Candidate Case Study

Brian Kao

5/31/19

Case Study Questions & Analysis

1)What day of week had the most average daily calls for service in July 2018?

2) What was the most common time of day for a DRUG/NARCOTIC incident to occur in 2017?

3)Analyze and describe the relationship between the time of day and volume and type of incidents

4)Create a model which predicts the weekly volume of incidents in 2018 by crime category type

Introduction

Case Study Instructions

Please complete the following dataset challenge over the next week and prepare for a 30-minute presentation with 15 minutes of Q&A.

Create a short presentation derived from analyzing the dataset:
- In developing your slides, assume the intended audience is the Mayor of San Francisco
- The presentation should tackle the following areas to the best of your ability:

Profile the provided data sets (e.g. datatypes, data distribution, missing values etc.)
- What day of week had the most average daily calls for service in July 2018?
- What was the most common time of day for a DRUG/NARCOTIC incident to occur in 2017?
- Analyze and describe the relationship between the time of day and volume and type of incidents

Create a model which predicts the weekly volume of incidents in 2018 by crime category type
- Be prepared to explain your methodology
- Feel free to use any means to obtain the answers (Python, R, etc.) and be prepared to share your work
  You do not need to limit your presentation to the answers to the questions above

Methodology:

Tools: Jupyter Notebook, R Studio, MS Excel
- Data Wrangling/Mining**:** Python(numpy, Pandas, sklearn, matplotlib, seaborn), MS Excel
- Data Science: R(forecast(arima, tbats, nnetar), ggplot)
- Data Source: https://www.kaggle.com/san-francisco/sf-police-calls-for-service-and-incidents
  *police-department-calls-for-service.csv (*data does not join to incidents.csv, did not find any accurate keys to join on).*
  police-department-incidents.csv
- Code: Link to .ipynb and .R file used for analysis: https://github.com/kbyuan/apple/

Assumptions:

- Context (Based on Data Samples and Case Study Questions):

1. It is August 1st 2018, and the Mayor wants to review the latest July 2018 Call Volumes and Weekly Trends.
2. The Major wants to Forecast 2018 Incidents for the year. Incidents data lags behind Call Volume data due to the amount of time it takes to record the event. Therefore the latest available data for Incidents is May 2018. *Furthermore, since the 2018 Incidents data shows a large YoY decline from Jan to May 2018, I am not using this in forecasting, as I am assuming it is still being gathered.
3. Not all slides presented today would be intended for the Mayor in this situation (i.e. in-depth forecast model slides).

Case Study Questions & Analysis

1)What day of week had the most average daily calls for service in July 2018?

July 2018 Call Volumes

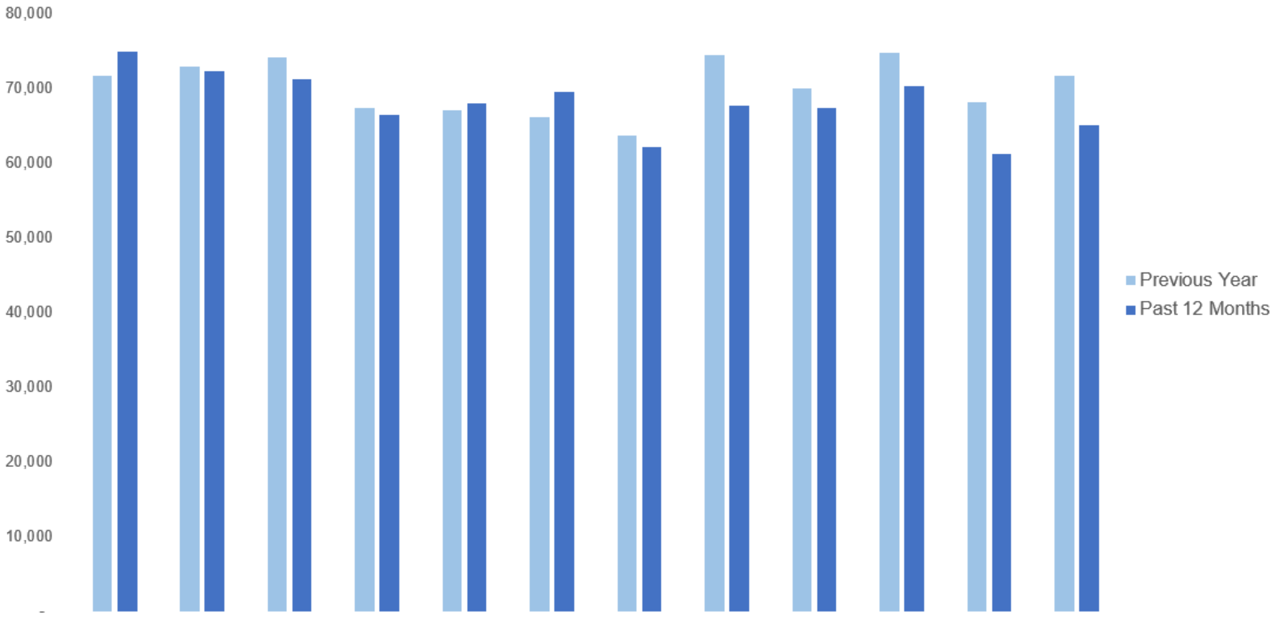
Monthly Call Volume – Past 12 Months

Monthly Call Volume:

Overall Monthly Calls were down -9.3% YOY in July 2018.

Monthly Call Volumes have dropped YoY for 5 straight months.

Monthly Call Volumes do seem to have a seasonal trend, as class in certain months seem tobe consistentl higher/lower than others. (i.e. October typically has a higher number of incidents than November)



	Aug-17	Sep-17	Oct-17	Nov-17	Dec-17	Jan-18	Feb-18	Mar-18	Apr-18	May-18	Jun-18	Jul-18
Monthly Calls	74,894	72,342	71,187	66,511	68,038	69,533	62,229	67,672	67,335	70,341	61,306	**65,111**
YoY% Var	4.5%	-0.9%	-4.1%	-1.2%	1.3%	5.0%	-2.4%	-9.2%	-3.7%	-5.9%	-10.0%	**-9.3%**
MoM % Var	4.4%	-3.4%	-1.6%	-6.6%	2.3%	2.2%	-10.5%	8.7%	-0.5%	4.5%	-12.8%	**6.2%**

Monthly Call Volume - July 2018

Call Volume by Category:

Traffic Stops were up +15% YoY, and was the top Call type In July.

Passing Calls dropped -8%, a trend that has continued over the past 5 months.

Homeless, Muni Inspection, and ParkingCalls also saw large YoY declines, falling -36%, -48%, and -57% YOY, respectively. A trend that began in April 2018.

*Note: it looks like some agencies may have changed Call Category Codes over the past 5 months. Therefore, we have cleaned the data (i.e. prefix: traf = Traffic Stop, Homeless combined with Trespassing).

July 2018 Calls: By Top 2 Categories

	July 2018 Calls	% to Total Calls	YoY % Var
Traffic Stop	9,090	14.0%	15%
Passing Call	8,813	13.5%	-8%
Homeless /Trespasser Complaint	3,514	5.4%	-31%
Suspicious Person	3,373	5.2%	6%
Muni Inspection	2,242	3.4%	-33%
Audible Alarm	2,106	3.2%	-5%
Suspicious Vehicle	1,751	2.7%	14%
Well Being Check	1,695	2.6%	3%
Noise Nuisance	1,505	2.3%	3%
22500e (Parking)	1,484	2.3%	-54%
Fight No Weapon	1,442	2.2%	0%
Auto Boost / Strip	1,281	2.0%	-11%
Poss	1,033	1.6%	1%
Mentally Disturbed	1,007	1.5%	5%
Assault / Battery	890	1.4%	-9%
Petty Theft	850	1.3%	9%
Meet W/citizen	788	1.2%	-14%
Drugs	779	1.2%	39%
Total Calls - Top 20 Categories	**43,643**	**67.0%**	

YoY% Call Volume Trend: Top 20 Categories

	Mar-18	Apr-18	May-18	Jun-18	Jul-18
Traffic Stop	-30%	-8%	1%	-3%	14%
Passing Call	-15%	-11%	-12%	-11%	-9%
Homeless /Trespasser Complaint	15%	19%	-2%	-21%	-36%
Suspicious Person	-11%	-7%	-2%	4%	7%
Muni Inspection	28%	3%	-19%	-19%	-48%
Audible Alarm	-3%	-11%	-7%	1%	-5%
Suspicious Vehicle	0%	13%	8%	14%	16%
Well Being Check	6%	7%	15%	13%	3%
Noise Nuisance	7%	6%	7%	10%	4%
22500e (Parking)	7%	-10%	-21%	-45%	-57%
Fight No Weapon	5%	2%	0%	3%	0%
Auto Boost / Strip	-25%	-23%	-32%	-34%	-11%
Poss	13%	17%	19%	16%	1%
Mentally Disturbed	21%	15%	25%	24%	5%
Assault / Battery	0%	-5%	-2%	9%	-10%
Petty Theft	-13%	17%	9%	18%	10%
Meet W/citizen	-12%	-11%	-18%	-17%	-13%
Drugs	18%	15%	66%	86%	51%

Avg Calls Per Week

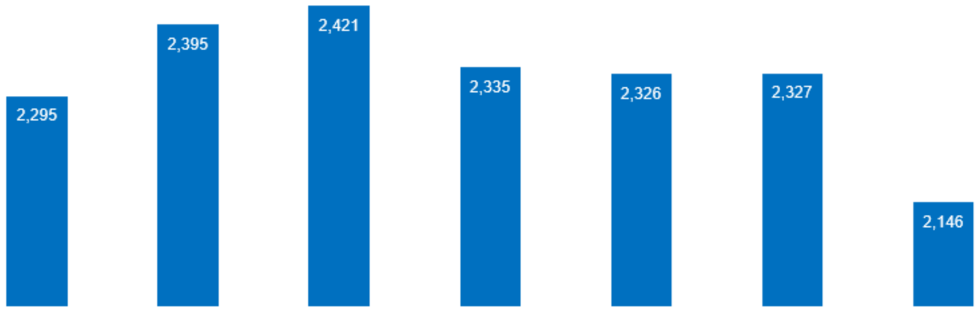
Call Volume by Day of Week:

Wednesday had the most average calls in July 2018, of 2,421.

Typically Tuesday's to Friday have the highest overall Avg Calls.

Sunday's typically have the least, with ~10% less than the top day of the week.

*Heatmap is based on benchmarking each day based on the highest day of the week. (i.e. the highest day is 100%, and the next highest day will be a percentage of the highest days value).



	Monday	Tuesday	Wednesday	Thursday	Friday	Saturday	Sunday
Jul-18	94%	98%	Top Day (100%)	94%	94%	93%	87%
Jun-18	96%	99.6%	99.6%	Top Day (100%)	97%	98%	88%
May-18	95%	95%	97%	96%	Top Day (100%)	99%	88%
Apr-18	99%	Top Day (100%)	99%	99%	97%	96%	88%
Mar-18	99%	94%	95%	97%	Top Day (100%)	97%	90%
Feb-18	92%	96%	98%	Top Day (100%)	99%	92%	87%
Jan-18	92%	99%	Top Day (100%)	99%	99%	94%	86%
Dec-17	87%	92%	Top Day (100%)	99%	95%	93%	83%
Nov-17	95%	Top Day (100%)	97%	90%	94%	93%	87%
Oct-17	95%	Top Day (100%)	96%	94%	100%	98%	92%
Sep-17	99%	97%	Top Day (100%)	98%	98%	92%	89%
Aug-17	91%	97%	97%	Top Day (100%)	98%	89%	87%

Case Study Questions & Analysis

4)Create a model which predicts the weekly volume of incidents in 2018 by crime category type

2017 Incidents Trends

2017 Incidents

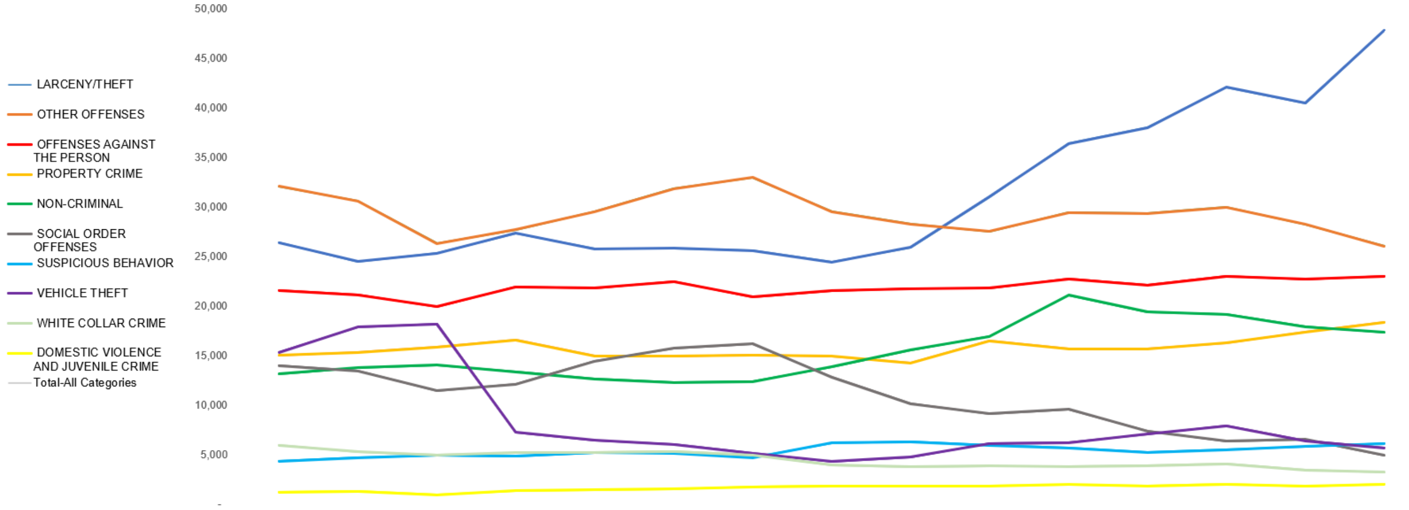
We had 154,773 Incidents in 2017, which was +3% YoY.

Larceny continues to be the Top Incident, and has grown +18% YoY.

Social Order Offenses are down significantly, a trend that began ~2010. July 2018 incidents were down -24% YoY. Most of this is due to a drop in Narcotic and Drug related incidents.

*Note: Incident Types were Categorized into 10 Groups for analysis and modeling purposes.

Annual Incidents - 2017



	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
Total Annual Incidents	149,176	148,148	142,186	137,853	137,639	141,311	139,860	133,523	132,697	140,847	152,806	150,128	156,526	150,883	**154,773**
YoY %Var		*-1%*	*-4%*	*-3%*	*0%*	*3%*	*-1%*	*-5%*	*-1%*	*6%*	*8%*	*-2%*	*4%*	*-4%*	***3%***

	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
LARCENY/THEFT		-7%	3%	8%	-6%	0%	-1%	-4%	6%	20%	18%	4%	11%	-4%	18%
OTHER OFFENSES		-5%	-14%	5%	7%	8%	4%	-10%	-4%	-3%	7%	0%	2%	-6%	-8%
OFFENSES AGAINST THE PERSON		-2%	-6%	10%	-1%	3%	-7%	3%	1%	0%	4%	-3%	4%	-1%	1%
PROPERTY CRIME		2%	3%	5%	-10%	0%	1%	-1%	-4%	16%	-5%	0%	4%	7%	6%
NON-CRIMINAL		5%	2%	-5%	-5%	-3%	1%	12%	12%	9%	24%	-8%	-1%	-7%	-3%
SOCIAL ORDER OFFENSES		-4%	-15%	6%	19%	9%	3%	-21%	-21%	-10%	5%	-23%	-14%	3%	-24%
SUSPICIOUS BEHAVIOR		8%	5%	-2%	7%	-1%	-8%	30%	2%	-5%	-4%	-8%	5%	6%	5%
VEHICLE THEFT		17%	2%	-60%	-11%	-6%	-14%	-16%	10%	30%	1%	14%	12%	-19%	-11%
WHITE COLLAR CRIME		-11%	-6%	4%	1%	1%	-7%	-21%	-4%	3%	-1%	2%	5%	-15%	-6%
DOMESTIC VIOLENCE AND JUVENILE		6%	-26%	37%	7%	7%	13%	5%	-3%	3%	7%	-6%	9%	-9%	10%

2017 Incidents by Category & District

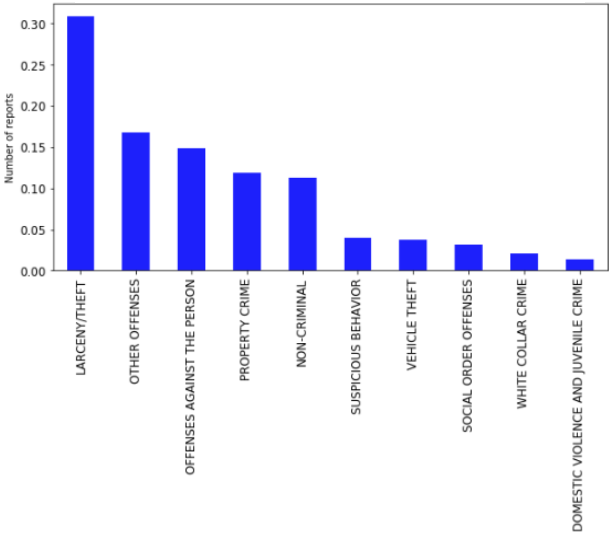
2017 Incidents by Category & Districts

Larceny was the largest category in 2017, representing 31%of all incidents, and growing +18% YoY. Social Order Offenses dropped -24%, largely due to a drop in Narcotic/Drug related incidents.

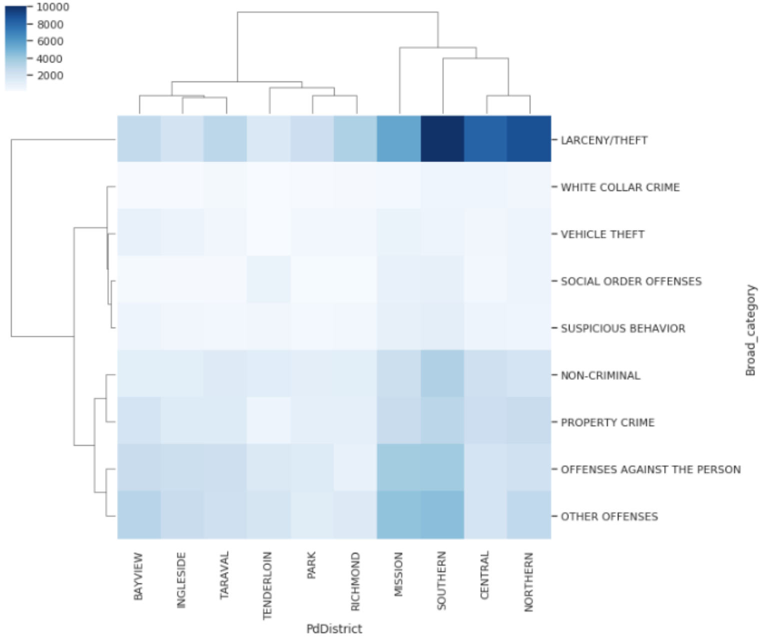
When comparing incidents by District, we can see that 'High Incident' Districts can have significantly different growth rates compared to 'Low Incident' Districts.

* High Incident Districts: Central, Mission, Northern, Southern

% to Total Incidents 2017 Incidents



Cluster Map 2017 Incidents – District vs. Category



	2017 incidents	YoY% Var	% To Total
LARCENY/THEFT	47,826	18.2%	30.9%
OTHER OFFENSES	26,036	-7.8%	16.8%
OFFENSES AGAINST THE PERSON	23,007	1.4%	14.9%
PROPERTY CRIME	18,382	5.7%	11.9%
NON-CRIMINAL	17,368	-3.1%	11.2%
SUSPICIOUS BEHAVIOR	6,152	5.2%	4.0%
VEHICLE THEFT	5,732	-10.7%	3.7%
SOCIAL ORDER OFFENSES	4,969	-24.4%	3.2%
DOMESTIC VIOLENCE & JUVEN	2,039	10.2%	1.3%
WHITE COLLAR CRIME	3,262	-6.4%	
TOTAL INCIDENTS	154,773	3%	

	"Low Incidents" Districts		"High Incidents" Districts	
	2017	YoY% Var	2017	YoY% Var
LARCENY/THEFT	15,246	9.4%	32,580	22.9%
OTHER OFFENSES	12,603	-4.2%	13,433	-11.0%
OFFENSES AGAINST THE PERSON	11,189	1.8%	11,818	0.9%
PROPERTY CRIME	7,677	2.9%	10,705	7.8%
NON-CRIMINAL	7,413	-5.7%	9,955	-1.0%
SUSPICIOUS BEHAVIOR	2,814	3.6%	3,338	6.6%
SOCIAL ORDER OFFENSES	2,020	-26.2%	2,949	-23.1%
VEHICLE THEFT	3,073	-15.9%	2,659	-3.9%
WHITE COLLAR CRIME	1,249	-10.8%	2,013	-3.4%
DOMESTIC VIOLENCE & JUVEN	1,209	16.8%	830	1.8%
TOTAL INCIDENTS	64,493	-1%	90,280	5%

Incidents by Hour and Day

2017 Incidents by Hour & Day

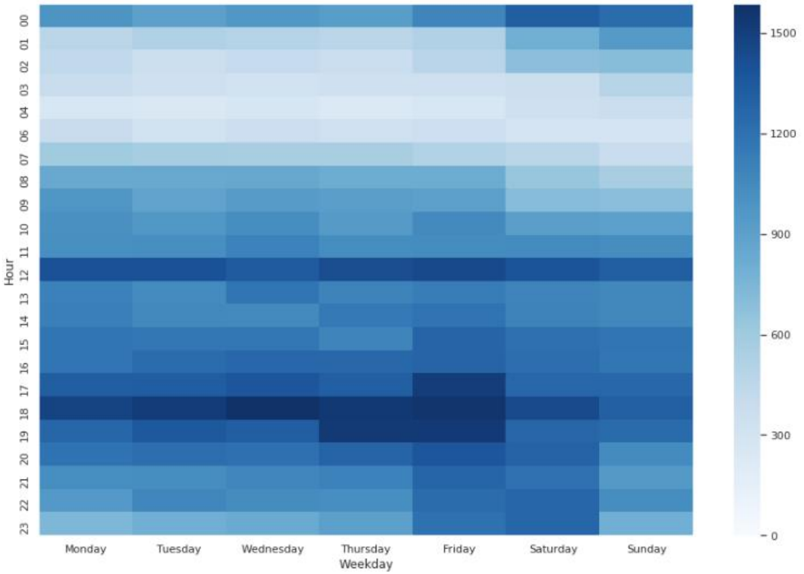
When looking at all incidents in 2017, 12pm and 6pm seem to have the most volume.

However, when we look at the incidents by Incident Category, we see that the Hour with the most volume can vary.

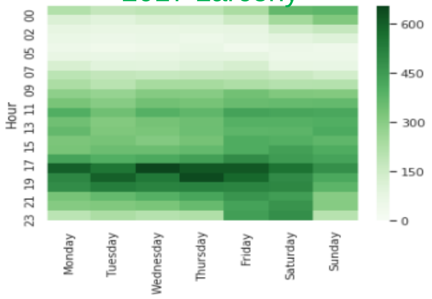
The Narcotics/Drug Category had the most incidents occur at 1PM, with 249 occurrences in 2017. 5PM is the next highest Hour at 246 incidents in 2017.

Offenses Against Other People incidents are distributed more evenly across the hours, with volume spread through Noon to around Midnight.

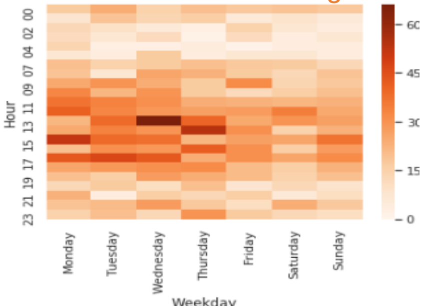
2017 All Incidents



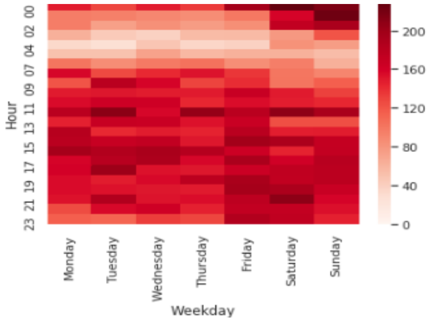
2017 Larceny



2017 Narcotics/Drugs



2017 Off Against People



Case Study Questions & Analysis

4)Create a model which predicts the weekly volume of incidents in 2018 by crime category type

2018 Incidents Forecasts

2018 Forecasted Incidents By Week

2018 Incidents Forecast

Our models forecast a slight YoY drop in incidents, falling -0.28% to 152K.

*The forecast is based on taking Weekly Incidents and forecasting through three separate time-series models.

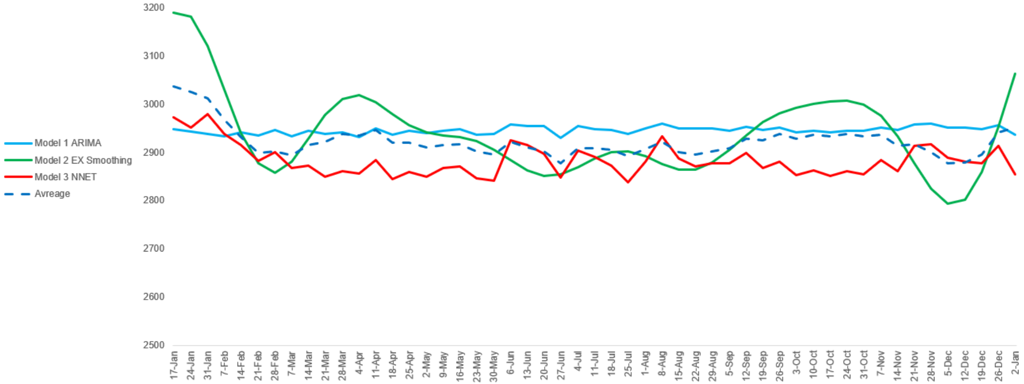
*We used a date range starting Jan 2015 after observing less variance in data from recent years. I am making some assumptions that data from earlier years may have been impacted by suboptimal data collection methods.

2018 Annual Incidents Forecast

2013	2014	2015	2016	2017	2018 (Forecasted)	YoY Var%
153,910	149,469	155,062	148,356	152,473	***152,049***	***-0.28%***

Model 1 ARIMA	*153,091*	*0.41%*
Model 2 Exp Smooth	*152,973*	*0.33%*
Model 1 NNET	*150,084*	*-1.57%*
Avg:	*152,049*	*-0.28%*

2018 Weekly Incidents Forecast



2018 Forecasted Incidents – By Week

Time Series Models

This slide compares 3 time series models from the R-forecast package: arima, tbats, nnetar.

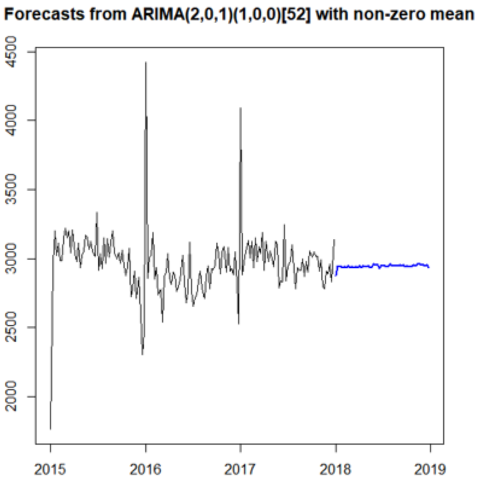
For each model, we use Weekly data beginning from Jan 2015 to Dec 2017. *Although we have weekly data going back to 2003, we choose to use only a subset of data due to observed trends and stationarity in more recent years.

The Test set is the latest 12 weeks for each data set.

The Training set is the entire data set after removing the latest 12 weeks.

Model 1

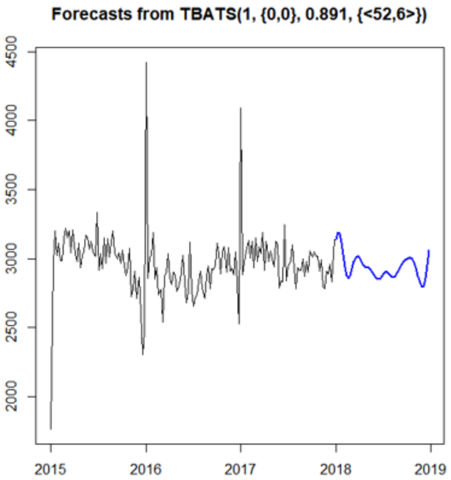
Weekly Avg.	Yearly Total	YoY% Forecast
2,944	153,091	0.41%



arima	MAPE	MPE
Train	4.78	-0.58
Test	3.09	-0.82

Model 2

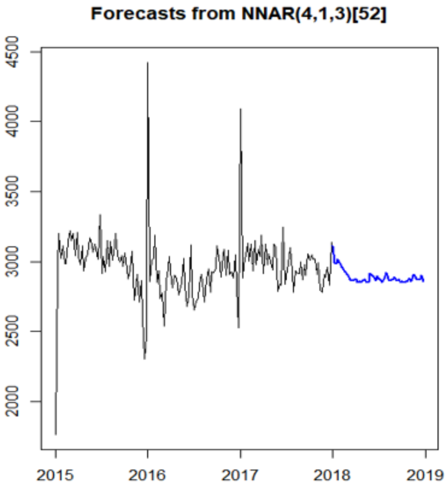
Weekly Avg.	Yearly Total	YoY% Forecast
2,942	152,973	0.33%



tbats	MAPE	MPE
Train	4.83	-0.38
Test	3.06	-0.99

Model 3

Weekly Avg.	Yearly Total	YoY% Forecast
2,886	150,084	-1.57%



nnetar	MAPE	MPE
Train	2.60	0.17
Test	3.17	-0.86

*MAPE: Mean Absolute Percentage Error

**MPE: Mean Percentage Error

Larceny Forecast 2018 – All Districts

Time Series Models: Larceny

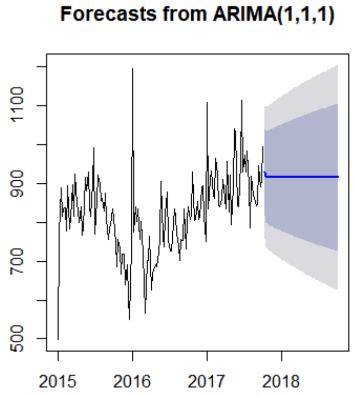
This slide compares Larceny incidents forecast based on 3 time series models from the R-forecast package: arima, tbats, nnetar.

Model 3(nnetar) seems to find a relatively strong trend compared to the other models. Nnetar also has the lowest test error of all models.

It is notable that there are material differences in the final forecasts of each model in terms of YoY % growth for 2018. (i.e. Model 3 predicts a 10% increase in Larceny, while Model 2 forecast a small -0.54% drop for 2018).

Model 1

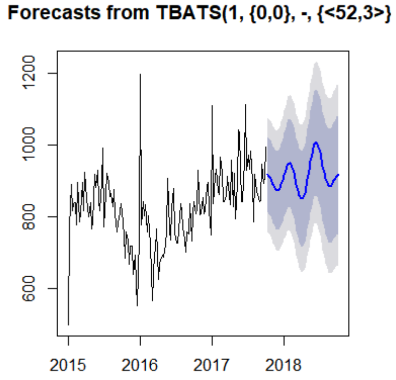
Weekly Avg.	Yearly Total	YoY% Forecas
963	50,055	4.66%



arima	MAPE	MPE
Train	7.22	0.27
Test	6.42	6.25

Model 2

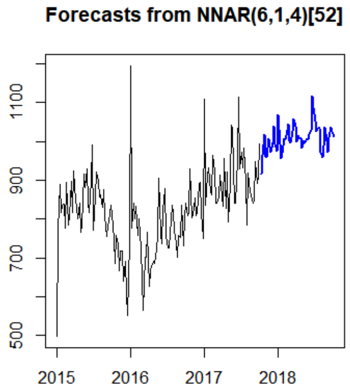
Weekly Avg.	Yearly Total	YoY% Forecast
889	47,566	-0.54%



tbats	MAPE	MPE
Train	6.83	-0.33
Test	8.84	8.84

Model 3

Weekly Avg.	Yearly Total	YoY% Forecast
1,007	52,387	9.54%



nnetar	MAPE	MPE
Train	3.55	-0.44
Test	4.62	0.54

Larceny Forecast-2018

Time Series Models: Larceny by District Type

When breaking out ‘High Incident’ and ‘Low Incident’ Districts, we can we see that the forecast for YoY% incidents growth can differ significantly.

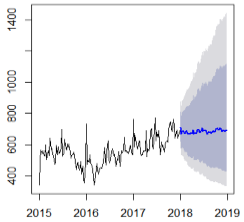
It is also notable that model performance can vary depending on the training data(i.e. Model 2 Test Errors on the High Incidents group are much higher than Test Errors for Low Incidents group.)

“High Incident” Districts

Model 1

Weekly Avg.	Yearly Total	YoY% Forecast
677	35,581	9.21%

Forecasts from ARIMA(3,1,0)(1,0,0)[52]

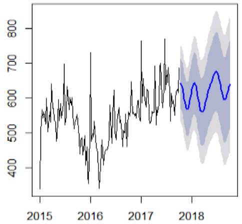


arima	MAPE	MPE
Train	7.74	0.13
Test	9.06	9.06

Model 2

Weekly Avg.	Yearly Total	YoY% Forecast
616	32,041	-1.65%

Forecasts from TBATS(1, {0,0}, -, {<52,4>}

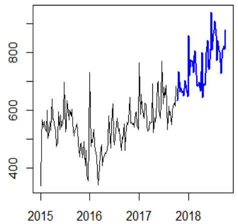


tbats	MAPE	MPE
Train	7.65	-0.43
Test	13.28	13.28

Model 3

Weekly Avg.	Yearly Total	YoY% Forecast
755	39,265	20.52%

Forecasts from NNAR(4,1,3)[52]



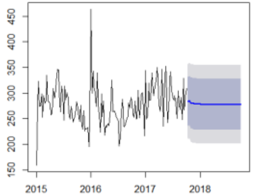
nnetar	MAPE	MPE
Train	5.17	-0.69
Test	7.41	6.03

“Low Incident” Districts

Model 1

Weekly Avg.	Yearly Total	YoY% Forecast
279	14,483	-5.01%

Forecasts from ARIMA(3,0,0) with non-zero mean

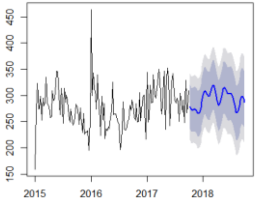


arima	MAPE	MPE
Train	4.78	-0.58
Test	3.09	-0.82

Model 2

Weekly Avg.	Yearly Total	YoY% Forecast
292	15,199	-0.31%

Forecasts from TBATS(1, {0,0}, -, {<52,6>})

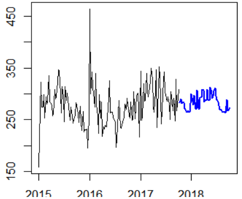


tbats	MAPE	MPE
Train	4.83	-0.38
Test	3.06	-0.99

Model 3

Weekly Avg.	Yearly Total	YoY% Forecast
283	14,695	-3.61%

Forecasts from NNAR(3,1,2)[52]



nnetar	MAPE	MPE
Train	2.60	0.17
Test	3.17	-0.86

Incidents Forecast Model: Next Steps & Considerations

Forecast Model Findings:

Model 1: arima

This model does not seem to find a strong trend in the data we tested. The forecasted weekly incident data has less weekly variance.

Model 2: tbats

This method allows for multiple seasonality, which seems appropriate for our use case. The model does find some trends, however the test error was the highest across all models.

Model 3: nnetar (best model)

This model seemed to do the best in finding trends, without over-fitting too much. This method works well with non-linear trends and large amounts of data. The next step would be to add other variables and increase the time range.

Model Refinements:

Data Governance:

- Standardize data collection and timestamp logging
- Revisit and define Incident Categories
- Clean data to remove/smooth outliers

Add Data Attributes:

- Incorporate more Data attributes into the model: Holidays, Weather, Economics, etc.
- Create Key to join to Call Data

Model Optimization:

- Create and engineer features that may help improve model accuracy (i.e. smoothing or weighting data based on off-line data).
- Continue to test different models to improve accuracy.