# Point72 Analytics Case Study

Hari Kumar
October 2023

# Table of Contents

# Abstract

- Tasked to **predict daily call volume** based on weather + ancillary data in NYC
- Had large set of historical call and weather data that I took creative strategies in loading for analysis
- Cut up the aforementioned call & weather data to **determine most valuable features** to predict using (built relevant geographic + time-series analyses)
- Engineered relevant features to incorporate into a **Random Forest model** for call volume prediction
  - *Final Model had r^2 = 0.89 and MSE = 0.6*
- Model was used to get total call volume for 7 days in 2019 (accuracy pending feedback)
- Believe that this is a **strong start**, but there are **more improvements that can be done** in regards to both analysis + predictive pieces

*Final Goal: Predictions of Daily 311 Inbound Calls for next 7 Days*

| Date | Predicted Num of Calls |
|---|---|
| 2019-01-01 | 5439.0 |
| 2019-01-02 | 6198.0 |
| 2019-01-03 | 5926.0 |
| 2019-01-04 | 5782.0 |
| 2019-01-05 | 8363.0 |
| 2019-01-06 | 6470.0 |
| 2019-01-07 | 6448.0 |

# Overview of Problem

- New Yorkers contact 3-1-1 when they have a non-emergency City service that needs tending to (e.g. street noise, water leak).
- We have reasons to believe that 3-1-1 calls are influenced by weather patterns from the area around them
- We want to see if we can predict the total # of 311 calls per day for the next week, given some metainformation about the call (e.g. location, type of problem) and the weather patterns

# Notes on Data Ingestion

- Given 2 Dataframes: **Weather** and **311** (Requests). Weather is of a smaller size and can be manipulated quickly with Pandas
- Requests is more challenging to load in to the notebook, given its size (4.5 GB)
  - *Used Dask to load it in instead of Pandas, because its parallel processing makes it faster*
  - *However, it also lazily evaluates the columns of its chunks, which led to massive mismatching.*
  - *Final solve for this exercise: Use a sample from Requests in Pandas to do data analysis, and apply results to entire dataframe*
- Would spend more time in Dask and other parallel computing resources to ingest the full dataframe into notebook moving forward

# EDA Insights from 3-1-1 Records

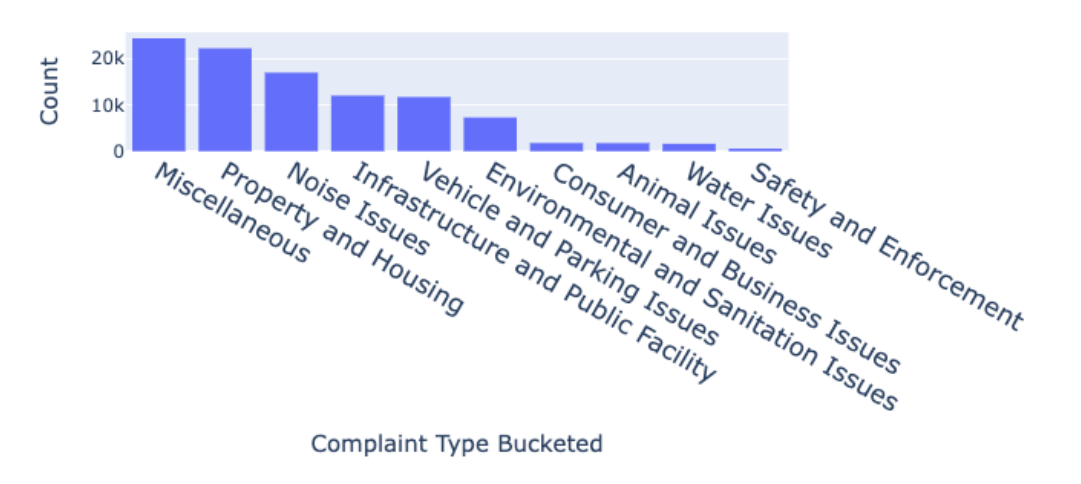
- 3-1-1 record data came at the per-request level, with information on:
  - *Created date + closed date*
  - *Geographic location (Borough, Lat/Long)*
  - *Spatial location (Location Type, Facility Type, Landmark)*
  - *Who the request was for (Agency Name, Complaint Type)*
- To anticipate importance, the **temporal** and **geographic** information was the first thing to look at.

| Agency | Facility Type | Borough | Day of Request | Day of Week of Request | Complaint Type Bucketed | Location Types Bucketed |
|---|---|---|---|---|---|---|
| dot | missing | unspecified | 2018-01-17 | wednesday | Infrastructure and Public Facility | Streets and Sidewalks |
| nypd | precinct | queens | 2018-01-17 | wednesday | Noise Issues | Residential Areas |
| dob | missing | brooklyn | 2018-01-17 | wednesday | Miscellaneous | Miscellaneous |
| dpr | missing | bronx | 2018-01-17 | wednesday | Miscellaneous | Streets and Sidewalks |
| nypd | precinct | brooklyn | 2018-01-17 | wednesday | Noise Issues | Residential Areas |

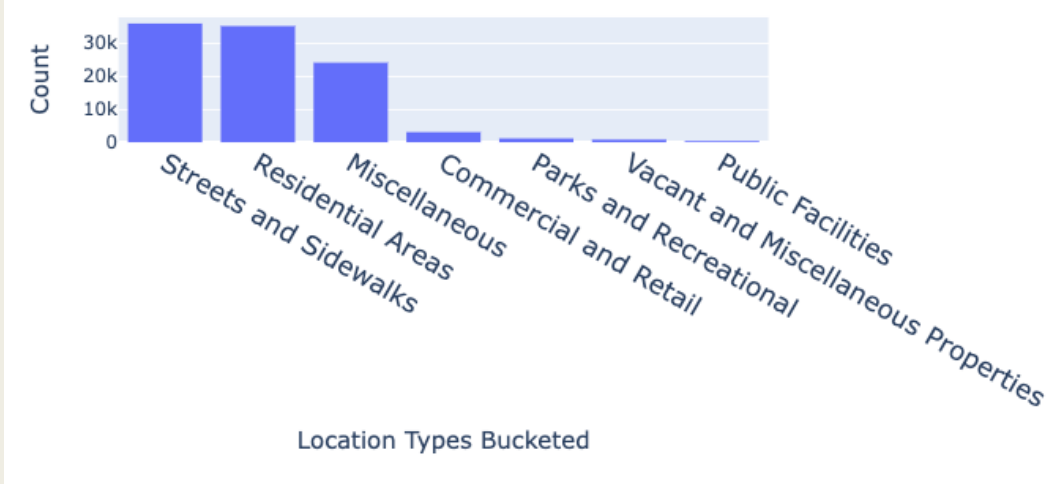
Key Information from 3-1-1 Table (post-cleaning)

# EDA Insights from 3-1-1 Records: General

- Complaint Type indicated that most were regarding Property/Housing Issues as well as Noise Problems




- The location of the 3-1-1 problem was equally likely to be a residential area as it was to be out in a street/sidewalk
  - *Rate of Park issues was low*

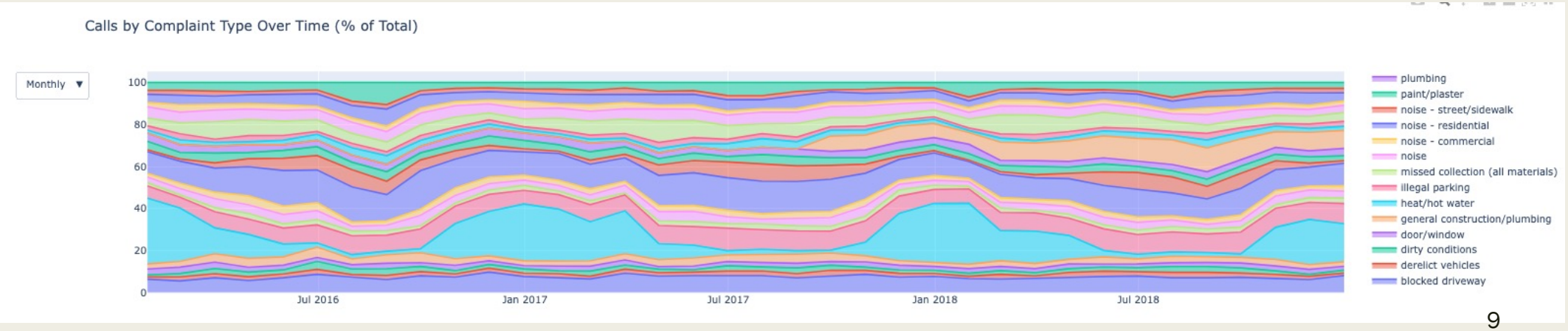


# EDA Insights from 3-1-1 Records: Temporal View

- The base view of our output variable (Daily Call Frequency) indicates that it has risen over the 3 year span, with a sharp spike at the beginning of 2018, and a recent tapering.
  - *There are no obvious seasonality implications here, but viewing from a daily/weekly perspective is also possible here if we want to fine tune our understanding.*
- It requires a deeper look at how the call data specifically has changed over time.

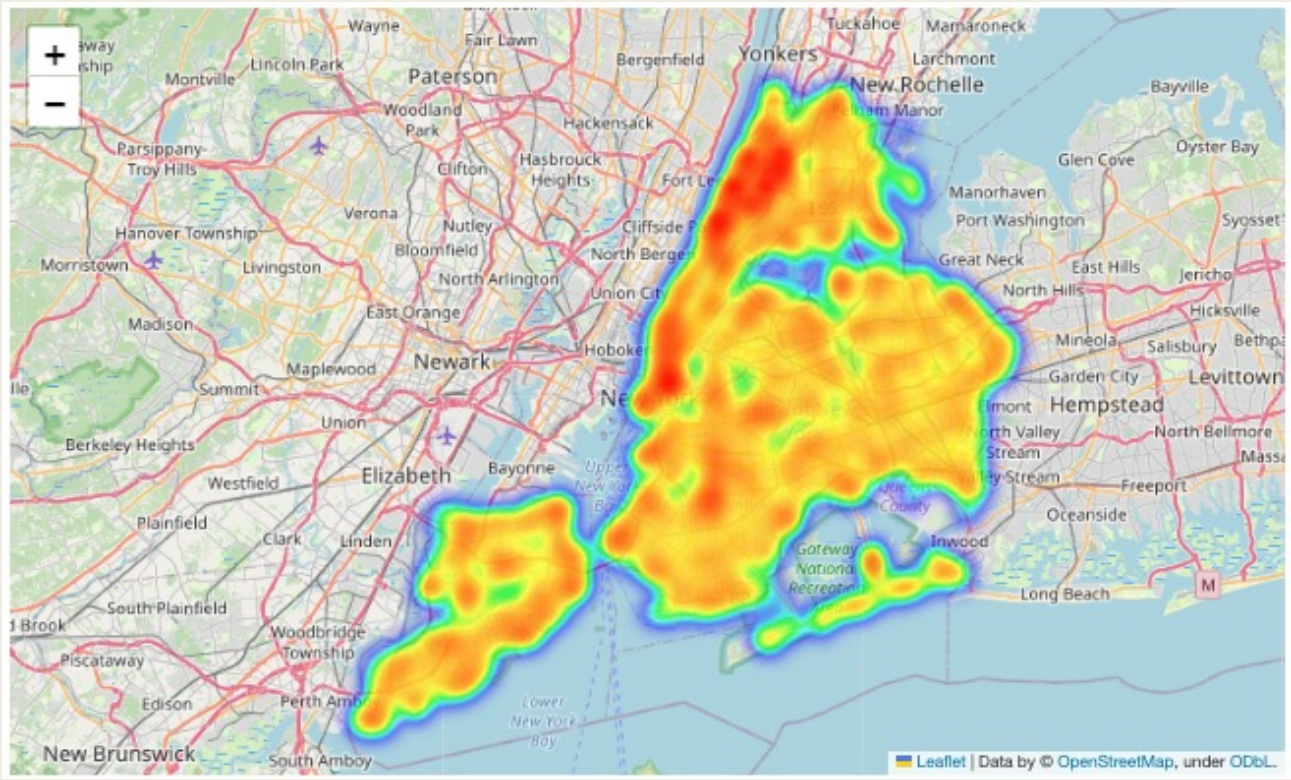
# EDA Insights from 3-1-1 Records: Temporal View

- **Temporal Data** of different parts of the 3-1-1 data from 2016-2018 were also illuminating to aid in future predictions
  - *Viewing complaint type over time, the Heat/Hot Water problems (blue area) spiked in the winter months most commonly, indicating seasonality concerns.*
  - *Issues popped up as new in the middle of our timeline, like "Requests for Large Bulky Item Collection" (orange area) that only started in 2017.*
  - *We can toggle by Day/Week/Month here for even more granularity.*



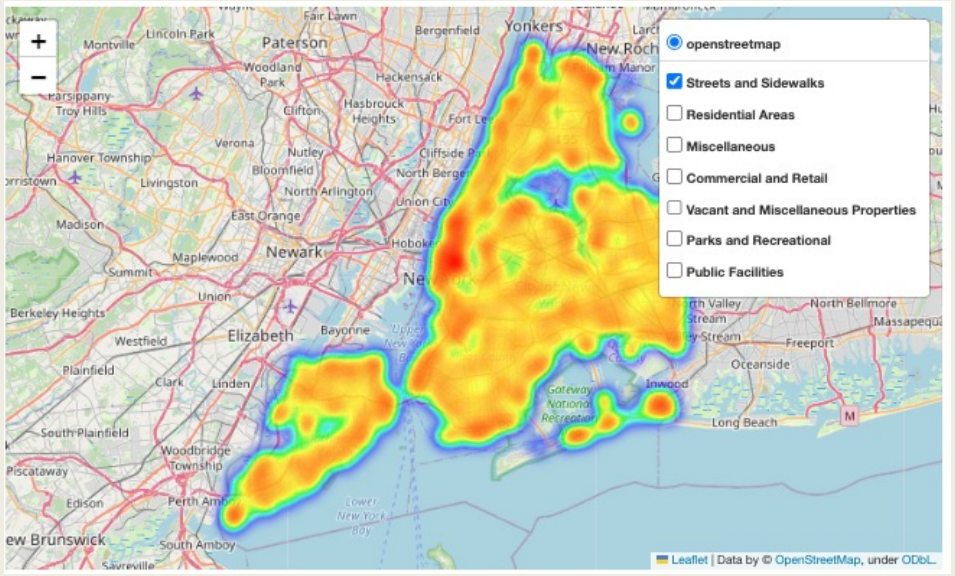
# EDA Insights from 3-1-1 Records: Geographic View

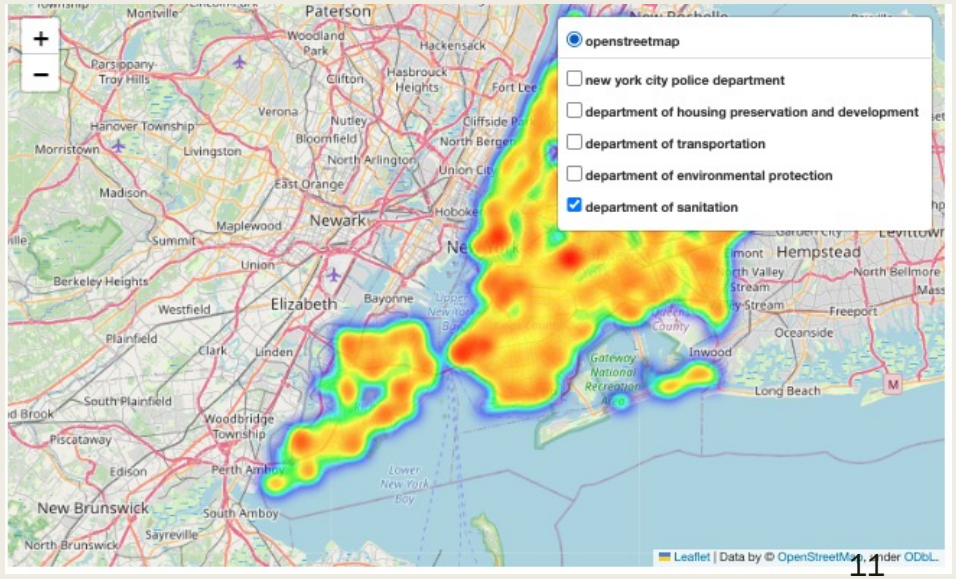
- In terms of quantity of 3-1-1 requests, they are most frequent in Upper West/Central Manhattan, as well as Lower Manhattan
  - *However, there is more to be discerned about the specificity of requests from a geographic perspective.*



# EDA Insights from 3-1-1 Records: Geographic View

- Initial analysis indicates that Lower Manhattan and South Brooklyn are more likely to complain to the Department of Sanitation, and Lower Manhattan has more issues with Street/Sidewalk behavior compared to Residential Areas
- These spatial insights are key in future steps to making accurate predictions of 3-1-1 data based on the location.



# EDA Insights from Weather Records

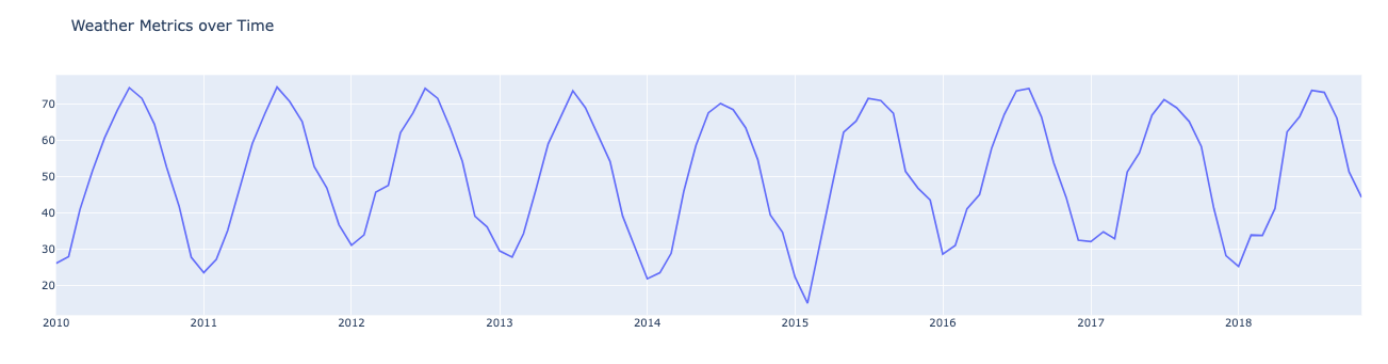
- Weather record data came at the daily level from 2010-2018, based on the USAF/WBAN Station that weather stats were recorded at:
  - *Temperature Data (Daily Min/Max/Mean Temp)*
  - *Binary Weather Data (Rain Indicator, Snow Indicator)*
  - *Continuous Weather Data (Amount of Rain/Snow, Wind Speed)*
- This data came from certain stations (~56 across New York state), so the first step was to clean the relevant data to what was useful and understand basic trends.

| | USAF | WBAN | StationName | State | Latitude | Longitude | MeanTemp | MinTemp | MaxTemp | DewPoint | Percipitation | WindSpeed | MaxSustainedWind | Gust | Rain | SnowDepth | SnowIce | Year | Month | Day |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 48579 | 725150 | 4725 | GREATER BINGHAMTON/E A LINK FIELD AP | NY | 42.207 | -75.980 | 10.2 | 6.1 | 30.0 | 3.8 | 0.00 | 7.0 | 12.0 | NaN | 0 | 3.1 | 1 | 2015 | 1 | 26 |
| 150119 | 997271 | 99999 | THE BATTERY | NY | 40.701 | -74.014 | 70.8 | 65.8 | 81.1 | NaN | 0.00 | 0.0 | NaN | NaN | 0 | NaN | 0 | 2015 | 6 | 15 |
| 120128 | 726225 | 64776 | PLATTSBURGH INTERNATIONAL AIRPORT | NY | 44.650 | -73.467 | 37.8 | 28.9 | 43.0 | 27.7 | 0.00 | 6.0 | 8.9 | NaN | 0 | NaN | 0 | 2010 | 11 | 2 |
| 148297 | 725190 | 14771 | SYRACUSE HANCOCK INTERNATIONAL AP | NY | 43.111 | -76.104 | 78.2 | 70.0 | 91.0 | 73.0 | 0.16 | 5.1 | 15.0 | 27.0 | 1 | NaN | 0 | 2018 | 9 | 3 |
| 83531 | 999999 | 64756 | MILLBROOK 3 W | NY | 41.786 | -73.742 | 8.4 | -8.1 | 23.0 | NaN | 0.80 | NaN | NaN | NaN | 0 | NaN | 0 | 2015 | 2 | 3 |

Key Information from Weather Table (at ingestion)

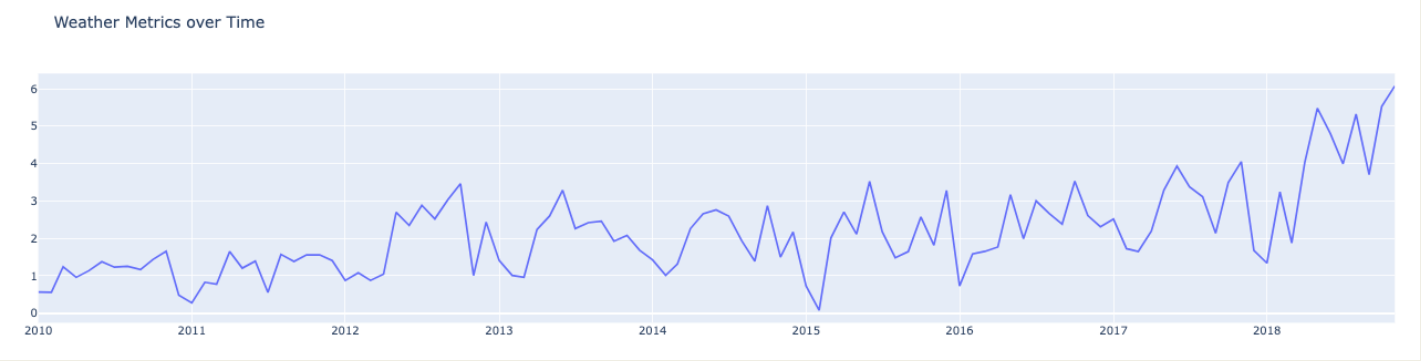
# EDA Insights from Weather Records: Temporal

- As expected, weather data shows great seasonality for many of the variables, as shown below with average temperature peaking in the summers and wind speed peaking in the winters.




*Mean Temp peaks in the summers*

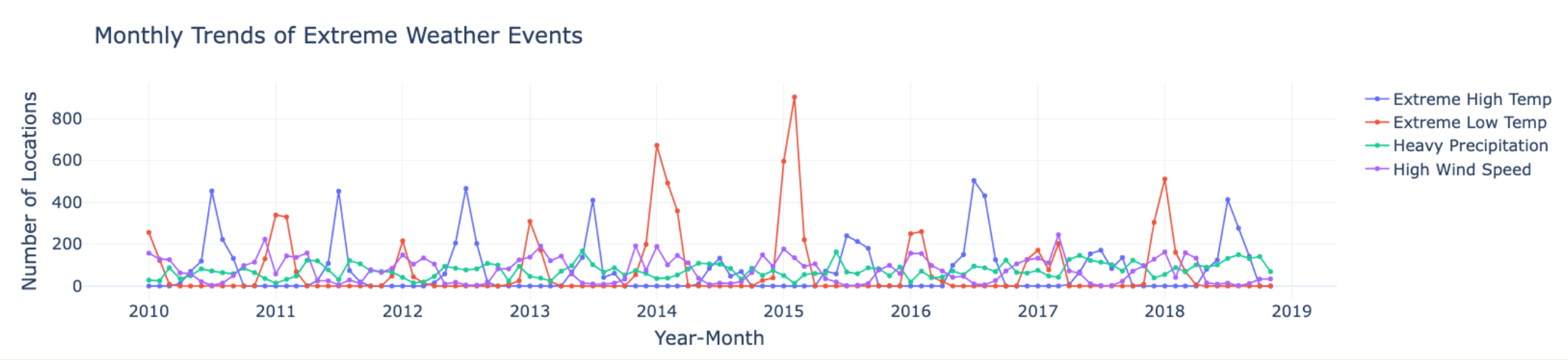
- Interestingly, however, a variable like precipitation shows a steady increase in later years




*Some seasonality with precipitation, but not nearly the main trend*

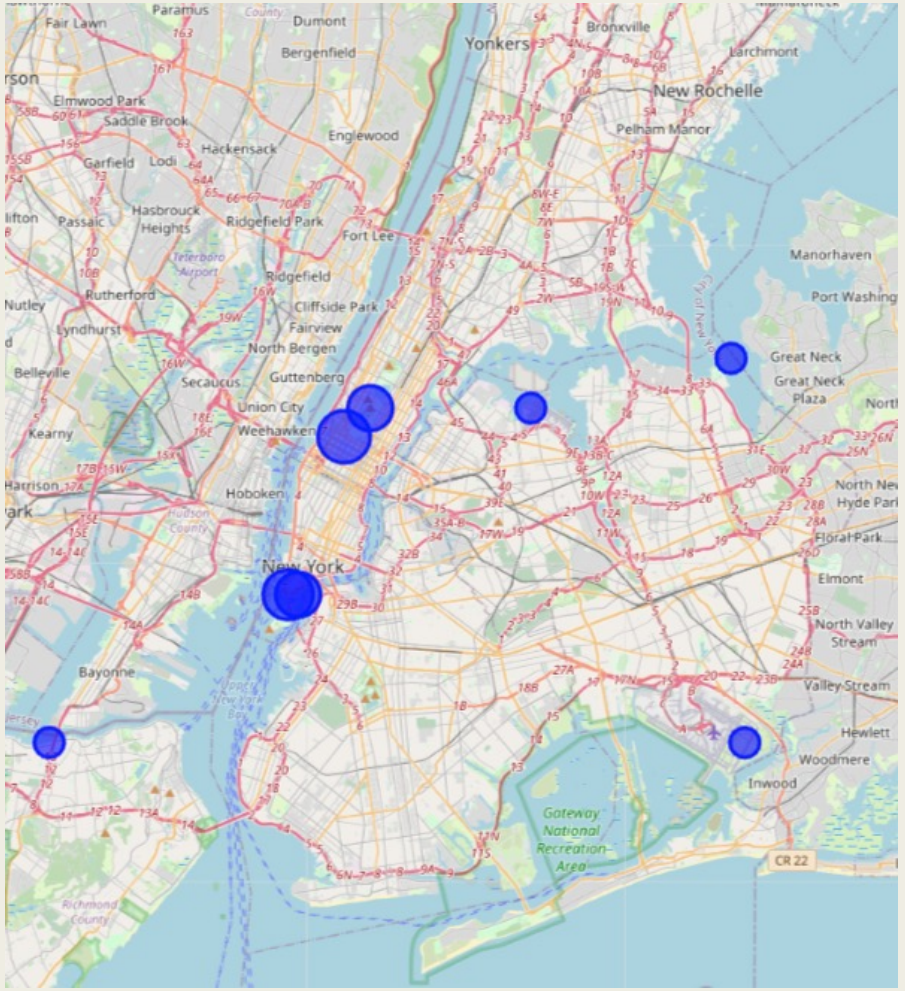
# EDA Insights from Weather Records: Temporal

- What may be more important for 3-1-1 Calls is extreme weather events, as issues like high precipitation can cause flooding and strong winds can damage trees.
  - *Monthly trends of some extreme weather events indicate that there is definitely seasonality at play (Extreme Low Temps in Winter; Extreme High Temps in Summer)*
  - *In our dataset time period (2016-2018), we see abnormally high Extremely High Temperatures in 2016/2017, so that it to watch out for.*



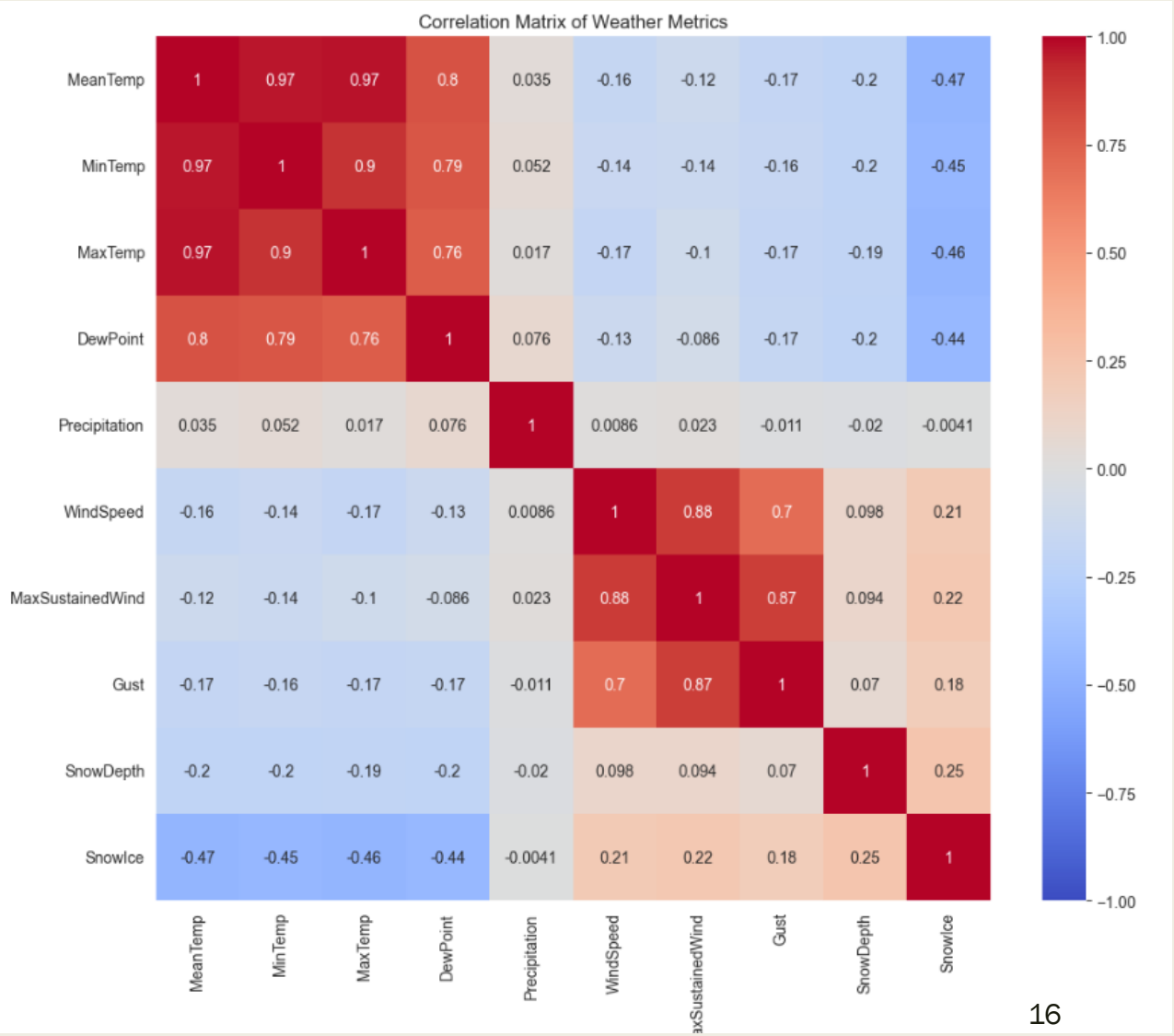
# EDA Insights from Weather Records: Geographic

- Of the 56 Weather Stations in New York State, around 10 of them are in the New York City area (with bubble size indicating the amount of data they have from 2016-2018).
- This indicates that when joining our Weather Data to Call Data, we should try to correspond the location of the 3-1-1 request to the closest weather station to be most accurate.



# EDA Insights from Weather Records: Correlation

- Indications from the Weather records show that we are likely to have high collinearity between many of our different variables
  - *E.g. Wind Speed tracks MaxSustainedWind highly (corr = .88)*
- Indication to be watchful of what weather features we are using together in the future.

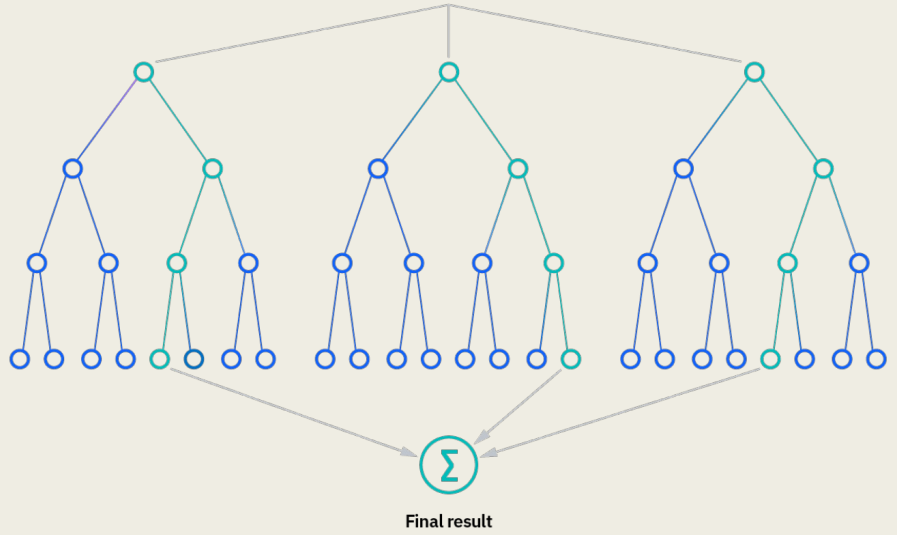


# Approach to Feature Selection

- Given EDA, I found it important to incorporate info from both Requests & Weather
- In addition to features mentioned prior, the biggest key to incorporating the 2 dataframes together is thinking **what seems valuable as a predictor**:
  - *Weather: Given seasonality & time-to-impact, add season, lagging weather metrics, and the extreme temperature indicators*
  - *Requests: Determine the closest Weather station to the location (if given), add Day of Week of request, and Buckets of Complaint + Location types*
    - One-Hot Encode features (e.g. borough, buckets) that are valuable to keep
- Want to ensure that we **don't overcrowd the model** (highly collinear features), and notice that the dataframe is crowded, so remove unnecessary, highly NaN columns as well
  - *Imputation by interpolation or ffill when it is reasonable to do so*

# Approach to Modeling + Output

- I ultimately chose to use a Random Forest Model for this predictive task. It felt like an appropriate choice given the data constraints + goals because:
  - *We think there will be a **non-linear relationship** between weather + calls and RFs can capture such non-linear relationships without the need for complex data transformations*
  - *We have both continuous and categorical data, so the **mixed data types** are better in an RF*
  - *RFs tend to be better in **capturing interactions** between features which is common in weather data (e.g. high wind with high rain)*
  - *We have outliers (random day with a lot of wind, and generally long tailed distributions), and RFs are solidly **robust to outliers***
- Evaluation Metrics are R^2 and RMSE, given we are predicting total # of calls per day (continuous)



# Approach to Modeling + Output

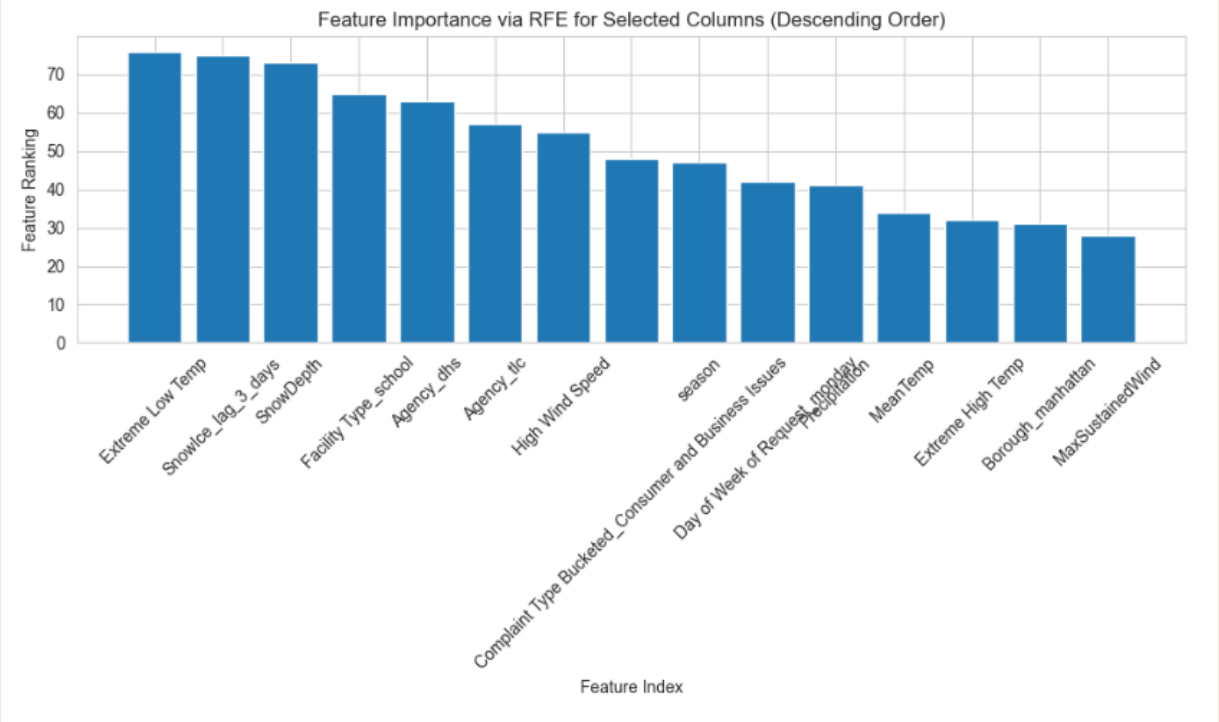
- I had to iterate through the model as the first attempt (with all of our features) had unbelievable R^2 (~1) and RMSE > 1, so it was overfitting
  - *Resolved this by implementing **Recursive Feature Elimination (RFE)** to pick out top features*
  - *This led to more believable results in R^2 and Mean Squared Error*
- I tried approaches incorporating all features, and just the Weather-related ones. Ultimately, chose to pick top 15 features from RFE and run Random Forest on it given its performance and feature variety

```
Mean Absolute Error: 0.348188524590164
Mean Squared Error: 0.6097281147540984
R-squared: 0.8864919899028622
```

Metrics for Model that incorporated top 15 predictive final features

# Approach to Modeling + Output

- The features ultimately involved in predictions were spread across multiple facets, but had some properties in common that made them understandably predictive:
  - *Extreme Conditions (Low Temp, High Temp, Wind Speed)*
  - *Season*
  - *High volume of data (Certain Buckets of Complaint Types, Day of Week of Request, Agency Called)*




*Importance Ranking of the Top 15 Features used in Modeling*

# Approach to Modeling + Output

- Our final output was to predict the total # of calls on the next 7 days
  - **I took that to mean the first 7 days of 2019*
  - *Used Historical Data to get relevant features for those dates*
  - *Applied the predictive model we built + scaled up for the total size of the 311 data*
- Resulted in the predicted total calls per day for the week
- Will look to the Point72 team to validate the correct answers and can assess how far these predictions were

| Date | Predicted Num of Calls |
|---|---|
| 2019-01-01 | 5439.0 |
| 2019-01-02 | 6198.0 |
| 2019-01-03 | 5926.0 |
| 2019-01-04 | 5782.0 |
| 2019-01-05 | 8363.0 |
| 2019-01-06 | 6470.0 |
| 2019-01-07 | 6448.0 |

*Model Predictions for Final Task*

# Takeaways + Next Steps

## Main Takeaways

- Was able to use Weather + Request Data to predict future frequency of Requests
- Random Forest looks to be a good approach for this task with iteration
- EDA is key to get proper visualizations + cuts of data to make useful features
- Ability to use alternative data to make predictions on caller data highlights its success (as seen with high predictive metrics)

## How to Improve upon this

- Build data ingestion process to work with more than a sample
  - *Test out other technologies than dask (e.g. Kafka, concurrent.futures)*
- Work through different modeling approaches + how to improve metrics even further
  - *Could have improved feature selection and made more combined features*
- Merging process can be cleaner and use a Left Join in future vs. Inner (had to given the high null issue)