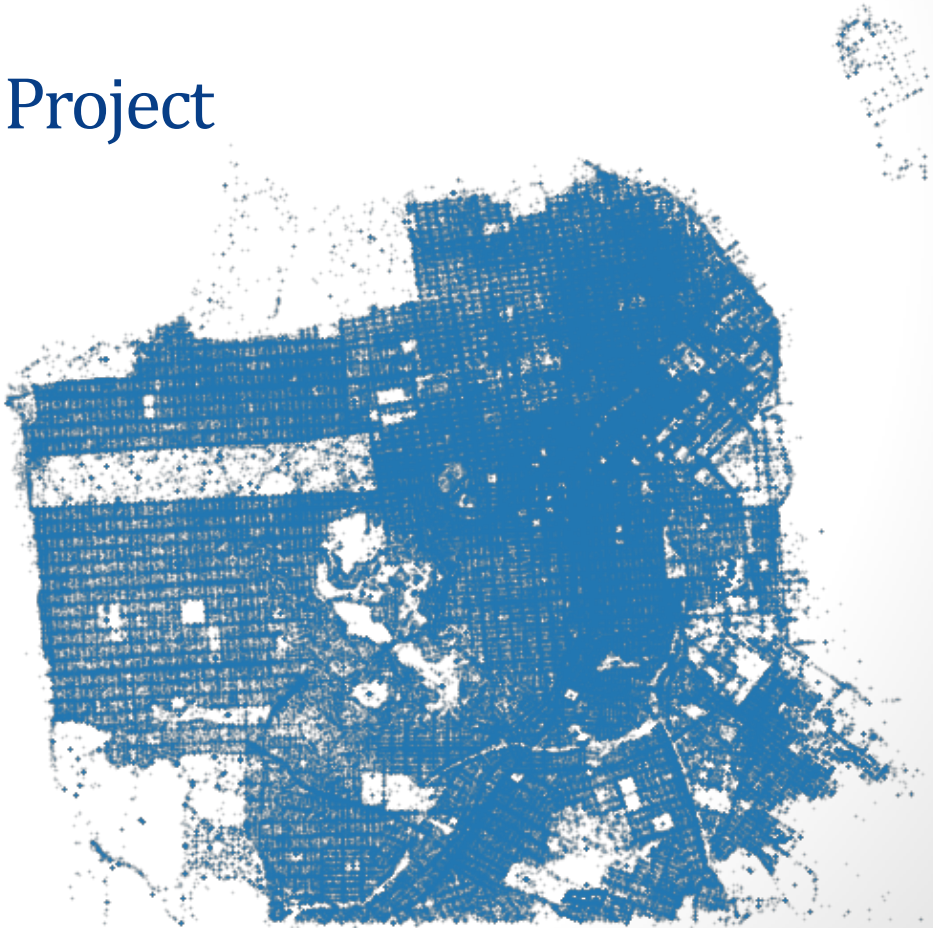
San Francisco 311 Data Predictions

Springboard Capstone #2 Project

Pawandeep Jandir



Springboard

Overview

- Introduction
  - Background and problem statement
- Data Wrangling
  - Data sources used, cleaning performed and full list of variables
- Data Analysis
  - Exploring the data with closer look at service request Categories
- Data Modeling and Predicting
  - Preprocessing data and results of models
- Conclusion
  - Further work

Introduction

- San Francisco operates a 311 program
  - Official platform for residents to get information, report problems, or submit service requests (non-emergencies only)
  - Available 24 hrs a day via phone, text, mobile app, website, and more
- 311 program is method of identifying how local government operates
  - Offers a look into how city services are provided and how people are interacting with the city
- Can we predict future service requests, in volume, in different areas of San Francisco?
  - Way to optimize the government and improve its efficiency and effectiveness for its residents
  - Can also use demographic & socio-economic data
- Project uses Python end-to-end

Data Sources

- 311 program data: From San Francisco's open data portal
  - Data available in multiple formats
  - Taken as single CSV file for this project
- US Census data: From the US Census Bureau FactFinder website
  - Data available in multiple formats after searching for specifics
  - Taken as multiple CSV files for this project
- San Francisco boundaries: From San Francisco's open data portal
  - Shapefile available in multiple formats
  - Taken as a single GeoJSON file for this project

Data Cleaning

- 311 data has over 2.8 million records dating back to mid 2008
- In order to connect a record to a US Census Tract within SF, need to use US Census' API
  - API is slow—would take 16+ days to cover entire dataset
- Large reduction in data size is needed
  - Keep only the data that passes these rules:
    - Street address that the API can easily read
    - Within top 8 service request Category types (listed below)
    - Within SF's boundaries
    - Within the years 2015 and 2017
  - Eventually get to 550,000 records

- Street and Sidewalk Cleaning
- Graffiti
- Abandoned Vehicle
- Homeless Concerns
- Encampments
- Sewer Issues
- Streetlights
- Damaged Property

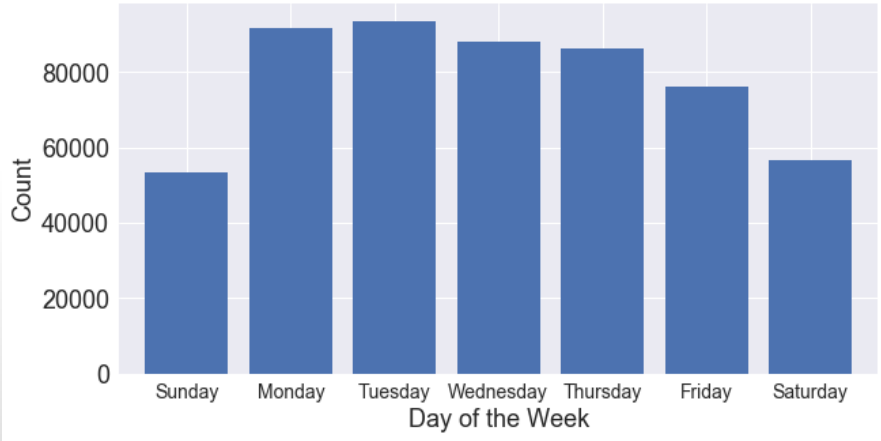
List of features in dataset

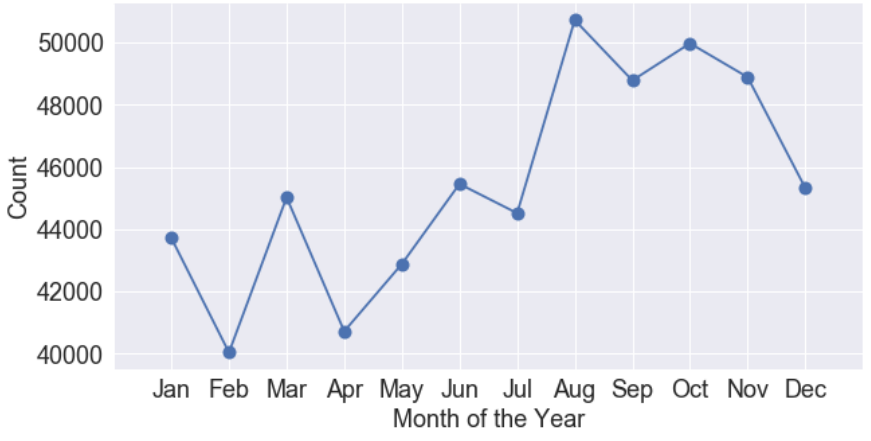
- The fully cleaned dataset contains 24 columns
- The table below lists these variables with a short description and any relevant notes

Name	Description	Notes
Date	Date of the service request	-
Category	Service request type	Eight categories
Supervisor_District	SF Supervisor District	Twelve categories
Police_District	SF Police District	Ten categories
Latitude	Latitude of service request	-
Longitude	Longitude of service request	-
Source	Ways to make service request	Seven categories
Has_Media_Link	If media was included in service request	Boolean
Is_Closed	If service request is now closed	Boolean
Tract	Given US Census Tract	199 categories
Poverty	Poverty ratio in percent	Range from 0 to 100
Income	Median income in dollars	Range from 19,440 to 250,000
Unemployment	Unemployment ratio in percent	Range from 0 to 100
Population	Total population	Range from 89 to 13,057
Percent_Male	Percentage of male population	Range from 0 to 100
Percent_Pop_0_14	Ratio of population between 0 and 14 years old	Range from 0 to 1
Percent_Pop_15_24	Ratio of population between 15 and 24 years old	Range from 0 to 1
Percent_Pop_25_64	Ratio of population between 25 and 65 years old	Range from 0 to 1
Percent_Pop_65_up	Ratio of population over 65 years old	Range from 0 to 1
Percent_Pop_White	Ratio of population identifying as white	Range from 0 to 1
Percent_Pop_Black	Ratio of population identifying as black	Range from 0 to 1
Percent_Pop_Hispanic	Ratio of population identifying as hispanic	Range from 0 to 1
Percent_Pop_Asian	Ratio of population identifying as asian	Range from 0 to 1
Percent_Pop_Other	Ratio of population identifying as other	Range from 0 to 1

Exploring Service Requests I

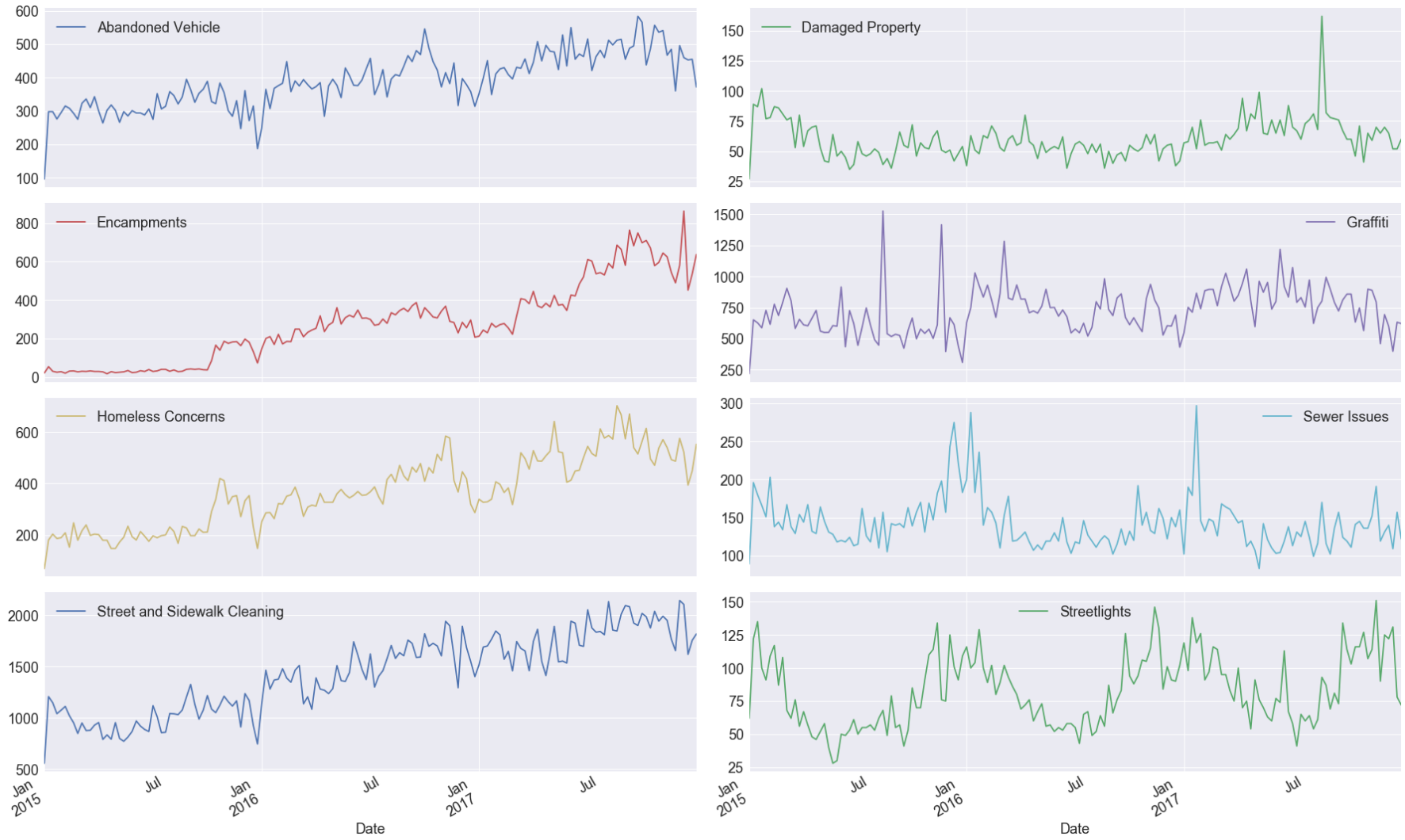
- Exploratory analysis of the data was particularly focused on the time (series) component of the service request data
- Graphs below show the total counts in different-sized bins across time
  - Left plot shows the total counts per day of the week
    - Less requests on the weekend
  - Right plot shows the total counts per month of the year
    - More requests in the second half of the year





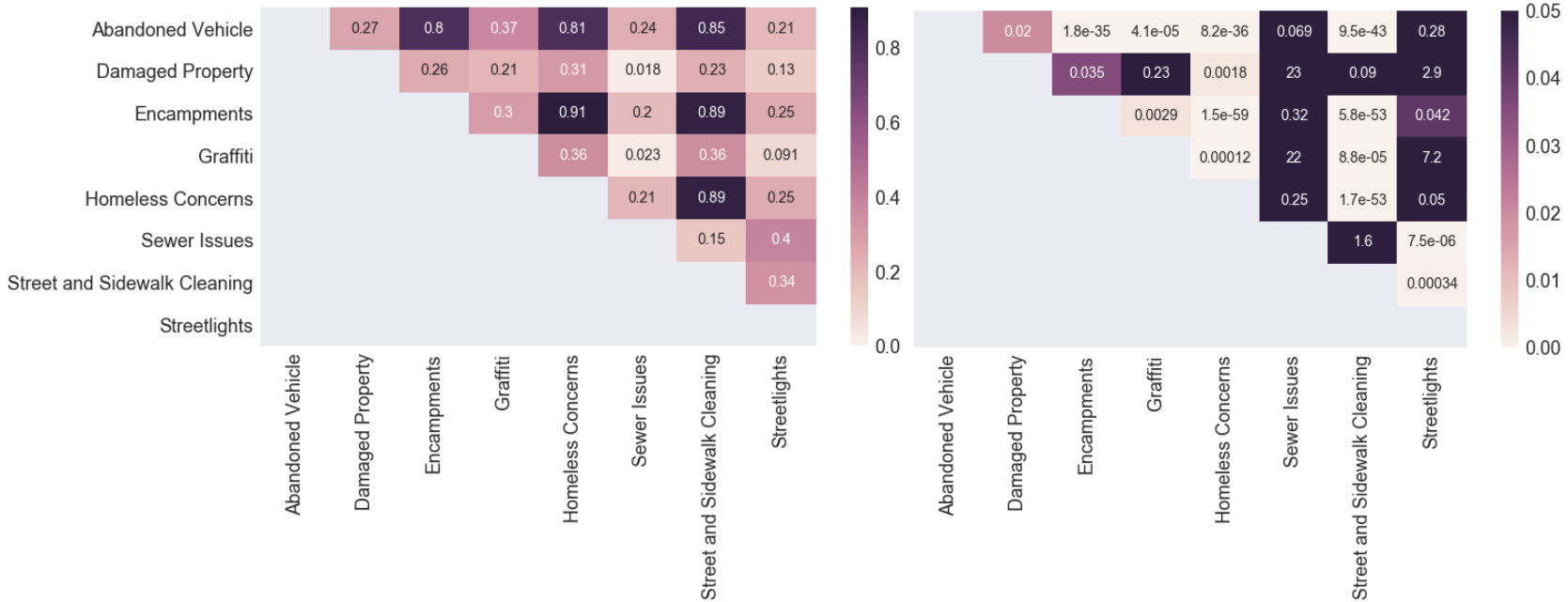
Exploring Service Requests II

- Plot weekly service request count for the eight Categories



Exploring Service Requests III

- Calculate the Pearson correlations between the eight weekly service request time series counts (see previous slide) on left and the associated *p*-values for significance on right
  - Mindful of multiple comparisons problem
    - Apply Bonferroni-like correction to the *p*-values
  - Half of the time series seem to share significant correlations
    - Makes sense: expect Encampments and Homeless Concerns to be related



Preprocessing Data

- Before building prediction models, need to preprocess and standardize the data
  - Recall list of variables on slide 6
- Standardization or normalization of data needed since different ranges exist in similar US Census features
  - Most of the US Census variables are ratios, ranging from 0 to 1
  - A few columns are ratios but range from 0 to 100
    - Scale them down
  - A few columns have arbitrary scales
    - Scale them down by maximum value of column

Time Series Modeling I

- Usual algorithm to model time series is ARIMA
  - Used ARIMA with additional Seasonal terms as well as eXogenous variables, the US Census columns: SARIMAX algorithm
- Goal is to model all 8 Categories in 195 US Census Tracts: 1,560 time series
  - Too many to identify SARIMAX parameters by hand or eye, like usual
  - Hyper-parameter tuning to get best model for each time series
- Get two different models for each time series
  - One with US Census variables (i.e. exogenous variables)
  - One without US Census variables
- A training-testing data split is made
  - Last 12 weeks of data is held out
  - Testing data is used to test and fairly compare models

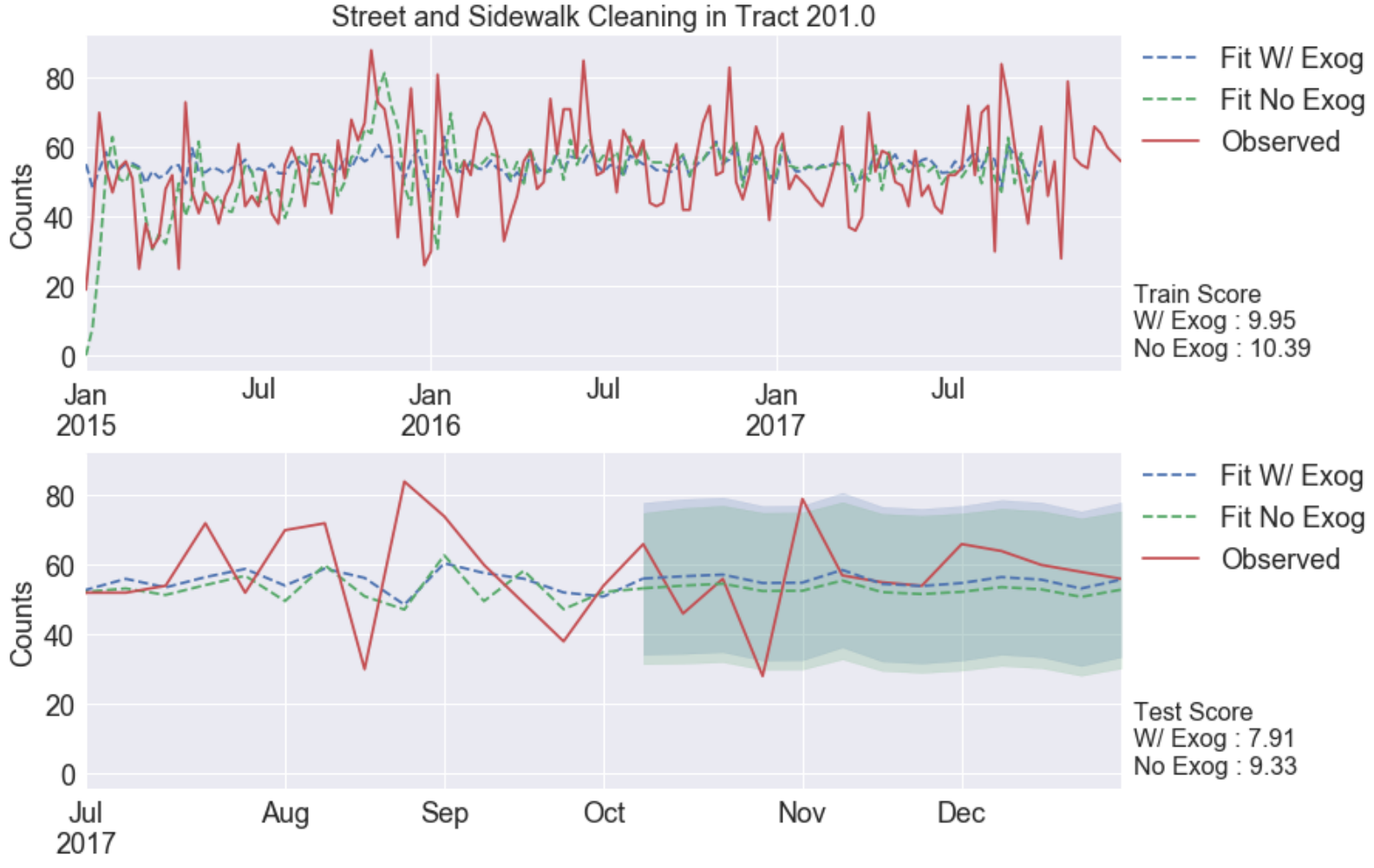
Time Series Modeling II

- Use mean absolute error as metric

- Additional concerns
  - Reduced dataset has led to many time series having null data
    - Weekly counts are up-sampled from seconds
    - In other words, many weeks without any service requests
    - Can interpolate to fill missing values
      - Introduces more bias
    - Require time series to have no null values
      - Only model 310 time series, out of 1,560
  - Local computation time per time series is 20-25 minutes
    - Have to limit hyper-parameter space we search over
    - Fit time series in batches and save to disk

Time Series Modeling III

- Example time series model with full model fit (top plot) and prediction (bottom plot)



Model Results

- Go through all 310 time series
  - Identify if models with US Census variables perform better, worse, or same ("Tie") as ones without US Census variables
  - Percentage of models listed at right
  - More often than not, US Census variables help
    - Good to see our idea to include them was justified
  - Also look at the results per Category, below
    - See similar results as the total

	Results [%]
Better	39.3
Tie	25.8
Worse	34.8

	Results [%]		
Category	Better	Tie	Worse
Abandoned Vehicle	36.4	22.7	40.9
Damaged Property	50.0	12.5	37.5
Encampments	22.2	66.7	11.1
Graffiti	33.3	24.1	42.6
Homeless Concerns	52.2	26.1	21.7
Sewer Issues	51.7	03.4	44.8
Street and Sidewalk Cleaning	38.3	32.3	29.3
Streetlights	40.0	0.0	60.0

Conclusion

- Construct a model to predict how many service requests will be made within one of 8 Categories and 195 US Census Tracts in San Francisco
  - Add US Census demographic data to SF 311 program data
  - Use SARIMAX algorithm
  - Addition of US Census data helps model more often than not
- Avenues for improvement
  - Don’t reduce dataset for additional predictive power
  - Cloud-based machines would reduce computation time and increase hyper-parameter grid search
  - Consider additional variables for inclusion to project
    - Data related to crime and weather might be useful
  - Explore additional algorithms to fit the data
    - Random Forests, Support Vector Machine, Neural Networks

Thank You

https://xkcd.com/1838/