Forecasting Fatalities in NYC

DSCT Capstone 1 – Data Story
Jonathan D. Williams

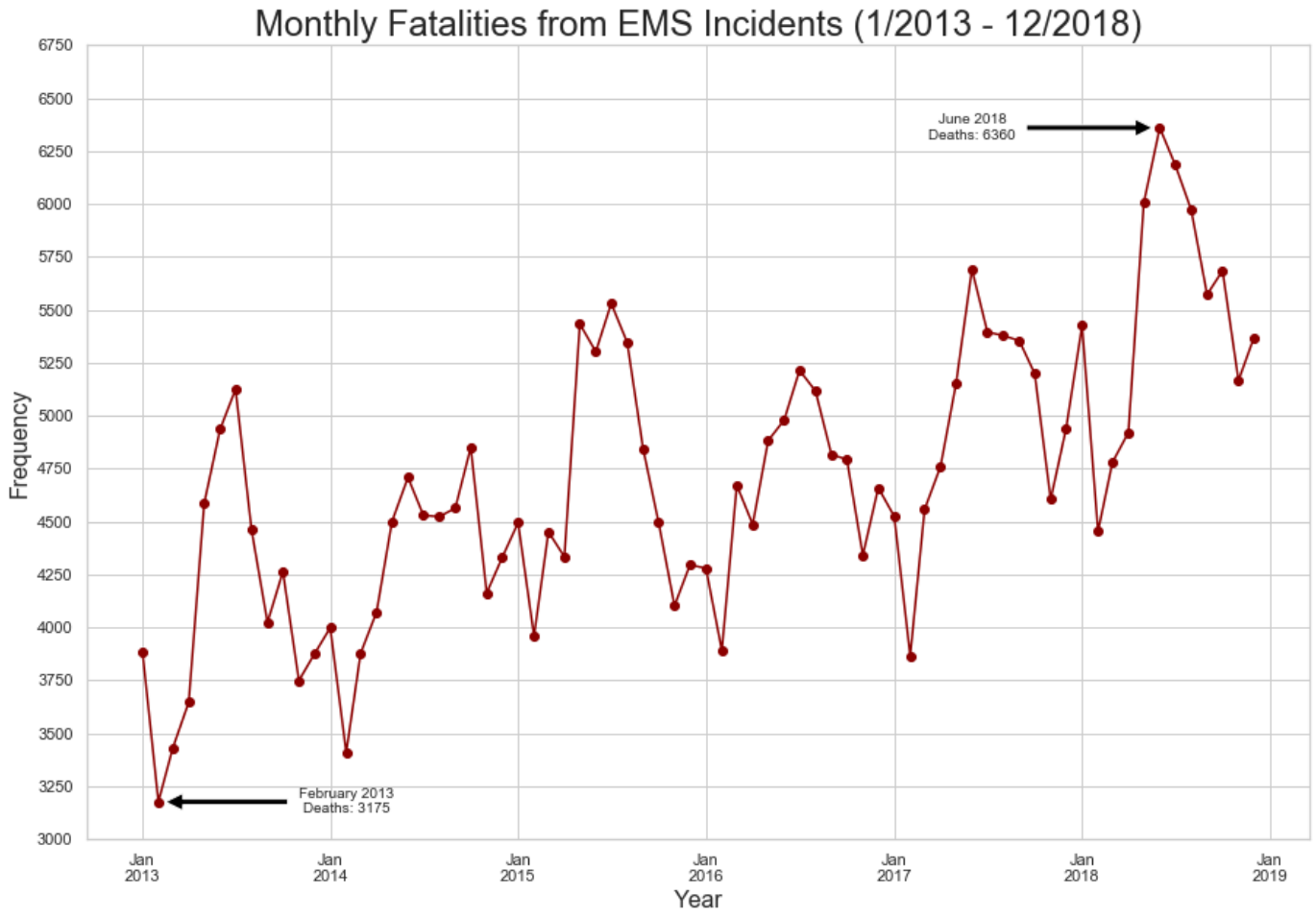
The Fire Department of the City of New York (FDNY) is the second largest fire department in the world, and the largest within the United States. More than 8.5 million residents and tourists within the five boroughs of New York City are protected on a daily basis by just under 15,000 uniformed personnel. This elite team of individuals is comprised of: firefighters, EMTs, paramedics, and administrative staff. The agency provides invaluable services to all within the city. Yet, in spite of their vigilance and bravery, the members of the FDNY cannot save everyone.

The goal of this project is to develop machine learning models that predict whether or not the outcome of an EMS incident will result in a fatality. This is a supervised, binary classification problem. Analyses will be performed on a collection of nearly 8.5 million records of documented incidents, which span the six year period from January 2013 through December 2018, and appropriate predictive models will be developed to achieve the primary objective. This dataset is robust and contains several feature variables that describe both various attributes of each incident as well as the responsive action taken by the FDNY. All of the aforementioned factors affect an individual's survivability once a response is initiated.

Exploratory data analysis was conducted prior to the development of any ML algorithms. The purpose of this process was to reveal trends that can help identify key feature variables and also provide contextual insights about the dataset. To this end, a natural starting point was to answer the following question: How many total recorded incidents resulted in a fatality, and how does this outcome vary over time? Of the 7,988,028 observations used in this analysis, only 338,684 resulted in a fatality.

EMS Incident Outcome	Value Count	Percentage
Survival	7988028	95.76%
Fatality	338684	4.24%

The time series below illustrates an overall upward trend in the number of fatalities that result from EMS incidents across the six-year observation period. In addition, the frequency of fatalities tend to spike during the middle of each year. This observation warrants a closer inspection to determine whether or not the incident month—or any other component of the incident time—is a deterministic factor of a fatality.



There are several feature variables within the dataset that are examined using exploratory data analysis (EDA)[1] methods. Various visualizations are used to highlight trends within the following categories:

- Fatalities by Time Period
- Fatalities by Geographic Region
- Incident Response Times
- Incident Call Types & Severity Levels

Inferential statistic techniques[2] are used to examine the dataset in greater depth beyond the limitations of EDA, and identify feature variables that are particularly significant to address the primary objective. Some questions that are investigated include, but are not limited to, the following:

- *Are the initial and final severity levels for EMS incidents the same?* Is there a strong correlation between the incident assessment determined from caller-provided information (initial_severity_level) and the assessment made by response personnel on-the-scene (final_severity_level)?

[1] Jupyter Notebook: Exploratory Data Analysis (CP1-02)
[2] Jupyter Notebook: Inferential Statistics (CP1-03)

- *Are survival rates for life-threatening and non-life-threatening severity levels the same?* Is there a significant difference between the survival rates for incidents that are assigned a *life-threatening* severity level (Code 1, 2, and 3) and those that are deemed non-life threatening (Code 4, 5, 6, 7, and 8)?
- *Are the response times for fatalities and survivals the same?* Is there a significant difference between the response times of incidents that result in fatalities and those where patients survive?
- *Are fatality rates highest during summer months?* Is there a significant difference between the fatality rates for incidents that occur during the summer months (June, July, and August) and those that occur during the rest of the year?

Ultimately, the insights gathered from this analysis will help identify the feature variables that have the most significant impact on the outcome of an EMS incident. The final results can be used by the FDNY to minimize the number of lives lost within New York City.