

NYC 311 Service Requests

Predicting the number of calls in the next week

Pranav Karnani

Executive Summary

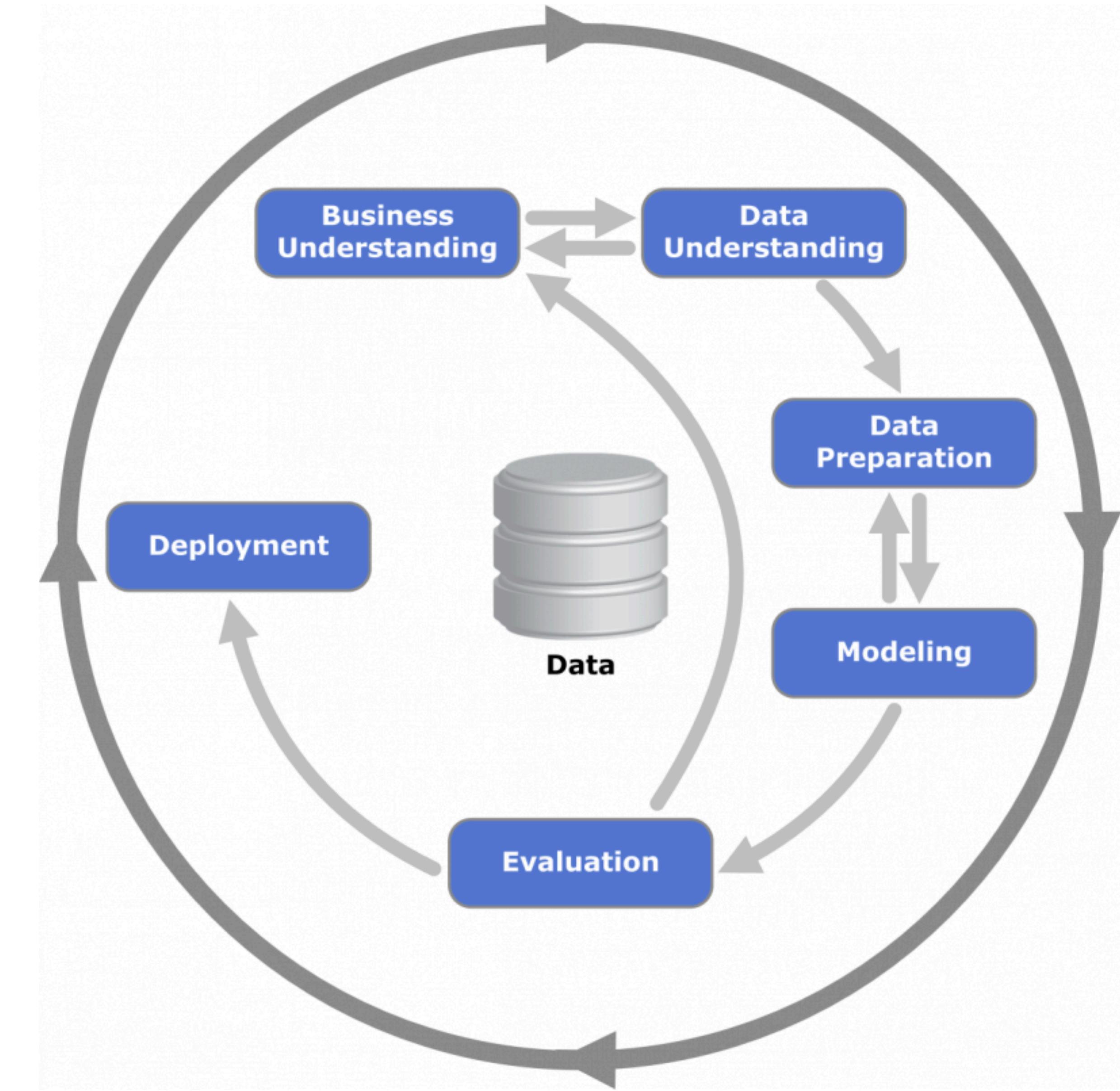
Problem Statement: Predict the number of 311 Service Request calls in the next week.

Deliverables	Method	Performance
A detailed analysis of the different types of requests based on historical data, ranging from location, complaint type etc	Simulation Plots, Distribution Plots, Word Clouds	
A time series model which predicts the temperature and other weather attributes	ARIMA	RMSE: 123.93 MAPE: 0.15
Causal tests to check for the causal effect of temperature and other weather attributes on the number of service requests raised.	Panel OLS	p-Value: 0.0464
A time series model, which uses weather as an exogenous variable to predict the number of calls in the upcoming week	SARIMA, Random Forest	RMSE: 120.30 MAPE: 0.17

Methodology

CRISP-DM Framework

- Business Understanding
- Data Understanding
- Data Preparation
- Modeling
- Evaluation
- Deployment



How to deal with such a large dataset ?

Removed features with:

- More than **80% null** values
- Exhibiting **1-1 correspondence** (Agency Name, Agency)
- More than **99% values** as Unspecified (Park Facility Name)
- With very **granular** information (Address, Street Name)
- Records where Closed Date is **before** the Created Date
- Subsetted the data on **Phone / Mobile** channels using Channel Type

How to deal with such a large dataset ?

Removed features with:

- More than **80% null** values
- Exhibiting **1-1 correspondence** (Agency Name, Agency)
- More than **99% values** as Unspecified (Park Facility Name)
- With very **granular** information (Address, Street Name)
- Records where Closed Date is **before** the Created Date
- Subsetted the data on **Phone / Mobile** channels using Channel Type

Reduced dataset size by **70%**

Data Understanding

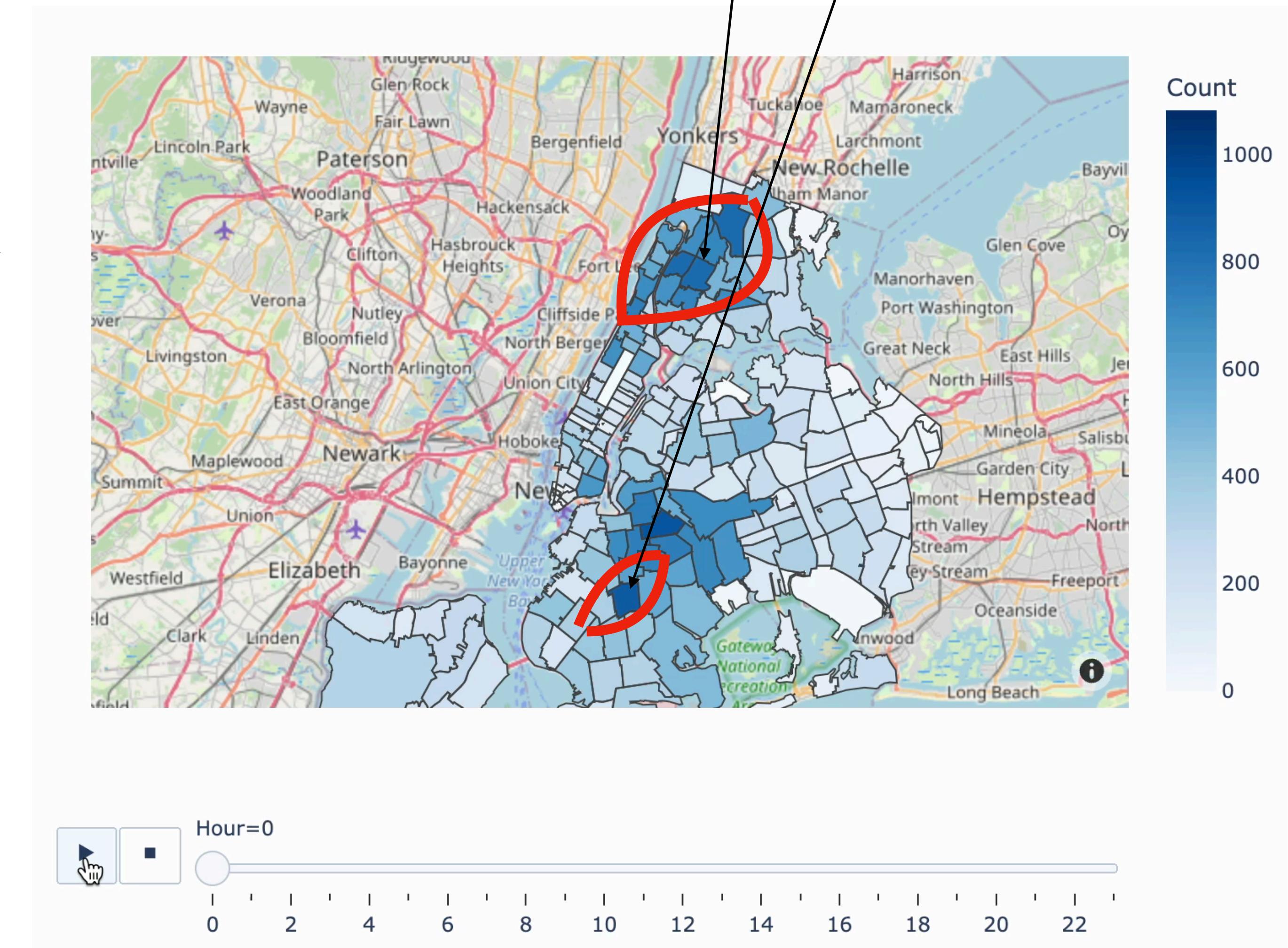
What we have ?

- **311 Service Requests from 2016 to 2018 (~7M+ records)**
 - Complaint Type (Noise, Heat, etc.)
 - Borough (Manhattan, Brooklyn etc.)
 - Location (Sidewalk, Residential etc.)
- **Weather Data collected from stations across the NY state from 2010 to 2018**
 - Snow and Precipitation
 - Temperature
 - Wind / Gust

Data Understanding Call Patterns

- Simulation plots which help get a better picture of **high activity zip-codes**.
- Plot on the right is a heat map of how calls on an **hourly basis**.

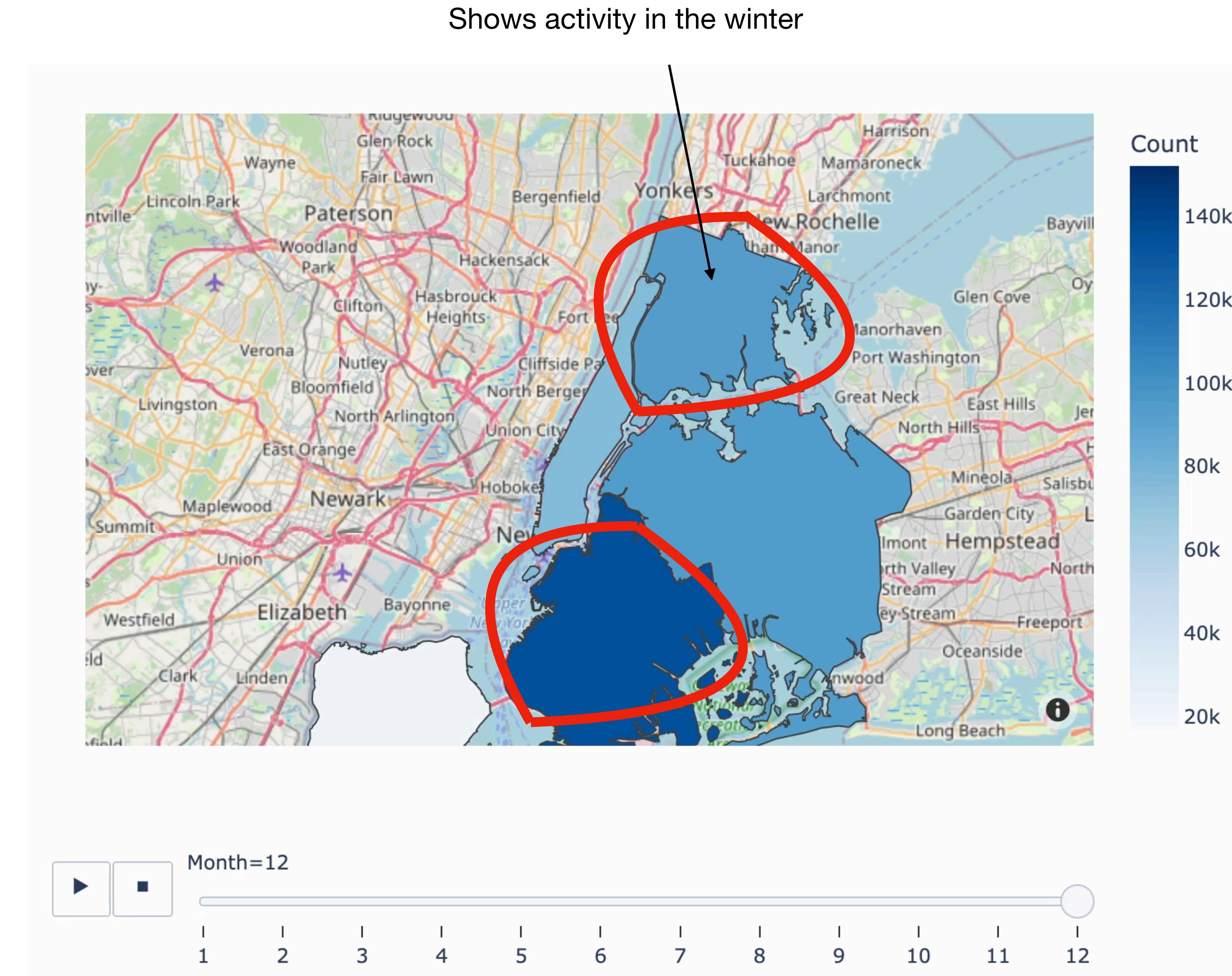
Shows most activity



Data Understanding

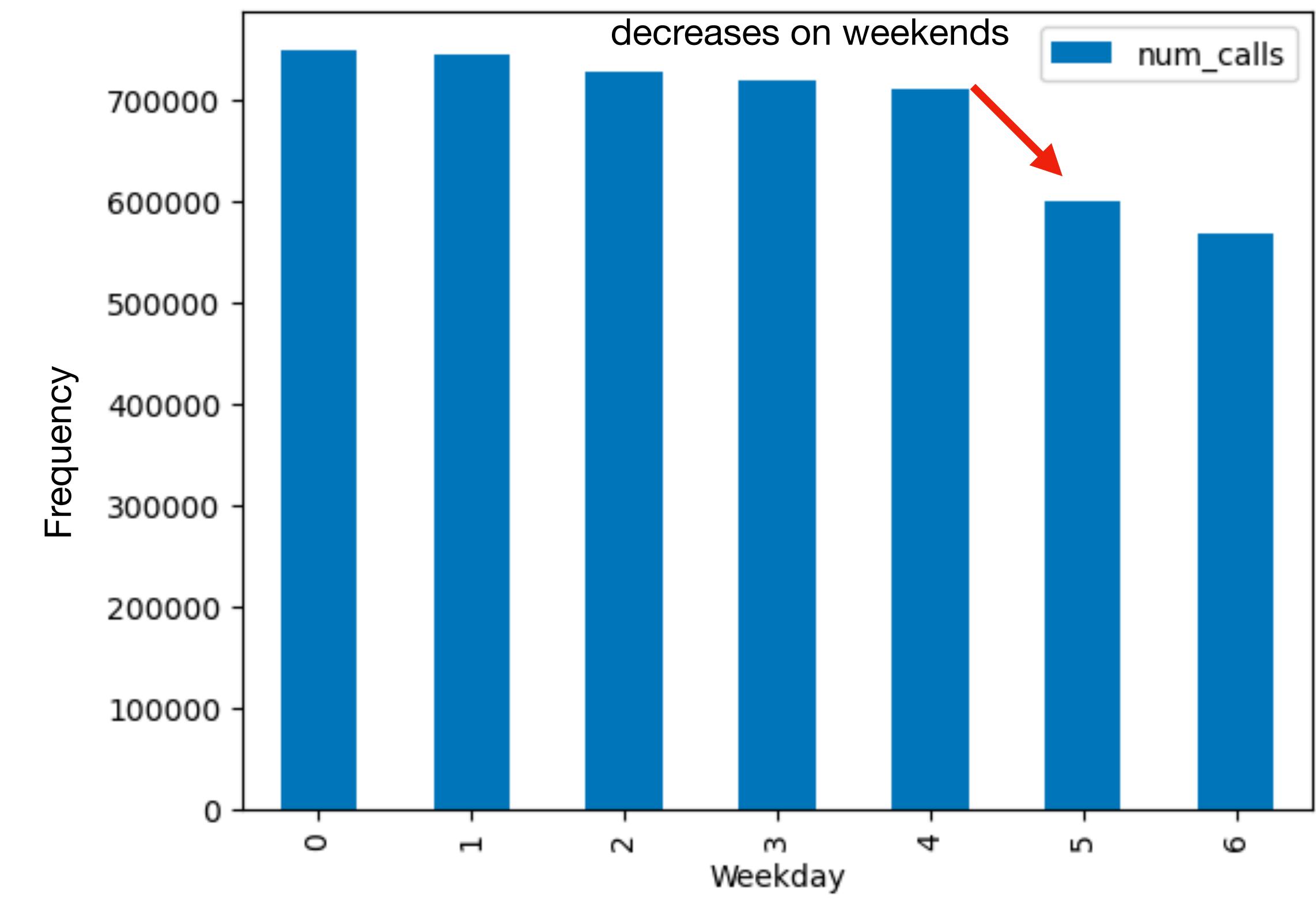
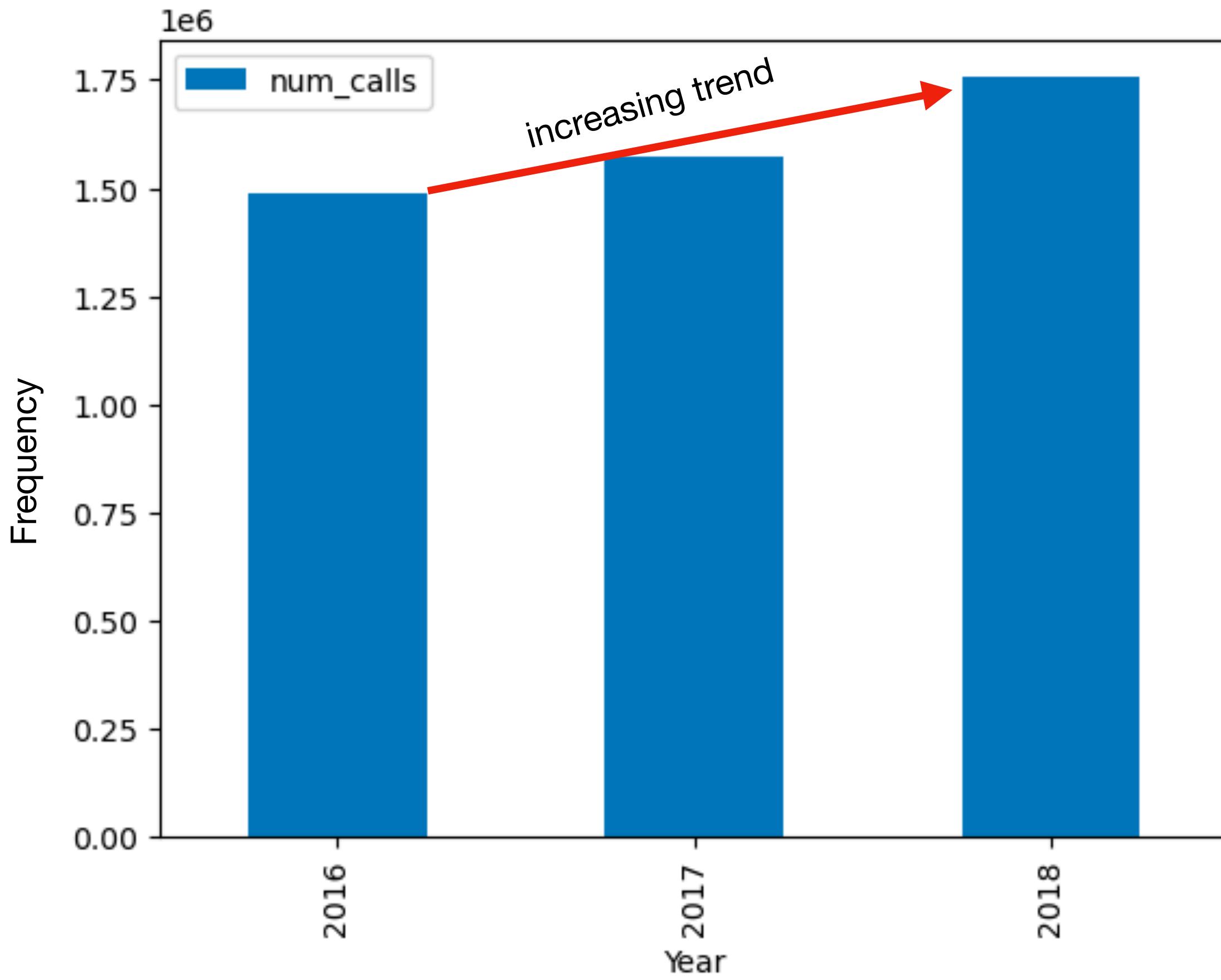
Call Patterns

- Simulation plots which help get a better picture of **high activity boroughs**.
- Plot on the right is a heat map of how calls on a **monthly basis**.
- Performed a similar analysis:
 - For select agencies
 - Complaint types
 - Varying level of granularity (weekdays / hourly / monthly)



Data Understanding

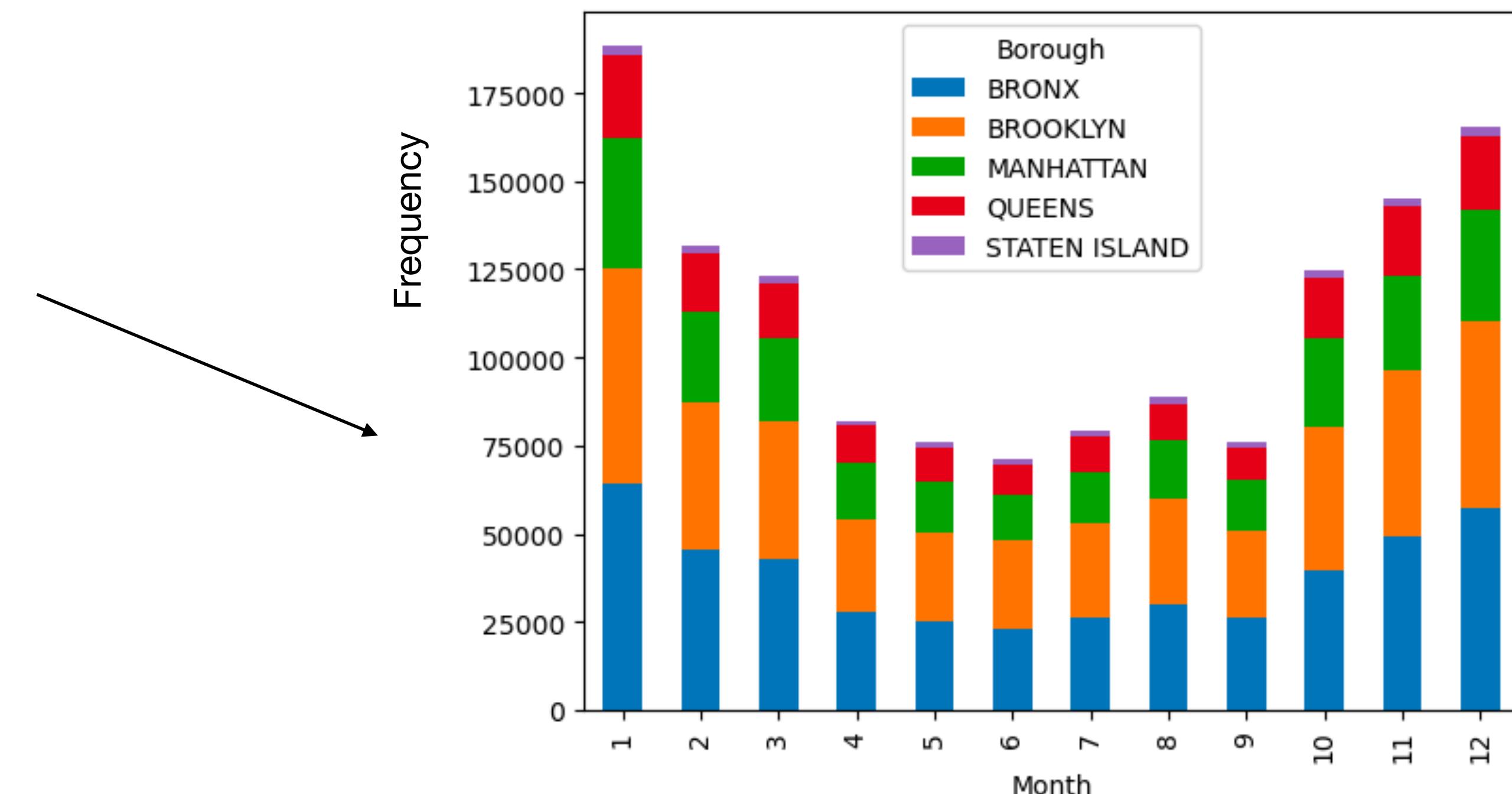
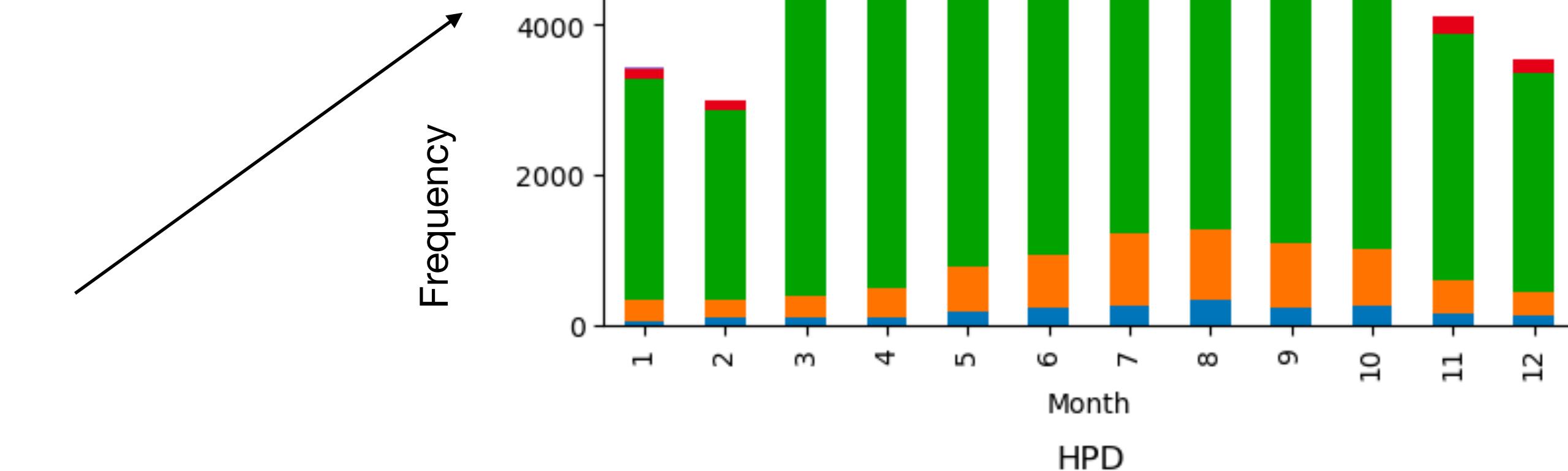
Call Patterns



Data Understanding

Call Patterns - Agency (over Months)

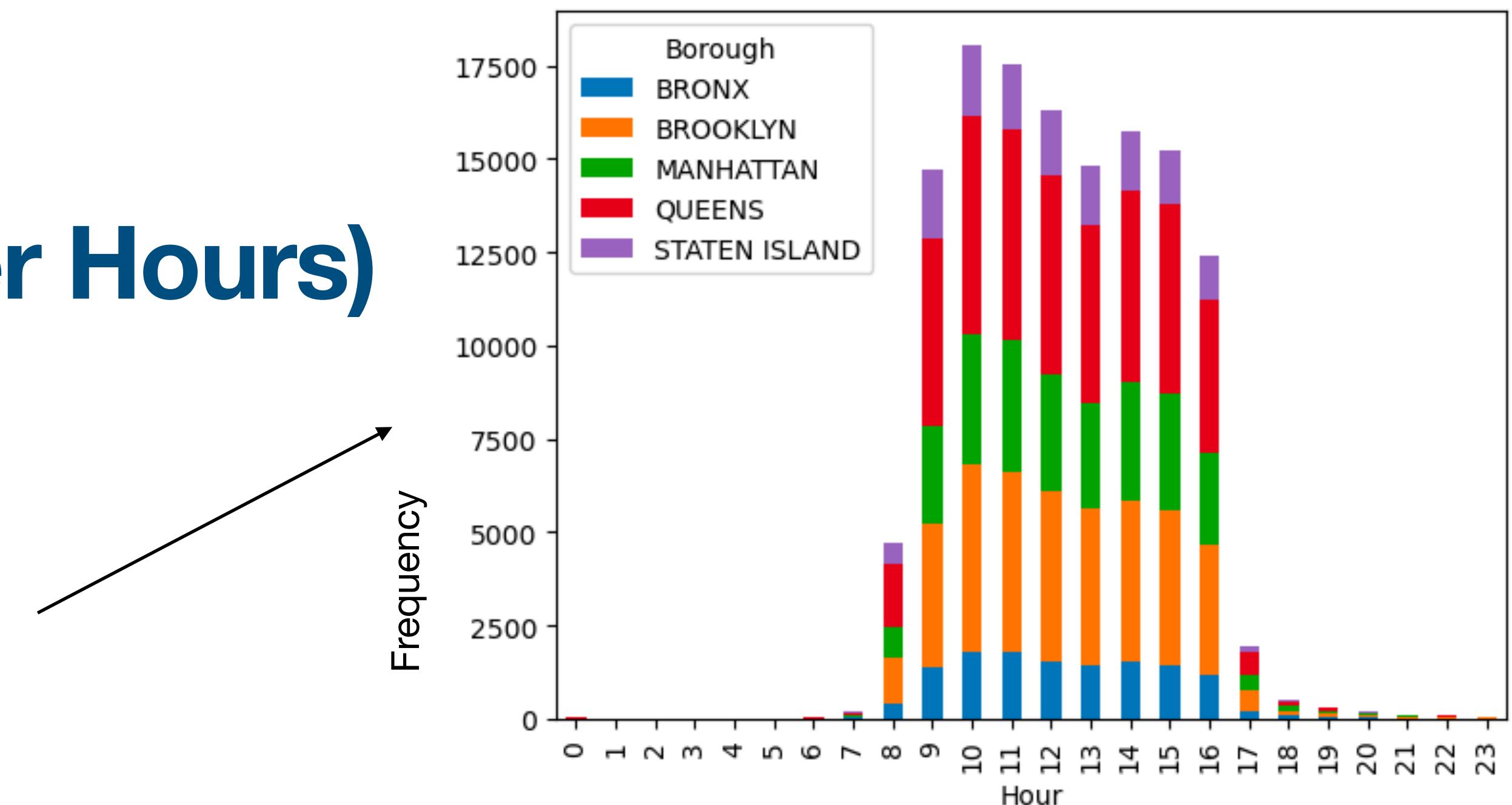
- DHS service requests mainly originate from Manhattan and is mostly dominant in the summer months
- HPD which receives the 2nd most number of calls, after NYPD is mostly witnessing requests in the winter.



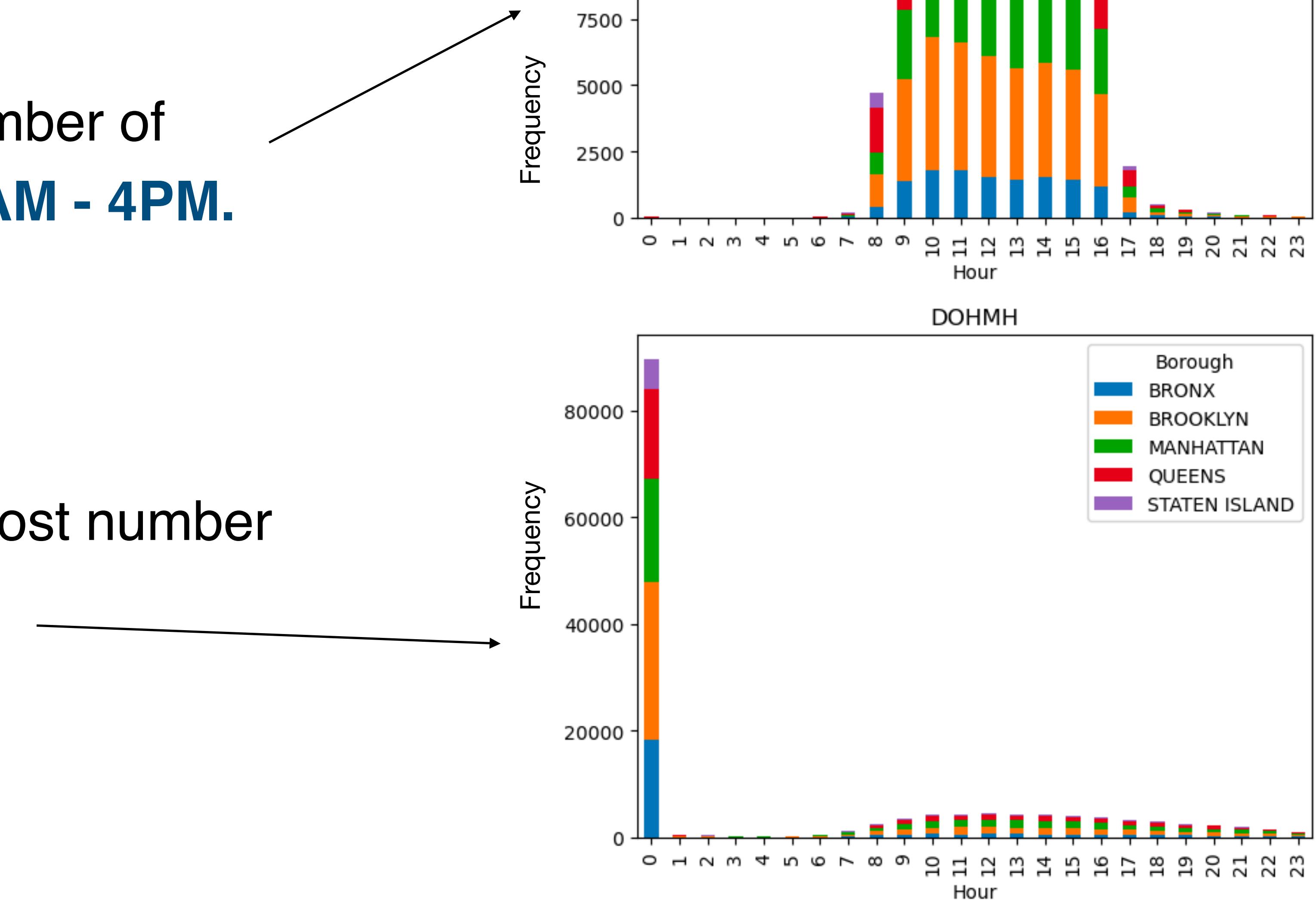
Data Understanding

Call Patterns - Agency (over Hours)

- **DOF** receives a high number of service requests from **9AM - 4PM**.

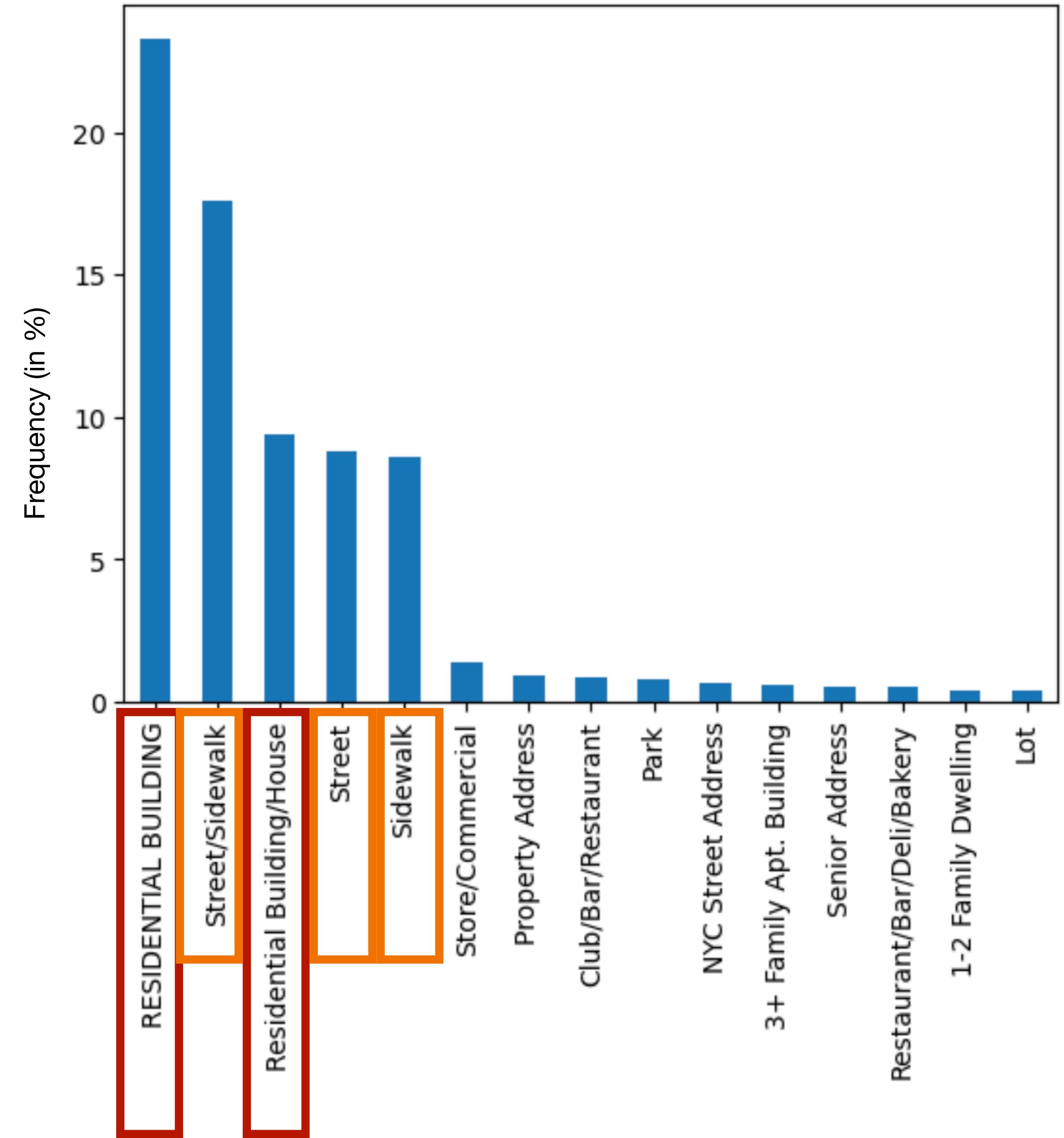


- **DOMHH**, receives the most number of calls **at midnight**.



Data Understanding

- Data grouping / categorization is **inconsistent**
- Used a string matching library: (`fuzzy_wuzzy`) to make more **robust categories**
- Found similar issues with:
 - **Location Type**
 - **Complaint Type**



Data Understanding

- Data grouping / categorization is **inconsistent**
- Used a string matching library: (`fuzzy_wuzzy`) to make more **robust categories**
- Found similar issues with:
 - **Location Type**
 - **Complaint Type**

Originally,

heat/hot water	499418
noise – residential	415754
blocked driveway	337988
illegal parking	330621
request large bulky item collection	226775

Now,

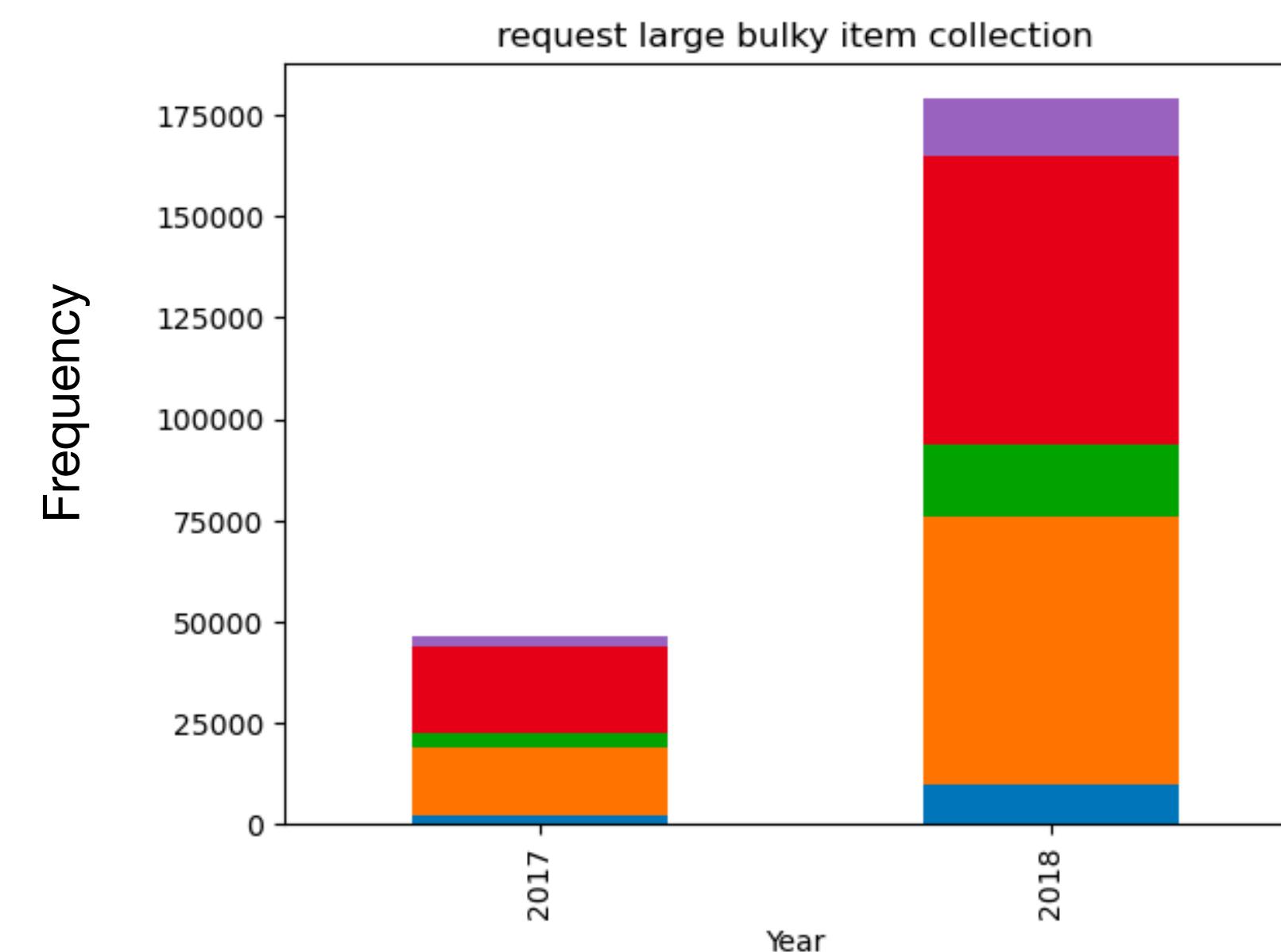
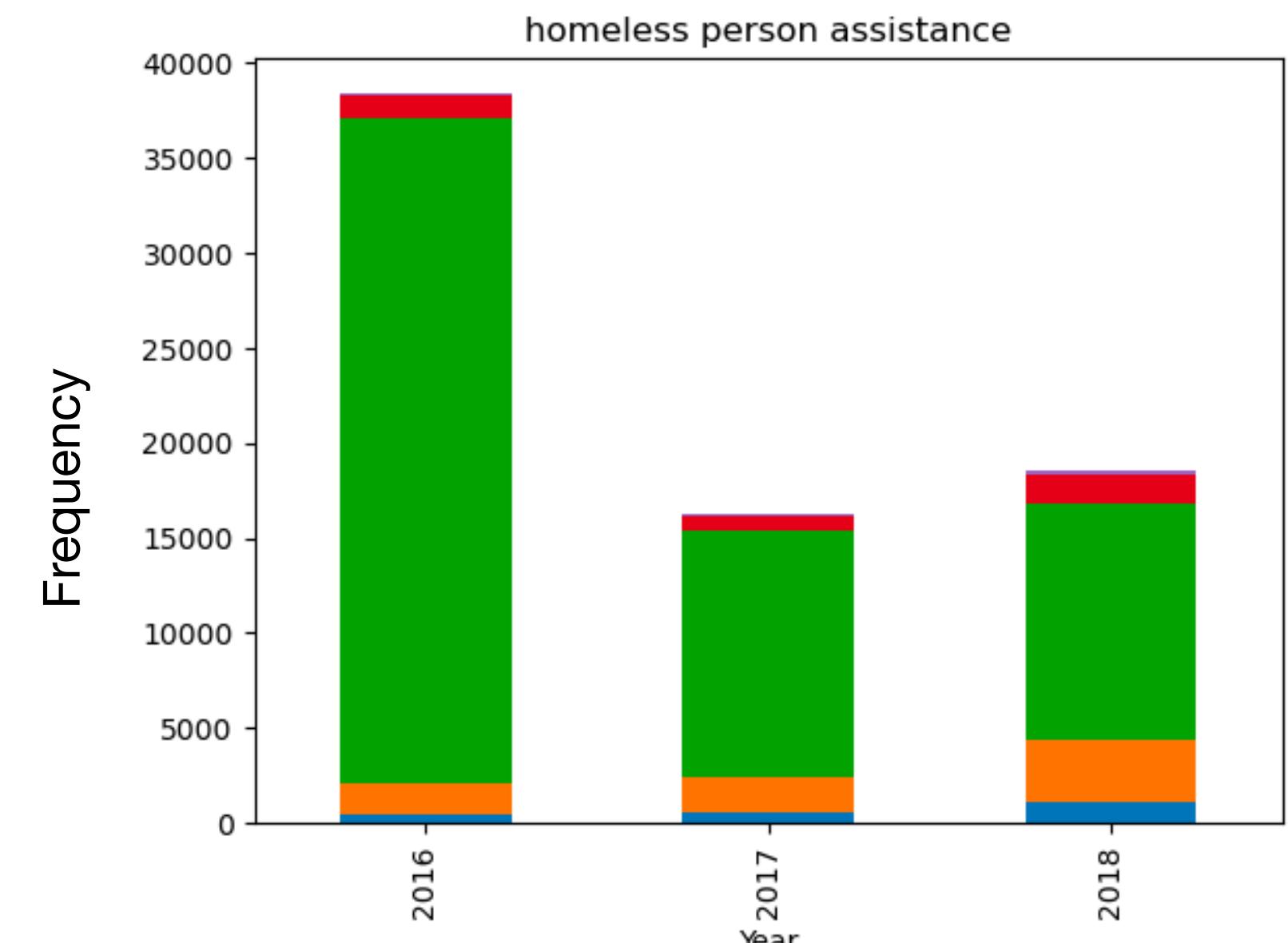


noise	823833
heat/hot water	499418
unsanitary condition	360490
blocked driveway	337988
illegal parking	330621

Data Understanding

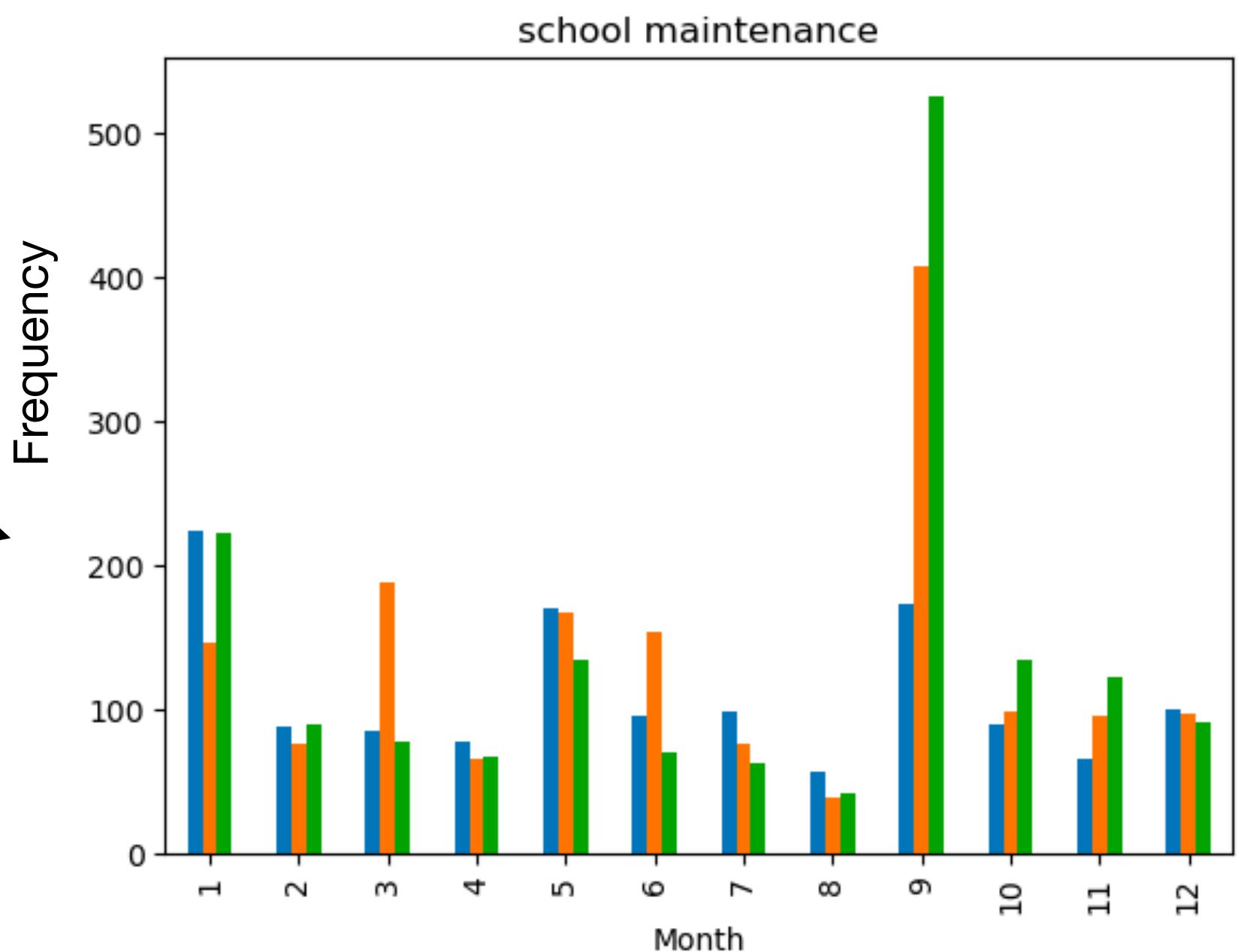
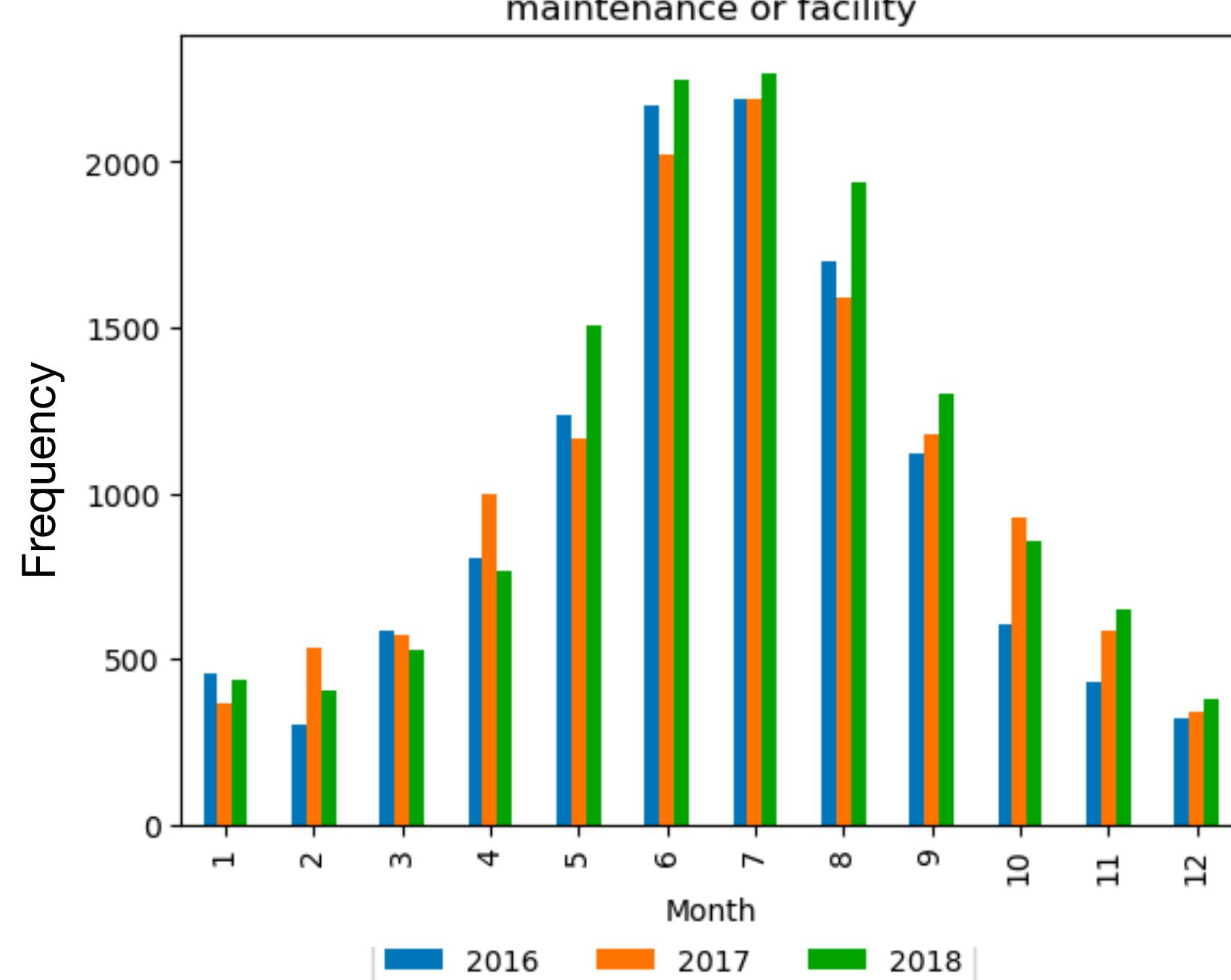
Call Patterns - Agency (over Hours)

- **Homeless person assistance** requests have decreased over the years, but is most dominant in **Manhattan**.
- **Request large bulky item collection** have increased exponentially from 2017. It also **did not exist** in 2016.



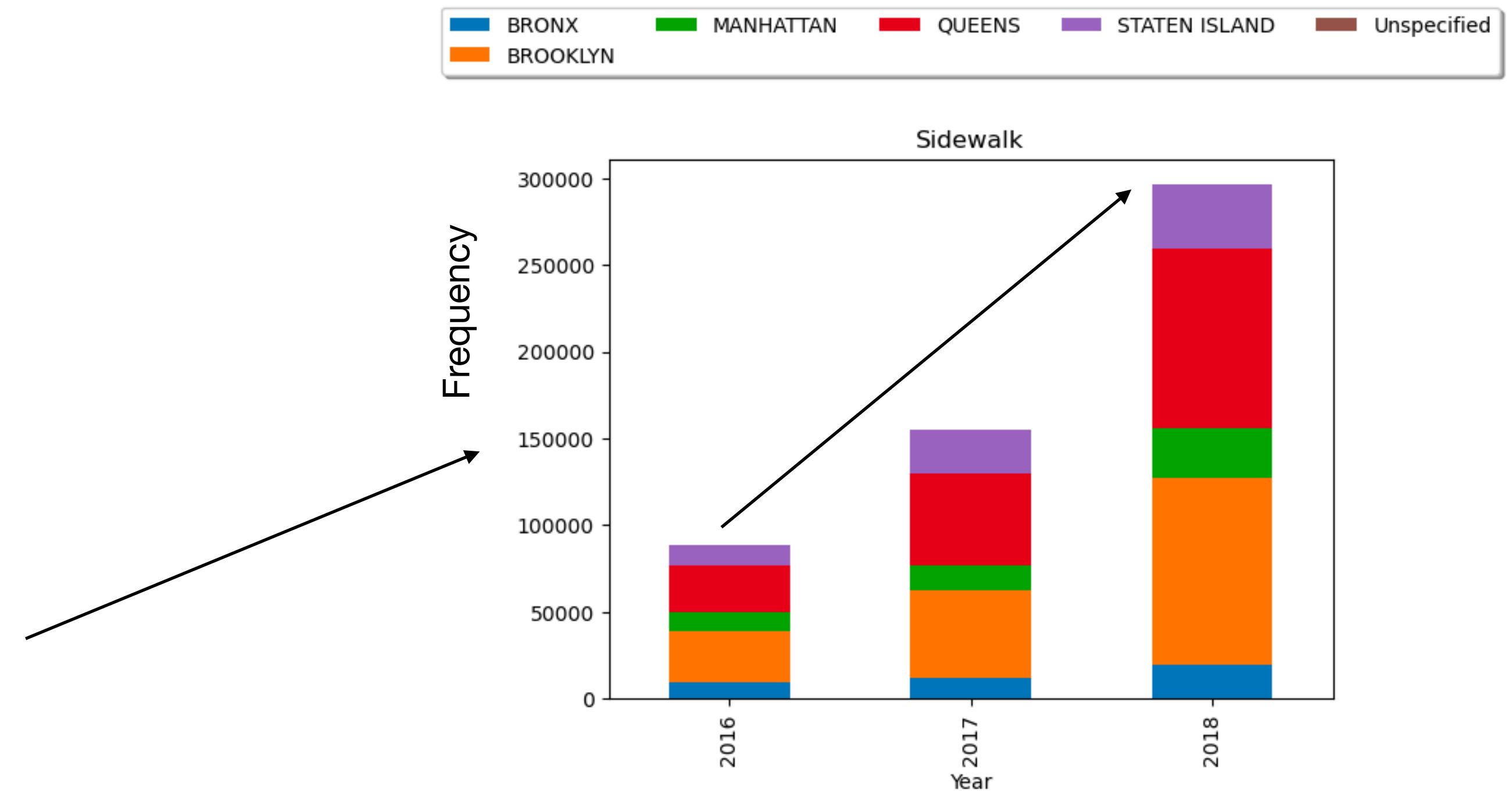
Data Understanding Call Patterns

- **Maintenance or facility** service requests increase until July and then decrease consistently.
- **School maintenance** peak in the month of **September**.

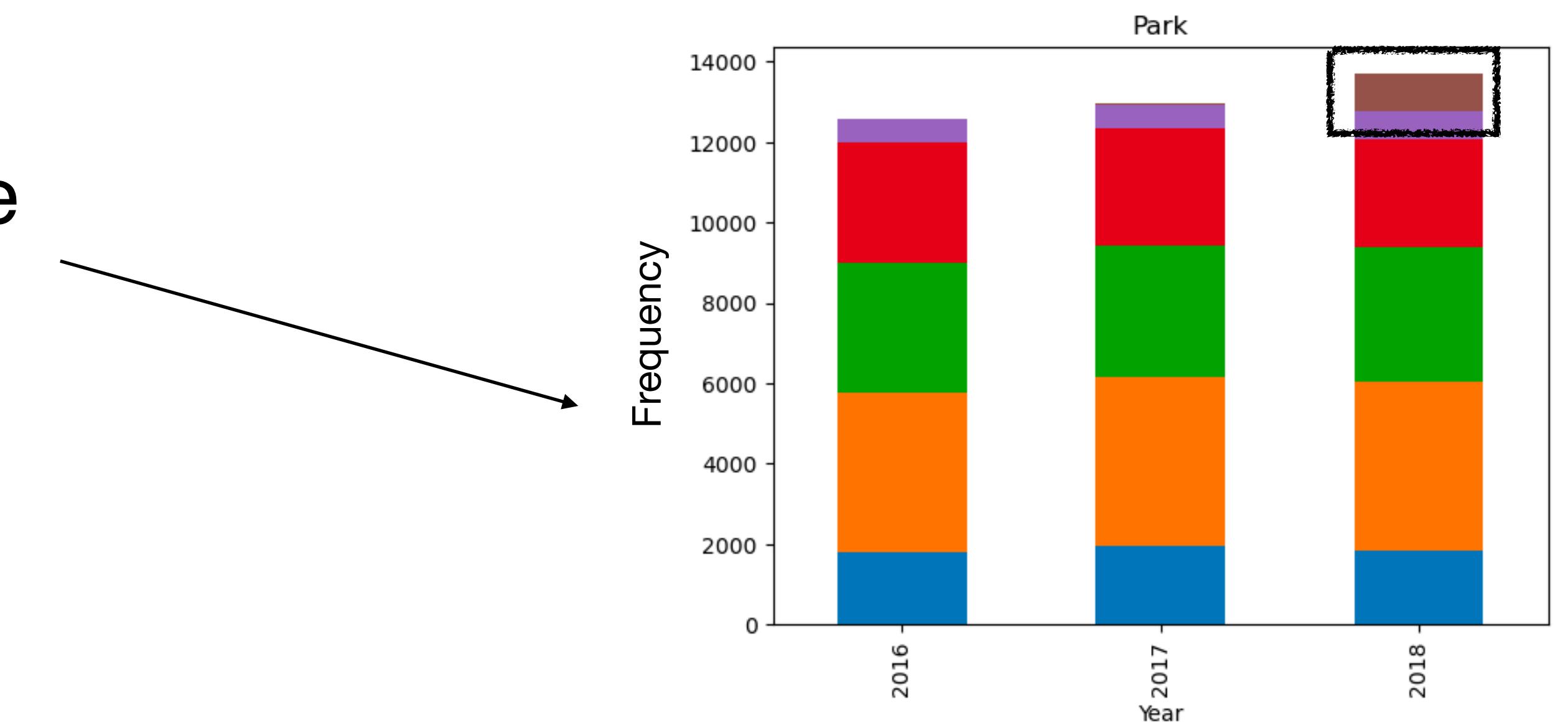


Data Understanding Call Patterns

- **Sidewalks** service requests have increased exponentially from 2016.

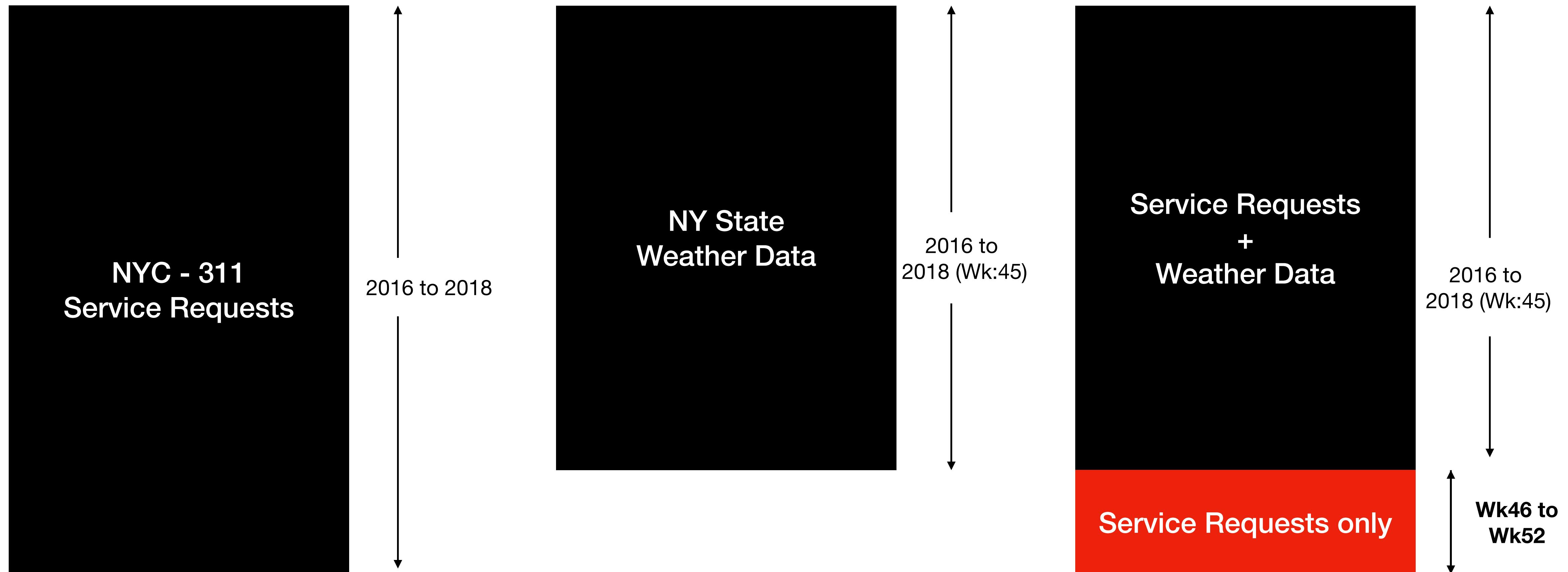


- In 2018, approximately 12.5% of the service requests at **parks** do not belong to a borough



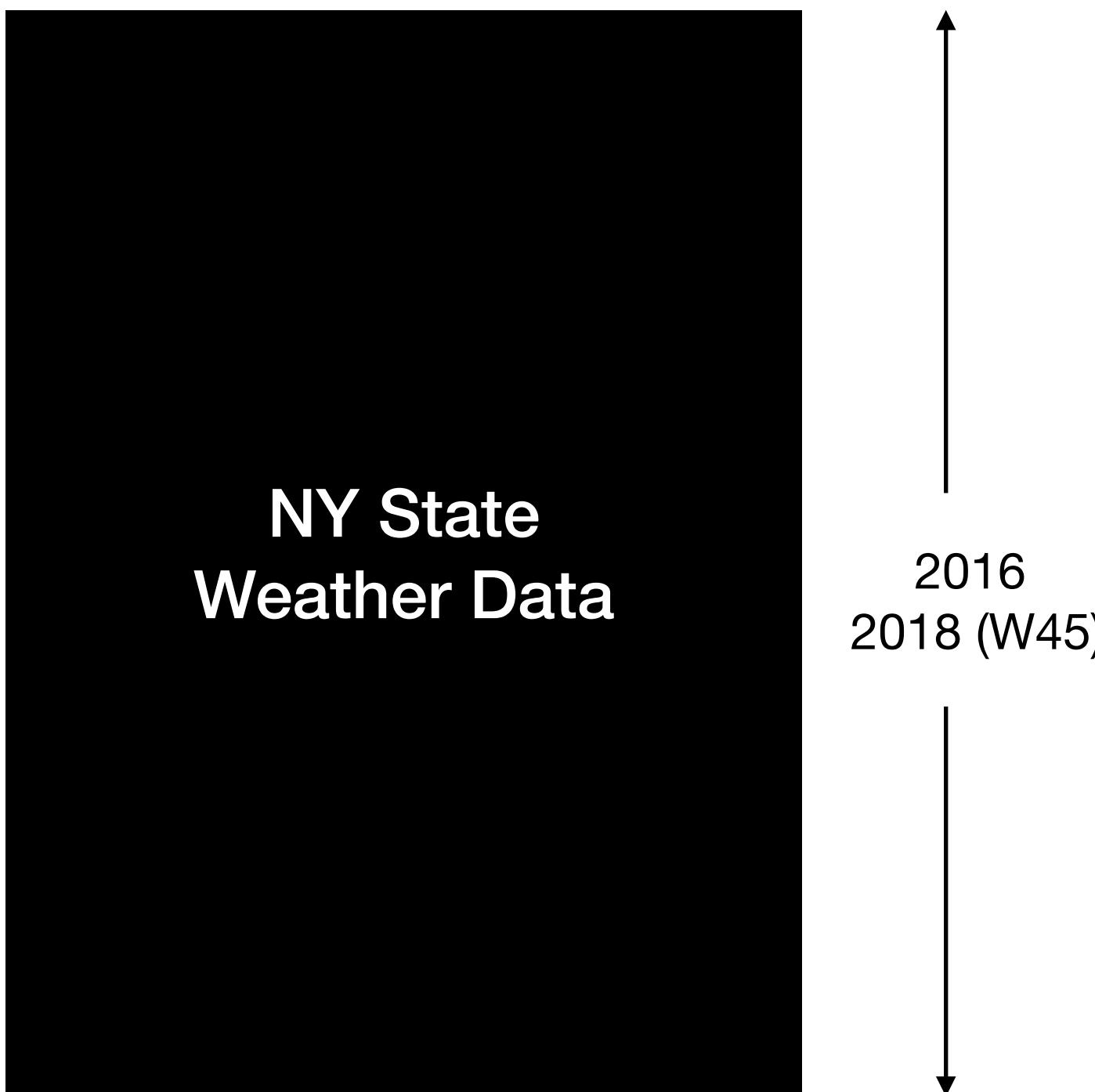
Data Preparation

Merging the two datasets



Data Preparation

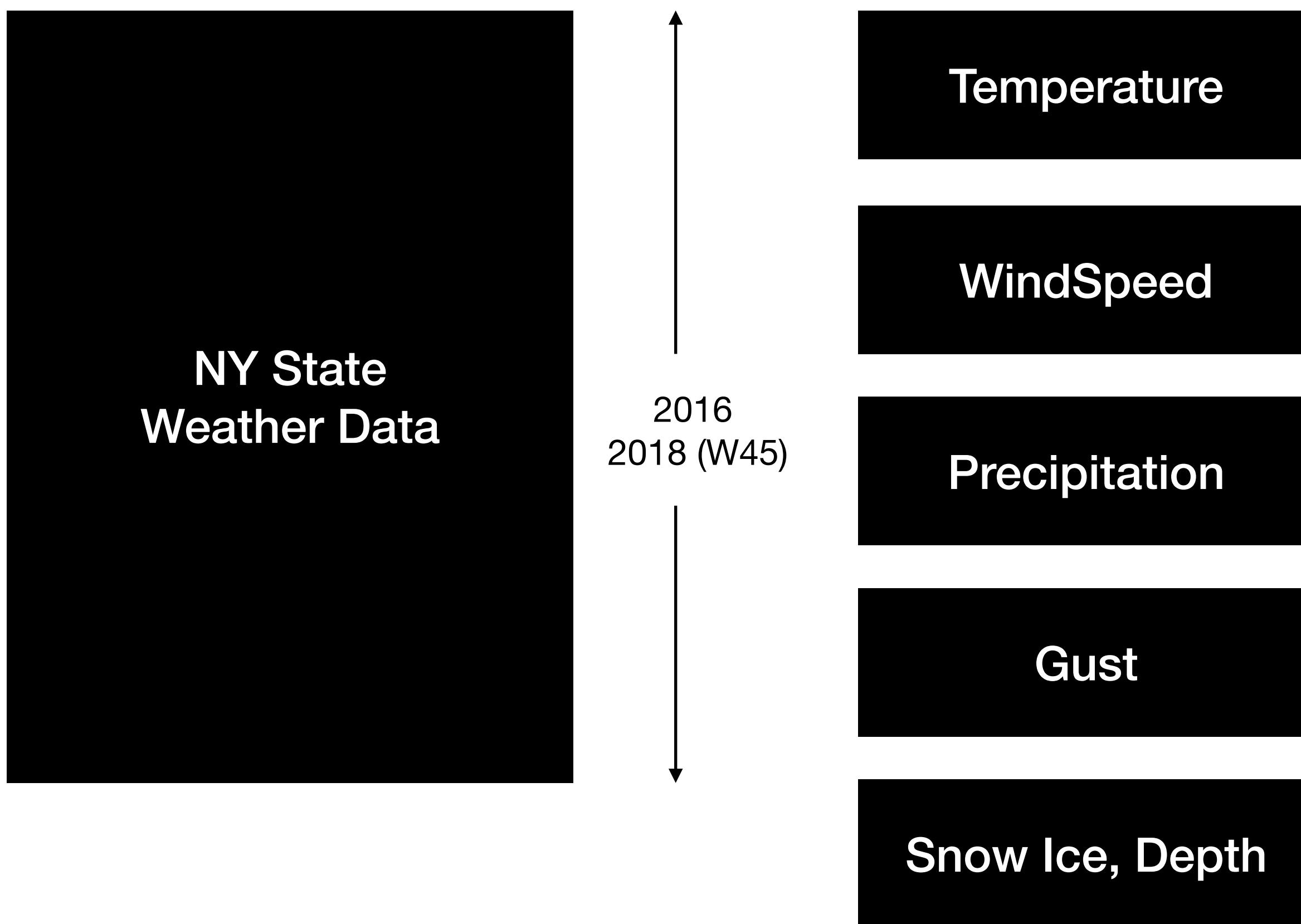
Understanding the Weather Data



- Average distance between stations:
~200kms
- We find the **max, min and mean (north, south and centre)** Latitude and Longitude for each borough in the dataset to get an idea of its bounds
- We use the geodesic distance to find the **5 nearest stations** to each borough and remove others

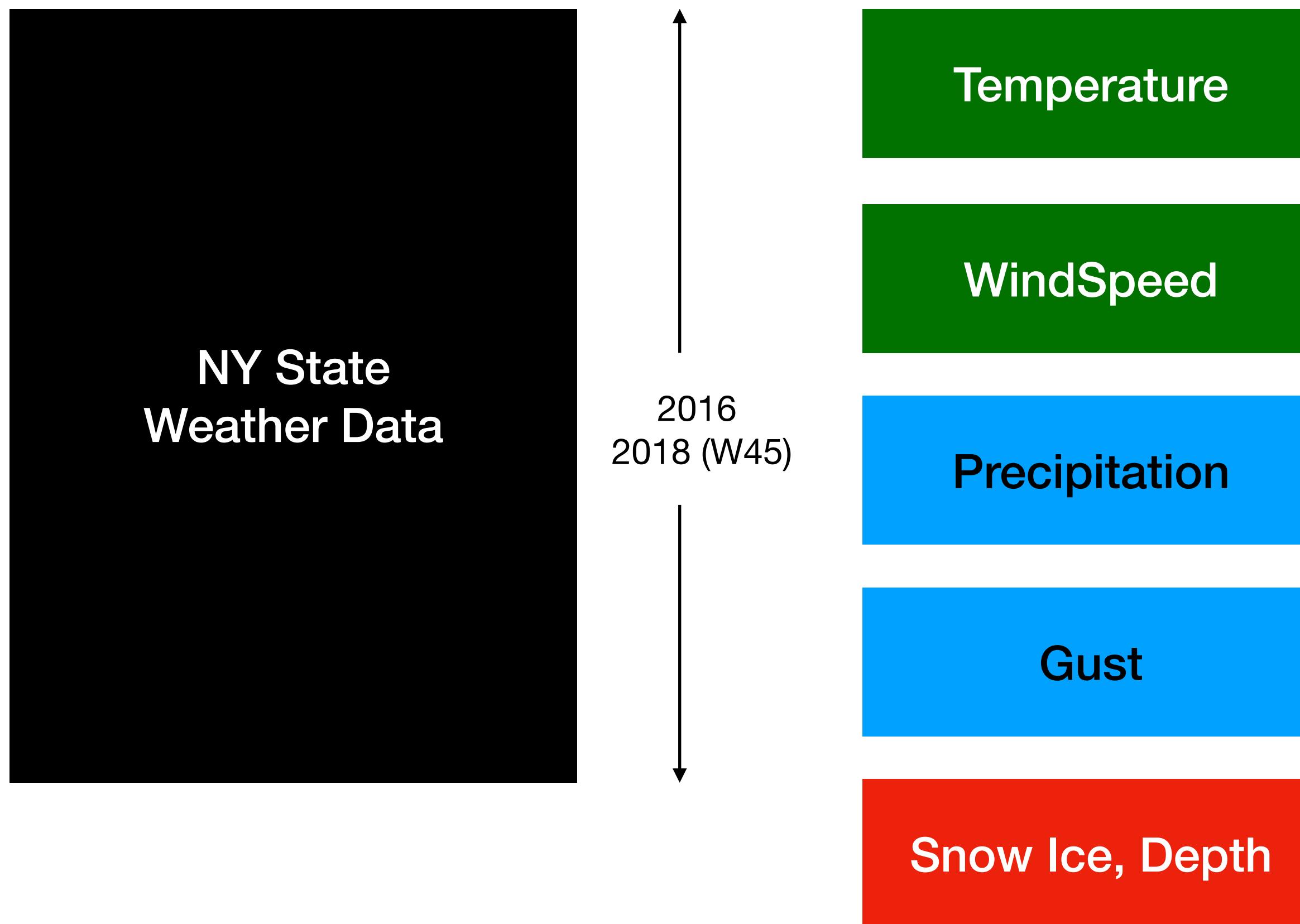
Data Preparation

Understanding the Weather Data



Data Preparation

Merging the two datasets



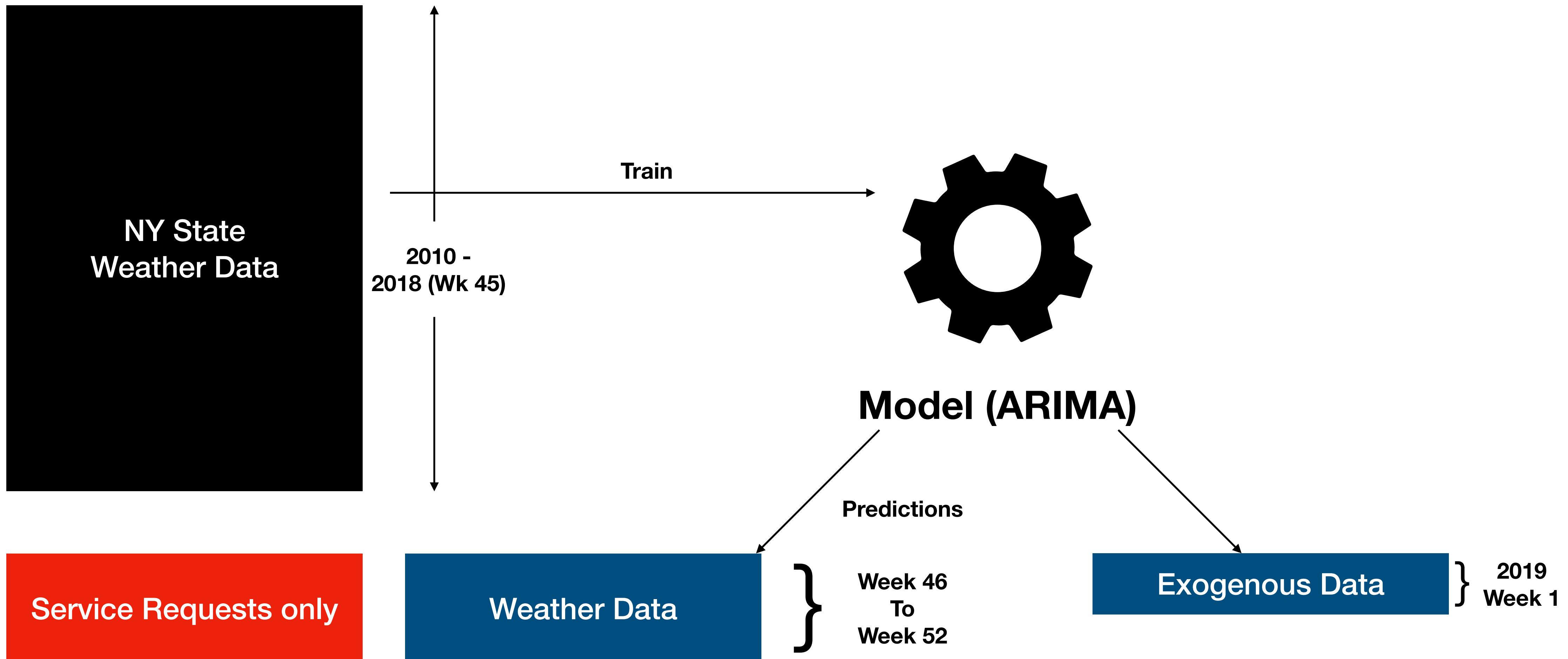
Feature Selection Basis:

Correlation + Variance Inflation Factor

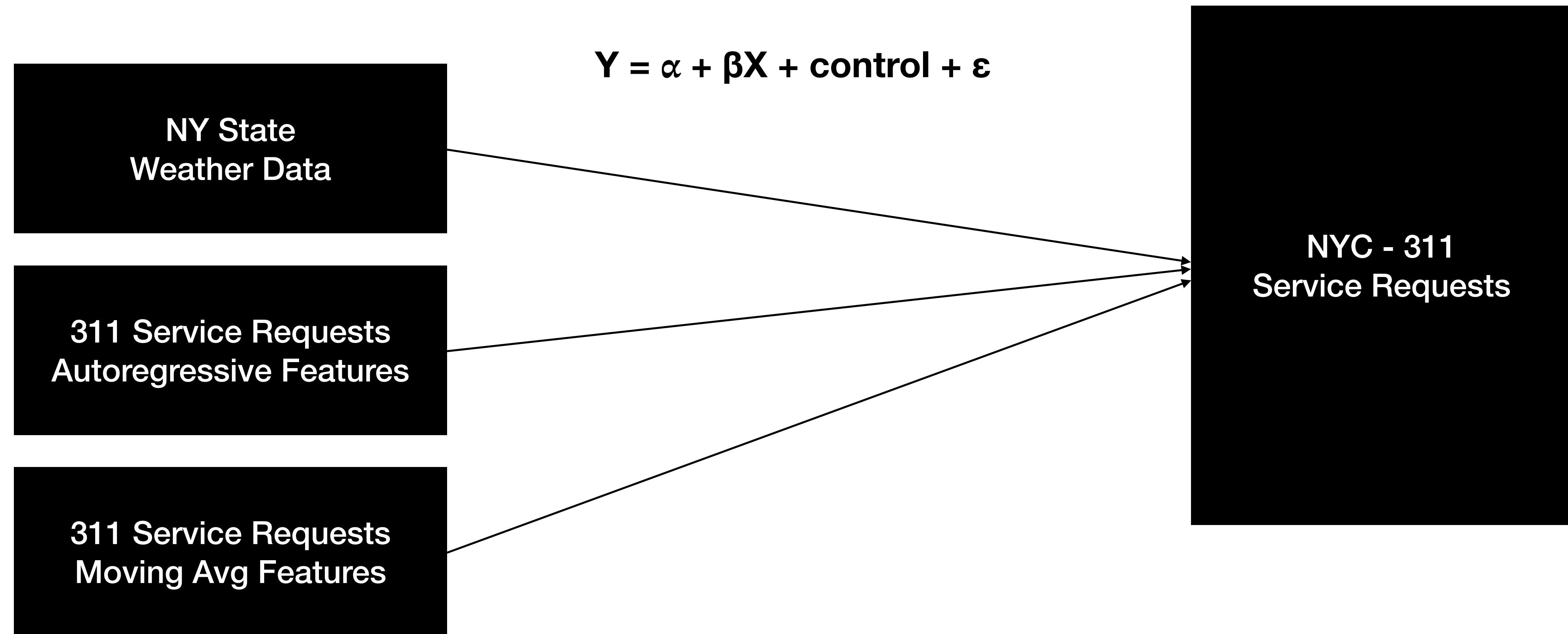
Noise

Data Preparation

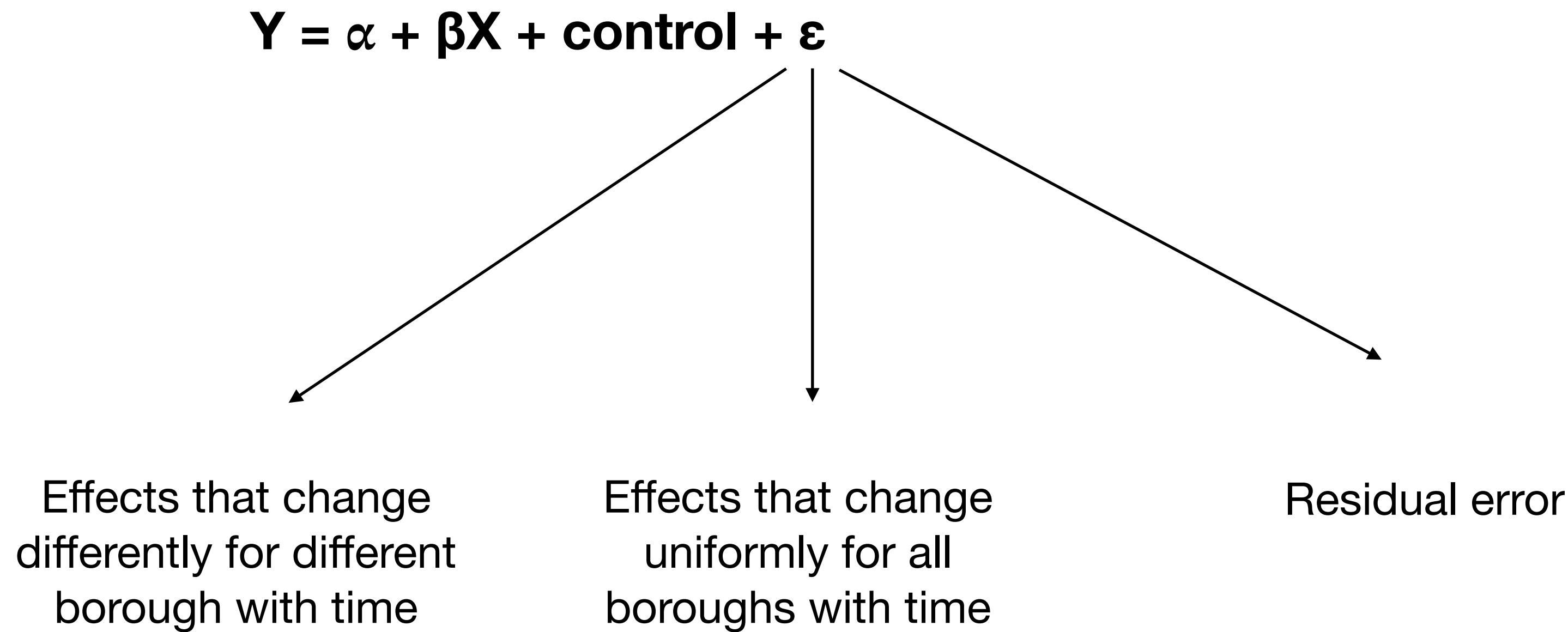
Prediction of Service Requests



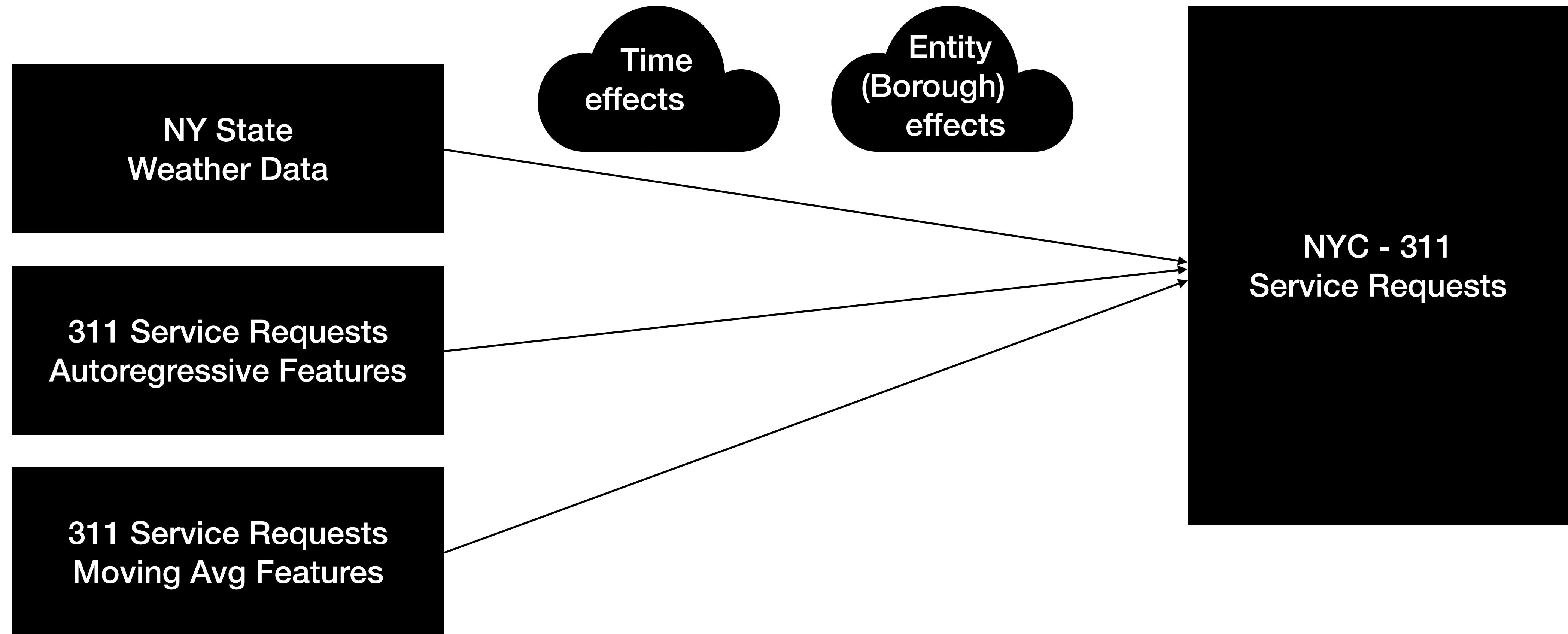
Data Validation Check for Causality



Data Validation Check for Causality

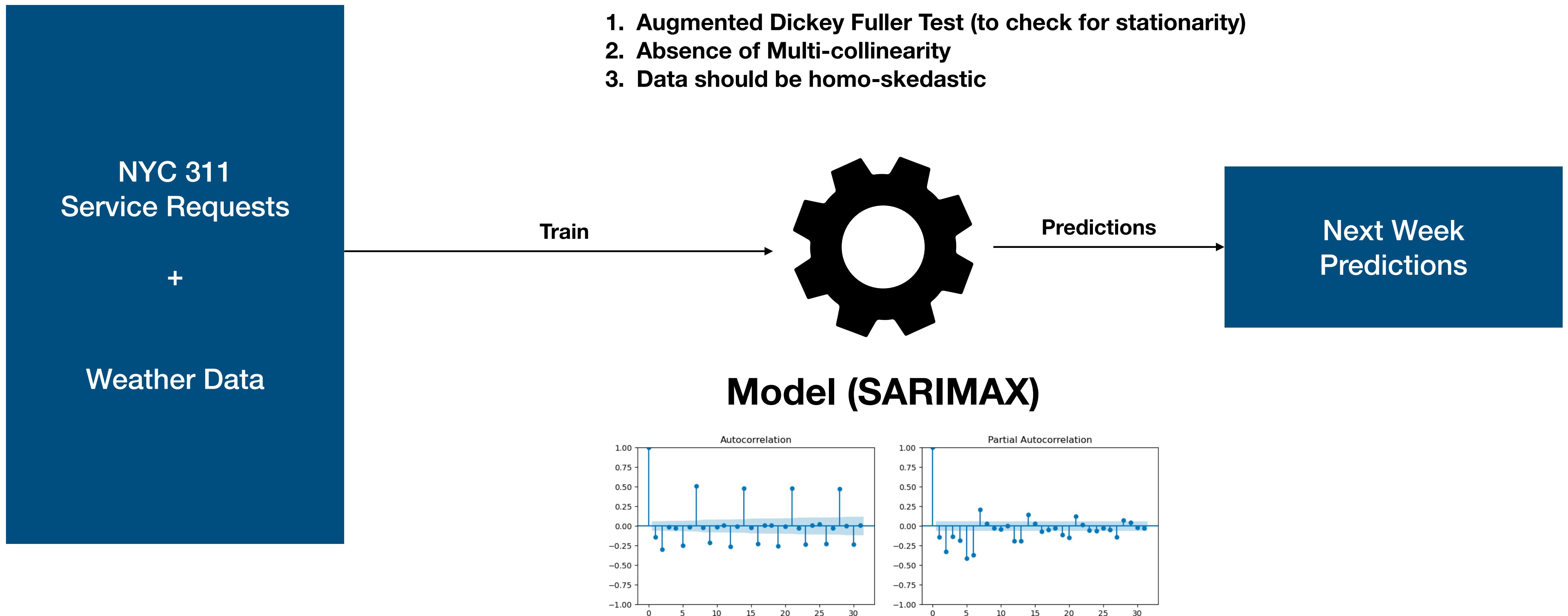


Data Validation Check for Causality - Panel OLS



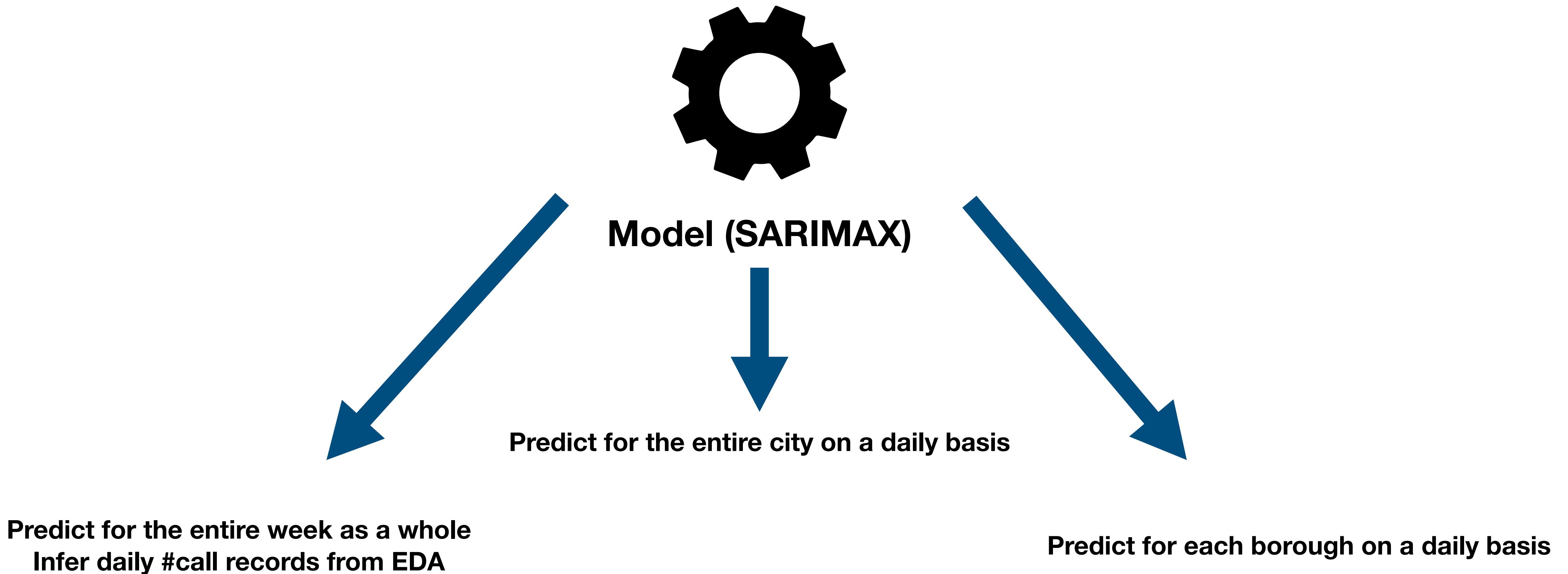
p-Value for Temperature = 0.04

Modelling Predicting the number of Service Requests



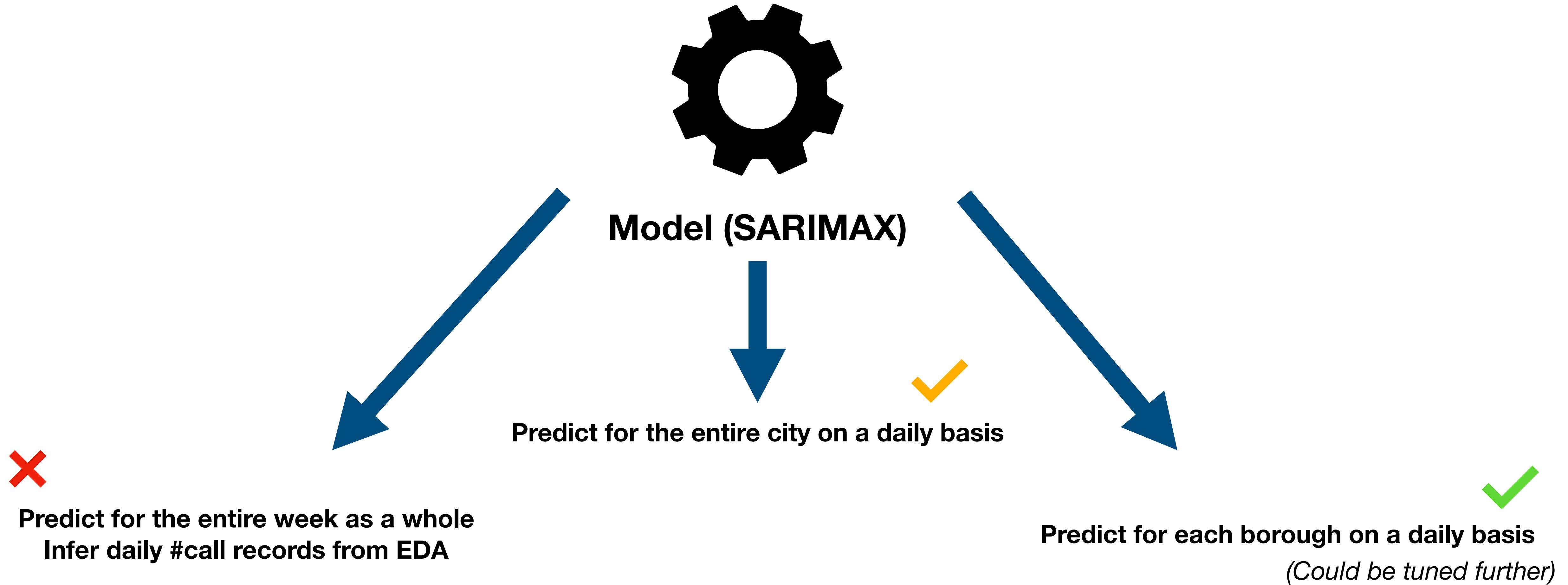
Modelling Approaches (Experimented)

Predicting the number of Service Requests



Modeling Approaches (Experimented)

Predicting the number of Service Requests

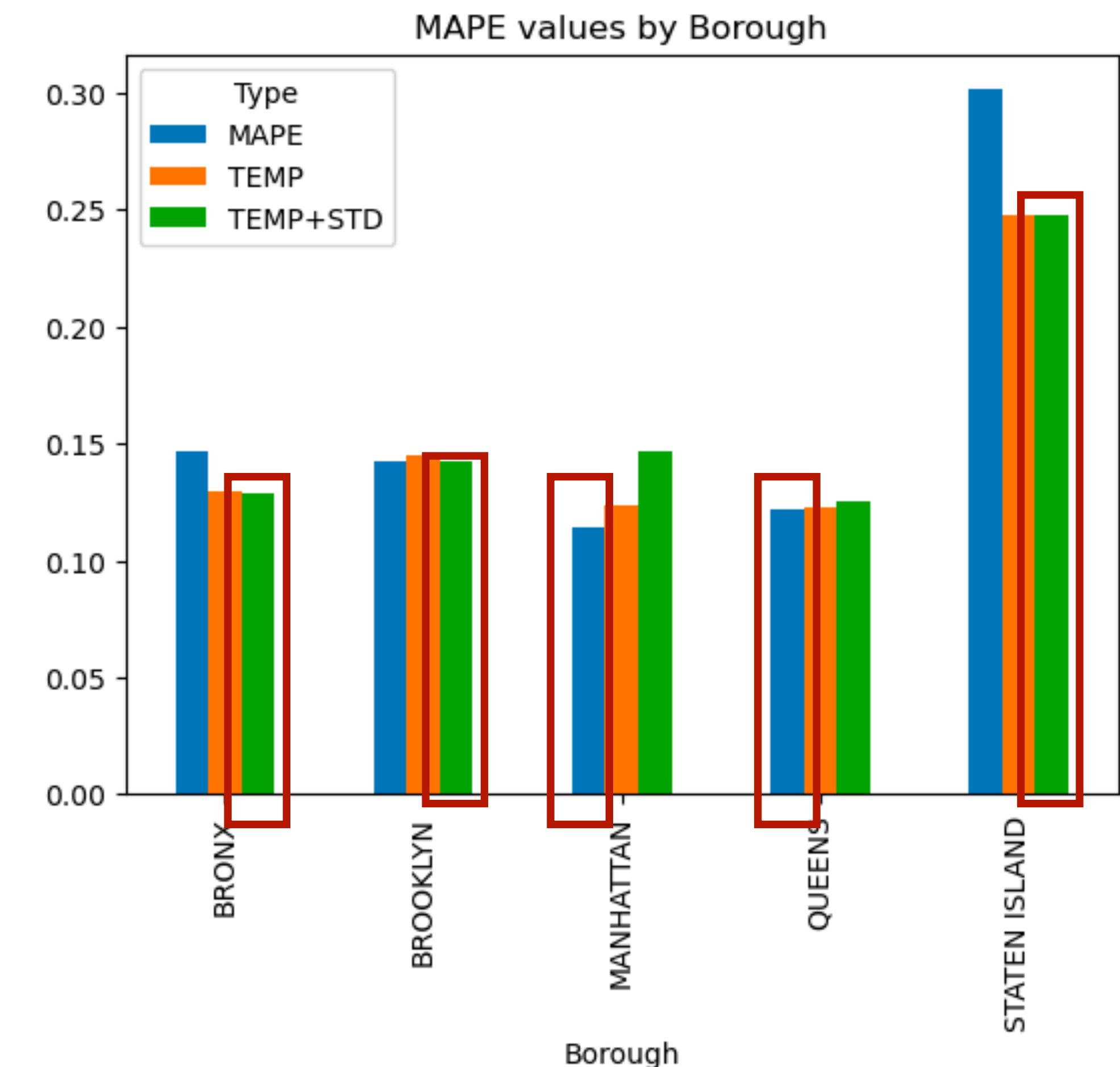


Modeling Approach 1

Predicting the number of Service Requests

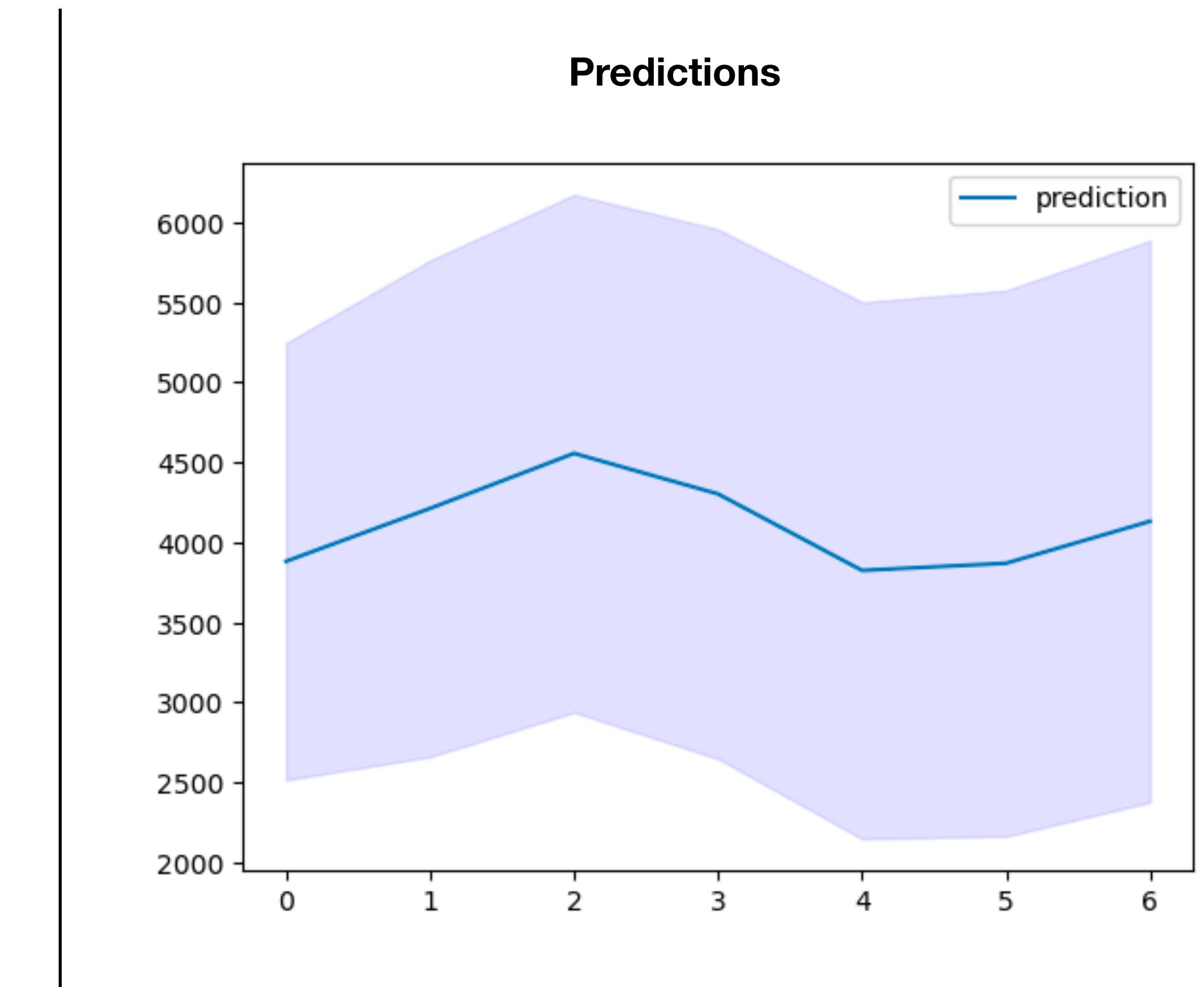
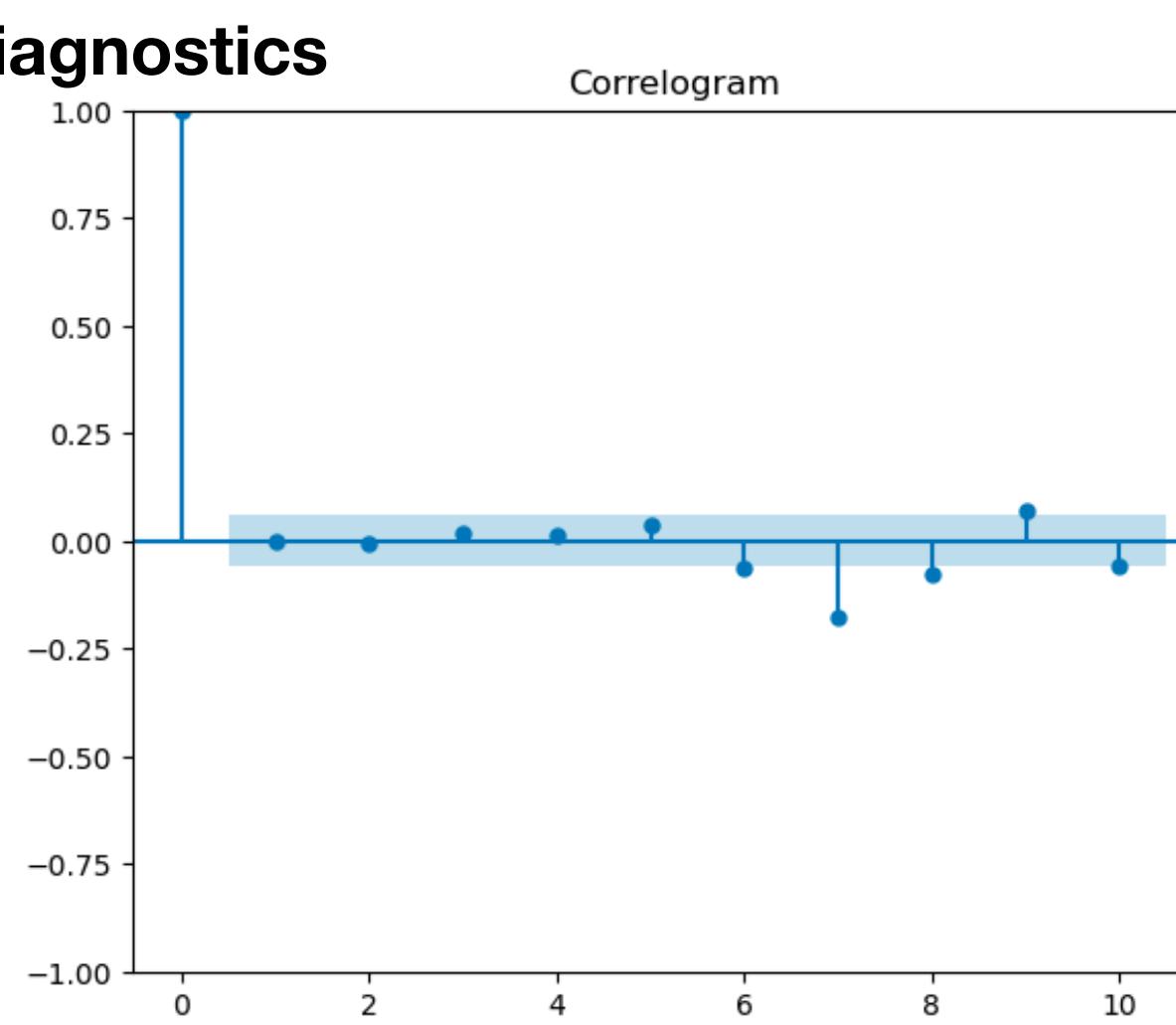
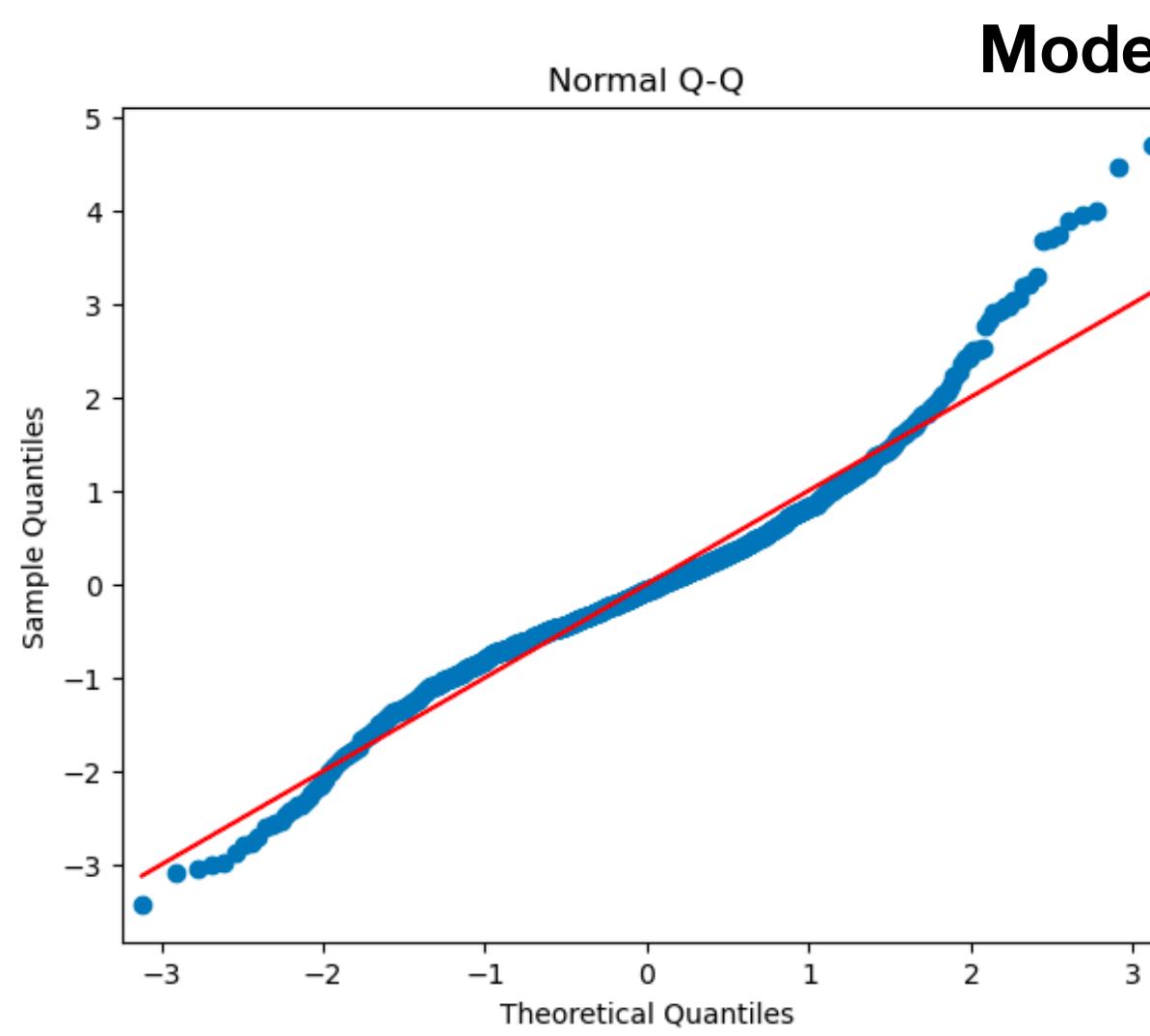
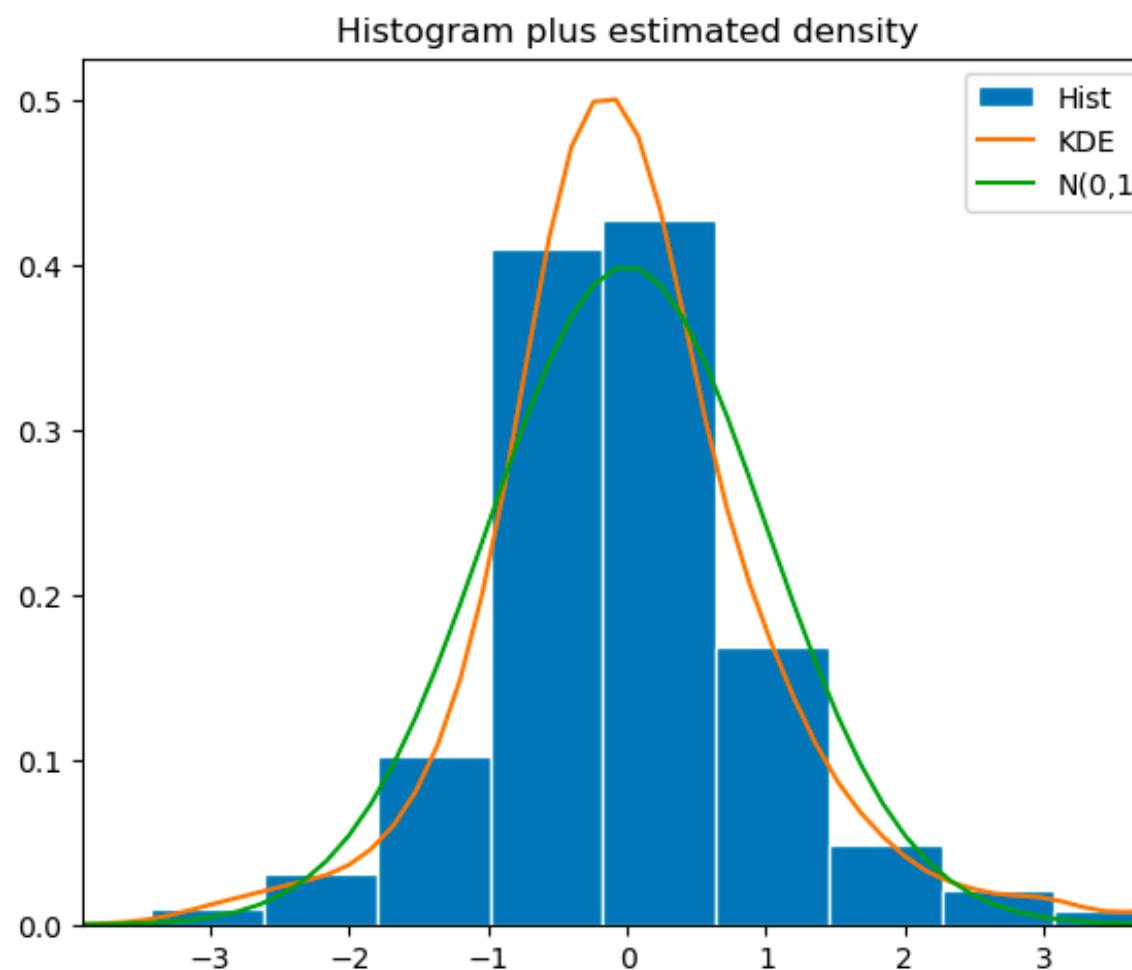
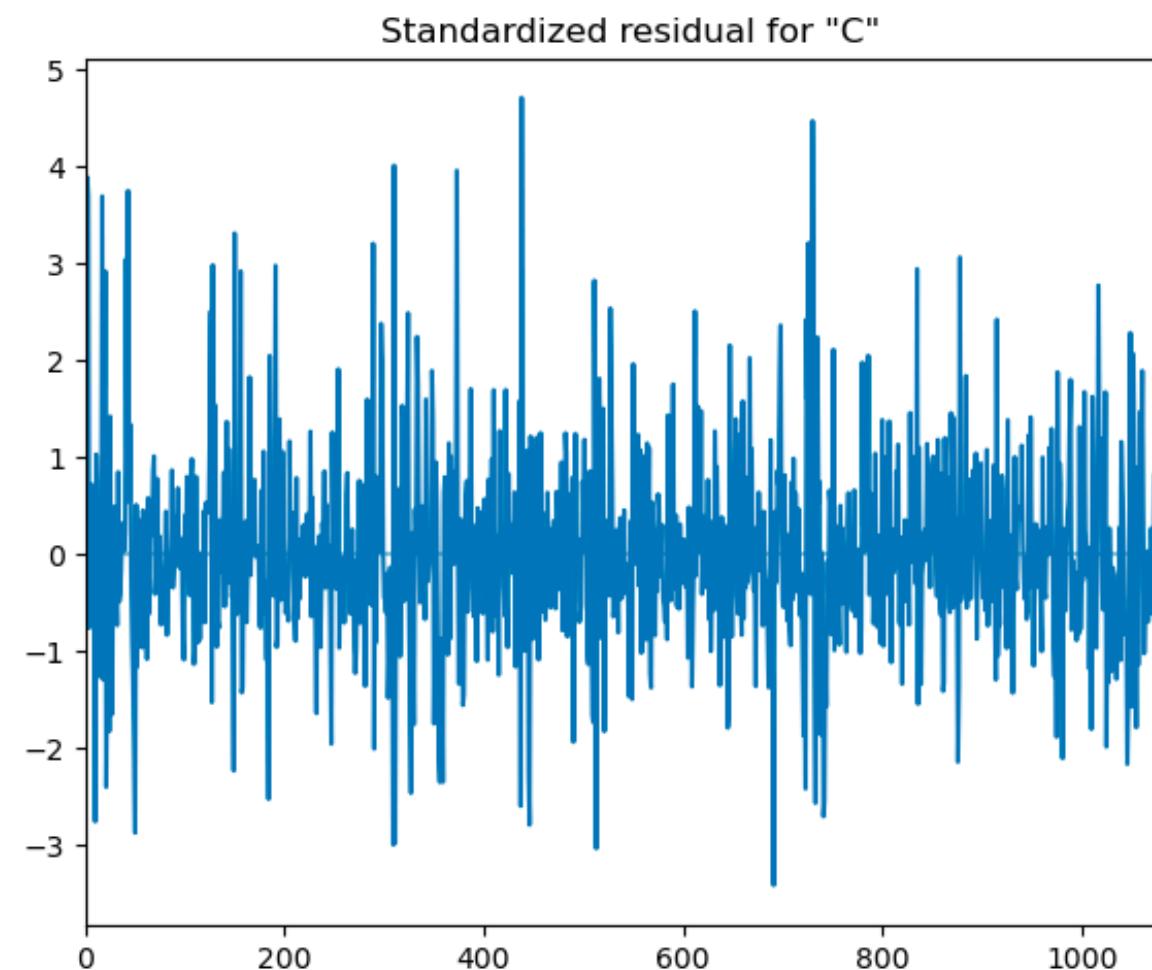
Predict for each borough on a daily basis

- MAPE: **Normal ARIMA** model
(with no exogenous variables)
 - TEMP: **SARIMAX** model
(with mean temperature as exogenous variable)
 - TEMP + STD: **SARIMAX** model
(with mean temperature and standard deviation as exogenous variable)
- MANHATTAN, QUEENS**
- ALL OTHERS**



Modeling

Analyzing the Model Performance



Deployment

Model Deployment

- Check for model drift and data drift.
- Use of data versioning tools to understand
- Develop data wrangling pipelines
- Performance drift across each borough's model with time
- Changes in the implied data distribution (service requests from boroughs)

Future Scope

- Alternative Datasets including:
 - Infrastructure Development / Spending
 - Crime Datasets
- Causality
 - Check for **heterogenous effects** across boroughs
- Other Models:
 - Prophet, Linear Regression with Exponential Smoothening
 - With Explainability measured using **SHAP and LIME**

Future Scope

- Modeling Approaches
 - Develop models for each borough and find exogenous variables that dominate in said boroughs. (Using **cross correlation**).
 - Use **intermittent models** to predict for particular seasons such as **iMAPA** and **Croston-TSB** to measure muted effects.

Thank You