Foundations of Data Science Final Project Proposal

Life Lane

Project Members:

- Kora Hughes (svh272)
- Swarali Dabhadkar (sd5664)
- Stuti Mishra (sm11538)

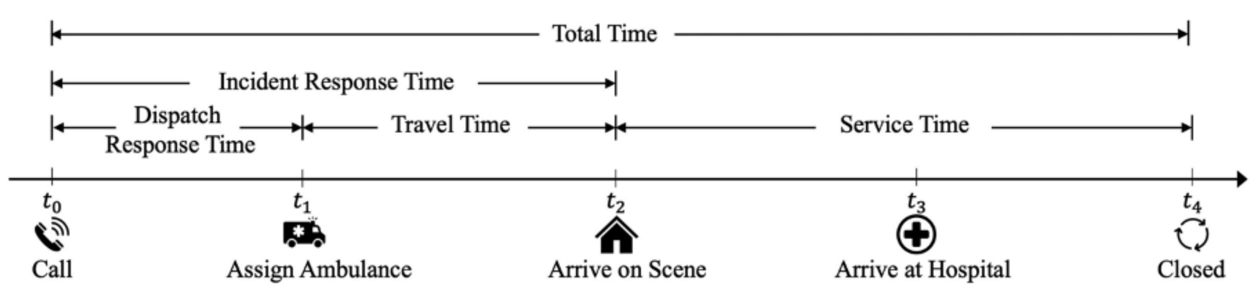
Project Questions:

What is the problem (including motivation and what is the specific outcome)?

Emergency medical system (EMS) response times in NYC are growing. According to the New York Post, average EMS response time increased by 20 seconds last year, totalling 9 minutes and 50 seconds.[1] This has led to an increase in deaths via fires and medical emergencies.

Thus, we want to take a data scientists' approach to optimizing this problem. While we can't easily restructure the current EMS dispatch sites, we can analyze EMS dispatch data and cross reference it with traffic patterns to optimize the placement of new dispatch sites and the staffing of current dispatch sites.

To accomplish this, we can create a regression problem by setting our target variable to INCIDENT_TRAVEL_TM_SECONDS_QY (As depicted by "Travel Time" in the figure below) in the EMS Dispatch Dataset which is the time elapsed in seconds between the first_assignment_datetime – time at which the emergency incident was assigned by the operator – and the first_on_scene_datetime – the time at which the team arrived at the incident location. Once optimized, we can reverse engineer this problem in order to find the times and locations where dispatch travel time is the worst. With this information, we can suggest new policies and procedures that specifically target these zones in order to minimize future dispatch time.



EMS response to incident timeline (Liu *et al*)[3]

[1] https://nypost.com/2023/09/16/nyc-response-times-to-fires-medical-emergencies-soaring/

How will you learn the background? (e.g. are there specific publications that discuss pertinent issues, is there a domain expert you will engage and what is their experience, etc.)

The literature regarding EMS response time has cited a variety of causes and solutions with respect to EMS time reduction. Griffin *et al* assess the variables most commonly associated with increased emergency response time as described by the opinions and views of EMS first responders.[2] The study concludes that traffic congestion is the most influential factor and proposes increased public education regarding EMS routing and access to pre-emptive green light devices for EMS personnel as potential solutions. Liu *et al* analyzes demographic and geographic patterns associated with the increase in response time due to the COVID-19 pandemic.[3] They cite remedial reallocation as a temporary solution to the EMS system with the existing ambulance capacity and validate their proposal via mathematical simulation. Though this provides an effective temporary solution, they do not address the long-term costs associated with resource allocation in ambulance relocation and suggest further research on more sustainable solutions.

Though the potential solutions remain widespread, the pressing need for a reduction in response time is virtually unanimous. Blackwell *et al* proposed a causal link between health outcomes like improved survival and EMS system response and travel time.[4] Similar studies can be found validating the use of EMS dispatch time for patient illness prediction and postulate that a reduction in response time could yield other health benefits as well.[[5]][[6]] The need for methods of reducing EMS response time has also been mirrored in mainstream media. According to Mayor Eric Adams' most recent management report, it took ambulances and firefighters 46 seconds longer on average to respond to “life-threatening medical emergencies” this fiscal year compared to last year.[7] It is with these major applications in mind that we pursue a project investigating the reduction of EMS time.

---

[2] Griffin R, McGwin G Jr. Emergency medical service providers' experiences with traffic congestion. J Emerg Med. 2013 Feb;44(2):398-405. doi: 10.1016/j.jemermed.2012.01.066. Epub 2012 Aug 9. PMID: 22883716.
[3] Liu, J., Ouyang, R., Chou, CA. et al. An Analytical Approach for Dispatch Operations of Emergency Medical Services: A Case Study of COVID-19. Oper. Res. Forum 4, 44 (2023). https://doi.org/10.1007/s43069-023-00218-3
[4] Blackwell, T. H., & Kaufman, J. S. (2002). Response time effectiveness: comparison of response time and survival in an urban emergency medical services system. Academic Emergency Medicine, 9(4), 288-295.
[5] Shah, M. N., Bishop, P., Lerner, E. B., Fairbanks, R. J., & Davis, E. A. (2005). Validation of using EMS dispatch codes to identify low-acuity patients. Prehospital Emergency Care, 9(1), 24-31.
[6] Hinchey, P., Myers, B., Zalkin, J., Lewis, R., & Garner Jr, D. (2007). Low acuity EMS dispatch criteria can reliably identify patients without high-acuity illness or injury. Prehospital emergency care, 11(1), 42-48.
[7] https://www.nyc.gov/office-of-the-mayor/news/673-22/mayor-adams-releases-mayor-s-management-report-fiscal-year-2022

What kinds of data will you use? (describe the data fully including it's temporal and spatial dimensions, features and their types and scales (e.g. numerical or text, ordinal or nominal, etc.))

We are planning to use two datasets from the NYC Open Datasets for our analysis. One is the EMS Incident Dispatch Data[8] and the other is Traffic Volume Counts.[9]

Data generation process for the source datasets:

- New York City Department of Transportation (NYC DOT) uses Automated Traffic Recorders (ATR) to collect traffic sample volume counts at bridge crossings and roadways. These counts do not cover the entire year, and the number of days counted per location may vary from year to year.
- The EMS Incident Dispatch Data file contains data that is generated by the EMS Computer Aided Dispatch System. The data spans from the time the incident is created in the system to the time the incident is closed in the system. It covers information about the incident as it relates to the assignment of resources and the Fire Department's response to the emergency. To protect personal identifying information in accordance with the Health Insurance Portability and Accountability Act (HIPAA), specific locations of incidents are not included and have been aggregated to a higher level of detail.

According to the documentation, privacy concerns have been addressed while collecting EMS data in order to protect PII data privacy of the individuals involved in the incidents.

Because of the accessibility of the data, we are focusing on NYC. Thus all of our datasets will need to have corresponding location like borough, zip code, area code, and or district number (nominal categorical) as well as temporal data such as date/time (continuous numerical).

EMS Data Description	
INCIDENT_DATETIME	The date and time the incident was created in the dispatch system
INCIDENT_TRAVEL_TM_SECONDS_QY	The time elapsed in seconds between the first_assignment_datetime and the first_on_scene_datetime.
ZIPCODE	The zip code of the incident.

After conducting an exploratory data analysis, we have identified 3 columns out of a total of 31 that are crucial to our analysis. These columns provide essential information and insights for our research. We have opted to reduce the dimensionality of our dataset by eliminating columns that do not provide significant value or can be derived from other existing columns. This step is aimed at streamlining our analysis, enhancing the manageability of the dataset, and improving processing efficiency.

[8] https://data.cityofnewyork.us/Public-Safety/EMS-Incident-Dispatch-Data/76xm-jjuj
[9] https://data.cityofnewyork.us/Transportation/Traffic-Volume-Counts/btm5-ppia

Traffic Volume Data Description	
Roadway name	Street name
Date	Date of the traffic count
12:00-1:00 AM	Count for the clock hour
1:00-2:00AM	Count for the clock hour
[…other hourly columns]	Count for the clock hour
11:00-12:00AM	Count for the clock hour

Similarly for the traffic volume dataset, we chose to exclude 5 columns out of a total of 31 columns that provide value to our analysis.

Database Schema:

EMS DATA: [response_time (INCIDENT_TRAVEL_TM_SECONDS_QY), zip code (ZIPCODE), datetime (INCIDENT_DATETIME)]

- Datetime → [date, time]
    - Time → time_block (every hour)
    - Date → day_of_week
        - Reduction by looking 7 days per month (each different week days)

TRAFFIC DATA: [street (roadway_name), time_block, date, traffic_volume]

- street → zip code
    - Aggregate on traffic_volume
- Date → day_of_week
    - Reduction by looking 7 days per month (each different week days)

~ after join operation ~

EMS_BY_TRAFFIC: [zip code, traffic_volume, day_of_week, time_block, dispatch_time]

In order to join our two major databases we transform the data in a few key ways. Firstly, we split the datetime objects into time blocks and days of the week. We aggregate time into blocks in order to merge with the inherent precision of the traffic-volume database, however the day of the week is extracted due to its intuitive affect traffic flow: we felt that day of the year was too precise to give quality insights and aggregated it to day of the week in order to account for the change in traffic flow due to periodic work schedules.

The target variable we will be looking at is EMS response time (dispatch_time) which is continuous numerical. Our main dependent variables are traffic volume (discrete numerical), zip code (nominal categorical), day of the week (ordinal categorical), and time block (ordinal categorical).

What kind of model will you build? (What approach will you take for solving the problem and why not any other approaches, including how data will be cleaned, what specific algorithm(s) and any parameters used, and how you will evaluate your approach – describe a figure/table used to illustrate the evaluation)

```
Basic Statistics:
       _12_00_1_00_am  _1_00_2_00am  _2_00_3_00am  _3_00_4_00am  _4_00_5_00am  \
count    42752.000000  42752.000000  42752.000000  42752.000000  42752.000000
mean       251.448423    178.591504    135.280318    117.619359    135.677980
std        407.435712    303.030296    242.091877    215.316979    249.763737
min          0.000000      0.000000      0.000000      0.000000      0.000000
25%         60.000000     38.000000     26.000000     22.000000     27.000000
50%        118.000000     79.000000     56.000000     47.000000     56.000000
75%        241.000000    171.000000    128.000000    110.000000    125.000000
max       4805.000000   4489.000000   4818.000000   4323.000000   4469.000000

       _5_00_6_00am  _6_00_7_00am  _7_00_8_00am  _8_00_9_00am  _9_00_10_00am  \
count  42752.000000   42752.00000  42752.000000  42752.000000   42752.000000
mean     206.655747     352.60051    491.184673    540.203605     524.477381
std      415.614033     621.37410    697.735616    703.696526     680.912007
min        0.000000       0.00000      0.000000      0.000000       0.000000
25%       42.000000      77.00000    133.000000    174.000000     184.000000
50%       85.000000     156.00000    270.000000    325.000000     317.000000
75%      182.000000     335.00000    538.000000    594.000000     555.000000
max     6456.000000    7513.00000   9226.330000   7899.000000    6766.000000

       ...  _2_00_3_00pm  _3_00_4_00pm  _4_00_5_00pm  _5_00_6_00pm  \
count  ...  42503.000000  42503.000000  42503.000000  42503.000000
mean   ...    630.179376    657.691175    661.120886    657.712538
std    ...    758.610055    779.552761    776.312508    770.034661
min    ...      0.000000      0.000000      0.000000      0.000000
25%    ...    250.000000    260.000000    261.000000    258.000000
50%    ...    406.000000    425.000000    428.000000    424.000000
75%    ...    679.000000    713.000000    724.000000    724.000000
max    ...   6996.000000   7524.000000   8683.000000   9762.000000

       _6_00_7_00pm  _7_00_8_00pm  _8_00_9_00pm  _9_00_10_00pm  \
count  42503.000000  42503.000000  42503.000000   42503.000000
mean     628.825612    570.346987    496.648166     428.457779
std      764.796252    732.905268    678.326096     614.480940
min        0.000000      0.000000      0.000000       0.000000
25%      237.000000    204.000000    167.000000     133.000000
50%      396.000000    344.000000    287.000000     235.000000
```

```
       _10_00_11_00pm  _11_00_12_00am
count    42503.000000    42503.000000
mean       376.555067      315.806014
std        565.197252      493.563207
min          0.000000        0.000000
25%        108.000000       82.000000
50%        196.000000      157.000000
75%        366.000000      303.000000
max       5460.000000     5027.000000

[8 rows x 24 columns]

Variance:
_12_00_1_00_am    166003.859238
_1_00_2_00am       91827.360577
_2_00_3_00am       58608.476719
_3_00_4_00am       46361.401429
_4_00_5_00am       62381.924487
_5_00_6_00am      172735.024110
_6_00_7_00am      386105.772593
_7_00_8_00am      486834.989702
_8_00_9_00am      495188.800168
_9_00_10_00am     463641.161800
_10_00_11_00am    454133.381204
_11_00_12_00pm    468611.431926
_12_00_1_00pm     490904.320014
_1_00_2_00pm      520505.316811
_2_00_3_00pm      575489.216000
_3_00_4_00pm      607702.507749
_4_00_5_00pm      602661.109851
_5_00_6_00pm      592953.378613
_6_00_7_00pm      584913.306466
_7_00_8_00pm      537150.132150
_8_00_9_00pm      460126.292896
_9_00_10_00pm     377586.825608
_10_00_11_00pm    319447.933153
_11_00_12_00am    243604.639631
```

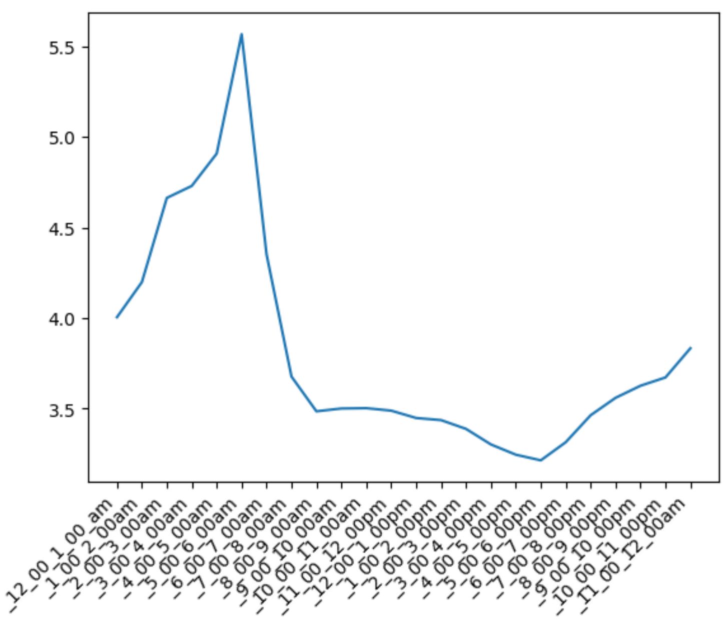
Traffic volume data statistics

```
Basic Statistics:
       dispatch_response_seconds_qy  incident_response_seconds_qy  \
count                 243979.000000                 233955.000000
mean                     140.121699                    713.033643
std                      636.607211                    818.179541
min                        0.000000                      0.000000
25%                       12.000000                    361.000000
50%                       25.000000                    524.000000
75%                       60.000000                    788.000000
max                    31597.000000                  32197.000000

       incident_travel_tm_seconds_qy
count                  233988.000000
mean                      576.080252
std                       476.026302
min                         0.000000
25%                       322.000000
50%                       469.000000
75%                       686.000000
max                     34110.000000

Variance:
dispatch_response_seconds_qy     405268.741274
incident_response_seconds_qy     669417.760791
incident_travel_tm_seconds_qy    226601.040088
dtype: float64
```

EMS Response Data Statistics



Skewness of Traffic Volume Data

```
1 null_df_ems = df_ems.apply(lambda x: sum(x.isnull())).to_frame('count')
2 print(null_df_ems)
```

```
                                count
cad_incident_id                     0
incident_datetime                   0
initial_call_type                   0
initial_severity_level_code         0
final_call_type                     0
final_severity_level_code           0
first_assignment_datetime        2402
valid_dispatch_rspns_time_indc      0
dispatch_response_seconds_qy        0
first_activation_datetime        2789
first_on_scene_datetime          9991
valid_incident_rspns_time_indc      0
incident_response_seconds_qy    10024
incident_travel_tm_seconds_qy    9991
first_to_hosp_datetime          89351
first_hosp_arrival_datetime     90091
incident_close_datetime            52
held_indicator                      0
incident_disposition_code        2776
borough                             0
incident_dispatch_area              0
zipcode                          2198
policeprecinct                   2195
citycouncildistrict              2195
communitydistrict                2196
communityschooldistrict          2365
congressionaldistrict            2195
reopen_indicator                    0
special_event_indicator             0
standby_indicator                   0
transfer_indicator                  0
```

Null values in EMS Dispatch Data

```
1 null_df = df_traffic.apply(lambda x: sum(x.isnull())).to_frame('count')
2 print(null_df)
```

```
                count
id                  0
segmentid           0
roadway_name        0
from                0
to                  0
direction           0
date                0
_12_00_1_00_am      4
_1_00_2_00am        4
_2_00_3_00am        4
_3_00_4_00am        4
_4_00_5_00am        4
_5_00_6_00am        4
_6_00_7_00am        4
_7_00_8_00am        4
_8_00_9_00am        4
_9_00_10_00am       4
_10_00_11_00am      3
_11_00_12_00pm      1
_12_00_1_00pm     253
_1_00_2_00pm      253
_2_00_3_00pm      253
_3_00_4_00pm      253
_4_00_5_00pm      253
_5_00_6_00pm      253
_6_00_7_00pm      253
_7_00_8_00pm      253
_8_00_9_00pm      253
_9_00_10_00pm     253
_10_00_11_00pm    253
_11_00_12_00am    253
```

Null values in Traffic Volume Data

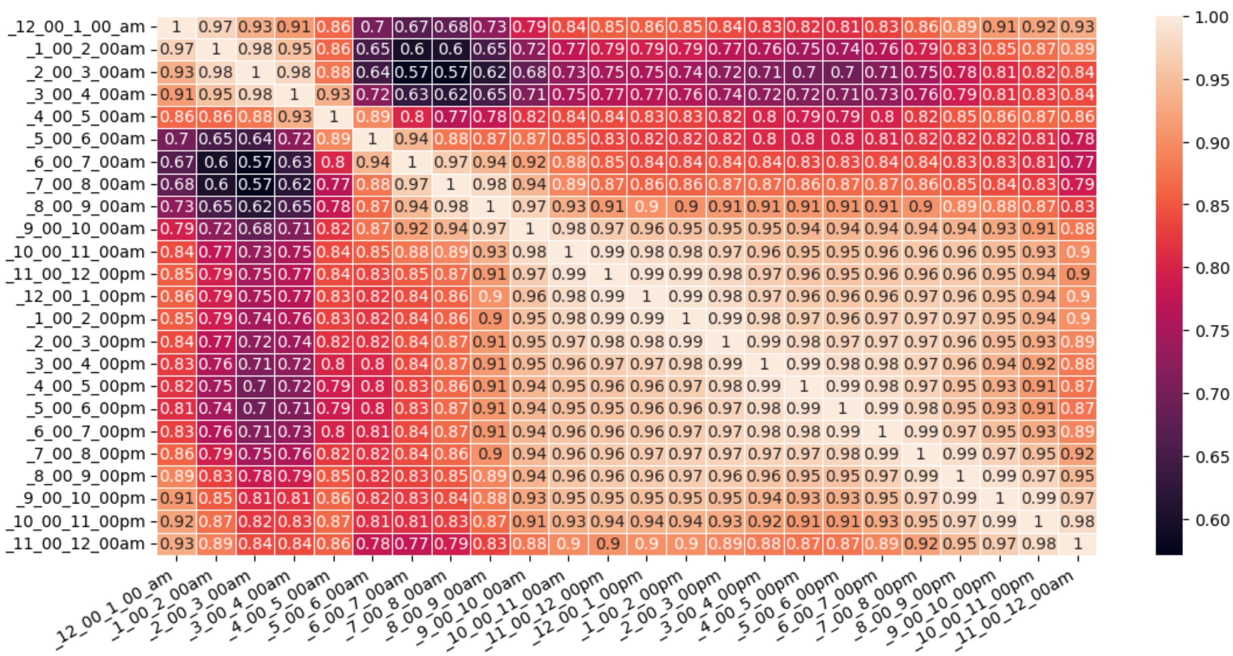
Cleaning and Preprocessing: Observing the presence of 9991 null values in the EMS response time dataset (size 24.4M) and 253 null values in the Traffic dataset (size 42.8K), we have chosen to proceed by removing these records. Given the substantial size of our dataset, eliminating these records is not expected to have a considerable impact on the dataset's overall size or affect our analytical processes.

Regression Analysis:

Performing regression analysis on NYC traffic data and EMS response times can provide valuable insights into the factors influencing response times and help optimize emergency services to improve public safety and save lives.

- $\text{EMS\_Response\_Time} = \beta_0 + \beta_1 * \text{Traffic\_Volume} + \varepsilon$

ElasticNet, Lasso, and Ridge Regression are the choices for our regression problem aimed at optimizing EMS response times in New York City.



Heatmap of the Correlation Matrix of Traffic Volume Data

As shown in the above figure, our traffic response time data has a high degree of multicollinearity. Additionally, we suspect traffic flow data to be correlated with time blocks and days of the week, indicating potential correlation between dependent variables. To mitigate this, we have chosen to evaluate a series regularization metrics throughout our regression analysis.

Firstly, we intend to evaluate the performance of ridge regression in our models since it is valuable when dealing with multicollinearity. This is done via L2 regularization, which distributes the impact of correlated variables more evenly. In our case, it can help manage correlated factors that might affect both traffic patterns and EMS response times. Ridge also improves the stability of regression estimates. It ensures that small changes in the data don't lead to significant changes in the model coefficients. This is essential for robustness in analysis, given that traffic patterns can be subject to fluctuations over time.

Secondly, we intend to evaluate Lasso regression. Lasso is particularly effective at feature selection. It drives some coefficients to zero, effectively eliminating less relevant features from the model. This is beneficial for our analysis as it may help our pruning of the time-related attributes and better identify the factors that have the most significant impact on EMS response times.

Lastly, we intend to evaluate ElasticNet, which combines the strengths of Lasso (L1 regularization) and Ridge (L2 regularization) regression. This is especially important when dealing with complex data and real-world noise – an aspect of all of our data, particularly dispatch time.

Overall, in our EMS response time optimization problem, we have identified the importance of traffic patterns and incident locations, both of which can involve multiple features. Using ElasticNet, Lasso, or Ridge Regression will help us build models that are less prone to overfitting, provide feature selection capabilities to identify key factors, and are interpretable. Moreover, these techniques handle multicollinearity and data noise effectively. By comparing the results from these models, we can gain insights into the impact of different features on EMS response times and potentially optimize dispatch site placement.

Decision Trees:

- Decision trees are also a strong choice for optimizing EMS response times in New York City. Their interpretability allows for a clear understanding of the key factors influencing response times. Moreover, decision trees can capture complex, non-linear relationships in data, such as the peak hours in traffic patterns based on time of day and incident locations. Decision trees can also more accurately make use of categorical variables such as days of the week without data transformations like one-hot encoding which will work to its strength in this case.

Ensemble Learning:

- Similarly, an ensemble of decision trees like Random Forest can capture complex relationships within the data, including traffic patterns, peak hours, and incident locations. It excels at handling categorical variables and can effectively classify weekdays, providing valuable insights into temporal patterns.
- Random Forest mitigates overfitting and offers robust performance through aggregation, enhancing predictive accuracy.
- Its interpretability allows for clear identification of influential factors, aiding in response time optimization. Overall, Random Forest's flexibility and predictive power make it a good choice.

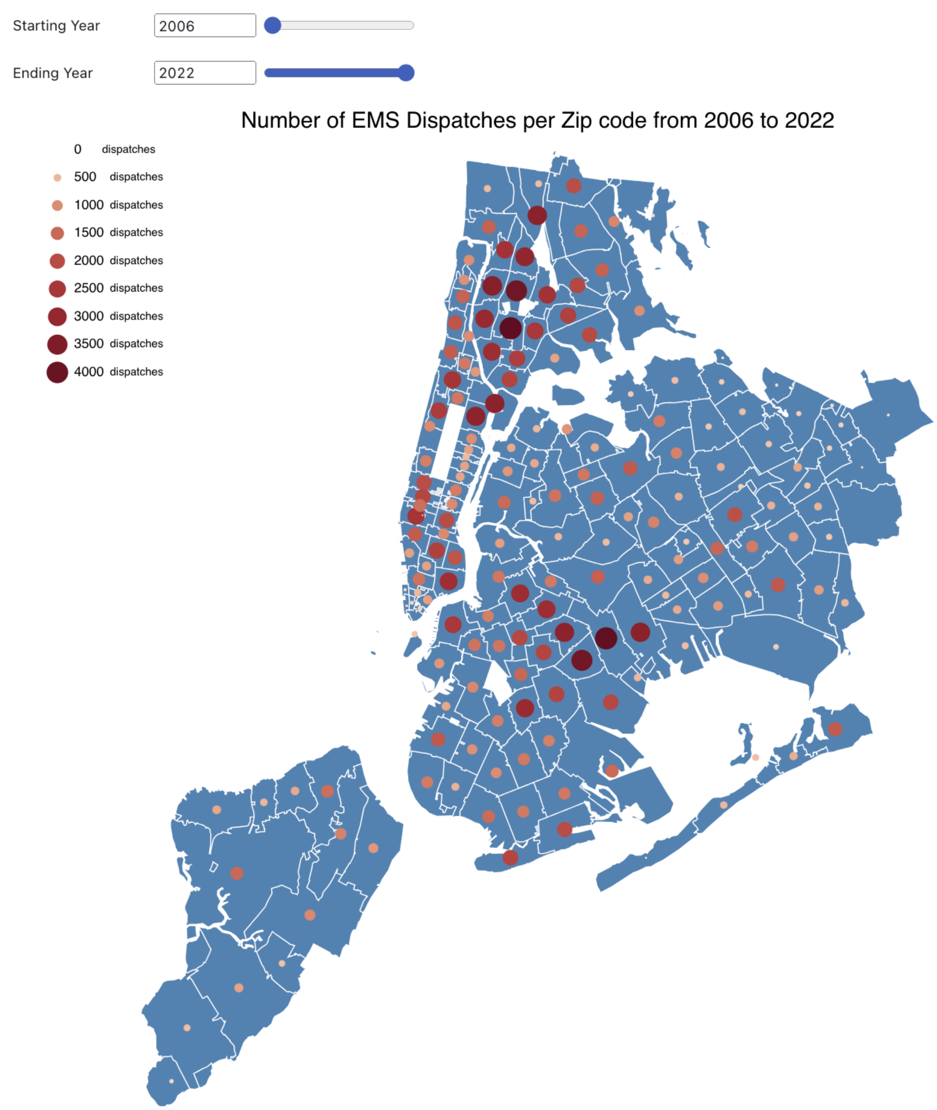
Evaluation:

We plan to use the evaluation methods mentioned below for our model:

- Adjusted R-squared is a modification of R-squared that adjusts for the number of predictors in the model. It penalizes the inclusion of unnecessary variables. It is useful when dealing with multiple regression models.
- MASE is a metric for time series regression that compares the performance of a model to the performance of a naive (e.g., simple moving average) model. It measures the relative accuracy of the model.

What assumptions are safe to make? (Explain clearly what the assumptions being made are and why that's okay, this could be in terms of features considered, potential confounding variables, variable types, etc.)

We make a few key assumptions in order to join databases. Firstly, we ignore the direction of traffic and simplify it to volume in order to facilitate the aggregation of street names into zip codes. Since we are analyzing the data over time, the direction of traffic flow is internalized by the system; however this is a generalization that could introduce a slight bias. Secondly, in order to aggregate dates into days of the week, we only take a subset of traffic data for each week day for every month/year in the dataset (ex. 1st monday, 3rd tuesday, etc.). Though this will similarly introduce bias and make the model more susceptible to outliers, we assume the randomness of looking at multiple months and years on the aggregate should minimize the potential harm. Finally, an underlying assumption of our model is that time of day and day of the week will give new information not already inherited by traffic volume. If the aforementioned ridge regression does not reduce bias enough, we may test additional model implementations omitting these variables to validate that they are significant to the model's accuracy and don't contribute to overfitting.



[Live Graph Link](Observable) (Observable)

- We will use similar graphs to further analyze traffic distribution and the combination of our dependent variables in our final analysis.

Points to be added in the final report:

1. Street names and zipcode mapping - we are assuming that a particular street is mapped to a single zipcode