Forecasting Fatalities in NYC

DSCT Capstone 1 – Milestone Report
November 2019
Jonathan D. Williams

Problem Statement

The Fire Department of the City of New York (FDNY) is the second largest fire department in the world, and the largest within the United States. More than 8.5 million residents--and tourists--within the five boroughs of New York City are protected on a daily basis by just under 15,000 uniformed personnel, which are comprised of: firefighters, EMTs, and paramedics. This agency provides invaluable services to all within the city, but its efforts are often subject to several constraints.

For this project, I have examined a dataset of incident records generated by the EMS Computer Aided Dispatch System that spans a six-year period (2013 through 2018) in order to gain insights on patterns of fatalities within the City of New York. The goals of this project are to answer the following questions:

1. Can the outcome of an EMS incident be predicted based on its attributes (e.g. time of occurrence, call type, geographic area, etc.) and the assessments made by emergency personnel (e.g. call type, severity level, etc.)?
2. Are there specific neighborhoods within the City of New York that should receive priority services in order to minimize the loss of life?

In addition to the FDNY, several stakeholders involved with the policy-making, administration, and delivery of emergency response services throughout the City of New York will benefit from this project. Such parties include, but are not limited to: the Office of the Mayor, the New York City Council, and the New York City Health and Hospitals Corporation (HHC). Findings from the subsequent analyses performed can help improve emergency response protocols and allow responders to better serve all individuals within the City of New York.

Description of Dataset

The dataset for this project is the NYC EMS Incident Dispatch Data, which is made publicly available via NYC Open Data. It contains information about the incident as it pertains to the assignment of resources and the Fire Department's response to the emergency. Specific locations of incidents are not included in order to protect personal identifying information in accordance with the Health Insurance Portability and Accountability Act (HIPAA).

Data Wrangling

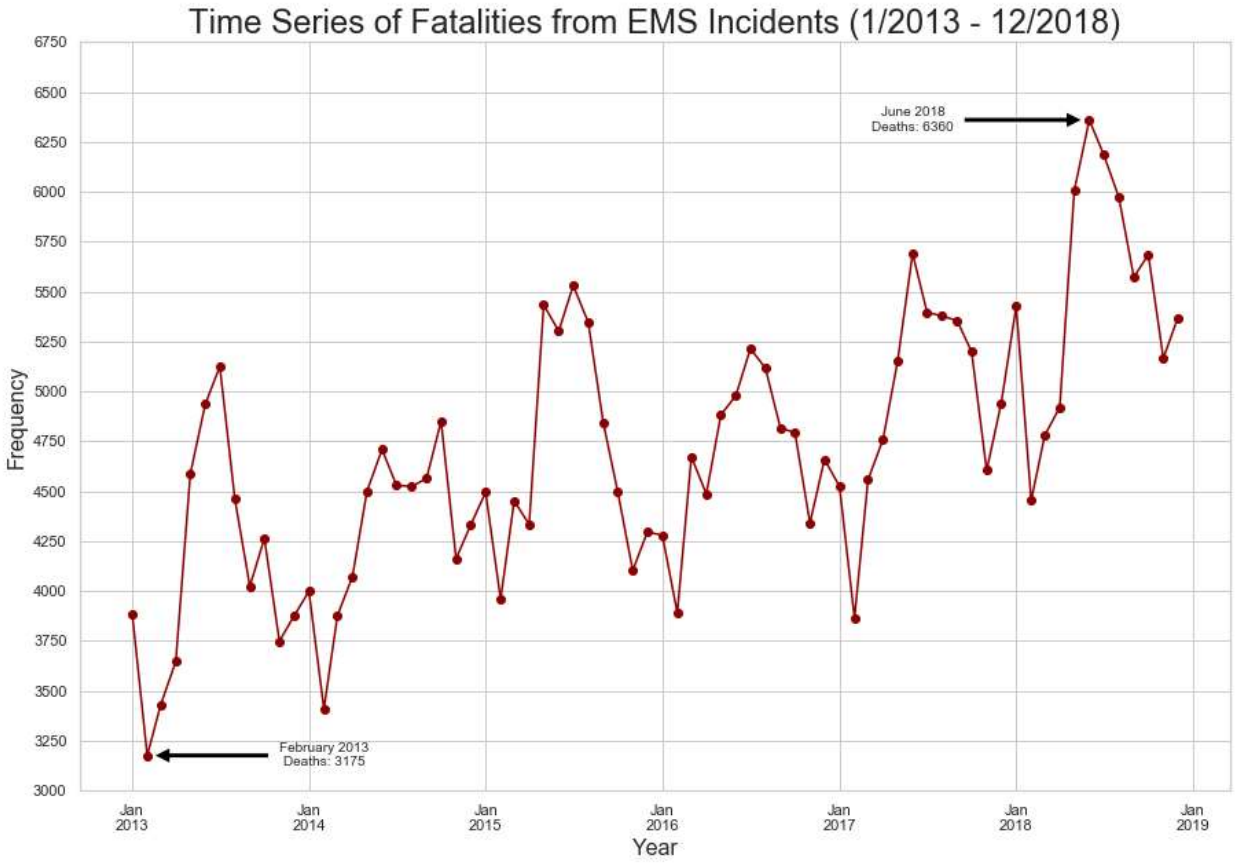
The principal dataset was imported via the Socrata Open Data API (SODA) as batches of JSON files that were converted to Pandas dataframe objects, appended to a single list object, and then concatenated to form the core EMS dataframe. Alternatively, I was able to download the dataset as a CSV file (2 GB) whenever I needed to perform analyses on different machines and was constrained by throttling limits imposed by the API. Irrespective of the acquisition method, the resulting dataset contains over 8 million records of EMS incidents. Additional wrangling techniques that were applied included:

- Removing incidents with missing values for: incident outcome, ZIP code, and response time
- Removing incidents that pertained to special events (e.g. facility transport, marathons, parades, etc)
- Removing incidents with indicated calculation errors for duration metrics
- Removing columns that pertain to redundant geographic data
- Parsing the incident datetime field to generate additional fields for select time components (e.g `year`, `month`, `hour`, and `weekday`)
- Modify the data types of several feature variables in order to drastically reduce the size of the dataframe object in memory
- Generate a boolean series from the incident outcome feature to create an explicit field for the target variable: `fatality`

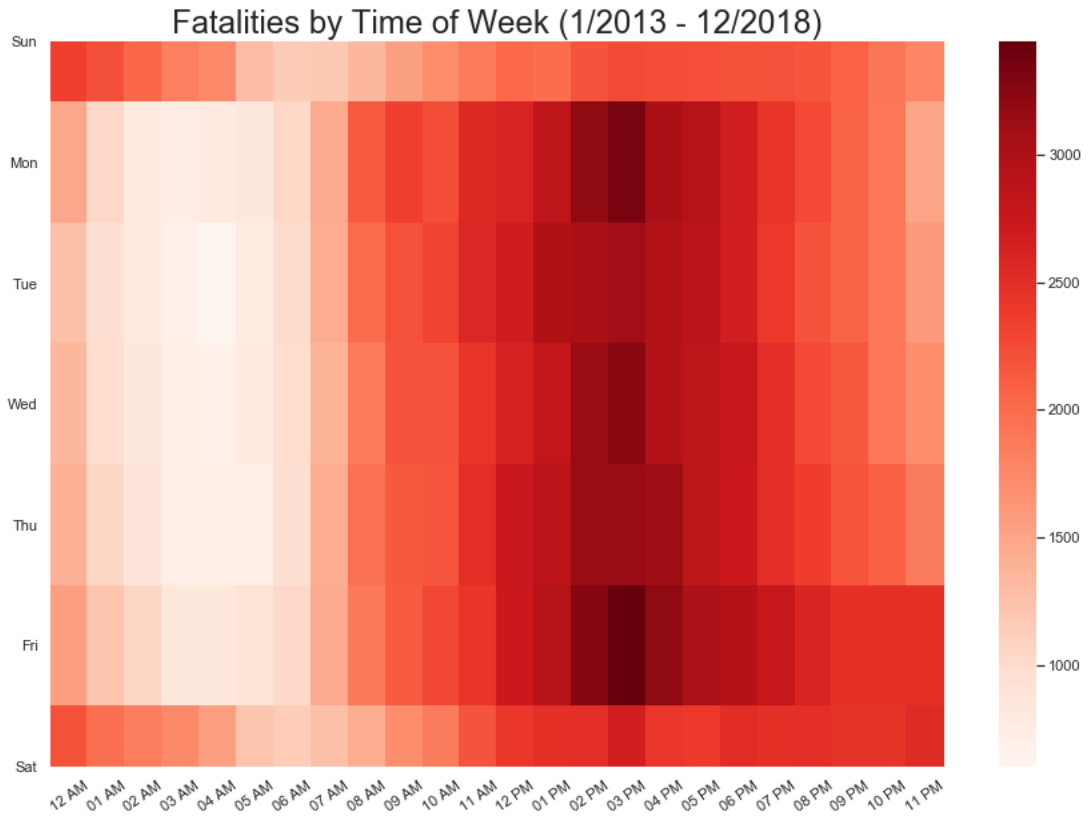
Exploratory Analysis

Visualizing the dataset revealed several patterns and other findings.

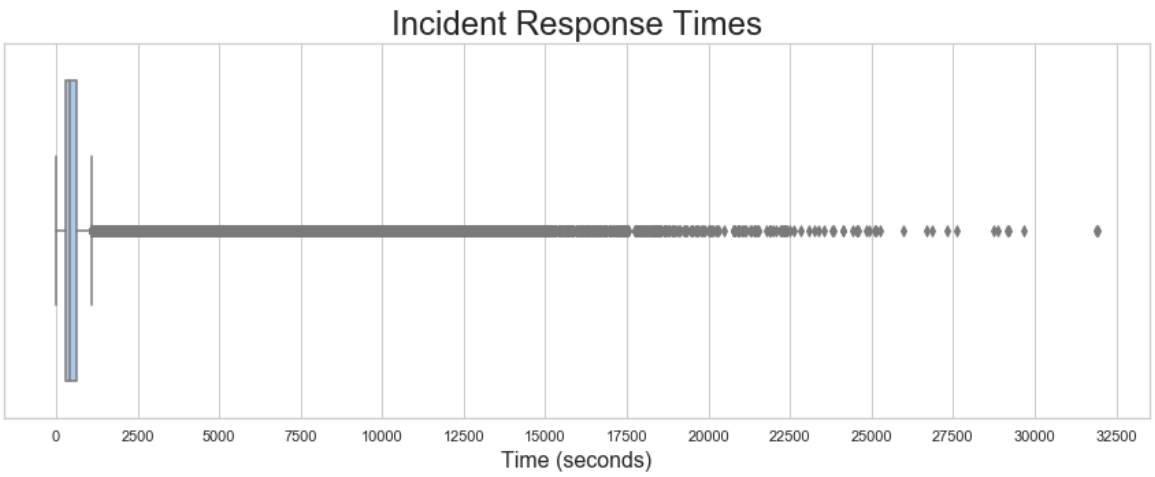
1. There has been an upward trend in the number of fatalities that result from EMS incidents across the six-year observation period (2013 through 2018). With the exception of 2014, fatalities tend to peak during the summertime in either June or July.



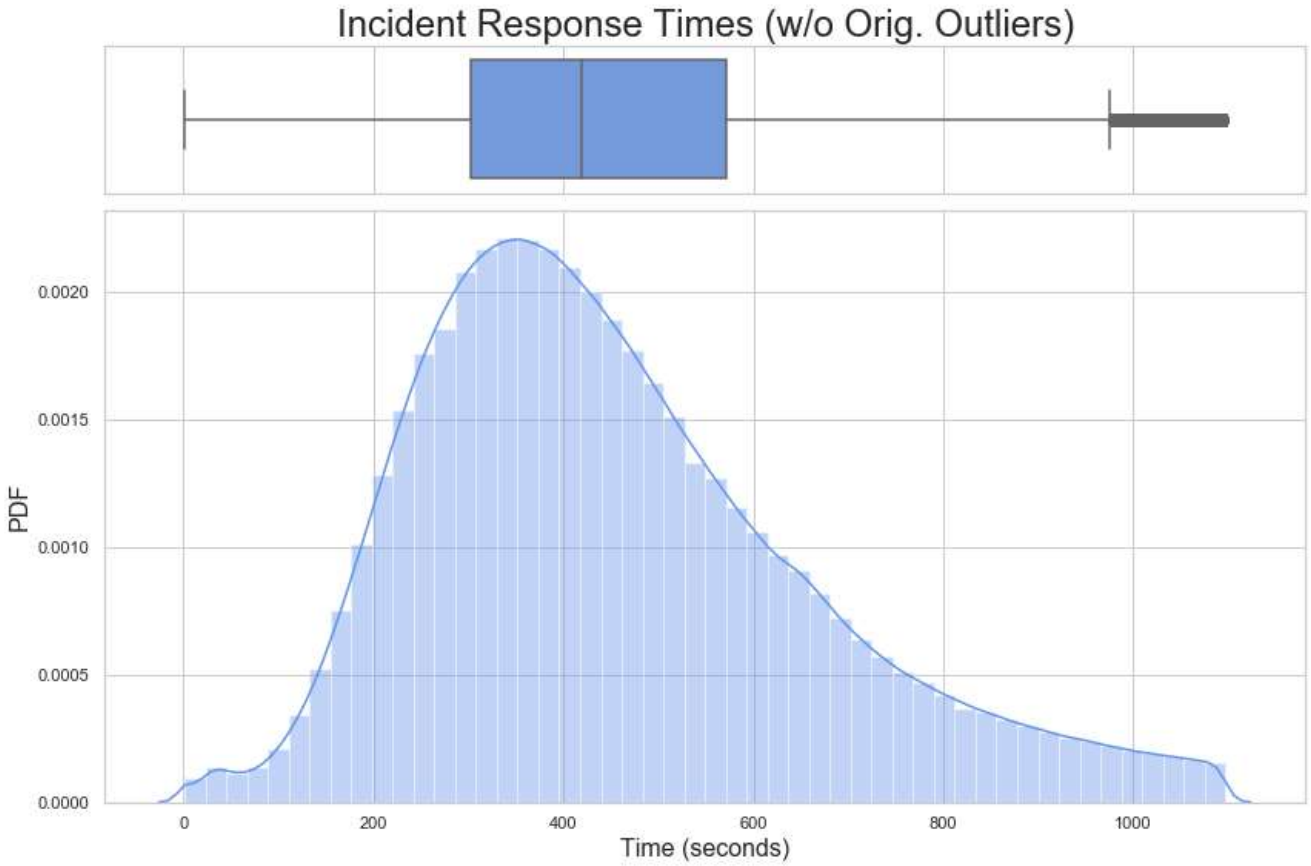
2. The greatest number of fatalities during any given week tend to occur during the afternoon. The heatmap below illustrates the sum of fatalities per weekday for each one-hour period beginning at midnight. It indicates that a high frequency occurs between 3:00 pm and 4:00 pm (21,207), with noticeable peaks on Mondays (3,346) and Wednesdays (3,442).



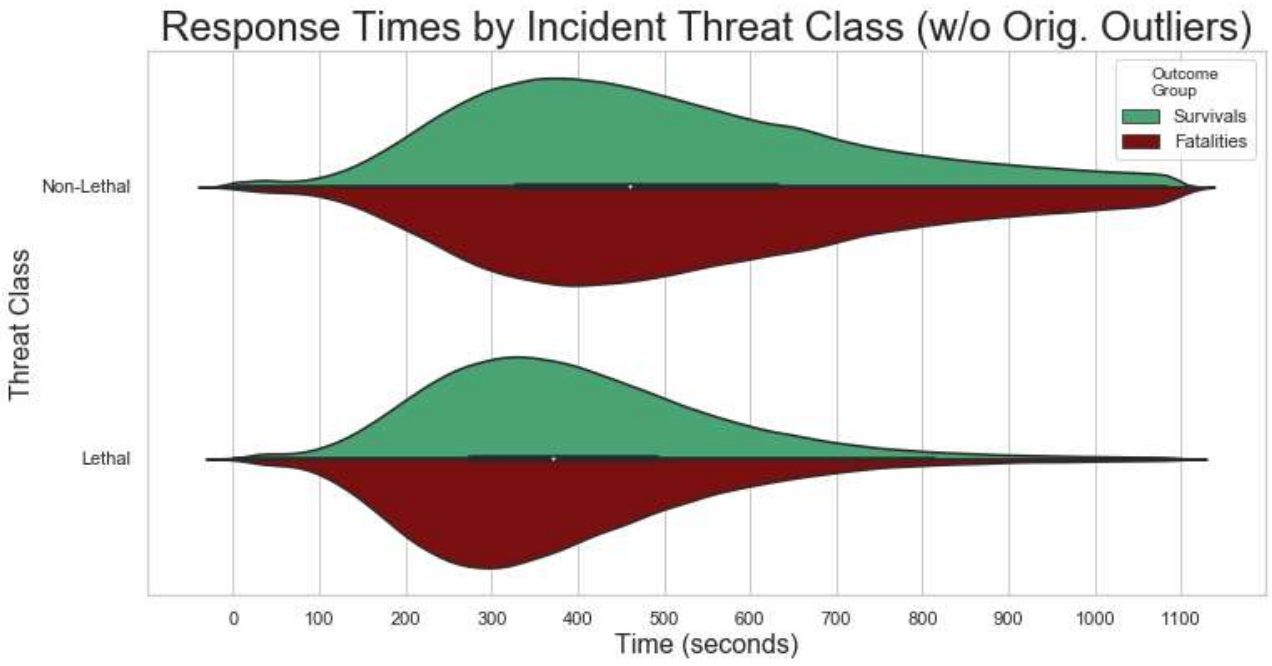
3. There is a wide spread of incident response times due to a large volume of outliers—with respect only to this variable—in the original dataset.



The distribution of response times appears positively-skewed when the outliers are removed from the original dataset.



4. The FDNY categorizes all EMS incidents into two threat classes according to the severity level code assigned to the incident: **Lethal** (Severity Levels 1, 2, and 3) and **Non-Lethal** (Severity Levels 4, 5, 6, 7, and 8). While the distribution of response times between outcome groups in the **Non-Lethal** threat class are very similar, there are noticeable difference between those in the **Lethal** threat class.



Inferential Statistics

Statistical measures were used to identify relationships amongst the variables contained within this dataset, as well as to perform hypothesis testing. Numerical predictor variables with high correlations to one another were identified using Pearson's correlation coefficients. Pairs that were examined include:

- `initial_severity_level` vs. `final_severity_level`
- `dispatch_time` vs. `response_time`
- `travel_time` vs. `response_time`

Correlation Statistics	
Initial vs. Final Severity Level:	$\rho = 0.9352$
Dispatch vs. Response Times:	$\rho = 0.7731$
Travel vs. Response Times:	$\rho = 0.7476$

The `initial_severity_level`, `dispatch_time`, and `travel_time` were all identified as redundant, numerical predictor variables. Similarly, categorical predictor variables with a high degree of association were identified using Cramér's V statistic. Two key pairs of variables examined by this measure were the `initial_call_type` and the `final_call_type`, and also `incident_disposition_code` and `fatality`.

Association Statistics (Cramér's V)	
Initial vs. Final Call Type:	$V = 0.7552$
Disposition Code vs. Fatality:	$V = 1.0000$

The `initial_call_type` was deemed a redundant categorical predictor variable given its moderately high degree of association with the `final_call_type`. The boolean target `fatality` was derived from `incident_disposition_code`. Thus, the latter variable is a *descriptor* of the incident outcome rather than a predictor of it and should be omitted from analyses. The exclusion of these redundant variables will, potentially, allow for the development of an accurate binary classifier for the dataset in future work.

Next Steps

The findings uncovered through both EDA and inferential statistics revealed several interesting patterns within the core dataset. Future work will include developing predictive models using Logistic Regression classifiers and applying hyper-parameter tuning as needed.