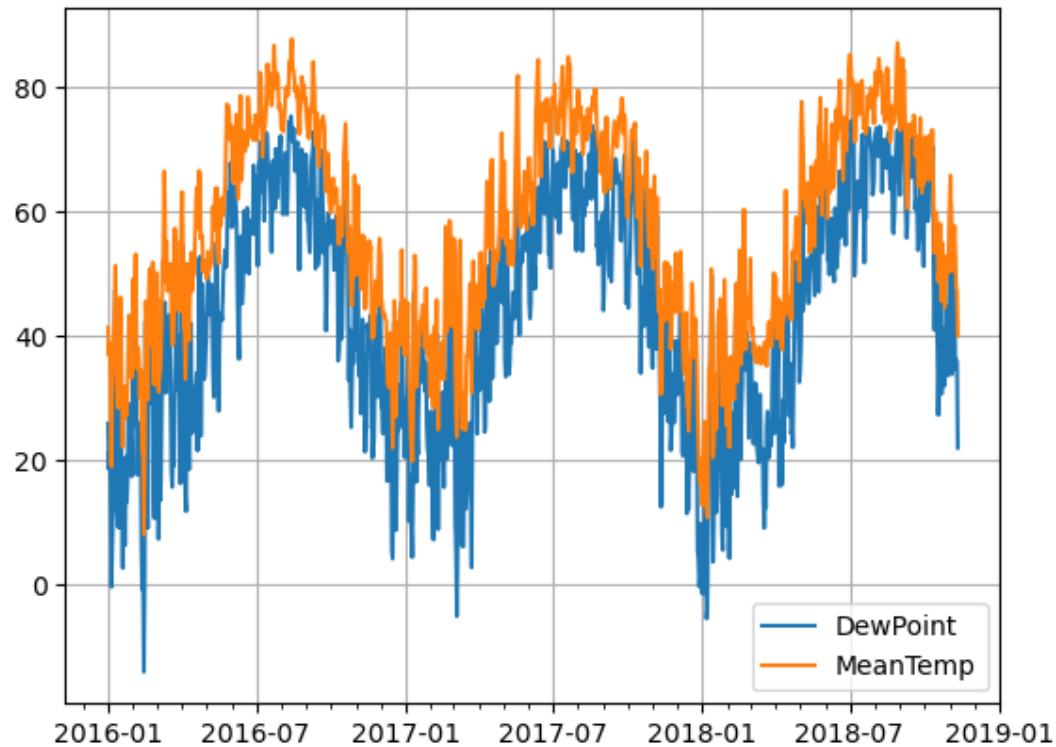
Exploration: Temperature data over time

- Min, Max, and Mean temperature have cyclic yearly pattern
 - Strongly correlated. Can likely include only one of these in modeling to improve interpretability.

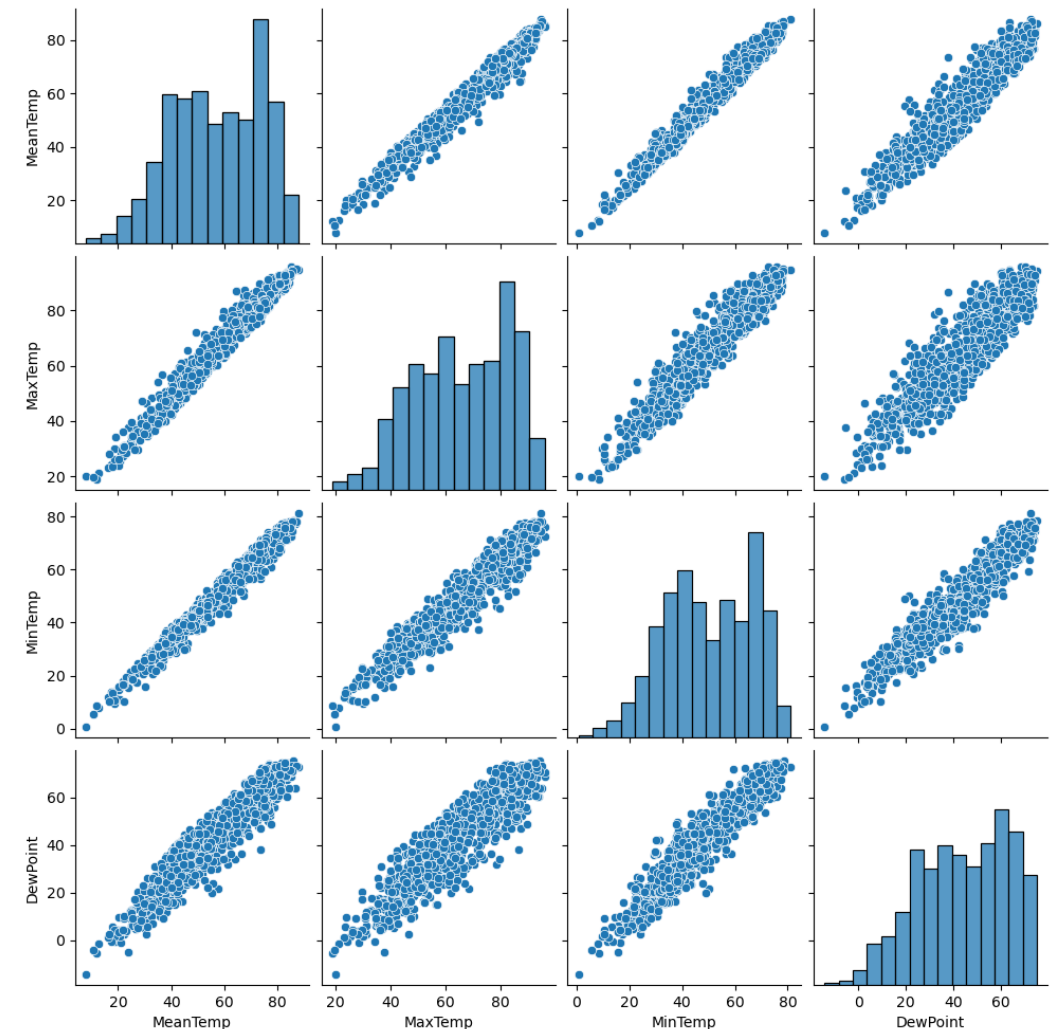


Exploration: DewPoint and MeanTemp over time

- DewPoint provides some information about relative humidity and heat index.
 - Could be included as a model feature
 - Still strongly correlated with temperature so could be left out of a model

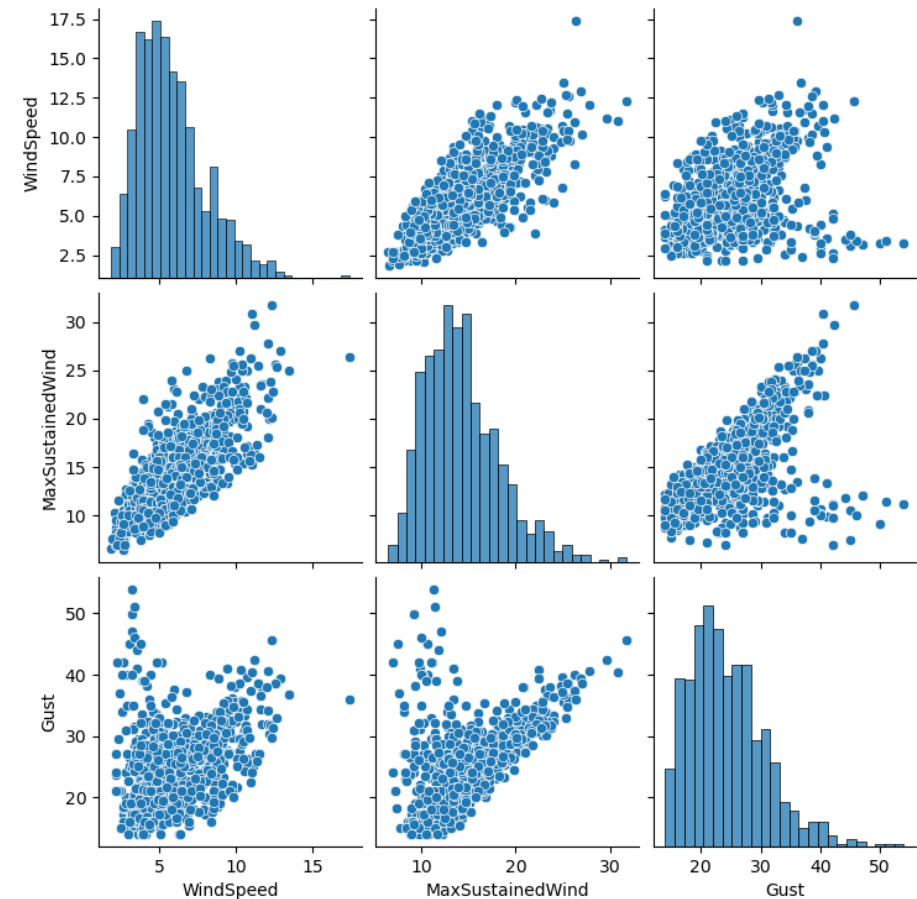
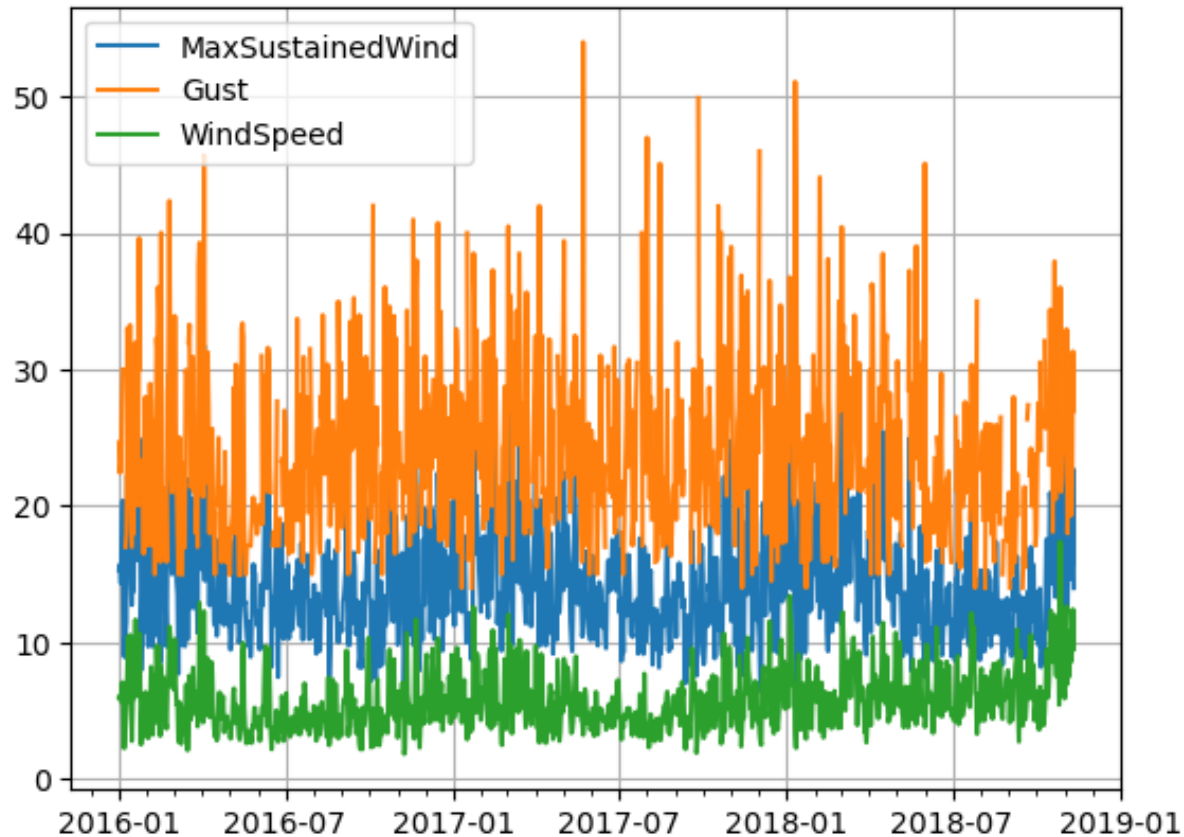


Temperature data correlations plot



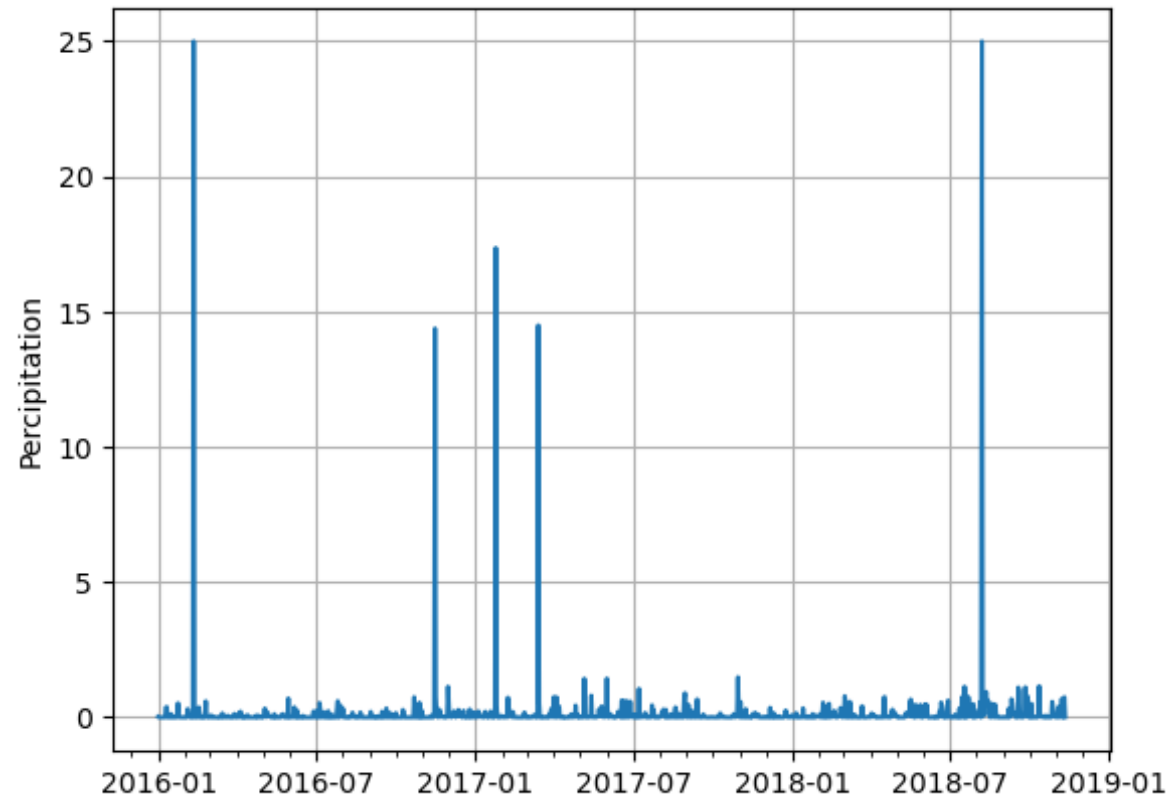
Exploration: Wind data over time

- WindSpeed, Gust, and MaxSustainedWind
 - Also quite correlated, though with some differences in Gust



Exploration: Precipitation over time

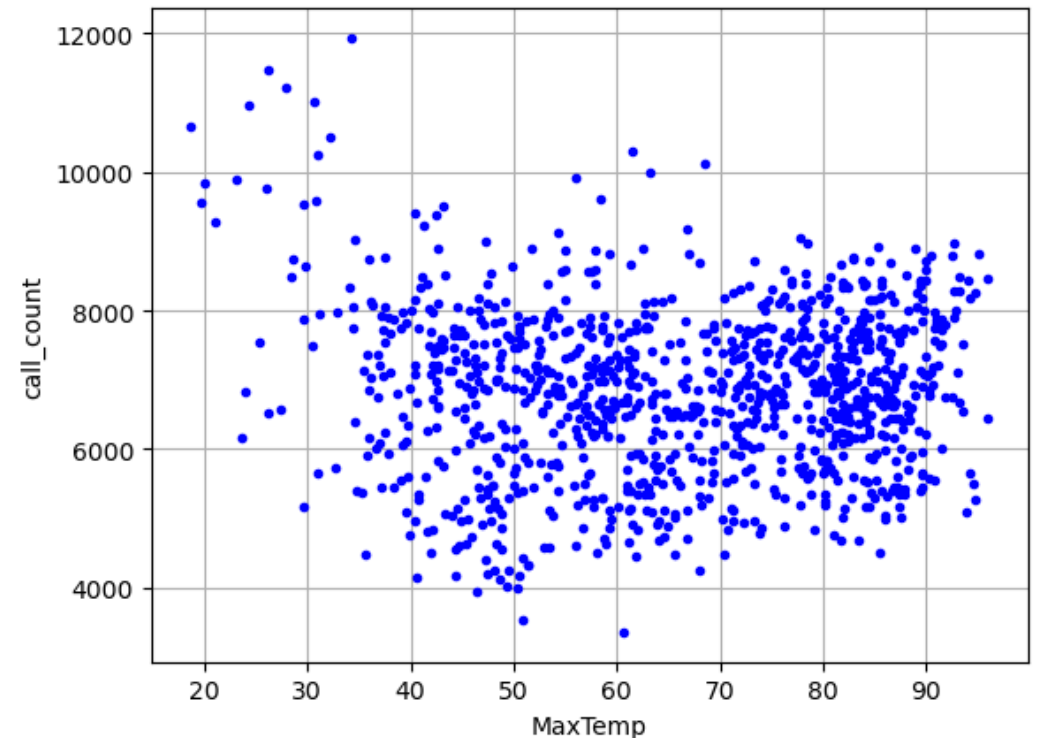
- Precipitation data has some enormous spikes
 - May be interesting to compare with 311 calls



I have a reasonable understanding of the weather data. Time to make a model!

311 call modeling (including weather data)

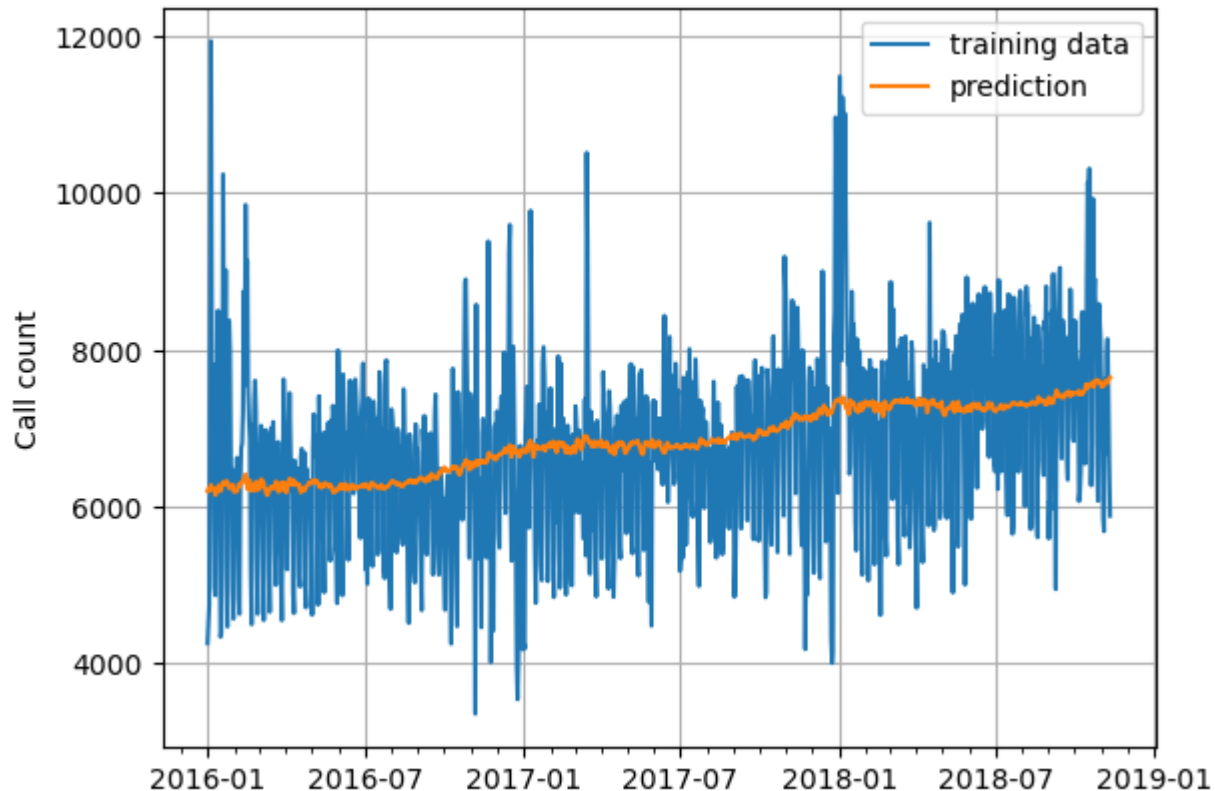
- Goal: Predict daily inbound 311 calls for the next 7 days
- Model Choice: Linear Regression
 - Relatively simple modelling method to predict a numerical value for each day
 - Coefficients can be interpretable for understanding model and interpreting feature impacts
- Combine daily 311 call counts and daily weather data (as organized in data exploration)
- Look at daily call count correlations with weather variables to help determine which features to include.
 - MaxTemp shows higher call counts when temperature is very low. (corr = -0.0214)



Model 1: Very simple linear regression

Start with a few very simple models to get an impression of the data and what type of model makes sense

- scikit-learn LinearRegression
- Features: MaxTemp, date_iter (count of days since 2016-01-01)
- Output: daily call count

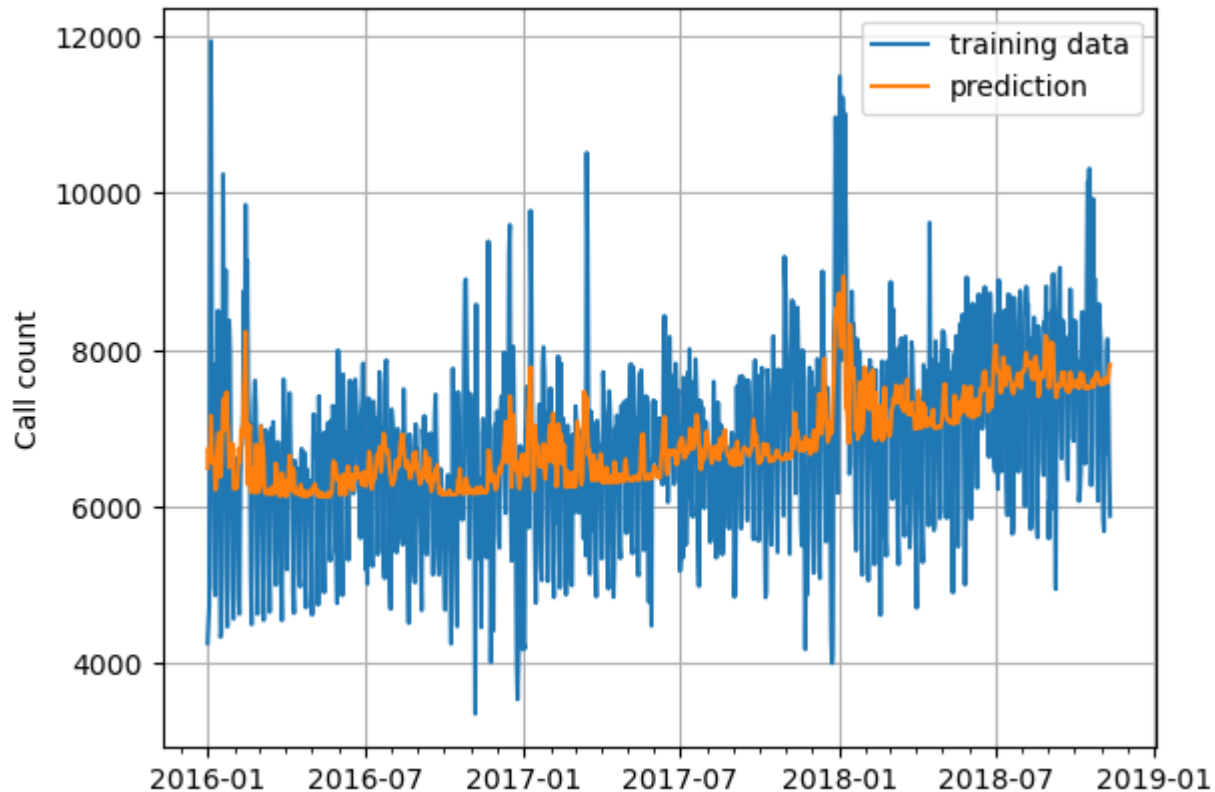


- Not a great fit. Missing peaks and most of seasonal variation.
- What about higher order terms?

Model 2: Very simple polynomial regression

Try polynomial terms

- scikit-learn LinearRegression using polynomial terms
 - `PolynomialFeatures(degree=2, include_bias=False)`
- Features: MaxTemp, date_iter, 2nd order combinations of the two
- Output: daily call count

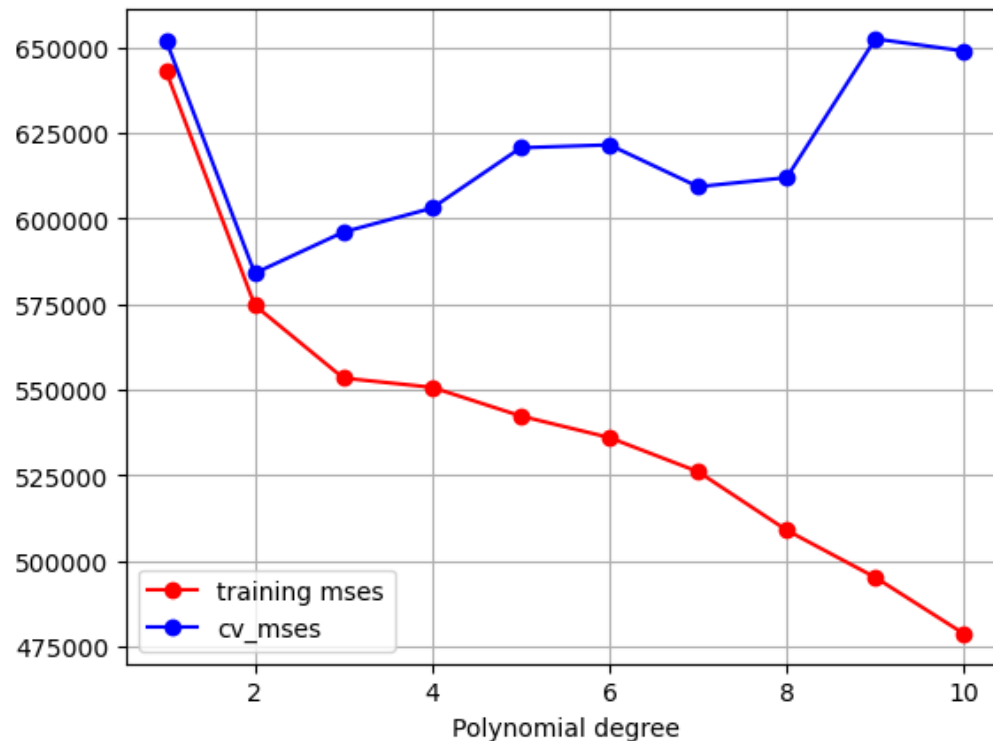


- Still rudimentary, but an improvement over pure linear regression.
- Temperature and date_iter capture some of the peaks represented in the training data

Model 2: Very simple polynomial regression

Try higher polynomial degrees

- scikit-learn LinearRegression using polynomial terms
 - `PolynomialFeatures(degree=n, include_bias=False)`
- Features: MaxTemp, date_iter, n^{th} order combinations of the two
- Output: daily call count

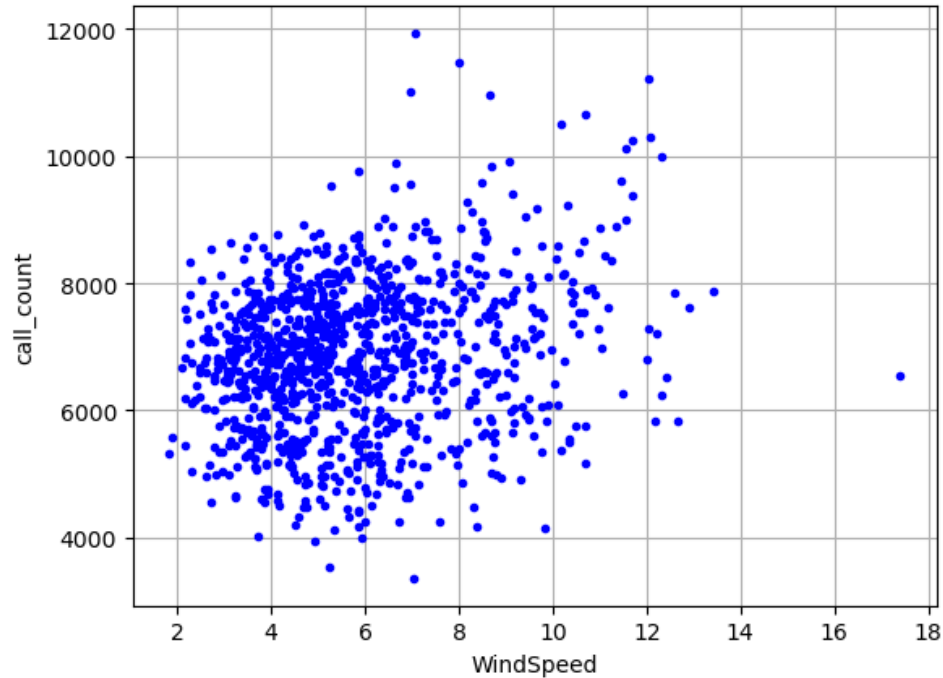


Increasing polynomial degree beyond 2 leads to overfitting (Cross Validation error increases)

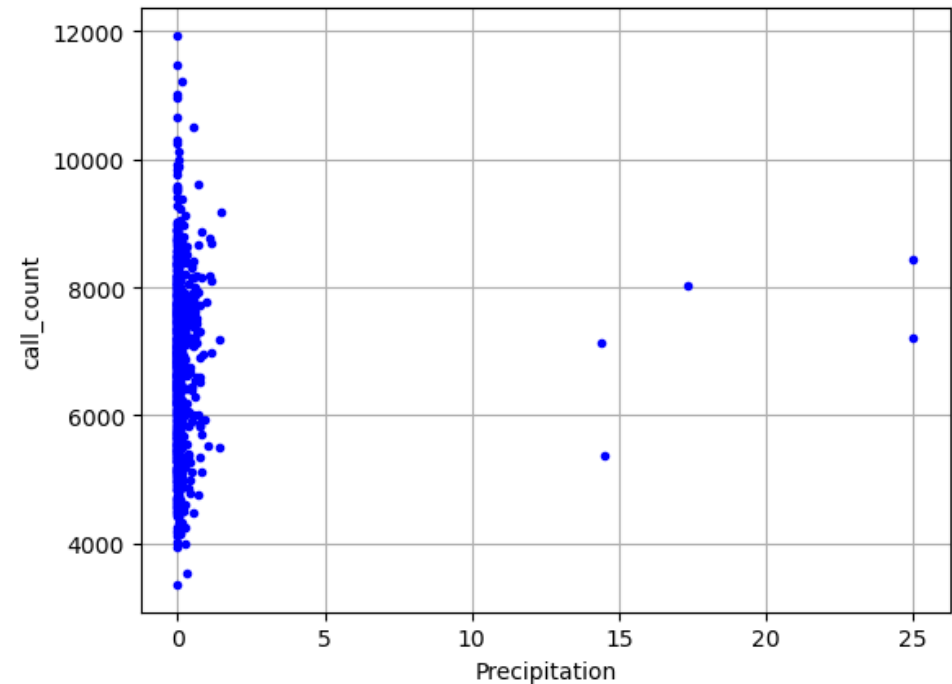
Next step: Add additional features to the model

Choosing additional model features

- WindSpeed is reasonably correlated with call_count (corr= 0.221)



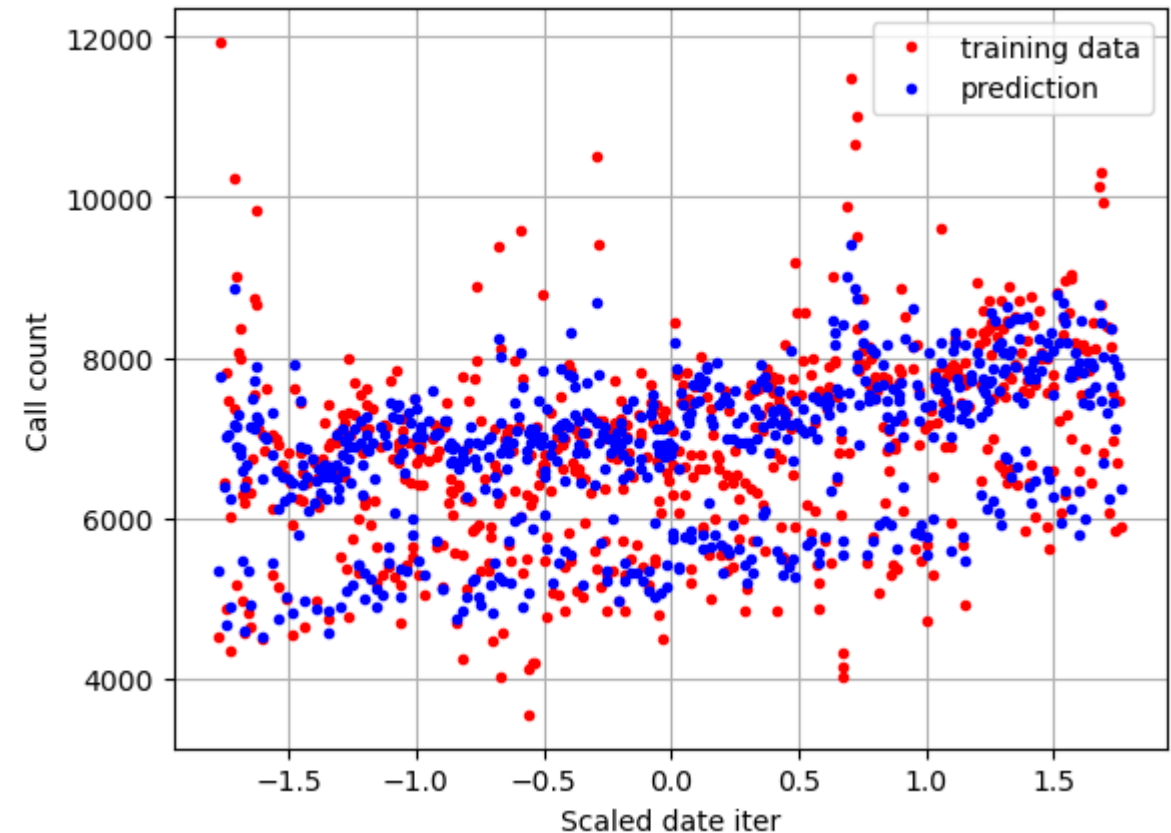
- Precipitation is only slightly correlated with call_count (corr= 0.040)
 - Don't see huge call spikes with large precipitation.



- Weekends are strongly negatively correlated with call_count. Include weekday as one-hot encoding.

Model 3: Linear Regression with some polynomial terms

- `feature_vars = ['date_iter', 'MaxTemp', 'MaxTemp**2', 'Precipitation', 'WindSpeed', 'WindSpeed**2', 'MO', 'TU', 'WE', 'FR', 'SA', 'SU']`
- Feature choice reasoning (through some testing)
 - Including all polynomial combinations with degree=2 led to overfitting (high CV MSE)
 - Pure linear regression misses peaks
 - Include some quadratic terms that improve model



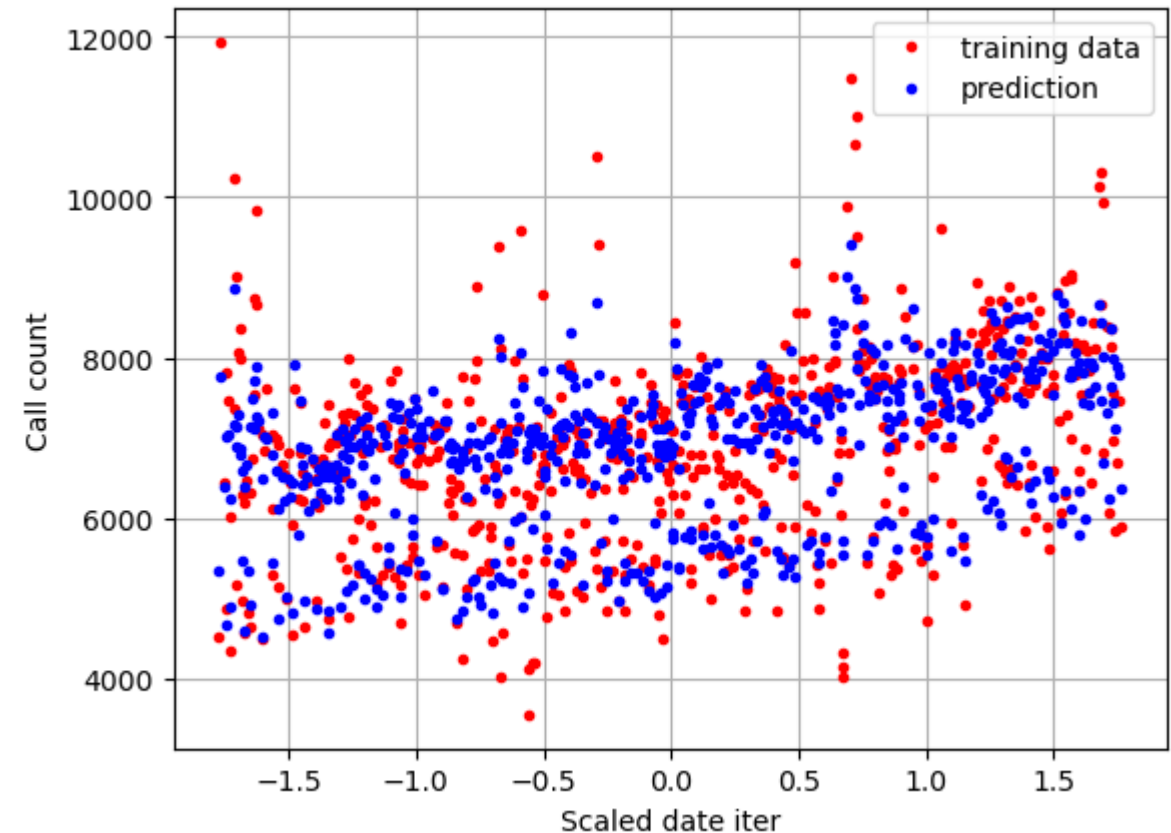
Model 3: Linear regression with some polynomial terms

- feature_vars = ['date_iter', 'MaxTemp', 'MaxTemp**2', 'Precipitation', 'WindSpeed', 'WindSpeed**2', 'MO', 'TU', 'WE', 'FR', 'SA', 'SU']

Reasonable fit. Captures spikes, weekday variation, increasing trend with time.

No sign of overfitting in CV or test error.

training MSE:	272500.15
CV MSE:	281340.34
test MSE:	258056.89



Model 3: Coefficients and interpretation

Model coefficients:

<u>Feature</u>	<u>Coefficient</u>
date_iter	353.82
MaxTemp	-2763.54
MaxTemp**2	2747.86
Precipitation	-85.30
WindSpeed	-347.92
WindSpeed**2	542.41
MO	60.68
TU	129.13
WE	62.92
FR	-77.73
SA	-528.73
SU	-572.56

Interpreting coefficients (applied to scaled features):

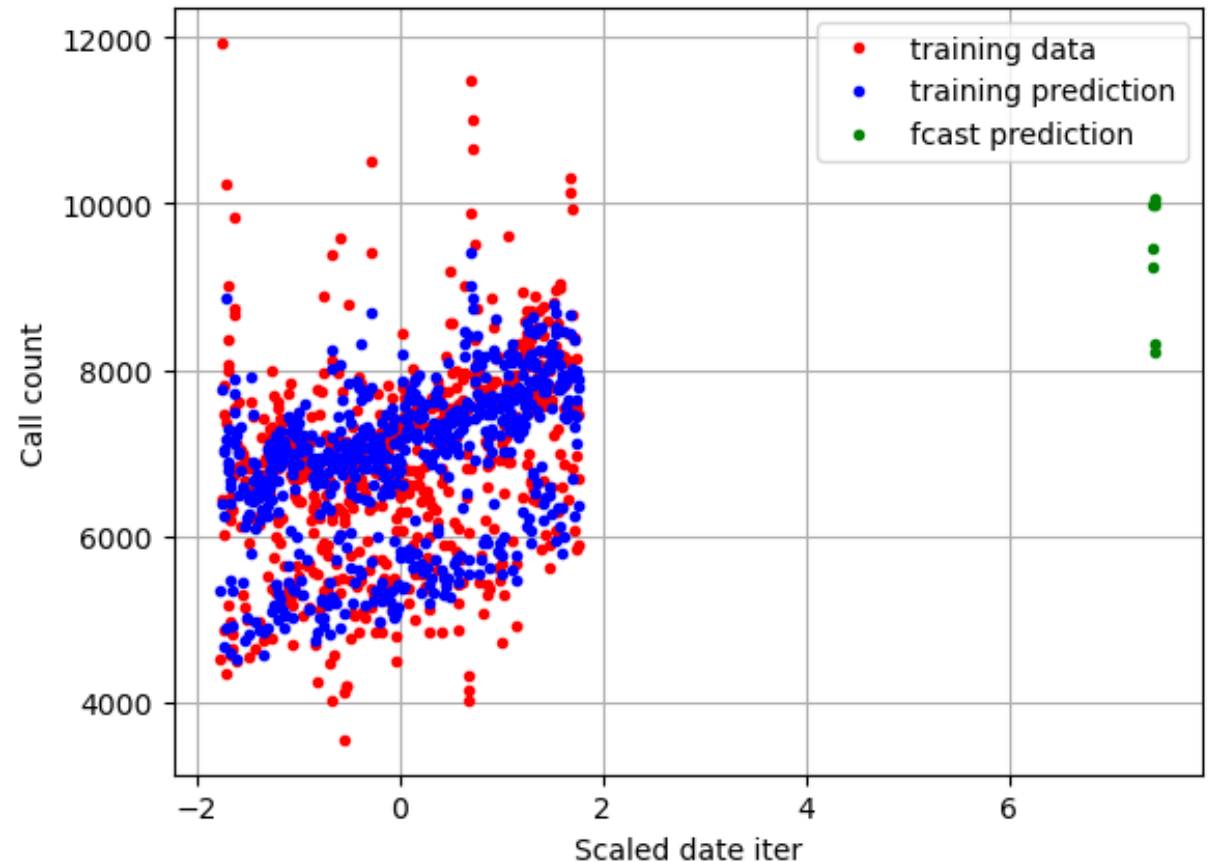
- Date is positive: call counts increase over time
- MaxTemp is negative: low temperatures lead to more calls
- MaxTemp**2 is positive: extremes lead to more calls
- Precipitation is (slightly) negative: precipitation is associated with slightly fewer calls
- WindSpeed is negative: not very meaningful since it turns out WindSpeed increased with date (collinearity)
- WindSpeed**2 is positive: extremes lead to more calls
- Weekdays are positive and weekends are negative: fewer calls occur on weekends

Model 3: Predict daily 311 calls for next week

Goal: Predict daily inbound 311 calls for the next 7 days

- Get weather forecast from wunderground.com and weather.com for week starting from Wednesday 2023-06-07 (MaxTemp, Precipitation, WindSpeed)
- Apply model to predict call counts starting from 2023-06-07
- Prediction looks in line with training data trends

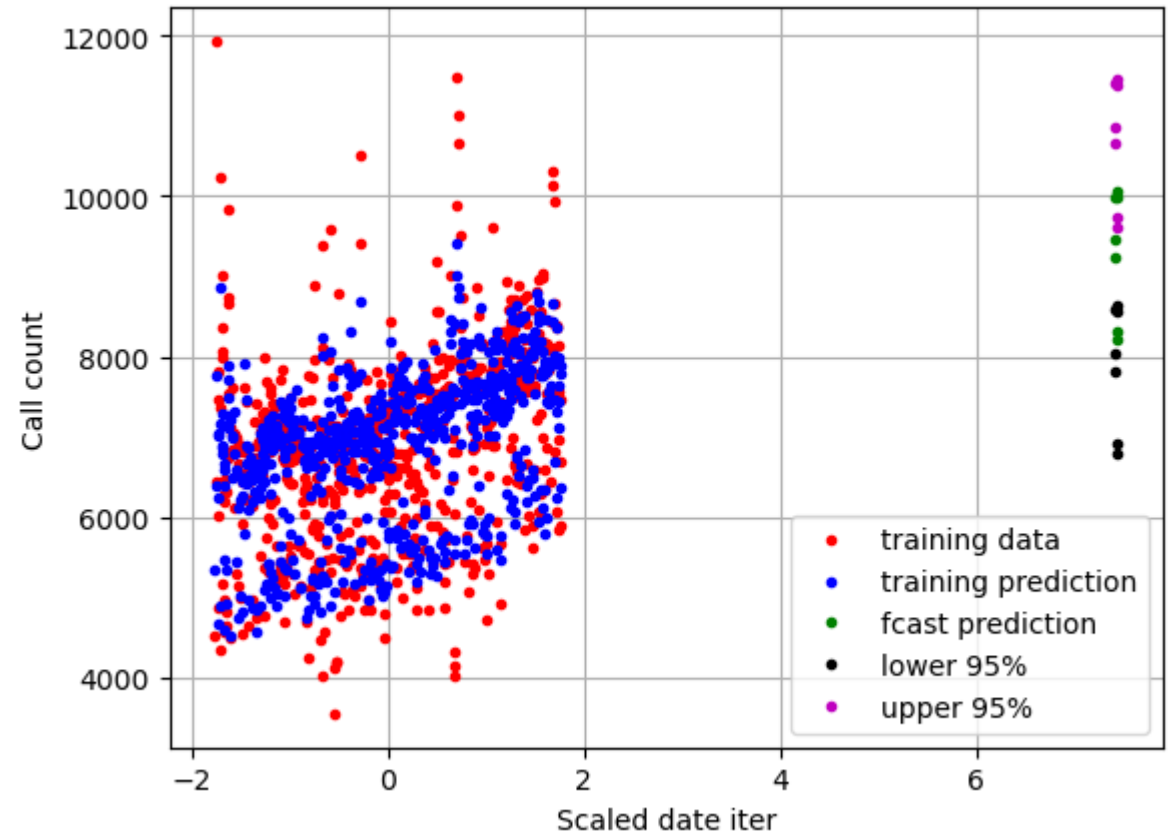
<u>Day</u>	<u>Call Count Prediction</u>
Wed	9993
Thurs	9459
Fri	9237
Sat	8207
Sun	8317
Mon	9982
Tues	10050



Model 3: Prediction interval estimate

- Use test error sum and standard deviation to estimate 95% prediction interval
 - Assumes no changes in patterns over a few years, which seems a bit optimistic

Date	Weekday	Call Count Prediction	Lower	Upper
2023-6-7	Wed	9992.66	8577.85	11407.47
2023-6-8	Thurs	9454.86	8040.05	10869.67
2023-6-9	Fri	9237.26	7822.45	10652.07
2023-6-10	Sat	8206.88	6792.07	9621.69
2023-6-11	Sun	8317.35	6902.54	9732.16
2023-6-12	Mon	9982.27	8567.46	11397.08
2023-6-13	Tues	10049.67	8634.86	11464.48



Conclusions and next steps

- Explored 311 call data from NYC and weather data from NY
- Fit linear regression model (with polynomial terms)
- Interpreted model coefficients based on data exploration
- Predicted 311 calls for the next week (and prediction interval)

Further improvements are certainly possible. Some reasonable next steps include:

- Apply a decision tree model. Perhaps GradientBoostingRegressor from scikit-learn. This step would likely improve prediction but decrease interpretability.
- Further refine polynomial model features and optimize regularization
- Explore Agency-based 311 call statistics and modeling

