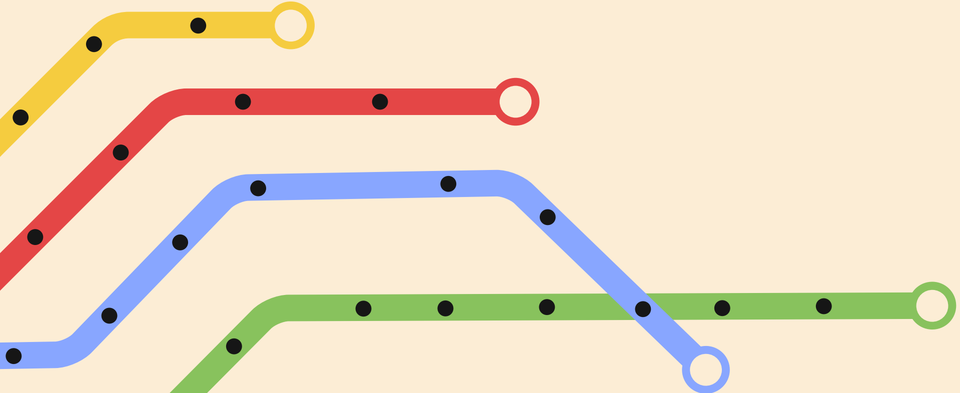
# Subway Delays and Crimes in NYC



# About Us

Fritz Grunert

- Colgate University
- B.A. in Computer Science

Mason Lonoff

- Wake Forest University
- B.S in Business Enterprise Management

Sara Douglas

- Syracuse University
- B.S. in Bioengineering

Susan

- CUNY Hunter College
  - M.A. in Physics
- CUNY John Jay
  - B.S. Forensic Science

# Project Background

- Our project examines the relationship between subway delays and crime in New York City
- We utilized three different datasets to gather information:
    - NYPD Historical Complaints
    - MTA Alerts Archive
    - NYC Transit Subway Map
- Our Hypothesis:
    - Subway delays influence crime in the surrounding area of affected stations

# Initial Questions

## Correlation

Are NYPD crime complaints and subway alerts/delays correlated?

## Proximity

Does the distance to the nearest subway station affect crime?

## Type of Crime

What kinds of crime occur most when there are subway delays?

## Covid-19 Impact

Has the pandemic had any impact on frequency of complaints and/or subway delays?

# Subway Delay Data

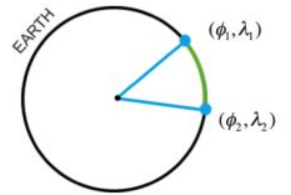
- Data Sources Used:
  - Subway Stops from MTA website
  - Historic alerts/delays were web scraped using Selenium (2018–2021)
    - Each parsed row of HTML is split into 5 columns, ~130,000 rows
    - Topic modeling was used to categorize the messages
      - Alerts were categorized using a LDA model – unsupervised learning method that sorts through text to find the underlying themes
  - Real-Time Alerts with Kafka

MTA New York City Subway

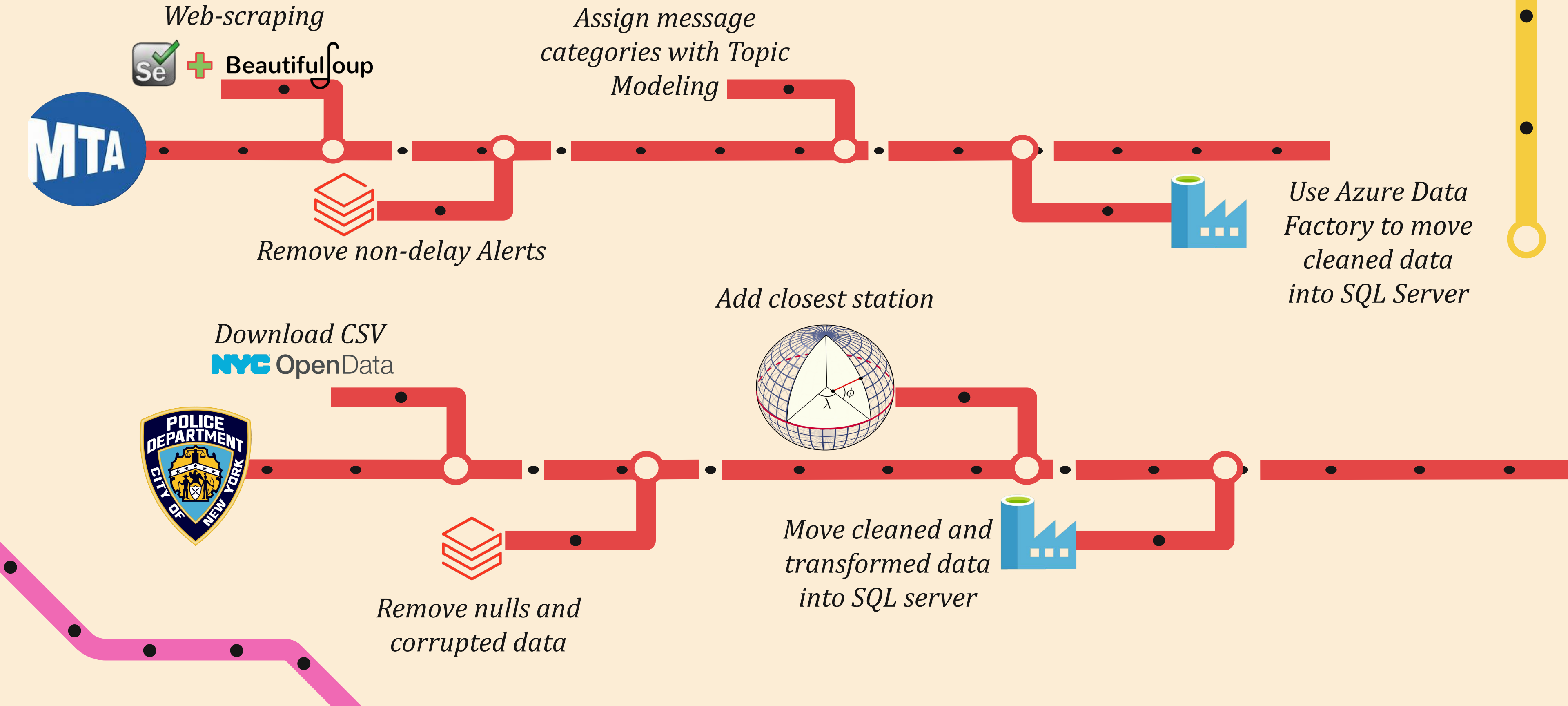
# Crime Data

- NYC OpenData CSV (2018–2021)
  - 35 columns, ~2 million rows
- Each complaint is linked by distance to a subway station
  - Coordinate proximity is determined using the Haversine formula, allowing consideration for the Earth's curvature
- Each complaint is linked temporally to delays
  - For each station that a delay affects, if a complaint is linked to that station within two hours after the delay, then that complaint is joined on the delay

$$\text{haversine}\left(\frac{d}{r}\right) = \text{haversine}(\phi_2 - \phi_1) + \cos(\phi_1)\cos(\phi_2)\text{haversine}(\lambda_2 - \lambda_1)$$



# Historic Data Pipeline



*Web-scraping*

*Remove non-delay Alerts*

*Assign message categories with Topic Modeling*

*Use Azure Data Factory to move cleaned data into SQL Server*

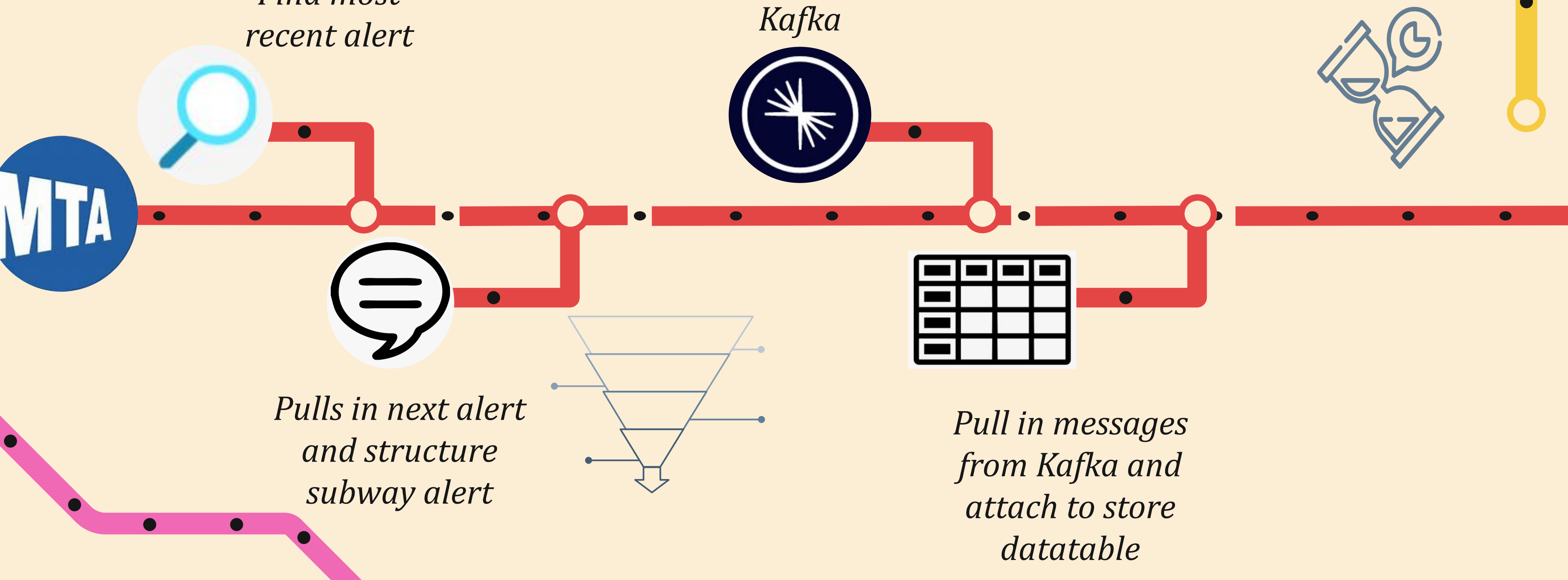
*Download CSV*

*Remove nulls and corrupted data*

*Add closest station*

*Move cleaned and transformed data into SQL server*

# Real-Time/Recent Alerts



*Find most recent alert*

*Pulls in next alert and structure subway alert*

*Send message to Kafka*

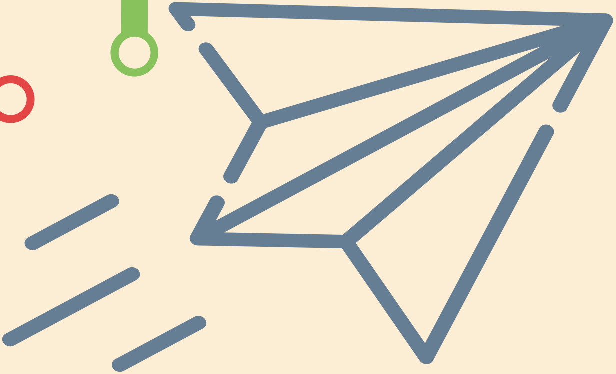
*Pull in messages from Kafka and attach to store datatable*

# Acknowledging Limitations in Data Cleaning: Subway Delay Data

- During exploration, some patterns became apparent:
    - Alerts that contains “train” within the title referred to unexpected delays
    - Alerts that contains “update” refer to a previous delay
    - Alerts contained the affected borough in the alert title
- Alerts that do not follow these patterns might be incorrectly filtered
- Misspelled borough within titles also could have resulted in loss of data since the filtering processed upset regular expressions.

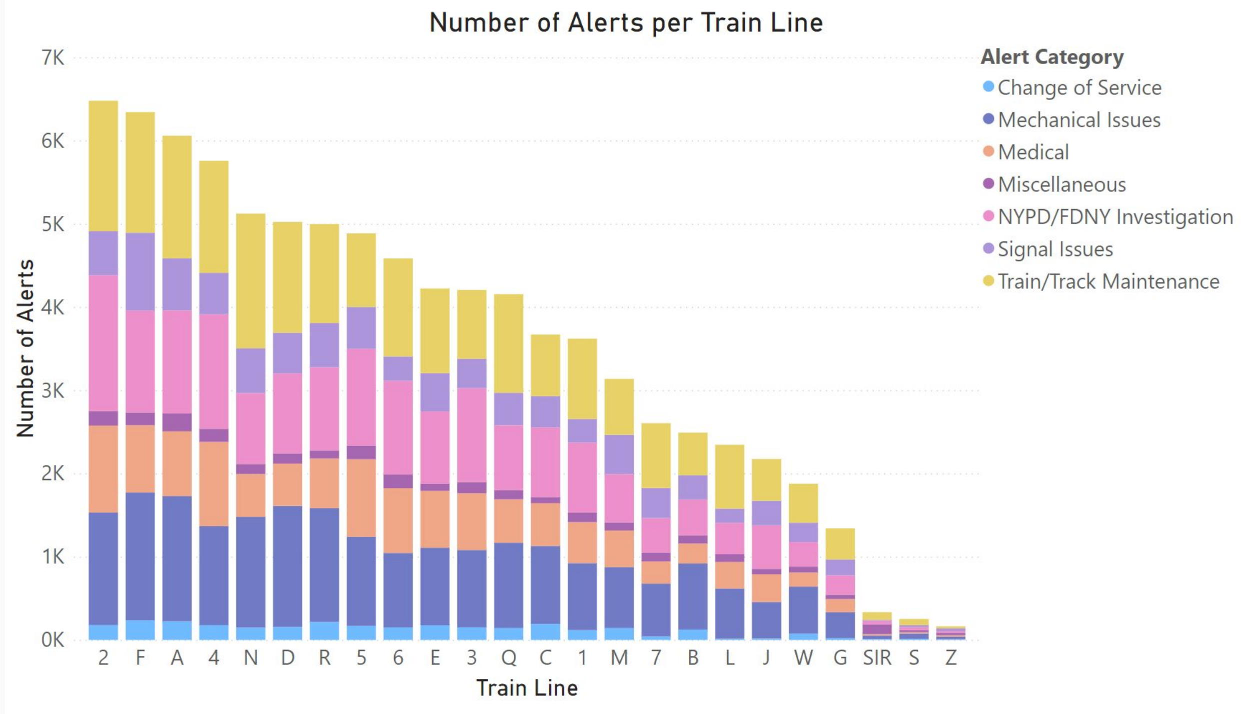
# Acknowledging Limitations in Data Cleaning: Crime Data

- Data Source used: NYPD Complaints Data
  - NYPD Complaint Reports contained some errors or blanks such as dates and ages of suspects and victims.
    - Missing dates and longitude and latitudes were removed.
    - Dates prior to the 2000 were corrected to the best of our abilities using contextual information
  - Corrupted Data
    - In uploading the Brooklyn crime dataset, the file was deemed to be corrupted and therefore, could not be analyzed or visualized.



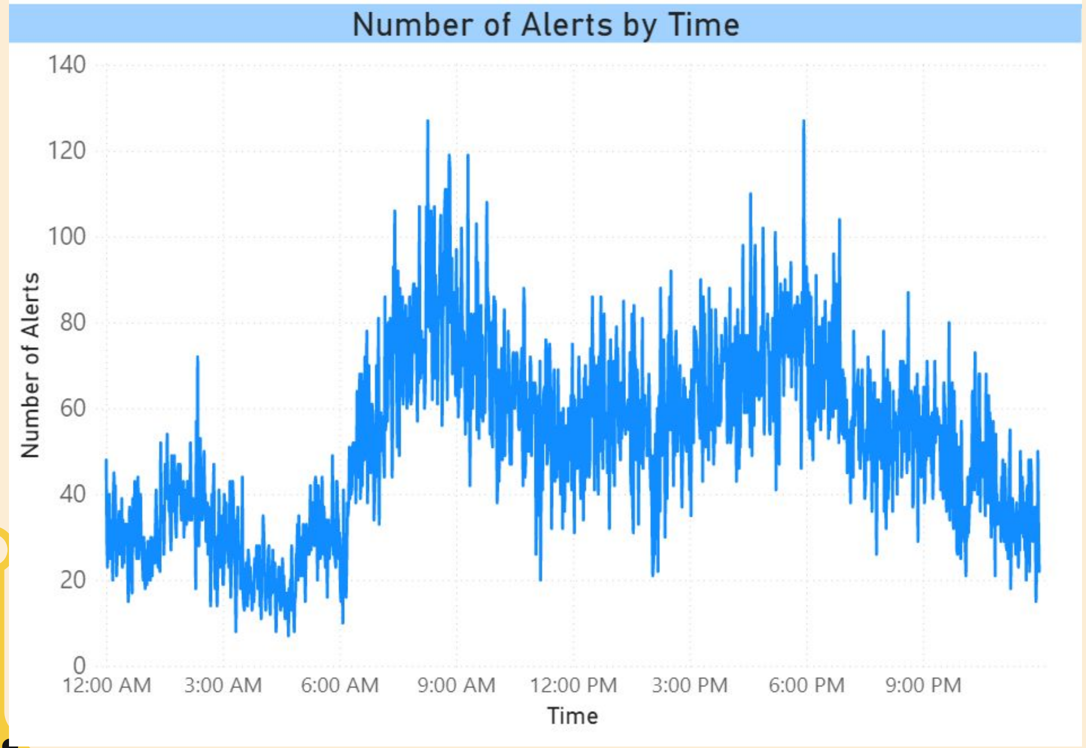
Now let’s bring you over to the dashboard ...

# Findings of Subway Data



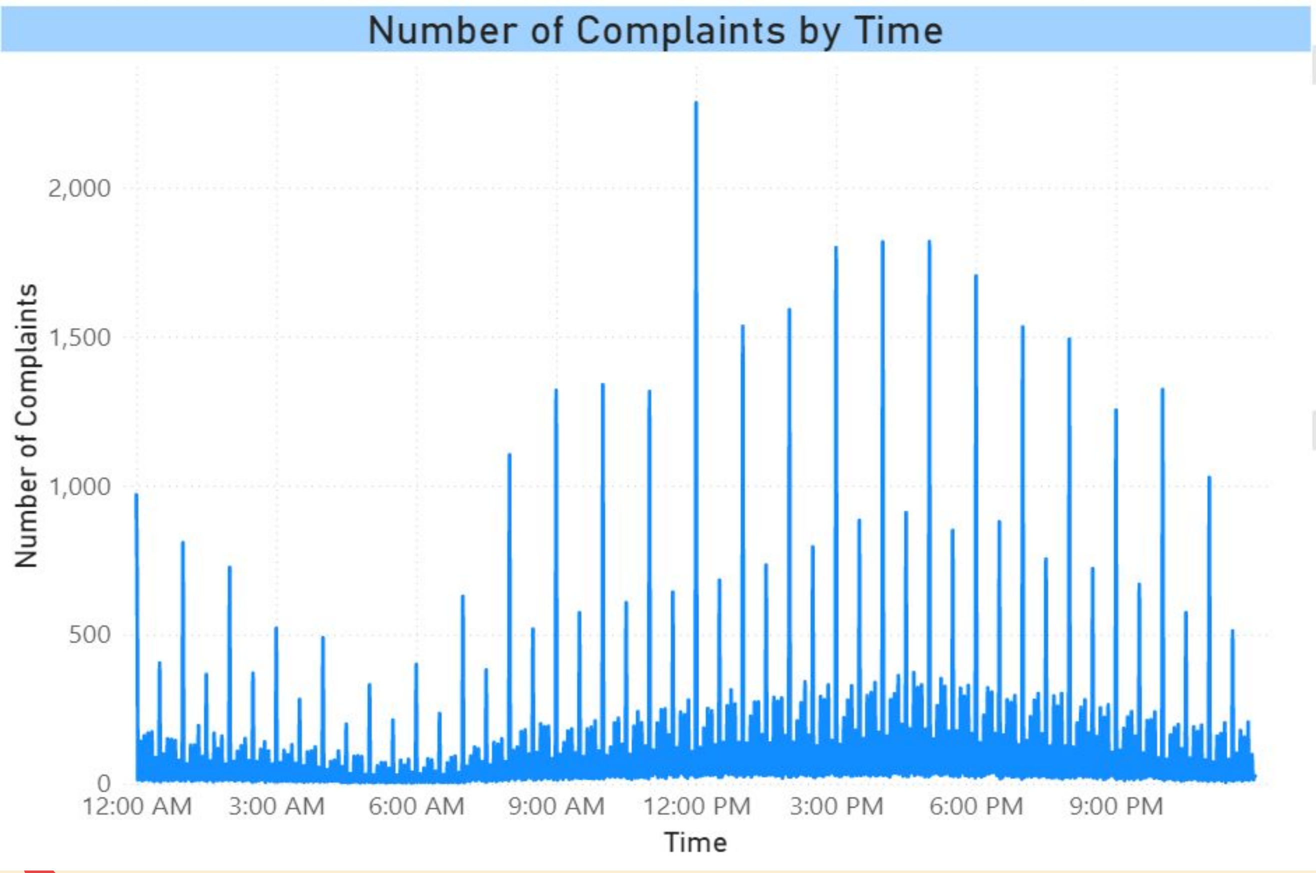
- The 2 Trains have the most number of unexpected delays, followed by the F Trains
- The reasons for delay are largely due to track maintenance and mechanical problems.

# Findings of Subway Data

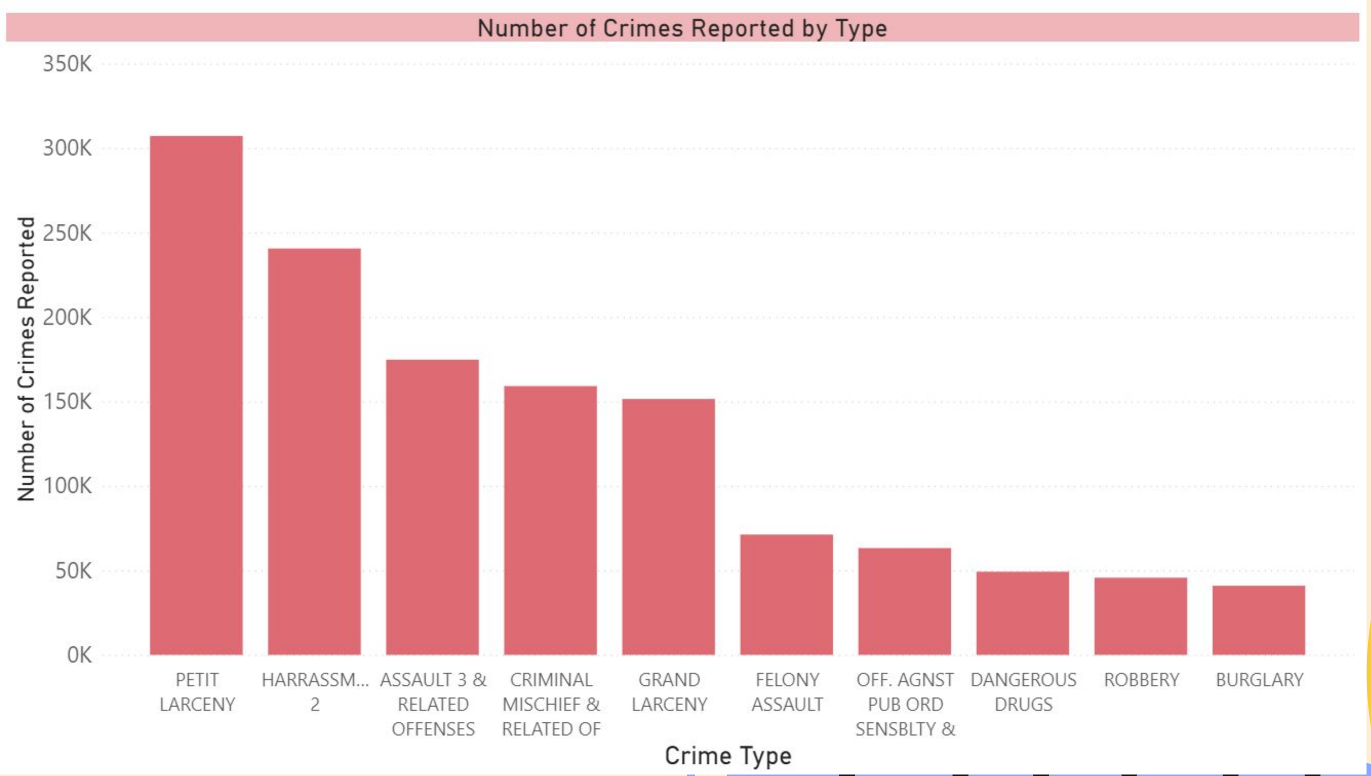


- The number of delays appear to be most common during rush hours, around 9 AM in the morning and 6 PM in the evening.

# Findings from Crime Data

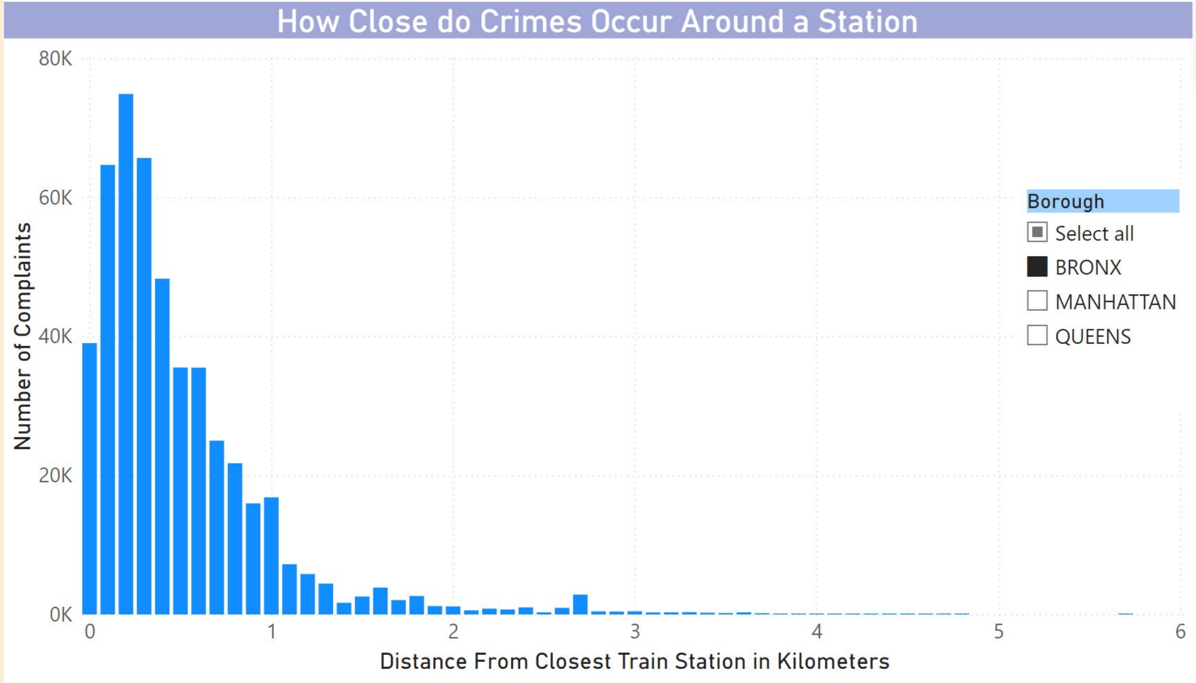


# Findings from Crime Data



# Findings: How close do crimes occur to Train Stations?

- It appears that crime most frequently occurs about 0.3 km away from a train station and then decreases as the location gets further away. On the right is an example of the statistics pulled from the Bronx.
- This can be used to argue that perhaps homes further away from a train station might be safer. However, areas near a train station might be more populated and therefore lead to more crime being reported.



# Does a Train Delay Correlate with Crime?

Parkchester Station subset testing:

Logistic Regression

- Prevalence of majority class (crime occurs) - 61%
- Accuracy score with datetime and delay categories - 73%
- Accuracy score with datetime and no delay categories - 62%

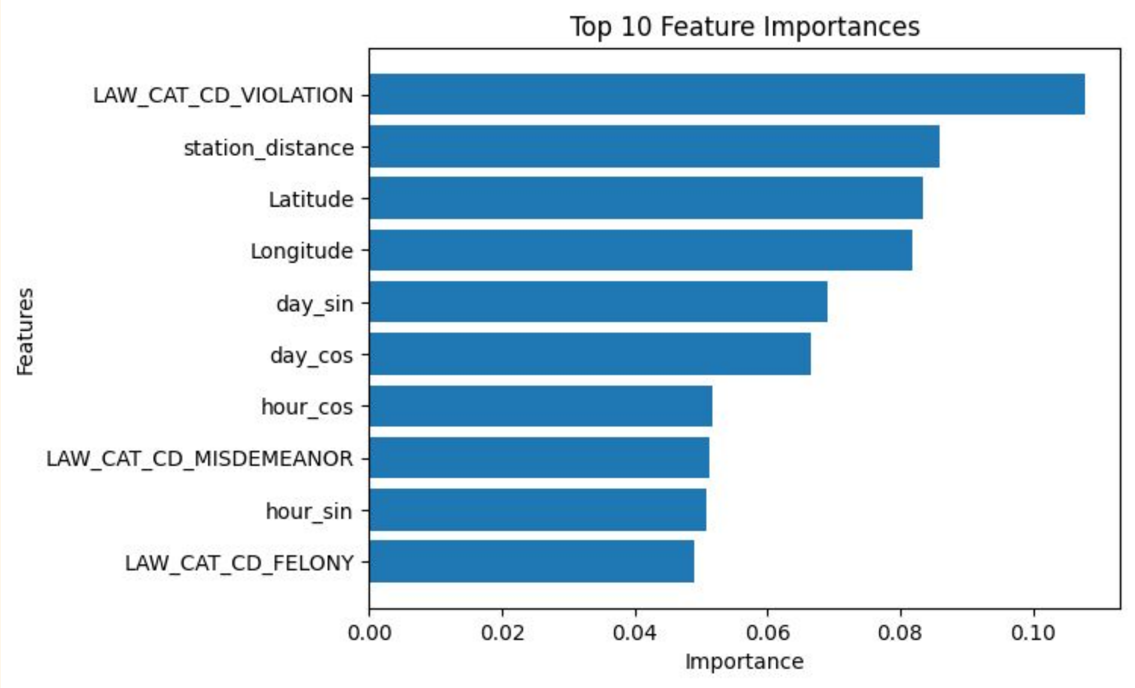
# Machine Learning – Random Forest

- We are predicting the type of crime that occurs
  - The original dataset had 53 crime categories; we narrowed that down to 5:
    - Violence-Related Crimes
    - Public Order Crimes
    - Property Crimes
    - Sex Crimes
    - Other Crimes
- Extra Processing Steps:
  - Encoding y-values
  - Cyclically encoding dates

# Machine Learning - Random Forest continued

Baseline score: 36%

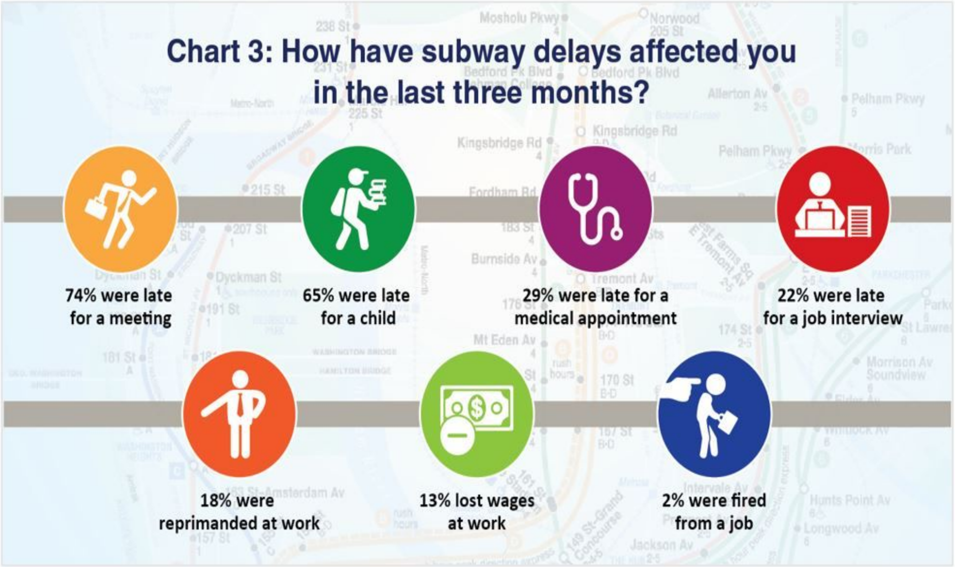
Accuracy score: 57.30%



Validation Accuracy: 57.01%

OOB score:  57.02%

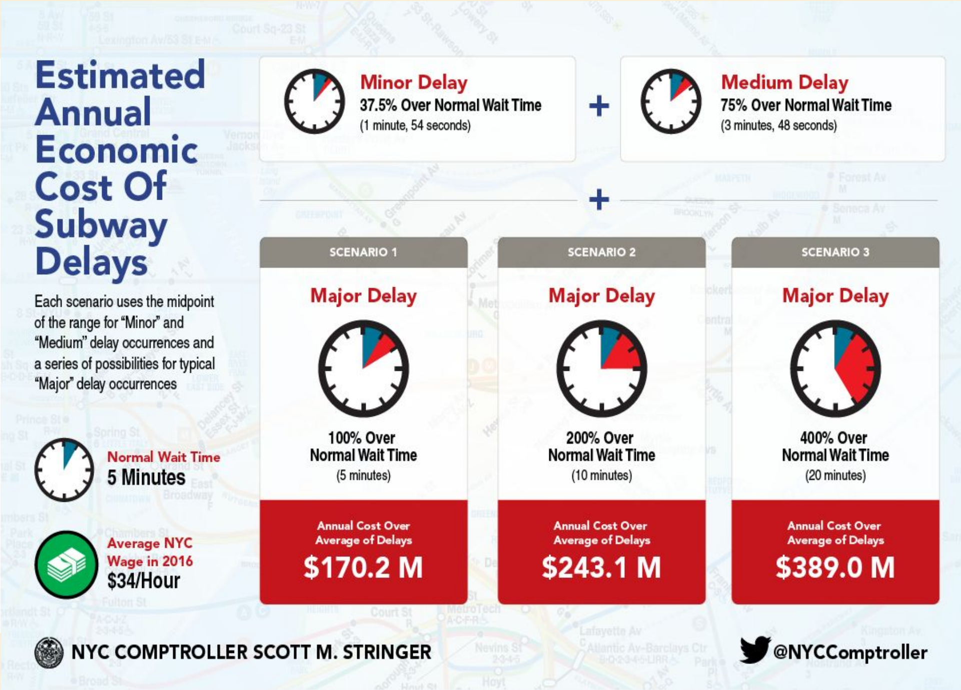
# Real-Word Applications/Implications



## Chart 3: How have subway delays affected you in the last three months?

- 74% were late for a meeting
- 65% were late for a child
- 29% were late for a medical appointment
- 22% were late for a job interview
- 18% were reprimanded at work
- 13% lost wages at work
- 2% were fired from a job

The Human Cost of Subway Delays: A Survey of New York City Riders : Office of the New York City Comptroller Brad Lander (nyc.gov)



## Estimated Annual Economic Cost Of Subway Delays

Each scenario uses the midpoint of the range for "Minor" and "Medium" delay occurrences and a series of possibilities for typical "Major" delay occurrences

Normal Wait Time
5 Minutes

Average NYC Wage in 2016
$34/Hour

Minor Delay
37.5% Over Normal Wait Time
(1 minute, 54 seconds)

+

Medium Delay
75% Over Normal Wait Time
(3 minutes, 48 seconds)

+

| SCENARIO 1 | SCENARIO 2 | SCENARIO 3 |
|---|---|---|
| Major Delay | Major Delay | Major Delay |
| 100% Over Normal Wait Time<br>(5 minutes) | 200% Over Normal Wait Time<br>(10 minutes) | 400% Over Normal Wait Time<br>(20 minutes) |
| Annual Cost Over Average of Delays<br>$170.2 M | Annual Cost Over Average of Delays<br>$243.1 M | Annual Cost Over Average of Delays<br>$389.0 M |

NYC COMPTROLLER SCOTT M. STRINGER

@NYCComptroller

Comptroller Stringer: Subway Delays Hit City Economy, Cost Workers and Business Nearly $400 Million Each Year : Office of the New York City Comptroller Brad Lander (nyc.gov)

# Thanks!

## Do you have any questions?

## Please, don’t hesitate to contact one of us if you’d like to discuss anything further.

Sara Douglas: sdouglas@dev-10.com
Fritz Grunert: fgrunert@dev-10.com
Mason Lonoff: mlonoff@dev-10.com
Susan Lu: slu@dev-10.com

**CREDITS:** This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**