Predicting the Invisible

MetroAnalytics - Team 2

Advanced Analytics Practicum - Spring 2025

By **Syed Hashim Raza** & **Derya Kirca**

Introduction (Hashim)

Human trafficking is a covert and deeply harmful global crime, affecting millions of individuals across labor, sexual, and domestic servitude sectors. According to the International Labour Organization, an estimated 27.6 million people were in situations of modern slavery globally in 2021, including 6.3 million victims of forced commercial sexual exploitation (ILO, 2022). In the United States, the National Human Trafficking Hotline received over 51,000 substantive signals in 2021, including tips, texts, and reports, resulting in 16,554 identified victims - a number widely considered an undercount (Polaris Project, 2023).

A key barrier to detection is the misclassification and underreporting of trafficking-related activity. Victims are often entangled in other crimes (e.g., undocumented labor, prostitution), manipulated by traffickers to distrust authorities, or afraid of retaliation (Farell et al., 2019). As a result, incidents involving trafficking are frequently recorded under broader crime categories such as assault, abduction, sex work, masking their true nature and complicating intervention.

This project was developed to address these challenges by using a multi-layered data science approach to model trafficking risk signals in New York City - especially around JFK International Airport, which handled over 61.6 million passengers in 2023, making it a high-risk corridor for transit-based exploitation (Port Authority of NY/NJ, 2024). Instead of relying solely on explicit trafficking case data, which is sparse, we analyzed indirect signals including crime trends, major city events, population movement, and spatial context. This project employed a multi-layered risk modeling framework to identify patterns of potential human trafficking activity in New York City.

The modeling architecture was designed to overcome the limitations of sparse or misclassified trafficking data by generating indirect risk signals, validating them across multiple perspectives, and ensuring the results remain interpretable for real-world use. The approach supports actionable insights for stakeholders such as public safety agencies, advocacy organizations, and transportation authorities to help prioritize monitoring, outreach, and resource allocation.

An overview of the modeling methodology - including supervised learning, unsupervised validation, and contextual interpretation - is provided in the next section, which outlines the technical foundation of the framework and its relevance to the project’s goals.

Background (Hashim)

This project was developed in collaboration with Metro Analytics, an urban planning and transportation consultancy that works with public and private stakeholders to address infrastructure, mobility, and regional development challenges. With a focus on data-driven solutions, Metro Analytics applies advanced analytics and modeling techniques to support long-term planning for cities, counties, and transit systems across the United States.

As part of their commitment to social impact and innovation, Metro Analytics partnered with Rutgers University through the MBS Externship Exchange Program to sponsor a data science initiative focused on human trafficking risk detection and prevention. The project was supervised by two expert mentors from Metro Analytics:

- **Dr. John Betak**, a senior operations and management consultant, a research fellow at Rutgers University's Infrastructure and Transportation (CAIT) department, and a senior research fellow at University of Texas at Austin Center for Risk Management and Insurance. He hails over 50 years of diverse international experience in freight and rail operations.
- **Dr. Felipe Aros-Vera** is an associate professor of Industrial and Systems Engineering (ISE) at Ohio University. He holds a Phd in Transportation Engineering from the Rensselaer Polytechnic institute (RPI), and an M.S. and B.S. in Engineering Sciences.

Their guidance was instrumental in shaping the problem framing, validating our approach, and ensuring real-world applicability. Within Rutgers University, the project was supported by **Professor Brian Petrus**, who provided academic mentorship, structured feedback, and ensured alignment with program standards and deliverables.

The objective of the project was to develop a data science solution that could identify spatio-temporal hotspots of human trafficking risk in urban environments, with a primary focus on the New York City area and JFK International Airport. The project explored how predictive modeling could assist public agencies and advocacy groups in prioritizing intervention efforts based on dynamically evolving patterns of activity.

Given the sensitive and underreported nature of human trafficking, this project employed a multi-layer modeling approach that allowed for flexible, interpretable, and resilient analysis. The modeling strategy integrated both supervised learning and unsupervised validation techniques, supported by contextual analysis.

Supervised models such as Long Short-Term Memory (LSTM) neural networks were used to capture temporal dependencies in location-based sequences of crimes, events, and mobility data. In parallel, gradient boosting models (e.g., XGBoost and LightGBM) were applied during early experimentation to benchmark predictive performance and assess feature importance.

Unsupervised methods - including clustering and hierarchical density-based spatial clustering - were used to surface spatial patterns and validate risk zones identified by supervised learning.

This combination allowed us to address key challenges such as sparse and delayed trafficking case data, misclassification of trafficking-related activity, and the need for proactive rather than reactive detection.

A deeper explanation of how each model was selected, structured, tuned, and evaluated is provided in the Project Methods section of this report.

Literature Review (Derya)

A study conducted by researchers from Universidad Andres Bello and Universitat Jaume I explored the spatio-temporal prediction of crime events in Baltimore using a hybrid deep learning model known as CLSTM-NN, which integrates Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) layers. The team utilized publicly available crime records from 2016 to 2018, focusing on high-frequency crimes like street robbery and larceny. Each day's crime data was converted into spatial grids (8×8 and 16×16), where each cell represented a count of incidents in that zone. These matrices were then used to train the model to forecast whether at least one crime would occur in each cell on future days. The study tested multiple scenarios, including different grid sizes and time windows (1 to 7 days), and evaluated performance using accuracy, AUC-ROC, and AUC-PR. The most effective configuration achieved 86% accuracy and an AUC-PR of 0.93 for larceny when aggregating seven days of data. The authors concluded that spatial and temporal patterns could significantly improve crime prediction and that CLSTM outperformed traditional models, especially when handling sparse event distributions. They also noted challenges with low-event-density areas and suggested integrating attention mechanisms and external variables (e.g., weather, demographics) in future research. This study directly informed our supervised approach to predicting human trafficking risks, validating our decision to use a grid-based LSTM model and multi-day input sequences to capture spatial and temporal dynamics.

Another study by Reddi, Kusuma, and Parvin (2024) conducted a comprehensive analysis of crime patterns in Los Angeles by integrating textual, spatial, and temporal dimensions. The researchers used machine learning algorithms such as Random Forest, Logistic Regression, and XGBoost for crime type classification, and applied clustering methods such as HDBSCAN, OPTICS, DBSCAN, and MiniBatch KMeans to identify spatial hotspots and cluster crimes based on latent topics extracted via LDA. Their temporal component employed an LSTM model to forecast future crime counts using historical trends. Clustering performance was evaluated using silhouette scores, with OPTICS and HDBSCAN showing the strongest separation. The study emphasized the importance of combining multiple data dimensions to improve public safety and strategic planning. This study's use of unsupervised clustering techniques to uncover high-risk areas without relying on labeled outcomes, directly supports our own approach of using KMeans and HDBSCAN to detect human trafficking hotspots and patterns based on natural spatial groupings.

The study titled "Semi-supervised learning for detecting human trafficking" by Alvari et al. (2017) focused on the impact that semi-supervised machine learning can have in identifying trafficking activity in online advertisements. In the early stages of our project, their approach closely aligned with our use of sentiment analysis and entity recognition in business reviews and social media. The authors introduced a model called S3VM-R, an enhanced Laplacian SVM that incorporated both labeled and unlabeled data along with domain-specific features tied to trafficking behavior. Their model successfully flagged suspicious ads on Backpage.com with strong validation from law enforcement experts. Their study guided our data-driven, risk-based approach by showing how minimal labeling combined with targeted feature engineering can uncover trafficking patterns, reinforcing our use of clustering to detect high-risk areas without relying on labeled outcomes.

Another related study that we were able to find regarding our project topic was titled "Sex Trafficking Detection with Ordinal Regression Neural Networks" by Wang et al. (2020). The study focused on improving the detection of sex trafficking in escort advertisements using a deep learning model tailored for ordinal classification. Trained on the expert-labeled Trafficking-10k dataset, their model incorporated word embeddings, gated-feedback RNNs, and a custom ordinal regression layer to improve consistency across prediction levels. They also used unsupervised techniques and t-SNE visualizations to identify subtle trafficking signals such as emojis and coded language within ad text. The model outperformed previous baselines, showing the power of targeted feature design. This work highlights the importance of feature selection and representation in improving model accuracy, which reinforced our own emphasis on feature engineering for business data analysis early in the project.

The fifth source of research is an analysis that was completed by researchers Dimas et al. (2022), who conducted a systematic review of 142 studies applying operations research (OR) and analytics to combat human trafficking. Their work categorized studies by methodology, trafficking type, data sources, and alignment with the anti-trafficking 4Ps framework, prevention, protection, prosecution, and partnership. Most studies focused on sex trafficking, relied on secondary data (especially online ads), and employed techniques like clustering, classification, and unsupervised learning. The authors also emphasized the need for more attention to labor trafficking, richer data diversity, and stronger cross-sector collaboration. This review reinforced our data-driven, risk-based approach by supporting the use of unsupervised methods like KMeans and HDBSCAN to identify trafficking patterns without relying on labeled outcomes.

A study conducted by Bermeo, Escobar, and Cuenca (2023) titled *"*Human Trafficking in Social Networks: A Review of Machine Learning Techniques" provides a comprehensive overview of how machine learning has been applied to detect human trafficking via social media platforms. The authors reviewed 23 key studies from 2016 to 2022, categorizing them by learning type; supervised, unsupervised, and semi-supervised and by tasks such as text classification, sentiment analysis, topic modeling, and anomaly detection. They also highlighted the use of social media APIs (particularly Twitter) for data extraction and outlined key preprocessing methods like tokenization and named entity recognition. In the early stages of our project, their findings helped guide our approach to social media scraping and anomaly detection, particularly in identifying suspicious business activity and behavioral patterns. The study further reinforced our use of unsupervised learning for detecting trafficking-related risks without relying on labeled outcomes.

The Data Overview (Hashim and Derya)

Supervised Models Data

The project relied on a multi-source, spatial-temporal dataset engineered specifically to detect potential human trafficking (HT) risk across New York City. Due to the underreporting and misclassifications of HT cases, the modeling strategy was built around proxy indicators drawn from diverse datasets, capturing patterns in crime, population movement, large-scale public events, and known HT reports. The goal was to create a unified structure that allowed daily-level risk prediction at a granular geographic resolution.

Data Sources and Collection

Data for this project was gathered from a combination of publicly available portals, governmental agencies, and restricted-access nonprofit sources. Public datasets included daily NYPD complaint records obtained from NYC OpenData, NYC large-scale event permitted records, and monthly inbound passenger volumes to JFK Airport sourced from the Port Authority of New York and New Jersey. Contextual data, such as borough boundaries, geospatial coordinates, U.S. holidays, and seasonal indicators, were sourced from public GIS datasets and government calendars to support spatial and temporal alignment.

In addition, a non-public dataset containing verified human trafficking case reports - TAHub Metro NY - was provided by Metro Analytics for use exclusively within this externship project. This dataset originated from trusted nonprofit advocacy sources and included confirmed incident dates and locations of trafficking reports in the greater NYC metro area. Due to its sensitive nature, the data was not publicly distributed and was handled with care under data-use agreements.

The overall time frame spanned from January 1, 2021, through December 31, 2024, offering over four years of daily activity data across all sources. The period was selected to capture both pre- and post-pandemic behavioral shifts, seasonal mobility trends, and sufficient historical overlap with verified HT incidents. All datasets except for the monthly JFK passenger volume were originally at a daily resolution. Passenger volume data was normalized and interpolated to fit the daily modeling structure. This ensured consistency across the full spatial-temporal grid used for predictive modeling.

The table below summarizes the data collected for this project alongside its sources and timeframe.

Domain	Dataset	Source	Timeframe
Crime	NYPD Complaint Data	NYC OpenData	Jan 2021 - Dec 2024
Verified HT Reports	TAHub Metro NY Reports	Metro Analytics	Jan 2021 - Dec 2024
Events	NYC Permitted Events Data	NYC OpenData	Jan 2021 - Dec 2024
Passenger Volume	JFK Monthly Passenger Volume	Port Authority NY/NJ	Jan 2021 - Dec 2024
Geography	NYC Borough and Grid Data	NYC GIS	Static
Calendar	US Holidays and Weekends	Public Calendar	Jan 2021 - Dec 2024

Data Instances and Attributes

The final merged dataset consisted of approximately 7.8 million rows, where each row represented a daily observation for a specific spatial grid cell within New York City. To ensure consistency in modeling risk over time and space, a custom geospatial grid was designed to divide NYC into uniform zones, each measuring approximately 0.25 square kilometers. This spatial framework served as the foundation for aggregating all data sources - allowing us to generate structured sequences of inputs per location for temporal modeling.

The purpose of creating this grid was twofold. First, it addressed the challenge of spatial irregularity in the original data, where incidents (e.g., crimes or HT reports) occurred at varying and often sparse coordinates across the city. By assigning these incidents to standardized grid cells using their latitude and longitude, we could ensure that each location had a fixed, traceable identity - a crucial requirement for fitting LSTM models, which learn patterns from sequences of data tied to consistent locations. Second, the grid structure enabled the aggregation of multi-source inputs (e.g., event density,

passenger flow, calendar features) to the same unit of space and time, making the dataset both interpretable and scalable.

While the original crime and TAHub datasets included raw latitude and longitude coordinates, these were used to map events to grid cells during preprocessing. Each grid cell was then associated with a centroid coordinate (lat/lon) to retain spatial reference, but model inputs and outputs were ultimately structured around the grid ID rather than raw coordinates. This also allowed for risk visualization and hotspot mapping at a uniform spatial scale.

Each row in the dataset contained a mix of temporal, spatial, and behavioral features. Temporal attributes included the observation date, binary indicators for weekends and the U.S. holidays, and a categorical variable for season (e.g., Winter, Spring). Spatial attributes consisted of grid IDs, borough names, and the grid cell's central coordinates. Behavioral features included:

- Daily crime counts, derived from NYPD complaint data
- Nighttime crime ratio, used as a proxy for hidden covert activity
- Number of large events occurring within a 1-mile radius, calculated using NYC's event permit data
- Scaled inbound passenger volume to JFK Airport, interpolated from monthly reports
- Presence of verified HT case, derived from the TAHub dataset and used as a binary evaluation label.

Additionally, we engineered temporal trend features such as 7-day rolling aggregates for crime and event indicators to help the model capture short-term escalation patterns. All features were either normalized or encoded as needed for model compatibility.

This structured setup ensured that each (date, location) pair had a comprehensive set of inputs for prediction and evaluation. It also allowed flexibility to train both deep learning and traditional machine learning models while preserving spatial-temporal interpretability for risk communication and visualization.

The table below summarizes the data attributes, instances, types, values, and sources.

Attribute Name	Data Type	Description	Range/Value	Source
Date	Date	Observation date	2021-2024	All Datasets
Location ID	Categorical	Unique ID for spatial grid cell	~2600+ unique values	Custom Spatial Grid
Borough	Categorical	NYC Borough where the grid cell is located	Manhattan, Bronx, Brooklyn, Queens, Staten Island	NYC GIS
Latitude/ Longitude	Float	Grid cell centroid coordinates	~40.5 to 40.9 (lat), ~-74.2 to -73.7(lon)	Custom Grid Centroids
Daily crime count	Integer	Total number of crimes reported within a grid cell	0 - 100+	NYPD Complaint Data
Night Crime Ratio	Float	Proportion of crimes occurring at night	0.0 - 1.0	NYPD Complaint Data
Event Count	Integer	Count of permitted events within 1-mile radius of the grid cell	0 - 10+	NYC OpenData - Permitted Events
Passenger volume scaled	Float	Scaled inbound passenger volume to JFK Airport	0.0 - 1.0	Port Authority of NY/NJ
Is weekend	Binary	Weekend indicator	0 = Weekday, 1 = Weekend	US Calendar
Is holiday	Binary	Federal holiday indicator	0 = No, 1 = Yes	US Federal Holiday Calendar
Season	Categorical	Season label for the date	Winter, Spring, Summer, Fall	Derived from date
Verified HT Case	Binary	Indicator for presence of a verified HT report in the grid on a given day	0 = No, 1 = Yes	TAHub (via Metro Analytics)
Rolling 7-day crime count	Integer	7-day rolling total of crimes in the grid cell	Varies by location	Engineered from NYPD data
Rolling 7-day event count	Integer	7-day rolling total of events near the grid cell	Varies by location	Engineered from events data

Data Cleaning and Preprocessing
Substantial cleaning and harmonization were required to align and integrate the different datasets. Dates across all sources were converted into a consistent format and validated to ensure completeness. Missing values - particularly for event and crime data - were either filled with zeros (where absence of reports implied no activity) or forward-filled in the case of monthly indicators. Duplicate entries were removed based on ID columns and fuzzy matching of descriptions and timestamps, especially in the TAHub case data.

Geospatial joins were conducted to match incidents and events to the appropriate grid cells based on their latitude and longitude. This was achieved using Python libraries such as GeoPandas and Shapely. Monthly JFK passenger volumes were normalized using MinMaxScaler and assigned to each day in the corresponding month. Categorical variables, such as boroughs and seasons, were either one-hot encoded or label-encoded depending on the model requirements.

Additional feature engineering steps included calculating night crime ratio as a proxy for hidden activity, estimating proximity-based event impacts, and constructing 7-day rolling trends to help models learn short-term temporal dynamics. The dataset was formatted to allow flexibility between tabular inputs for tree-based models and sequence inputs for deep learning models like LSTM.

Database Structure and Integration
To support large-scale data handling and model experimentation, we constructed a unified master table combining all sources at the date and location level. Intermediate tables were saved in csv (comma separated values) for efficient querying and manipulation. This setup enabled fast reshaping for model-specific input requirements - for instance, generating time-windowed sequences for LSTM or flattening temporal aggregates for XGBoost and LightGBM.

Each join operation was conducted carefully to preserve spatial and temporal integrity. Merges were performed using spatial containment logic (for events and crime) and time-matching functions. Whenever possible, comments were created within the code scripts to track and audit the flow of data from raw source to final model input.

Justification of Data Design
The selected data sources and engineered features were chosen based on their strong contextual relevance to known risk patterns. For example, large event density and passenger surges have been repeatedly cited in human trafficking risk literature as a

contributing factor to exploitation risk. Nighttime crime patterns and mobility changes around airports also served as important behavioral signals in prior urban crime studies.

Importantly, the entire framework was designed to generalize beyond NYC. While this project was localized, each data type - crime, mobility, events, and contextual overlays - can be adapted to other cities that have similar open data infrastructure, making the modeling pipeline scalable and transferable.

Unsupervised Models Data
Data Instances and Attributes

The final dataset used for clustering comprised 30,691 daily observations, with each row representing a unique (date, borough) combination across Manhattan, Brooklyn, Queens, Bronx, and Staten Island, spanning from January 1, 2021, to December 31, 2024. Unlike earlier grid-based approaches used in our LSTM models which divided the city into artificial spatial zones, this dataset retained the raw geographic coordinates (latitude and longitude) associated with each event. This decision was intentional: preserving raw spatial detail allowed unsupervised clustering algorithms like KMeans and HDBSCAN to detect natural groupings and spatial patterns of human trafficking (HT)-related activity, free from the constraints of predefined grids.

To construct this dataset, we merged multiple sources, each capturing a different dimension of risk or context:

- NYPD Complaint Data provided the baseline counts of HT-related crime reports across boroughs daily.
- TAHub Metro NY Reports offered verified human trafficking cases used for post-clustering validation, enabling us to assess the credibility of detected hotspots.
- NYC Permitted Events data was used to quantify daily large-scale public events potentially high-risk situations that draw transient populations.
- JFK Airport Passenger Volume was sourced monthly and linearly interpolated to estimate daily inbound flow, serving as a mobility proxy for population movement across boroughs.

Each row in the dataset was structured to contain a comprehensive mix of temporal, spatial, and behavioral features, optimized for pattern detection and hotspot analysis:

Temporal Features:

- Date (exact observation day)
- Day of the week, weekend indicator, and U.S. holiday flag
- Season label (Winter, Spring, Summer, Fall), derived from date

Spatial Features:

- Borough name (for interpretability)
- Latitude and Longitude, which anchored clustering analysis and visualization

Behavioral & Contextual Features:

- ht_related_crimes: Daily count of HT-related incidents reported
- ht_reports: Presence of verified HT case reports (binary or count)
- event_count: Number of city-permitted events occurring that day in the borough
- closure_count: Events that included street closures, acting as a proxy for higher-footfall zones
- jfk_passengers: Scaled passenger inflow estimates per borough for each day

While the ht_crime_occurred variable, a binary indicator of whether any HT-related incident occurred on a given day was also included in the merged dataset, it was not used in unsupervised clustering training. Instead, it was preserved for post-clustering evaluation, allowing us to assess whether identified clusters coincided with actual crime activity.

This dataset architecture enabled a flexible and interpretable approach to pattern discovery, allowing both KMeans and HDBSCAN to operate on rich, multi-source inputs while preserving the spatial and temporal integrity of each record. The resulting clusters were used not only for mapping potential hotspots, but also for deriving behavioral summaries across key risk indicators such as time of week, seasonality, and event influence supporting both spatial insight and preventative strategy development.

Attribute Name	Data Type	Description	Range/Value	Source
Date	Date	Observation date	2021–2024	All datasets
Borough	Categorical	NYC borough where the event occurred	Manhattan, Bronx, Brooklyn, Queens, Staten Island	NYC GIS
Latitude /Longitude	Float	Raw spatial coordinates of the reported crime/event	~40.5–40.9 (lat), ~-74.2 to -73.7 (lon)	NYPD, Event Data
ht_related_crimes	Integer	Count of HT-related crimes reported on that date and location	0–5+	NYPD Complaint Data
event_count	Integer	Number of large-scale events held	0–5+	NYC OpenData – Permitted Events

		in the borough on that day		
jfk_passengers	Float	Scaled inbound passenger volume for the borough (interpolated from monthly)	0.0 – 1.0	Port Authority of NY/NJ
day_of_week	Integer	Day of the week	0 (Monday) – 6 (Sunday)	Derived from date
is_weekend	Binary	Indicator if the date falls on a weekend	0 = Weekday, 1 = Weekend	US Calendar
month	Integer	Month of the observation	1–12	Derived from date
season	Categorical	Season label for the observation	Winter, Spring, Summer, Fall	Derived from date
area_cluster	Integer	Cluster label assigned by KMeans or HDBSCAN	Varies by model	Model output (unsupervised)
crime_ordered_cluster	Integer	Re-ranked cluster label based on total HT-related crimes	1–Top N	Derived during post-processing

Data Cleaning and Preprocessing (Clustering Methods)

Preparing the dataset for clustering involved simplifying the structure to focus on spatial and temporal integrity while preserving key behavioral indicators. All datasets were first standardized to a consistent daily format spanning January 1, 2021, to December 31, 2024. Dates were validated, and missing values were handled by assuming zero activity (e.g., zero events or crimes) where applicable.

Raw latitude and longitude coordinates were retained (unlike the grid-based transformation used for LSTM) to allow KMeans and HDBSCAN to detect natural groupings based on actual event locations. Duplicate entries were removed through ID checks and timestamp matching. Event data was filtered to include only large-scale, permitted events likely to attract high foot traffic, and monthly JFK passenger volumes were interpolated to daily values using linear scaling to align with the rest of the dataset.

Additional preprocessing included:

- Borough-level tagging to enable interpretability of clustered regions
- Temporal flags such as day of week, season, and weekend indicators
- Standardization and scaling for clustering inputs (e.g., MinMax scaling for continuous features if needed for distance-based clustering like KMeans)

Database Structure and Integration

For clustering, a flat, row-based structure was adopted, where each record represented a unique (date, borough) instance with attached spatial coordinates. This structure allowed direct clustering using latitude and longitude while associating each point with additional contextual variables like crime counts, event density, and mobility data.

All source datasets including NYPD complaints, TAHub HT reports, JFK passenger volumes, and NYC event permits were merged using either spatial joins (for location-matching) or date-based joins (for time alignment). The merged master table included both raw clustering inputs (like spatial coordinates) and reference features (e.g., ht_reports) used post hoc for validating the significance of discovered clusters.

To enable map-based visualization, each point in the dataset retained its geographic coordinate pair and cluster assignment label. These were later used to generate folium-based interactive maps, with cluster boundaries drawn using convex hulls for interpretability.

Justification of Data Design

The design of the clustering dataset was driven by the goal of discovering spatial patterns of risk without relying on labeled outcomes. By using raw coordinates instead of fixed spatial grids, we enabled unsupervised models like HDBSCAN to detect irregularly shaped, organically occurring hotspots, while KMeans provided a complementary view of symmetrical cluster distributions.

Behavioral and contextual features such as event counts, passenger volume, and crime density were chosen based on their relevance in human trafficking risk literature, which often highlights increased vulnerability during periods of high mobility and dense gatherings. These features were not directly used in clustering inputs but were attached for post-clustering interpretation and labeling.

Finally, the simplified structure and use of open-source geospatial libraries made the pipeline adaptable to other urban regions. The absence of grid constraints and reliance

on natural coordinates ensures that this methodology can scale to different cities, provided comparable datasets are available.

Project Methods (Hashim and Derya)

Supervised Learning Models (Hashim)

The modeling pipeline for this project was developed to address the challenge of predicting HT risk using indirect indicators across space and time. Due to the extreme sparsity of confirmed trafficking cases and the high potential for misclassification in public records, a multi-stage modeling strategy was employed. This included tabular machine learning models for benchmarking, followed by sequence-based deep learning for temporal risk prediction.

The modeling process began with baseline classifiers using LightGBM and XGBoost. These gradient boosting models were selected due to their efficiency on structured data, strong performance with imbalance classes when appropriately configured, and their ability to quickly surface feature importance for early validation. Each model was trained on daily-level features extracted from the unified dataset, where each row represented a unique date and location combination. The input features included a 7-day rolling crime count, night crime ratio, event density, scaled monthly passenger volumes to JFK Airport, and various temporal indicators such as day of week, holiday status, month, and season. Spatial features like boroughs were one-hot encoded. These models were used to establish baseline recall and AUC values and to assess the presence of a learnable signal in the data.

To mitigate the severe class imbalance, class weighting was applied during training. For LightGBM, ‘class_weight = balanced’ was used, while XGBoost utilized ‘scale_pos_weight’ proportional to the imbalance ratio. However, both models performed poorly in identifying the minority class(i.e., positive HT cases). The weighted LightGBM model achieved a recall of just 1.6% and an AUC of 0.5069. XGBoost performed slightly better, reaching a recall of 2.1% and an AUC of 0.5110. These results demonstrated the limitations of static, tabular models in capturing temporal escalations or behavioral patterns associated with trafficking.

Given these limitations, the focus shifted to temporal modeling using a Long Short-Term Memory (LSTM) neural network. LSTM models are well-suited for learning from sequences, allowing the system to identify risk patterns that unfold across consecutive days. The dataset was restructured to form 7-day rolling windows for each location. For each sequence, the label indicated whether a verified HT case occurred the following

day at the same location. This structure enabled the model to learn how crime, events, and mobility indicators evolve over time leading up to high-risk conditions.

All features were normalized using MinMaxScaler, and the sequences were created by location ID, maintaining their chronological order. The final LSTM model architecture consisted of a single LSTM layer with 64 units, followed by a dropout for regularization, a dense ReLU layer, and a sigmoid output layer to produce a risk probability between 0 and 1. The model was trained over five epochs with a batch size of 512, and a validation split of 20% was used. To address the highly imbalanced class distribution, a class weighting ratio or approximately 1:4609 (negative to positive) was applied during training.

Importantly, the LSTM model produced a continuous risk score for each location-day pair rather than a binary label. This allowed for flexible thresholding during evaluation. At the default threshold of 0.5, the model achieved a 51% recall, which significantly outperformed the baseline models. When the threshold was lowered to 0.2, the recall increased to 75%, indicating strong sensitivity to high-risk conditions. Although this came at the cost of precision, the model's primary value lies in its ability to prioritize surveillance or outreach efforts across thousands of locations, rather than deliver definitive classification. By treating the output as a risk probability rather than a strict yes/no decision, the system remains adaptable to real-world operational needs.

Unsupervised Learning Models (Derya)

To uncover hidden spatial patterns in human trafficking–related crime data, we used unsupervised learning methods, KMeans and HDBSCAN. These models helped us discover natural groupings in the data without relying on pre-labeled outcomes, making them especially useful for a problem like human trafficking where many cases go undetected or unreported. We began by applying KMeans clustering to the latitude and longitude coordinates of crime-related incidents. The goal was to identify the top five geographic hotspots across NYC not predefined by boroughs, but purely by spatial density of incidents. Each cluster was then labeled with the most common day of week, month, and season where crimes were concentrated. For instance, Friday emerged as the most common day, July as the peak month, and summer as the riskiest season although one cluster spiked specifically in October. These patterns provide actionable insights that can support targeted and timely intervention efforts by law enforcement, shifting strategy from reactive policing to proactive deployment based on both where and when risk is highest.

Our initial choice of five clusters was informed by NYC's five boroughs, though the resulting clusters did not align with those administrative regions. Instead, they reflected the most intense areas of human trafficking–related activity. To validate the quality of these clusters, we used the Silhouette Score, which measures how well-separated and cohesive the clusters are. All five KMeans clusters fell into the “moderate” range, indicating the groupings were reasonably distinct but not highly separated.

To further refine our analysis, we used HDBSCAN, a density-based clustering algorithm that identifies organically formed high-risk zones while filtering out low-density noise. Unlike KMeans, which requires a predefined number of clusters and assigns every data point to one, HDBSCAN identifies clusters only in areas where activity is dense and meaningful. Points that don’t meet the density threshold are treated as noise, ensuring that clusters reflect true hotspots. We used a minimum threshold of 15 crimes per cluster but noted that HDBSCAN’s flexibility allows for adjusting this value to explore additional patterns. Clusters were then re-ranked by total human trafficking–related crimes, providing a clearer view of priority areas. We also evaluated HDBSCAN using the same Silhouette Score metric. The top five clusters all achieved strong scores ranging from 0.91 to nearly 1.0, indicating tight, well-separated groups with minimal noise. These results demonstrated that HDBSCAN yielded cleaner, more distinct clusters than KMeans, making it a more effective tool for spatial hotspot detection in this context.

Ultimately, both clustering methods provided valuable insights. KMeans helped surface consistent patterns across hotspots, while HDBSCAN enabled more granular, noise-filtered detection of high-density crime zones. Together, they offer a robust foundation for identifying both spatial and temporal hotspots, supporting more focused intervention, smarter surveillance, and improved resource allocation in human trafficking prevention.

Results (Hashim and Derya)

Supervised Learning Results: (Hashim)

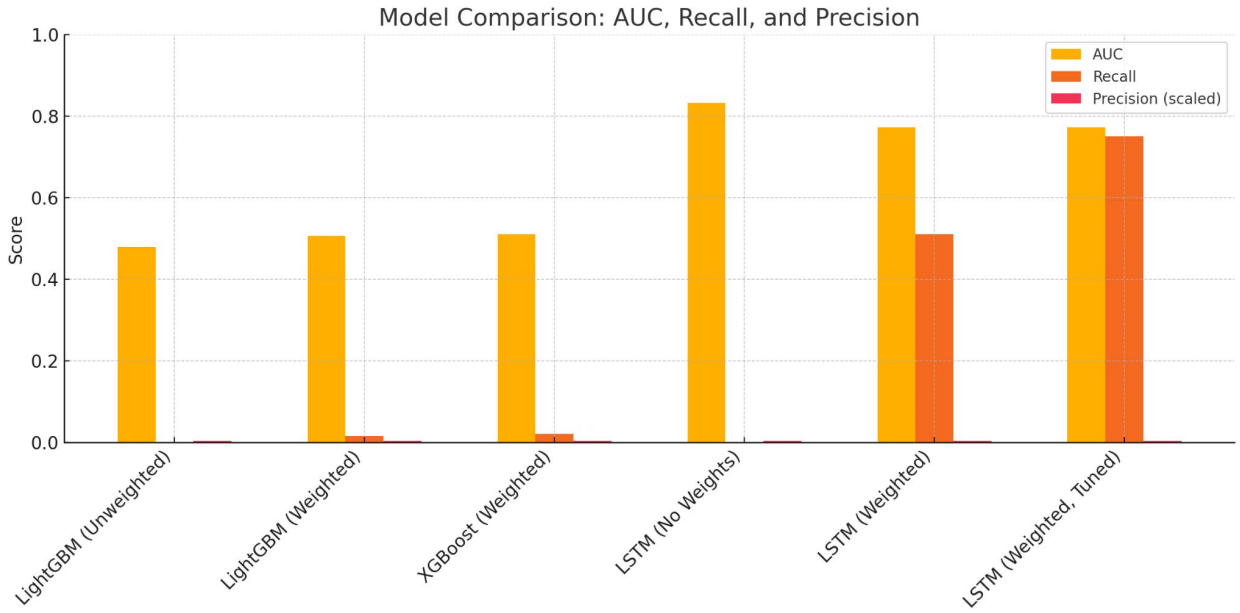


Figure: Comparison of Model Performance Across AUC, Recall, and Precision

This chart summarizes the performance of all models tested in the project across three key metrics: AUC, recall, and precision. The LSTM model with class weighting and threshold tuning (“LSTM Weighted, Tuned”) achieved the highest recall (75%) and a strong AUC (0.77), demonstrating its ability to detect rare human trafficking signals better than baseline models. While precision remained low across all models - a common outcome in rare-event classification - the LSTM’s ability to generate continuous risk scores enabled its use as a screening tool rather than a binary classifier.

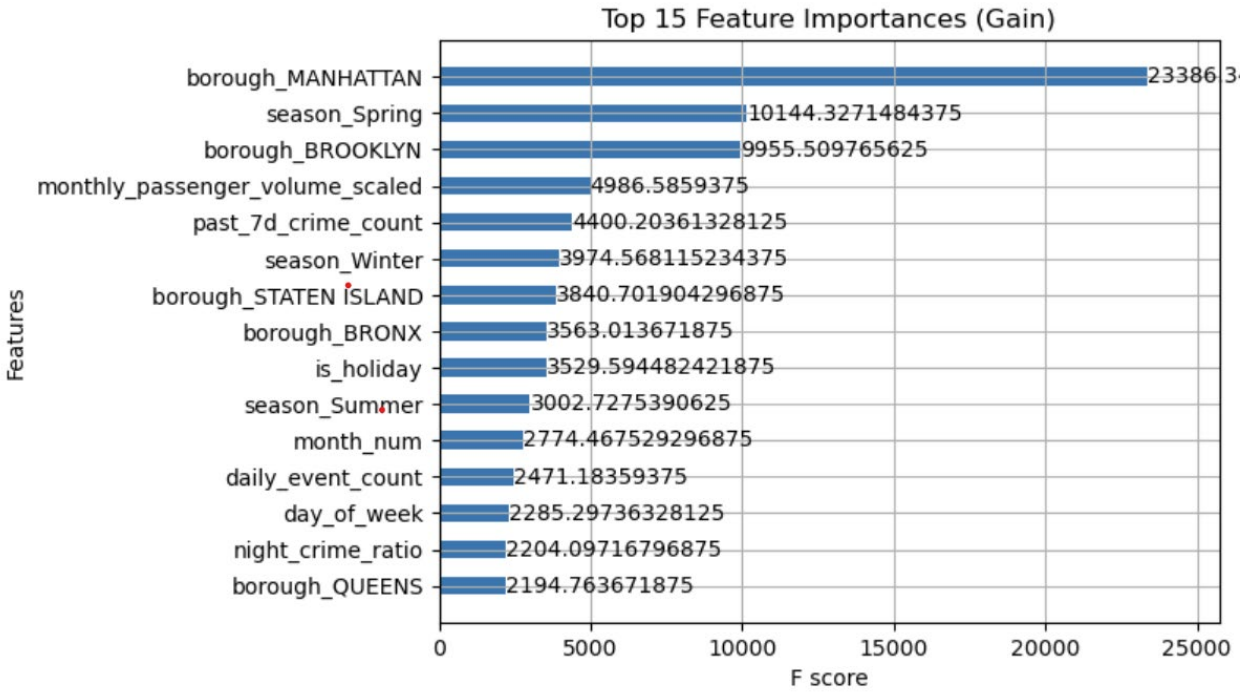
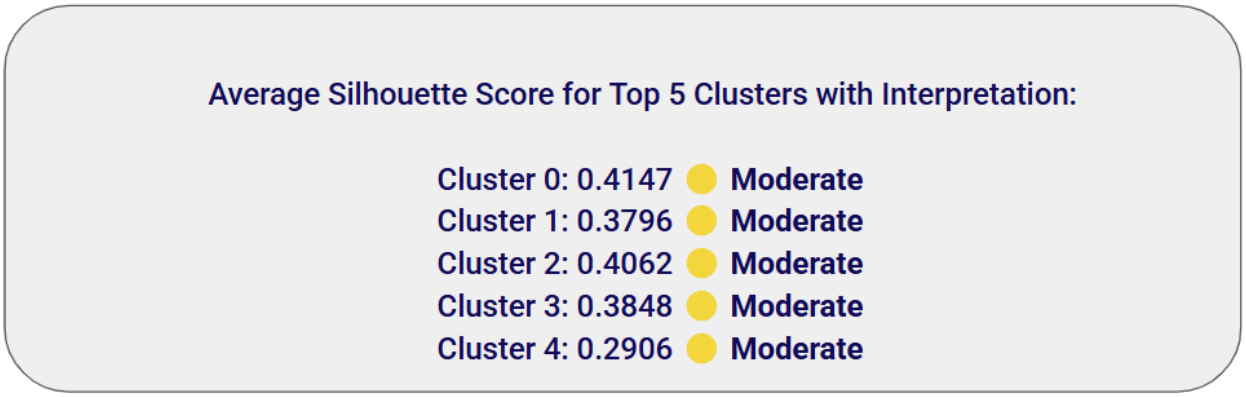


Figure: Top 15 Feature importance from XGBoost Model (Gain-Based)

The XGBoost baseline model, while limited in predictive power, provided useful early insights into the relative contribution of different features. The most influential variables included geographic identifiers such as borough: Manhattan, and borough: Brooklyn, followed by temporal indicators like seasons. Other notable features were the monthly passenger volumes, and past 7-day crime counts, which reflect the role of mobility and recent crime trends in potential trafficking risk.

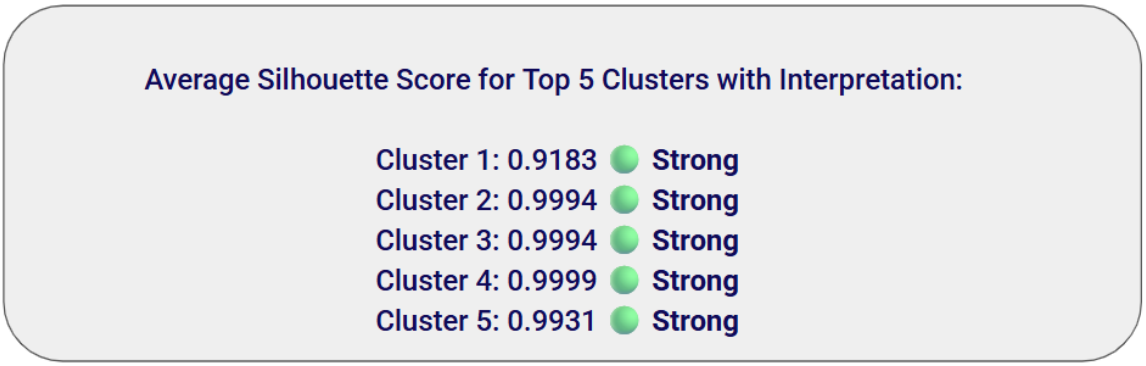
Unsupervised Learning Results: (Derya)



Silhouette Score Range	Interpretation
>0.5	**Strong** - clusters are well separated and dense
0.25 - 0.5	**Moderate** - clusters are somewhat well defined
0.0 - 0.25	**Weak** - possible overlap or uneven density
<0.0	**Poor** - Likely misclassified points

Figure: Clustering Quality Metric for KMeans using Silhouette Scores

The above figure represents the silhouette scores for the KMeans clustering model. Each of the five clusters displays a silhouette score ranging between 0.29 and 0.41, placing them in the moderate range. This indicates that the clusters identified by KMeans are reasonably well-defined but not strongly separated. These results suggest that KMeans was effective in capturing general spatial groupings of human trafficking-related activity; however, the moderate scores also highlight some limitations in cluster cohesion and separation. These scores also motivated us to explore another clustering method, HDBSCAN, to evaluate whether a density-based approach could produce stronger, more naturally separated clusters.



Silhouette Score Range	Interpretation
>0.5	**Strong** - clusters are well separated and dense
0.25 - 0.5	**Moderate** - clusters are somewhat well defined
0.0 - 0.25	**Weak** - possible overlap or uneven density
<0.0	**Poor** - Likely misclassified points

Figure: Clustering Quality Metric for HDBSCAN using Silhouette Scores

This figure presents the silhouette scores for the HDBSCAN clustering model, showcasing significantly stronger performance compared to KMeans. Each of the top five clusters achieved a silhouette score above 0.91, placing them firmly in the Strong category. This indicates that the clusters are well separated, internally cohesive, and effectively capture dense patterns of human trafficking–related activity. Unlike KMeans, which requires a fixed number of clusters and includes all data points (even low-density noise), HDBSCAN identifies natural groupings and filters out sparse or ambiguous regions. These results validate HDBSCAN as a more precise tool for hotspot detection, allowing us to highlight meaningful clusters while reducing noise and improving interpretability.

Interpretation of Results (Hashim and Derya)

Interpretation of Supervised Learning Results: (Hashim)
The LSTM model, particularly the version trained with class weighing and evaluated at a tuned threshold of 0.2, demonstrated significant capability in detecting potential human trafficking risk patterns. While the model's recall reached 75% and its AUC remained strong at 0.773, the precision remained extremely low (~0.00083). This outcome is not unexpected given the extreme imbalance and ambiguity inherent in human trafficking data, where confirmed cases are not only rare, but often delayed or hidden due to victim non-disclosure, misclassification, or systematic gaps in reporting.

Rather than serving as a binary decision-maker, the LSTM model functions as a risk scoring engine, producing a continuous probability of between 0 and 1 for each location-day pair. This allows the model to be applied flexibly - for instance, to rank locations by risk or filter the top 5% of zones needing further investigation. The low precision is therefore not a flaw in predictive logic, but a reflection of its role: it is intentionally tuned to maximize sensitivity, ensuring that high-risk areas are not missed. In practical terms, this positions the model as a screening tool, enabling human analysts, city agencies, or NGOs to prioritize review, rather than triggering direct intervention based on output alone.

To further address the low precision and avoid operational fatigue from false positives, the model's predictions can be post-processed by cross referencing with contextual signals (e.g., known hotspots, ongoing events, socioeconomic indicators, land-use type in the locations), hence our multi-layer approach to the problem. Additionally, setting tiered thresholds (e.g., only acting on top 5% risk scores) is another implication in terms of application.

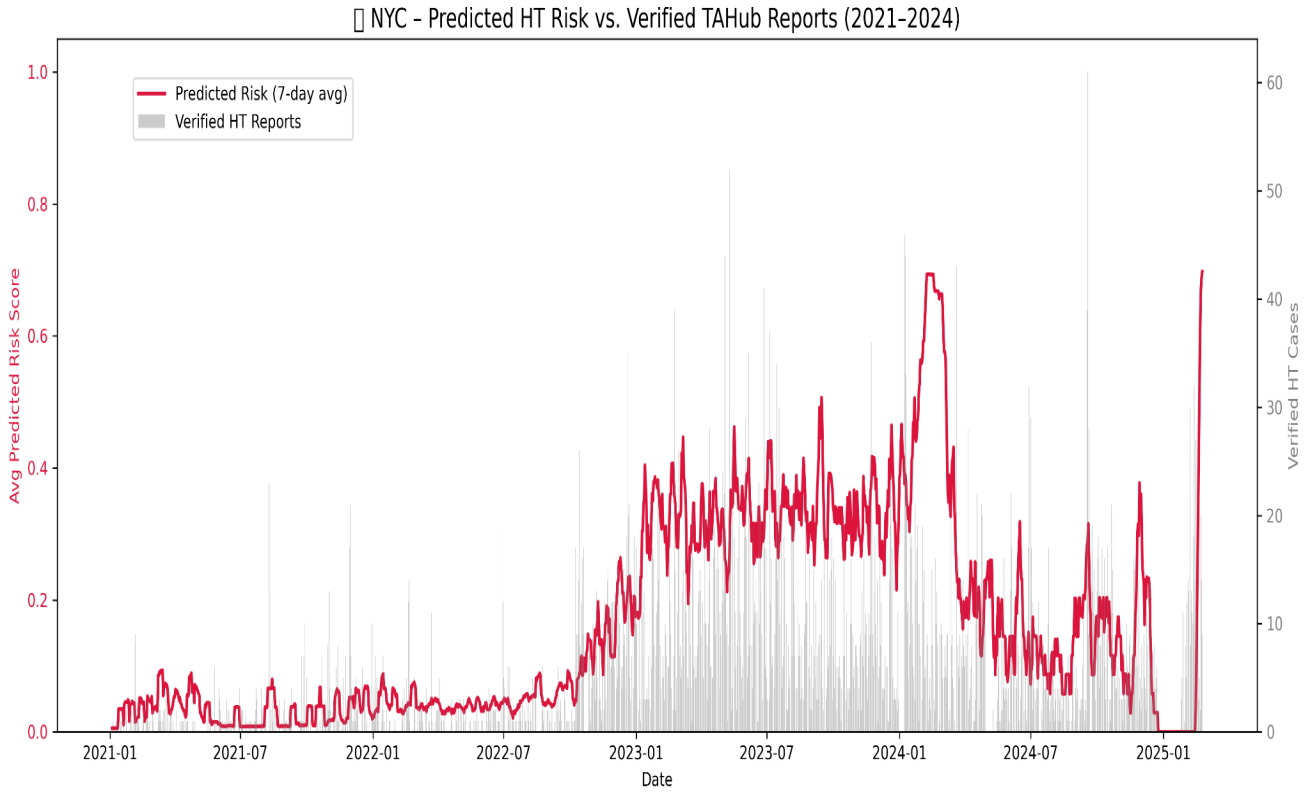


Figure: Predicted Risk vs Verified Human Trafficking Reports (2021-2024)

This time series plot compares the aggregated predicted daily human trafficking risk with the actual case reports from the TAHub Metro dataset. The alignment between predicted risk surges and real-world case spikes - particularly visible during mid-to-late 2023 and early 2024 - validates the model's temporal learning. Despite the sparse and noisy nature of labels, the LSTM was able to learn escalation patterns that precede actual trafficking case confirmations. This is a strong indication that the model captures real underlying temporal dynamics and is not simply overfitting noise.

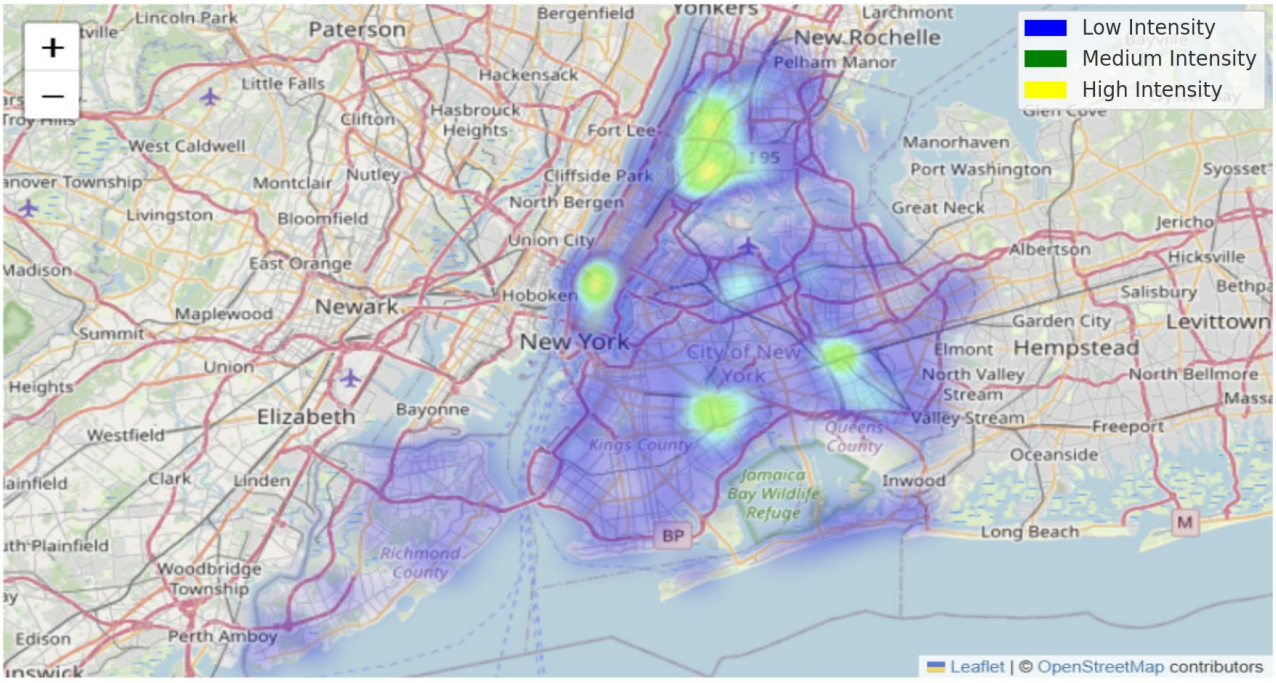


Figure: Daily Predicted Risk Heatmap from May 14, 2024

The spatial heatmap shows the model's predicted HT risk for a specific day, with color intensities representing low, medium, and high-risk zones across New York City. Notably, the model consistently flagged areas like Midtown Manhattan, downtown Brooklyn, and the vicinity of JFK airport as high-risk - aligning with known trafficking corridors and urban mobility hotspots. This spatial concentration further supports the model's validity and enhances its practical utility for citywide monitoring, especially in large jurisdictions where deploying field teams is resource intensive.

Interpretation of Unsupervised Learning Results (Derya)

To complement our predictive models, we applied two unsupervised learning techniques: KMeans and HDBSCAN to uncover spatial and temporal patterns of human trafficking (HT)-related crime across New York City. These models were not designed to predict a binary outcome, but instead to reveal natural groupings in the data that could guide proactive intervention, resource allocation, and deeper contextual understanding.

KMeans Clustering: Broad Risk Zones and Temporal Trends

The KMeans clustering model grouped geographic coordinates into five clusters, representing the top five areas with the highest concentration of HT-related crimes. These clusters were not aligned with boroughs but instead reflected true incident density. For instance, Hotspot #1 spanned both Manhattan and Brooklyn, demonstrating that high-risk zones often cross administrative boundaries.

Each cluster was annotated with its most common day of the week, peak month, and dominant season. Four out of five clusters showed a clear spike in summer, with Friday as the most frequent crime day. However, Cluster #4 stood out, showing an unusual peak in October (Fall), a pattern that may reflect localized risk factors or event-based variation. This points to potential heterogeneity within the cluster, where several neighborhoods with different risk timelines are grouped spatially, and a single region with strong fall activity may be driving the trend.

The choice of five clusters was exploratory based initially on the five boroughs and served as a baseline. Although not optimized, this starting point helped surface consistent temporal signals across the city. Notably, KMeans treats all data points equally, forcing all incidents into a cluster regardless of density. This rigidity, while useful for general mapping, may dilute the accuracy of true hotspot detection. To assess performance, we calculated Silhouette Scores for each cluster, which ranged from 0.29 to 0.41, falling in the moderate range. This indicates that the clusters were reasonably cohesive but lacked strong separation, prompting the need for a more flexible clustering method.

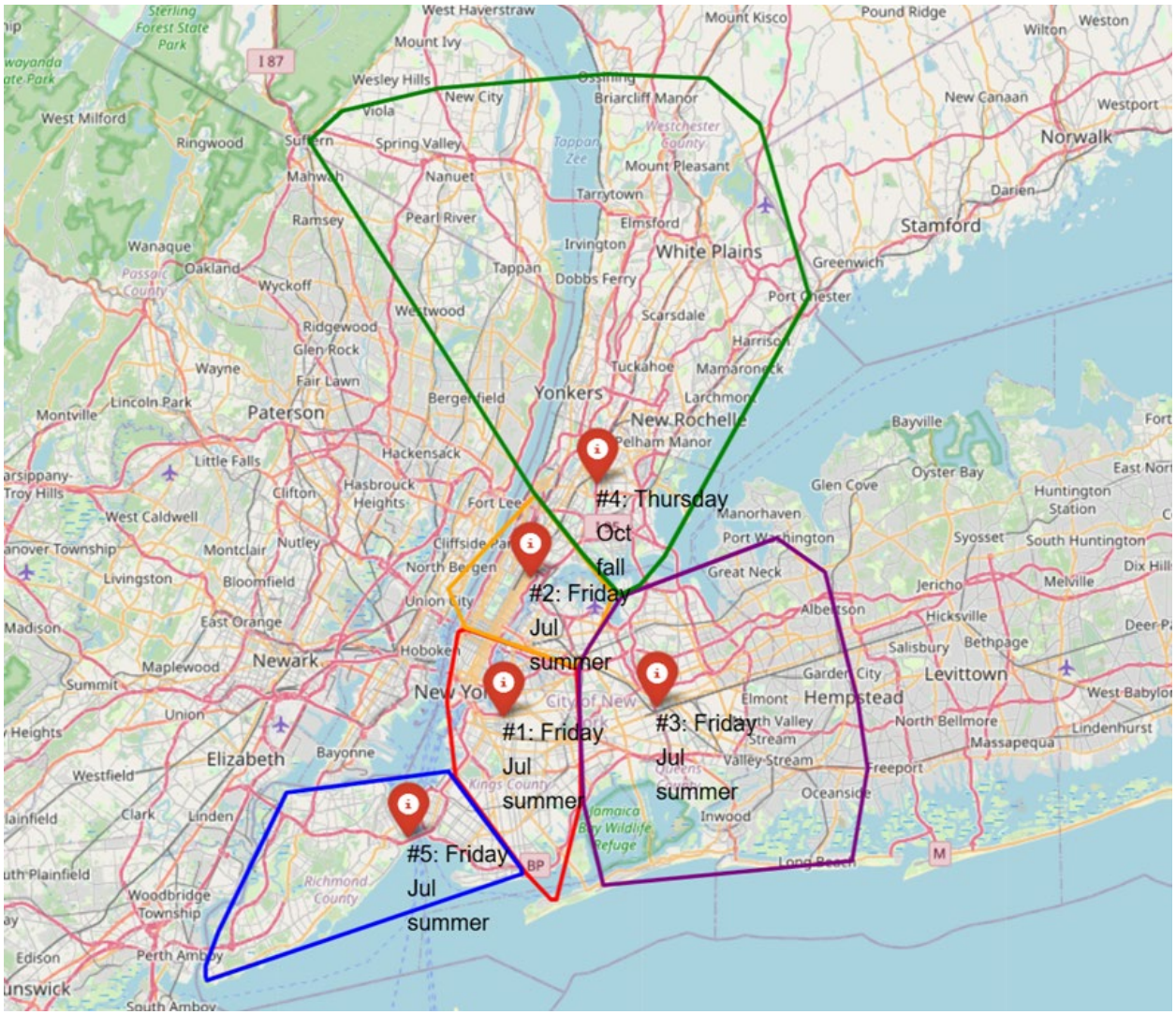


Figure: Top 5 High-Risk Human Trafficking Hotspots Identified by KMeans Clustering (2021–2024)

HDBSCAN Clustering: Precision Hotspot Detection

To improve spatial precision and filter out noise, we applied HDBSCAN, a density-based clustering method that does not require predefining the number of clusters. Instead, it forms clusters only in areas with dense activity and excludes scattered or low-volume incidents such as noise. The key advantages of HDBSCAN lie in its ability to identify natural groupings based purely on the density of incidents, without requiring a preset number of clusters. Unlike KMeans, it does not force all data points into clusters; sparse or scattered regions are excluded entirely as noise. This allows HDBSCAN to detect many smaller, more organically shaped clusters, providing a more accurate and flexible reflection of real-world crime patterns than large, rigidly defined zones. In our implementation, only clusters with 15 or more HT-related crimes were visualized to ensure statistical relevance. Regions that appeared empty on the map had either too few incidents or failed to meet HDBSCAN's density threshold. Clusters were re-ranked based on total HT-related crimes, and each was labeled with the most common day, month, season, and total crime count.

For example, Cluster #1 in Elmhurst, Queens, had over 800 crimes, with peaks on Fridays in June offering strong insight into both spatial and temporal targeting. Other clusters emerged in the South Bronx, downtown Manhattan, and East Brooklyn, capturing localized patterns that broader clustering would have obscured. HDBSCAN's effectiveness was confirmed by its Silhouette Scores, which ranged from 0.91 to 0.99, placing them in a strong range. These high scores reflected tight cohesion within clusters and clear separation between them making HDBSCAN highly effective for identifying actionable, high-confidence hotspots.

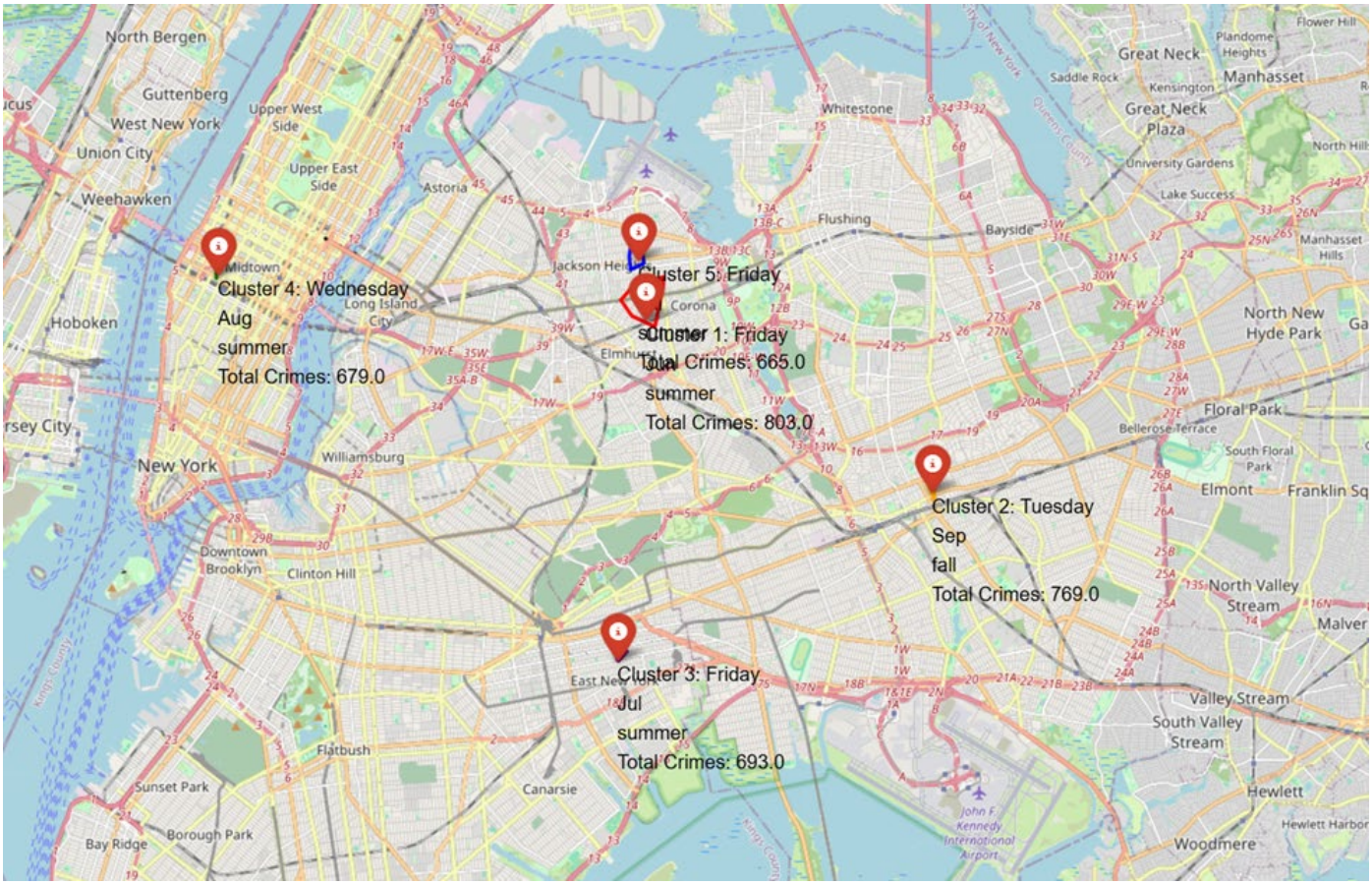


Figure: Spatial Hotspots Identified by HDBSCAN Clustering Method

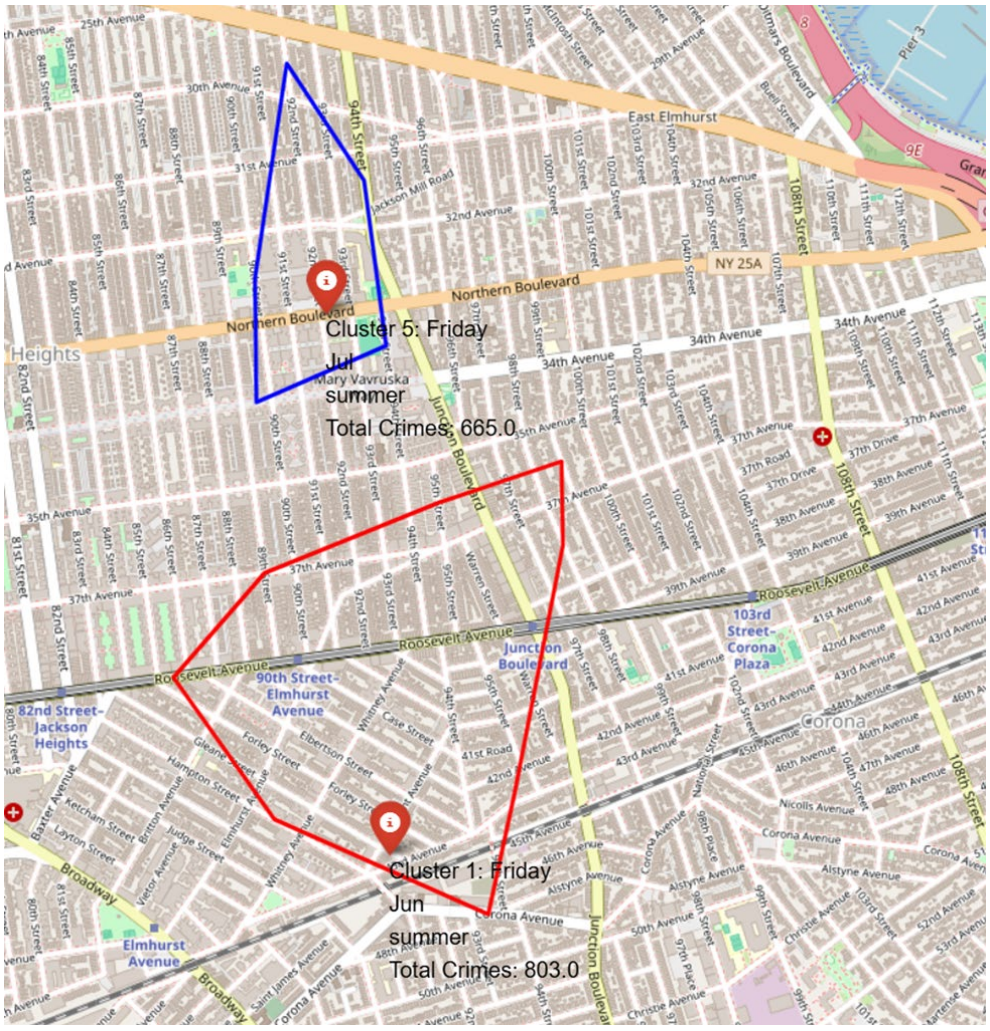


Figure: Close-up view of high-density risk zones within a compact area

Combined Value and Strategic Application

Taken together, the unsupervised models enabled a risk-aware and spatially contextualized view of urban geography, helping to surface patterns that are often hidden in raw data. By applying KMeans, we gained a broad understanding of where human trafficking-related activity clusters over time identifying consistent seasonal and weekly trends that cut across borough boundaries. This high-level perspective allowed us to detect temporal commonalities among different hotspots, such as the tendency for crimes to peak on Fridays in the summer months, offering a framework for macro-level planning and trend anticipation.

In contrast, HDBSCAN provided sharper granularity, highlighting localized zones with dense and repeated activity, while simultaneously filtering out noise and irregular

patterns that may not warrant intervention. This enabled us to pinpoint micro-level hotspots areas that might otherwise be overlooked using broader clustering methods and offered a more realistic representation of true risk concentrations within the city. These smaller, more precise clusters are especially valuable when resources are limited and need to be deployed with surgical precision.

Together, the two models work in tandem. KMeans tells us when and generally where trafficking-related activity tends to concentrate, while HDBSCAN tells us exactly where and how intensely. This multi-layered insight serves as a powerful complement to our supervised learning pipeline, not only enhancing our ability to predict risk, but also explaining the context behind it and informing data-driven prioritization. These spatial intelligence layers support a variety of real-world applications from patrol route optimization and resource allocation to public safety outreach, nonprofit engagement, and community-specific monitoring. By combining predictive signals with interpretable geographic clusters, our approach strengthens the operational bridge between machine learning insights and actionable field strategies.

Contextual Validation Results: (Derya)

After identifying spatial and temporal hotspots using our LSTM model, we conducted a contextual validation phase to better understand the underlying conditions driving elevated human trafficking (HT) risk. This step helped translate model predictions into actionable insight by examining the socioeconomic and environmental realities present in high-risk areas.

Globally, an estimated 50 million people are living in modern slavery, according to the Walk Free Foundation’s 2021 report. This includes 27.6 million in forced labor and 22 million in forced marriage representing nearly 1 in every 150 people worldwide. Many trafficking cases go unreported or misclassified, as victims often do not disclose their exploitation due to fear, coercion, or lack of support. Traffickers commonly prey on individuals who face poverty, limited education, and unstable living conditions, making socioeconomic context a critical factor in understanding risk.

In our local analysis of LSTM-flagged hotspots across NYC, we integrated data from public sources to evaluate poverty and unemployment rates, zoning regulations, and nearby business activity. To verify economic conditions such as unemployment, we referenced authoritative sources including the U.S. Census Bureau (2023), Table B23025 from the ACS 1-Year Estimates, ensuring that our validation was grounded in official labor statistics. These findings reinforced the idea that trafficking risk is not solely a function of isolated incidents but is shaped by systemic environmental conditions that can be observed and measured. Across many of these high-risk areas, we consistently found patterns such as elevated poverty and unemployment, the presence of high-risk

businesses (e.g., massage parlors and motels), and mixed-use zoning that allows such establishments to operate near residential areas. These environments often lacked visible community oversight and were home to populations especially vulnerable to exploitation. One notable example occurred at 123 Main St., Brooklyn, where our model assigned a risk score of 0.81. Field review showed three massage parlors within 0.3 miles of the location. Socioeconomic data revealed a 42% poverty rate and a 20% youth population, further emphasizing the community's vulnerability. These indicators reflect conditions that traffickers often exploit unstable employment, economic insecurity, and limited adult supervision.

By anchoring predictive analytics in real-world context, our validation process made the data more interpretable, human-centered, and operationally relevant. It highlighted not just where the risks are, but why those risks exist providing valuable guidance for law enforcement, social service agencies, and policymakers. This step bridged the gap between algorithmic prediction and meaningful intervention, reinforcing the importance of embedding contextual intelligence in data-driven anti-trafficking efforts.

Additional Work (Hashim & Derya)

Illicit Advertisement Scraping for Risk Signal Enrichment

In addition to the core spatial-temporal modeling efforts, and the unsupervised learning for validation, we undertook supplementary work to enhance the model's feature set by extracting real-time indicators of trafficking activity (e.g., sexual exploitation) from the web. Specifically, we built a custom web scraper using Python's BeautifulSoup library to collect illicit service advertisements from publicly accessible website 'escortalligator'. The objective was to engineer a daily advertisement frequency feature that could potentially serve as a proxy signal to human trafficking demand (e.g., sex work), and potentially correlate with crime trends and event activity.

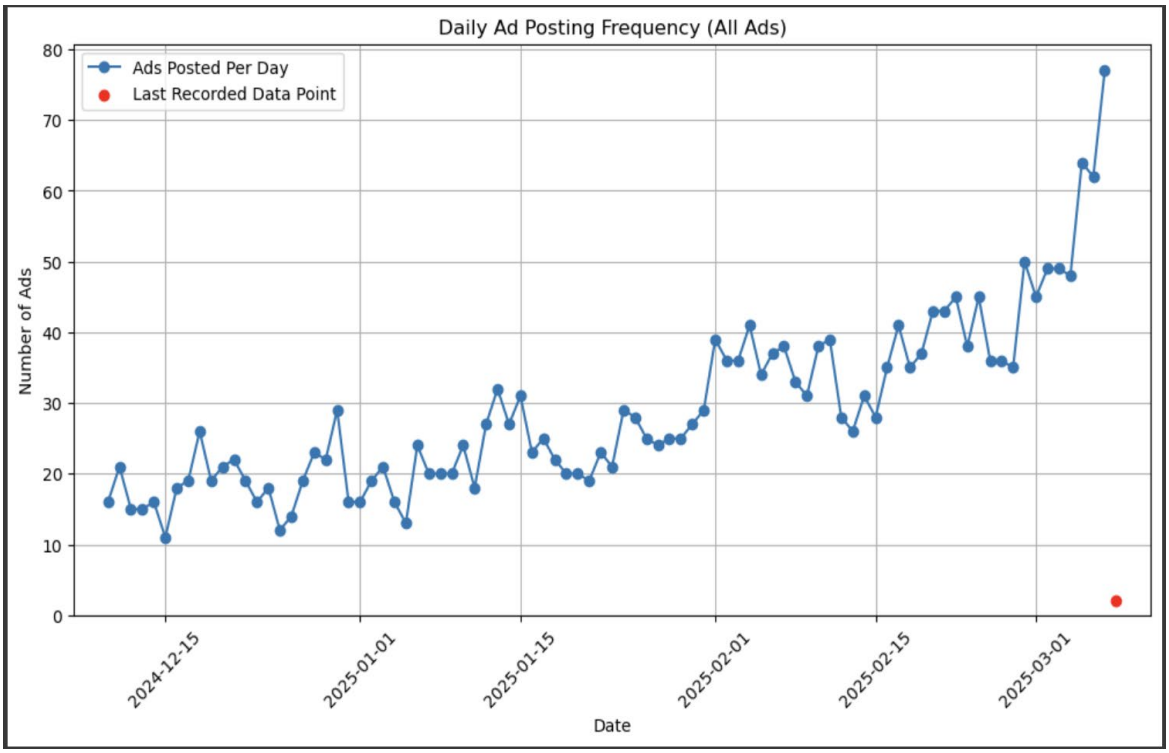


Figure: Daily frequency of illicit advertisements posted on 'escortalligator' from December 10, 2024, to March 8, 2025.

Over a 3-month period, we collected 2,551 advertisements posted in locations across Manhattan, Brooklyn, and Queens. These ads were parsed and time-stamped, allowing us to construct a daily ad frequency time series. As shown in the figure, we observed a clear spike in ad activity beginning in February and peaking in March - a trend that paralleled seasonal upticks in crime and mobility indicators in our main dataset.

This work was designed with the intention of integrating the daily ad volume as a temporal feature into the supervised modeling pipeline. However, the effort was ultimately paused due to insufficient data volume and a key limitation: the target website automatically deletes older ads, making long-term data collection unsustainable without continuous scraping.

Although this feature was not deployed in the final model due to data limitations, the analysis demonstrated both initiative and technical feasibility. The patterns revealed through this effort provide strong justification for future investment in building sustainable scraping infrastructure or collaborating with data vendors who offer archived, structured illicit ad datasets. The work showed clear alignment between ad frequency surges and other trafficking signals, suggesting this source serve as a valuable real-time indicator in future iterations of the model.

Sentiment Analysis of High-Risk Businesses for Contextual Validation

To enhance the real-world interpretability of the predicted high-risk zones, we conducted additional modeling work focused on contextual validation using location-specific business data. This involved extracting user reviews for potentially high-risk establishments - such as motels, massage parlors, strip clubs, late-night service businesses - located within a 30-mile radius of JFK Airport. The dataset was assembled by programmatically querying the Google Places API (Google API Services), filtering for business categories historically linked to trafficking vulnerability.

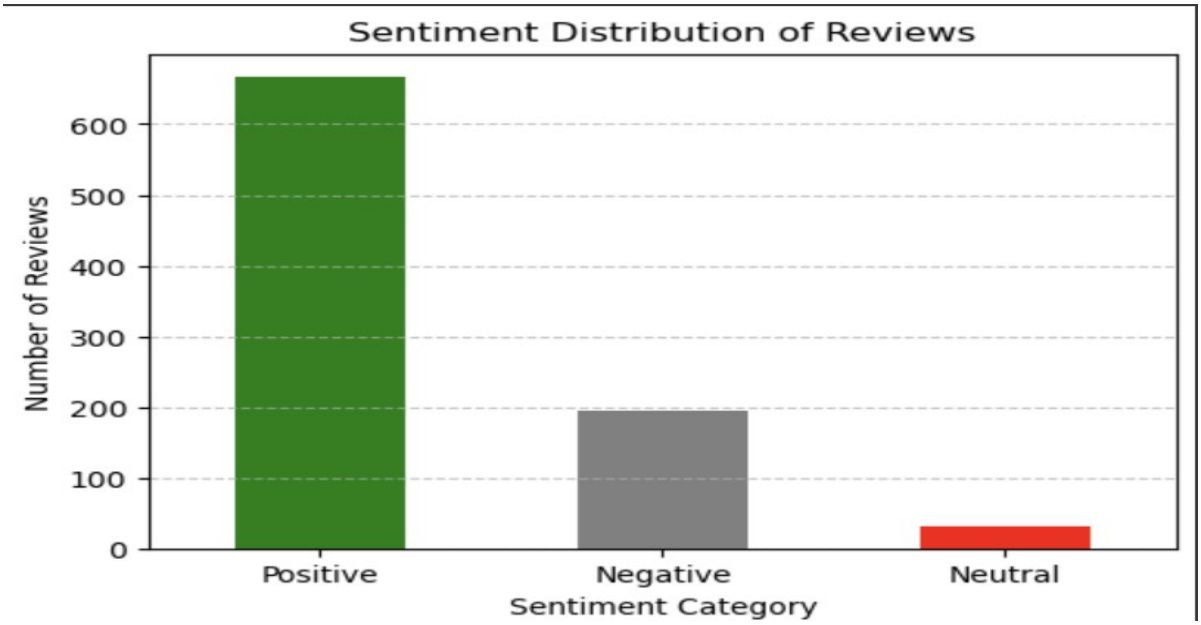


Figure: Sentiment Distribution of High-Risk businesses

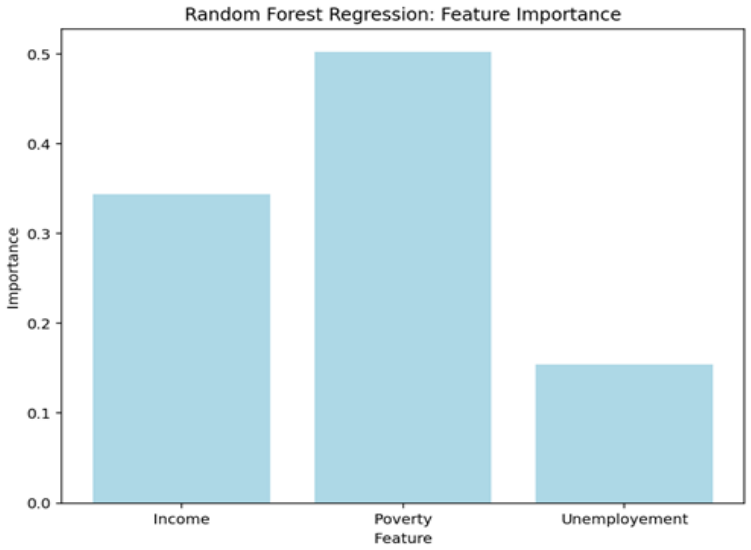
In total, we collected 180 high-risk business listings and scraped 893 reviews. These reviews were then processed through sentiment classification pipeline, categorizing them as positive, negative, or neutral. As visualized, many reviews skewed positive, but a substantial number exhibited negative sentiment - which may be indicative of illicit, unsafe, or suspicious activity in certain zones.

Although this sentiment data was not included directly in model training, it served as a valuable qualitative check on the model's outputs. In several instances, high predicted LSTM risk zones coincided with areas containing negatively reviewed establishments, strengthening confidence in the model's relevance to real-world trafficking risk. Furthermore, this additional data stream lays the foundation for building human-in-the-loop systems, where flagged locations can be validated through both structured outputs and external context.

This work demonstrates how sentiment-driven insights can be integrated into spatial risk modeling to support decision-making, field validation, and stakeholder communication. With further scaling and refinement - including natural language processing (NLP) to detect red flag language in reviews - this approach could evolve into a powerful auxiliary signal for prioritizing site-level intervention and resource deployment.

Repurposing Borough-Level Socio-Economic Data for Spatial Insight

In this project, we initially explored the use of borough-level socio-economic data, specifically income, poverty, and unemployment rates across New York City from 2021 to 2024, with the aim of incorporating these variables into our predictive modeling. The data was sourced from the U.S. Bureau of Economic Analysis via the FRED (Federal Reserve Economic Data) platform. Since many of the original datasets were available only on an annual basis, we converted income and poverty figures into monthly estimates to better align with our daily crime and trafficking-related activity data. In cases where 2024 data was not yet available, we duplicated 2023 estimates to maintain temporal coverage. Despite these preprocessing efforts, the socio-economic data proved too coarse and static to support daily-level predictions. When tested using models such as Random Forest, Multiple Linear Regression, and ARIMAX, these variables demonstrated minimal predictive value. Their limited granularity and temporal resolution made them ill-suited for detecting short-term trafficking risk patterns. As a result, we chose to exclude these features from the final supervised models. Instead, socio-economic indicators were repurposed as a contextual analysis layer used to interpret clustering results and understand borough-level vulnerabilities behind the spatial distribution of risk. This shift in strategy allowed us to preserve the relevance of socio-economic insights without compromising the precision of our modeling pipeline.




```
          Feature  Importance
0          Income    0.343683
1         Poverty    0.502359
2  Unemployement     0.153957
R-squared on Test Data: 0.9967310828327505
Mean Squared Error on Test Data: 73411.1261027232
```

Figure: Feature Importance – Bronx Socio-Economic Factors (Random Forest)

Recommendations for Future Work (Derya)

To further advance the scope and impact of this project, several strategic directions are recommended that would enhance the quality, interpretability, and operational usefulness of trafficking risk models. One of the most compelling opportunities lies in the continued monitoring and integration of illicit online advertisement data. Our preliminary scraping of *escortalligator* revealed this source's potential to act as a near real-time signal of trafficking-related activity, particularly in contexts where traditional crime data may be delayed, misclassified, or underreported. However, the limitations of short-term data retention and inconsistent scraping prevented this signal from being incorporated into our final model. Future iterations should focus on building a sustainable scraping pipeline that continuously archives and timestamps ad data or partnering with vendors that offer structured historical datasets. Doing so would enable long-term pattern recognition and strengthen the temporal dimension of risk detection by aligning spikes in ad activity with shifts in crime, mobility, and clustering outputs.

Another essential enhancement involves acquiring more granular mobility data, particularly at major transit nodes like JFK Airport. Although our model used monthly passenger volumes to estimate seasonal travel patterns, the lack of daily or terminal-specific data restricted our ability to detect finer-grained correlations between passenger flow and trafficking risk. Future efforts should prioritize obtaining daily flight counts, international arrivals and departures, and terminal-level breakdowns. Additionally, integrating taxi and rideshare route data would support the identification of high-frequency travel corridors often used in trafficking operations. Such data could uncover spatial linkages across boroughs, further contextualizing hotspot clusters with behavioral insights on how individuals may be moved across the city.

Socio-economic indicators while excluded from predictive modeling due to low explanatory power remain vital for interpreting model outputs and understanding broader structural vulnerabilities. In this project, borough-level data on income, poverty, and unemployment were collected from 2021 to 2024, primarily sourced from the U.S. Bureau of Economic Analysis via FRED. Because many indicators were available only annually, they were converted to monthly estimates for alignment with daily crime data, resulting in artificially smoothed signals. Future work should seek to leverage neighborhood-level or quarterly data where available, offering more granular and timely insights into socio-economic stressors. Variables such as housing instability, school dropout rates, or domestic violence statistics may provide stronger contextual layers when interpreting why certain clusters emerge as high-risk. These features should not be treated purely as predictors, but as diagnostic tools to support policy recommendations and explain model behavior.

In addition to improving data inputs, future work should deepen the interpretability of spatial clustering results. While both unsupervised models we used revealed high-risk zones, the next step involves mapping these clusters to real-world geographic and administrative contexts. This could include linking cluster centroids to zip codes, neighborhood boundaries, or zoning classifications, as well as reverse-geocoding to identify common business types within each cluster. Enhanced interpretability would allow for targeted intervention strategies whether through law enforcement patrols, community outreach, or zoning policy adjustments.

Finally, deploying model outputs in practice requires the development of validation frameworks that incorporate human judgment. Real-time dashboards combining predictive scores with contextual signals (e.g., nearby high-risk businesses, user-generated reviews, or Google Street View) can support analysts in confirming flagged areas before triggering interventions. This human-in-the-loop design ensures accountability while maximizing the utility of predictive insights. Further collaboration with public sector agencies and nonprofit organizations would also help translate

insights into meaningful action whether in the form of service provision, public education, or coordinated response planning. By pursuing these enhancements, future work can shift the paradigm from reactive detection to proactive risk mitigation enabling more timely, informed, and impactful efforts to combat human trafficking.

Conclusion (Derya)

This project provided a powerful, data-driven lens into the hidden patterns of human trafficking risk in urban environments. By integrating diverse datasets across time and geography, we developed a robust modeling pipeline that uncovered risk signals from both structured indicators and contextual cues. Our approach combined supervised models like LSTM with unsupervised clustering methods such as KMeans and HDBSCAN, allowing us to predict, interpret, and visualize hotspots of trafficking-related activity across New York City. To validate our LSTM predictions, we considered unsupervised learning as an independent check. When we closely examined the LSTM-generated risk heatmaps alongside the spatial clusters produced by the unsupervised models, we found a strong degree of alignment particularly with weekday patterns and summer surges. This overlap was especially notable in high-density areas flagged by both approaches. The consistency between supervised and unsupervised output strengthens our confidence that the LSTM wasn't just reacting to noise, but was highlighting genuine, recurring risk zones.

Through this process, we gained hands-on experience working with real-world data at scale over 7.8 million observations while sharpening our understanding of how spatial and temporal dynamics influence the occurrence of rare, high-impact events. The LSTM model captured escalation patterns leading to verified trafficking incidents, while clustering models revealed consistent spatial-temporal trends across boroughs. KMeans provided a broad overview of regional risk, while HDBSCAN offered greater granularity by filtering out noise and identifying tightly grouped, organically emerging hotspots. These tools enabled more than just detection, they supported meaningful interpretation and prioritization.

Beyond modeling, the project deepened our appreciation for data storytelling and its critical role in translating complex technical findings into strategies that public-sector organizations can act on. We didn't stop at predictions, we validated our findings with environmental and socioeconomic factors, reviewed business-level sentiment data, and explored external signals through scraped online advertisements. Each step reinforced the importance of layering domain context over predictive outputs.

Ultimately, this experience challenged us to think beyond algorithms. It pushed us to design systems that are interpretable, adaptable, and relevant to the real-world challenges of trafficking detection. By combining predictive modeling with contextual validation, we laid the groundwork for more responsive, location-aware interventions and gained valuable insight into how data can be used to surface risk where it often goes unseen.

References

1. Farrell, A., Owens, C., & McDevitt, J. (2019). New laws but few cases: Understanding the challenges to the investigation and prosecution of human trafficking cases. *Crime, Law and Social Change*, 61(2), 139–168. https://doi.org/10.1007/s10611-013-9442-1

2. International Labour Organization (ILO). (2022). *Global Estimates of Modern Slavery: Forced Labour and Forced Marriage*. https://www.ilo.org/global/publications/books/WCMS_854733/lang--en/index.htm

3. Polaris Project. (2023). *2022 U.S. National Human Trafficking Hotline Statistics*. https://polarisproject.org/wp-content/uploads/2023/06/2022-NTL-Stats-Pages.pdf

4. Port Authority of New York and New Jersey. (2024). *Airport Traffic Report: 2023*. https://www.panynj.gov/airports/en/statistics-general-info.html
5. Esquivel, N., Nicolis, O., Peralta, B., & Mateu, J. (2020). Spatio-temporal prediction of Baltimore crime events using CLSTM neural networks. *IEEE Access*, 8, Article 9251302. https://doi.org/10.1109/ACCESS.2020.3036715
6. Reddi, T., Kusuma, C., & Parvin, S. (2024). *Mapping crime dynamics: Integrating textual, spatial, and temporal perspectives*. Proceedings of the 15th Annual IEEE Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON). IEEE. https://ieeexplore.ieee.org/document/10754762
7. Dimas, J., Nguyen, P., Hupert, N., & Jordan, C. (2022). Operations research and analytics to combat human trafficking: A systematic review. *Computers & Industrial Engineering, 172*, 108560. https://doi.org/10.1016/j.cie.2022.108560
8. Alvari, H., Shakarian, P., & Park, J. H. (2017). Semi-supervised learning for detecting human trafficking. *Security Informatics, 6*(1), 1–11. https://doi.org/10.1186/s13388-017-0034-1
9. Wang, Y., Hahn, J., & Tavabi, N. (2020). Sex trafficking detection with ordinal regression neural networks. *arXiv preprint*. https://arxiv.org/abs/2004.03889
10. Bermeo, J. L., Escobar, F., & Cuenca, E. (2023). Human trafficking in social networks: A review of machine learning techniques. *ResearchGate*. https://www.researchgate.net/publication/374480817_Human_Trafficking_in_Social_Networks_A_Review_of_Machine_Learning_Techniques
11. Walk Free. (2022, September 12). 50 million people worldwide in modern slavery. https://www.walkfree.org/news/2022/50-million-people-worldwide-in-modern-slavery/
12. U.S. Census Bureau. (2023). Employment status for the population 16 years and over (Table B23025), 2023: ACS 1-Year Estimates Detailed Tables. American Community Survey. https://data.census.gov/table/ACSDT1Y2023.B23025?q=B23025

Appendix

1. TAHub_NYC_Metro_Cases.csv

ï»¿"MDB ID #"	Incident Reportir	Trafficking Type	Industry Sector	Recruitment M	Transportatic	Coercion Meth	Trafficker Ge	Trafficker Ag	Victim Ger	Victim Age	URL Sourc	Info Sourc	Data Sour	Country C	Country	Address/Loca	Location T	Coordinat	Latitude	Longitude
64636880ddfbf	3/7/2021	Sexual Exploitatio	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	Unknown	NGO	ngo	US	United Sta	Manhattan, N	Transit	40.78343,	40.78343	-73.9663
600ba3a6ffee3	1/1/2021	Child Trafficking,F	Personal Sexual	Relationship M	Aeroplane	Forced,Minor,	Female,Male	25-year-old,	Female	four-year-(	https://em	Internet/U	tah	US	United Sta	New York City	Target	40.71427,	40.71427	-74.006
600bab012574	1/3/2021	Child Sexual Explo	Child Marriage,D	Financial Induc	Car	Forced,Minor,	Unknown		Female		https://ww	Internet/U	tah	US	United Sta	New York City	Unknown	40.71427,	40.71427	-74.006
600bab012574	1/3/2021	Child Sexual Explo	Child Marriage,D	Financial Induc	Car	Forced,Minor,	Unknown		Female		https://ww	Internet/U	tah	US	United Sta	Financial Dist	Unknown	40.70789,	40.70789	-74.0086
600bab012574	1/3/2021	Child Sexual Explo	Child Marriage,D	Financial Induc	Car	Forced,Minor,	Unknown		Female		https://ww	Internet/U	tah	US	United Sta	Manhattan, N	Unknown	40.78343,	40.78343	-73.9663
600c4a3b0f86(	1/5/2021	Child Sexual Explo	Personal Sexual	Violence Threa	Unknown	Minor,Restrict	Unknown		Male	9 years old	https://ww	Internet/U	tah	US	United Sta	Bronx, New Y	Arrest	40.82732,	40.82732	-73.9236
600c5fde325b(	1/10/2021	Child Labour,Sexu	Domestic Work/	Violence Threa	Car	Minor,Restrict	Unknown			14-year-ol	https://ww	Internet/U	tah	US	United Sta	Brooklyn, Kin	Unknown	40.6501,-7	40.6501	-73.9496
5fffaa517e11fc	1/12/2021	Child Trafficking,C	Illicit Activities,P	Social Media	Bus	Minor	Unknown				https://stn	Internet/U	tah	US	United Sta	New York City	Unknown	40.71427,	40.71427	-74.006
600cce19075a	1/16/2021	Child Trafficking,C	Domestic Work/	Financial Induc	Car	Minor,Psychol	Unknown			under 10,u	https://ww	Internet/U	tah	US	United Sta	New York City	Unknown	40.71427,	40.71427	-74.006
600f6bc5c0bda	1/24/2021	Child Trafficking,C	Personal Sexual	Servitude,Servit	Seaport	Minor,Restrict	Unknown				https://ww	Internet/U	tah	US	United Sta	Riverdale, Bro	Unknown	40.90056,	40.90056	-73.9064
600d0e6e3da5	1/17/2021		Personal Sexual	Violence Threa	Unknown	Forced,Minor	Unknown		Female		https://nyp	Internet/U	tah	US	United Sta	Astoria, Quee	Arrest	40.77205,	40.77205	-73.9301
600f41752f593	1/20/2021	Child Trafficking,C	Domestic Work/	Violence Threa	Aeroplane	Minor,Physica	Female	82-year-old,	Female,Mi	16,newbor	https://ms	Internet/U	tah	US	United Sta	Manhattan, N	Trafficker_	40.78343,	40.78343	-73.9663
600f6bc5c0bda	1/24/2021	Child Trafficking,C	Personal Sexual	Servitude,Servit	Unknown	Minor	Unknown				https://ww	Internet/U	tah	US	United Sta	Riverdale, Bro	Unknown	40.90056,	40.90056	-73.9064
6010c405d2f1(	1/25/2021	Child Sexual Explo	Manufacturing/F	Relationship M	Boat	Minor,Sexual /	Unknown	age of 31,73-	Female	six years o	https://ww	Internet/U	tah	US	United Sta	Bronx, New Y	Trafficker_	40.82732,	40.82732	-73.9236
60160995a84e	1/29/2021	Child Trafficking,C	Personal Sexual	Financial Induc	Aeroplane,Ca	Forced,Lie,Mi	Male		Female	minors,un	https://the	Internet/U	tah	US	United Sta	Manhattan, N	Unknown	40.78343,	40.78343	-73.9663
601601d77670	1/29/2021	Child Sexual Explo	Personal Sexual	Social Media	Unknown	Minor	Unknown	26-year-old,26		17 years o	https://ww	Internet/U	tah	US	United Sta	East New Yor	Unknown	40.66677,	40.66677	-73.8824
601caa37a24f9	2/3/2021	Child Sexual Explo	Personal Sexual	Violence Threa	Unknown	Forced,Minor,	Unknown				https://nyp	Internet/U	tah	US	United Sta	Manhattan, N	Unknown	40.78343,	40.78343	-73.9663
601f3c66d41f0	2/5/2021	Sexual Exploitatio	Personal Sexual	Social Media	Unknown	Minor	Male	32,24,34,18	Female		https://hu	Internet/U	tah	US	United Sta	West New Yo	Trafficker_	40.78788,	40.78788	-74.0143
601f3c66d41f0	2/5/2021	Sexual Exploitatio	Personal Sexual	Social Media	Unknown	Minor	Male	32,24,34,18	Female		https://hu	Internet/U	tah	US	United Sta	Rockland Cou	Source	41.15243,	41.15243	-74.0241
601f3c66d41f0	2/5/2021	Sexual Exploitatio	Personal Sexual	Social Media	Unknown	Minor	Male	32,24,34,18	Female		https://hu	Internet/U	tah	US	United Sta	Queens Coun	Unknown	40.65749,	40.65749	-73.8388
601f4d410c61(	2/5/2021	Forced Criminality	Personal Sexual	Financial Induc	Unknown	Forced,Minor	Unknown		Female	between tl	https://ww	Internet/U	tah	US	United Sta	New York City	Arrest	40.71427,	40.71427	-74.006
601def891d18(	2/4/2021	Forced Criminality	Personal Sexual	Financial Induc	Unknown	Forced	Female,Male	61,49,39,28,	Female		https://ww	Internet/U	tah	US	United Sta	Rockland Cou	Arrest	41.15243,	41.15243	-74.0241
601def891d18(	2/4/2021	Forced Criminality	Personal Sexual	Financial Induc	Unknown	Forced	Female,Male	61,49,39,28,	Female		https://ww	Internet/U	tah	US	United Sta	East Elmhurst	Unknown	40.76121,	40.76121	-73.8651
601def891d18(	2/4/2021	Forced Criminality	Personal Sexual	Financial Induc	Unknown	Forced	Female,Male	61,49,39,28,	Female		https://ww	Internet/U	tah	US	United Sta	Queens Coun	Arrest	40.65749,	40.65749	-73.8388
601def891d18(	2/4/2021	Forced Criminality	Personal Sexual	Financial Induc	Unknown	Forced	Female,Male	61,49,39,28,	Female		https://ww	Internet/U	tah	US	United Sta	Corona, Que	Unknown	40.74705,	40.74705	-73.8601
602095fe72eb1	2/6/2021	Child Sexual Explo	Arts and Entertai	Relationship M	Car	Forced,Minor,	Unknown				http://axis	Internet/U	tah	US	United Sta	New York City	Unknown	40.71427,	40.71427	-74.006
602095fe72eb1	2/6/2021	Child Sexual Explo	Arts and Entertai	Relationship M	Car	Forced,Minor,	Unknown				http://axis	Internet/U	tah	US	United Sta	Brooklyn, Kin	Unknown	40.6501,-7	40.6501	-73.9496
60272d95aa54	2/11/2021	Child Sexual Explo	Personal Sexual	Violence Threa	Unknown	Minor	Unknown		Female	23,1-year-	https://im	Internet/U	tah	US	United Sta	New York City	Arrest	40.71427,	40.71427	-74.006
601def891d18(	2/4/2021	Sexual Exploitatio	Personal Sexual	Financial Induc	Unknown		Unknown	61,59,39,28	Female	age from 1	https://pal	Internet/U	tah	US	United Sta	Queens, Que	Unknown	40.68149,	40.68149	-73.8365

2. HT_RelatedCrime_2021_2024.csv

CMPLNT_I	CMPLNT_I	CMPLNT_I	CMPLNT_TO	CMPLNT_TI	ADDR_PC	RPT_DT	KY_CD	OFNS_DES	PD_CD	PD_DESC	CRM_ATPT	LAW_CAT_	BORO_NM	LOC_OF_C	PREM_TYP	JURIS_DES	JURISD	PARKS_NM	HADEVEL(	HOUSING_	X_COORD	Y_COORD	SUSP_AGE	SUSP_RAC	SUSP_SEX	Latitude	Longitude	Lat_Lon
2.3E+08	######	1:00:00	5/31/2021	9:20:00	102	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	QUEENS	FRONT OF	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1025247	191527	25-44	WHITE HIS	M	40.6923	-73.8522	(40.69228
2.3E+08	######	9:00:00	3/16/2021	9:20:00	77	######	233	SEX CRIME	681	CHILD, EN	COMPLETI	MISDEME/	BROOKLYI	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1003509	185018	25-44	WHITE HIS	F	40.6745	-73.9306	(40.67449
2.4E+08	######	23:59:00	10/20/2021	17:00:00	47	######	104	RAPE	157	RAPE 1	COMPLETI	FELONY	BRONX	INSIDE	HOTEL/MC	N.Y. POLI(	0	(null)	(null)	(null)	1026480	262584	UNKNOWI	BLACK	M	40.8873	-73.8473	(40.88731
2.4E+08	######	18:40:00	12/22/2021	19:04:00	49	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	BRONX	OPPOSITE	STREET	N.Y. POLI(	0	(null)	(null)	(null)	1026773	256451	18-24	BLACK	M	40.8705	-73.8462	(40.87047
2.2E+08	######	20:30:00	1/12/2021	2:00:00	47	######	233	SEX CRIME	170	SEXUAL M	COMPLETI	MISDEME/	BRONX	INSIDE	HOTEL/MC	N.Y. POLI(	0	(null)	(null)	(null)	1026480	262584	UNKNOWI	UNKNOWI	U	40.8873	-73.8473	(40.88731
2.4E+08	2/7/2022	14:30:00		(null)	44	######	104	RAPE	153	RAPE 3	COMPLETI	FELONY	BRONX	FRONT OF	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1006490	244533	45-64	BLACK	F	40.8378	-73.9196	(40.83784
2.4E+08	######	22:00:00	10/28/2021	23:00:00	43	######	116	SEX CRIME	164	SODOMY :	COMPLETI	FELONY	BRONX	REAR OF	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1020219	239110	25-44	BLACK HIS	M	40.8229	-73.87	(40.82291
2.3E+08	######	22:25:00	6/22/2021	23:13:00	106	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	QUEENS	REAR OF	COMMERC	N.Y. POLI(	0	(null)	(null)	(null)	1023342	186822	25-44	WHITE HIS	M	40.6794	-73.8591	(40.67937
2.4E+08	######	23:00:00	12/15/2021	4:00:00	47	######	104	RAPE	157	RAPE 1	COMPLETI	FELONY	BRONX	INSIDE	HOTEL/MC	N.Y. POLI(	0	(null)	(null)	(null)	1026480	262584	45-64	BLACK	M	40.8873	-73.8473	(40.88731
2.4E+08	######	12:00:00	12/11/2021	17:33:00	47	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	BRONX	INSIDE	STREET	N.Y. POLI(	0	(null)	(null)	(null)	1022728	256756	(null)	(null)	(null)	40.8713	-73.8609	(40.87132
2.3E+08	6/4/2021	14:30:00	6/4/2021	14:30:00	75	6/4/2021	343	THEFT OF	475	UNAUTH. :	COMPLETI	MISDEME/	BROOKLYI	(null)	TRANSIT -	N.Y. TRAN	1	(null)	(null)	(null)	1021568	185710	45-64	BLACK	M	40.6763	-73.8655	(40.67632
2.3E+08	######	18:15:00	10/14/2021	18:30:00	47	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	BRONX	(null)	STREET	N.Y. POLI(	0	(null)	(null)	(null)	1023648	256736	UNKNOWI	BLACK	M	40.8713	-73.8575	(40.87126
2.2E+08	######	11:20:00	2/27/2021	11:35:00	77	######	233	SEX CRIME	681	CHILD, EN	COMPLETI	MISDEME/	BROOKLYI	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1003509	185018	25-44	BLACK	M	40.6745	-73.9306	(40.67449
2.4E+08	######	17:00:00	11/24/2021	17:30:00	47	######	116	SEX CRIME	177	SEXUAL AI	COMPLETI	FELONY	BRONX	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1026480	262584	UNKNOWI	UNKNOWI	M	40.8873	-73.8473	(40.88731
2.4E+08	######	9:00:00	11/1/2021	13:00:00	49	######	233	SEX CRIME	170	SEXUAL M	COMPLETI	MISDEME/	BRONX	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1027434	251387	<18	UNKNOWI	M	40.8566	-73.8439	(40.85657
2.6E+08	######	2:00:00	1/20/2023	2:30:00	43	######	104	RAPE	157	RAPE 1	COMPLETI	FELONY	BRONX	INSIDE	HOMELES:	N.Y. POLI(	0	(null)	(null)	(null)	1020219	239110	UNKNOWI	BLACK	M	40.8229	-73.87	(40.82291
2.4E+08	######	19:30:00	8/15/2021	19:46:00	115	######	343	THEFT OF	475	UNAUTH. :	COMPLETI	MISDEME/	QUEENS	(null)	TRANSIT -	N.Y. TRAN	1	(null)	(null)	(null)	1022285	212504	45-64	BLACK	M	40.7499	-73.8627	(40.74986
2.3E+08	######	22:00:00	3/14/2021	14:41:00	49	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	BRONX	FRONT OF	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1026215	256827	25-44	BLACK	M	40.8715	-73.8483	(40.87150
2.4E+08	######	23:30:00	10/16/2021	23:45:00	110	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	QUEENS	FRONT OF	STREET	N.Y. POLI(	0	(null)	(null)	(null)	1022564	207129	25-44	WHITE HIS	F	40.7351	-73.8617	(40.73511
2.3E+08	######	13:45:00	7/21/2021	13:50:00	49	######	104	RAPE	157	RAPE 1	ATTEMPTE	FELONY	BRONX	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1027434	251387	25-44	BLACK HIS	M	40.8566	-73.8439	(40.85657
2.3E+08	4/5/2021	23:05:00	4/5/2021	23:15:00	110	4/5/2021	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	QUEENS	FRONT OF	STREET	N.Y. POLI(	0	(null)	(null)	(null)	1023172	212159	25-44	BLACK	M	40.7489	-73.8595	(40.74891
2.4E+08	######	1:15:00	11/11/2021	1:20:00	77	######	233	SEX CRIME	681	CHILD, EN	COMPLETI	MISDEME/	BROOKLYI	INSIDE	RESIDENC	N.Y. HOUS	2	(null)	(null)	365	1003509	185018	UNKNOWI	BLACK	M	40.6745	-73.9306	(40.67449
2.3E+08	######	10:30:00	8/24/2021	14:30:00	112	######	116	SEX CRIME	177	SEXUAL AI	COMPLETI	FELONY	QUEENS	INSIDE	OTHER	N.Y. POLI(	0	(null)	(null)	(null)	1025401	202586	<18	UNKNOWI	M	40.7226	-73.8515	(40.72264
2.3E+08	######	1:00:00	8/25/2021	1:28:00	49	######	104	RAPE	157	RAPE 1	COMPLETI	FELONY	BRONX	INSIDE	RESIDENC	N.Y. HOUS	2	(null)	(null)	34132	1027434	251387	45-64	BLACK	M	40.8566	-73.8439	(40.85657
2.4E+08	######	18:00:00	11/25/2021	19:00:00	49	######	116	SEX CRIME	177	SEXUAL AI	COMPLETI	FELONY	BRONX	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1027434	251387	<18	UNKNOWI	U	40.8566	-73.8439	(40.85657
2.3E+08	7/9/2021	4:00:00	7/9/2021	4:30:00	47	######	104	RAPE	157	RAPE 1	COMPLETI	FELONY	BRONX	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1026480	262584	45-64	BLACK	M	40.8873	-73.8473	(40.88731
2.4E+08	######	15:00:00	10/20/2021	15:43:00	77	######	233	SEX CRIME	681	CHILD, EN	COMPLETI	MISDEME/	BROOKLYI	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1003509	185018	25-44	BLACK	M	40.6745	-73.9306	(40.67449
2.2E+08	######	1:22:00	1/19/2021	1:26:00	47	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	BRONX	FRONT OF	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1027055	260514	18-24	BLACK HIS	M	40.8816	-73.8452	(40.88162
2.3E+08	######	17:00:00	4/18/2021	20:00:00	34	######	124	KIDNAPPI	187	KIDNAPPI	COMPLETI	FELONY	MANHATT,	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1002792	249893	(null)	(null)	(null)	40.8526	-73.933	(40.85255
2.3E+08	######	21:44:00	9/22/2021	21:50:00	47	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	BRONX	FRONT OF	RESIDENC	N.Y. HOUS	2	(null)	(null)	873	1026970	261137	UNKNOWI	BLACK	U	40.8833	-73.8455	(40.88333
2.3E+08	######	20:34:00	7/14/2021	20:45:00	102	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	QUEENS	FRONT OF	STREET	N.Y. POLI(	0	(null)	(null)	(null)	1027085	189355	UNKNOWI	UNKNOWI	U	40.6863	-73.8455	(40.68631
2.3E+08	8/1/2021	4:00:00		(null)	47	######	116	SEX CRIME	177	SEXUAL AI	COMPLETI	FELONY	BRONX	INSIDE	RESIDENC	N.Y. POLI(	0	(null)	(null)	(null)	1026480	262584	25-44	WHITE HIS	M	40.8873	-73.8473	(40.88731
2.4E+08	######	20:00:00	11/3/2021	20:15:00	110	######	353	UNAUTHO	462	UNAUTHO	COMPLETI	MISDEME/	QUEENS	(null)	STREET	N.Y. POLI(	0	(null)	(null)	(null)	1023974	211314	UNKNOWI	WHITE HIS	M	40.7466	-73.8566	(40.74659

3. **Passenger_volume_JFKinbound_2021_2024.csv**

Activity Period	Revenue Passenger Volume	Freight Volume	Mail Volume	Total Flights
1/1/2021	611259	61865.0235	3521	7584
2/1/2021	490626	59431.428	2764.8265	6651
3/1/2021	564526	71477.8475	2577.019	7458
4/1/2021	954353	77971.272	3391.678	9806
5/1/2021	1116420	79335.923	3478.53	11690
6/1/2021	1461523	75854.9135	2979.8065	13122
7/1/2021	1751383	82669.18	3421.713	14231
8/1/2021	1927236	83060.437	2697.0445	14363
9/1/2021	1546092	85006.304	2895.3625	13356
10/1/2021	1488641	88585.251	3225.385	13605
11/1/2021	1747351	81529.2165	2935.967	16130
12/1/2021	1800794	86229.778	3721.0765	16555
1/1/2022	1507015	73120.9455	2786.0365	16373
2/1/2022	1536943	65406.0205	2681.7125	15804
3/1/2022	2168453	80145.737	2718.751	18400
4/1/2022	2289310	77814.161	2726.3305	18044
5/1/2022	2429078	78493.582	2651.456	19080
6/1/2022	2502614	77657.903	2935.694	19124
7/1/2022	2764798	78824.14	2470.389	20291
8/1/2022	2794191	79786.602	2504.255	20191
9/1/2022	2549927	77073.938	2815.246	19315
10/1/2022	2542057	84018.3585	2993.7765	19718
11/1/2022	2315851	78780.2555	2797.4505	19036
12/1/2022	2320458	82190.7725	3111.2495	18939
1/1/2023	2273636	67744.841	2084.8075	18832
2/1/2023	2007887	65215.319	2455.4995	17556
3/1/2023	2597725	82079.979	2494.6455	20555
4/1/2023	2666065	80137.4775	2206.3445	20144
5/1/2023	2741721	83701.621	2310.5455	21067

4. Filtered_events_v2.csv

event id	event name	event agency	event type	event borough	event location	event street side	street closure type	communit	police pre	start_date	start_time	end_date	end_time
761335	kindness carnival	Parks Department	Special Event	Queens	Cambria Playground/Cabbell Park: Lawn			13,	105,	10/5/2024	12:00:00	10/5/2024	17:00:00
758116	senator addabbo's paper sl	Parks Department	Special Event	Queens	Forest Park: Bandshell Parking Lot			82,	102,	9/29/2024	10:00:00	9/29/2024	14:00:00
739132	rockaway beach santa suit	Parks Department	Special Event	Queens	Rockaway Beach and Boardwalk: Beach 94th Street			14,	100,	12/14/2024	10:00:00	12/14/2024	12:00:00
805705	sss cpw96 rudin lawn fall 2	Parks Department	Special Event	Manhattan	Central Park: Rudin Playground Lawn			64,	22,	9/30/2024	13:30:00	9/30/2024	17:00:00
612499	davis cup	Parks Department	Special Event	Brooklyn	Marcy Playground: Flagpole			3,	79,	7/30/2022	10:00:00	7/30/2022	13:00:00
591320	kids program	Parks Department	Special Event	Brooklyn	Prospect Park: Picnic House North			55,	78,	12/16/2021	15:00:00	12/16/2021	17:00:00
804548	tango in the park at sutton	Parks Department	Special Event	Manhattan	Sutton Place Park: 56th Street Plaza ,Sutton Place Park: Pig Plaza 57th ,Sutton Place P			6,	17,	10/1/2024	17:00:00	10/1/2024	18:30:00
801518	miscellaneous	Parks Department	Special Event	Brooklyn	Van Voorhees Playground: Softball-01			06,	76,	12/5/2024	12:00:00	12/5/2024	15:00:00
607714	chanuka 2021	Parks Department	Special Event	Manhattan	Herald Square: Herald Square			5,	14,	12/6/2021	17:00:00	12/6/2021	23:00:00
804535	2024.09.26 idlewild park fo	Parks Department	Special Event	Queens	Idlewild Park: Idlewild Park			13,	105,	9/26/2024	9:00:00	9/26/2024	12:00:00
776074	776 track event	Parks Department	Special Event	Manhattan	Randall's Island Park: Icahn Stadium ,Randall's Island Park: Soccer-10			11,	25,	9/27/2024	11:00:00	9/27/2024	23:00:00
806521	sf bbq	Parks Department	Special Event	Manhattan	Randall's Island Park: East River Picnic Area Orange			11,	25,	9/15/2024	12:00:00	9/15/2024	16:00:00
782456	14th annual nyc pizza run -	Parks Department	Special Event	Brooklyn	Fort Greene Park: Mrytle and Washington Park Entrance			2,	88,	9/15/2024	11:00:00	9/15/2024	11:30:00
757115	global gratitude march	Parks Department	Special Event	Queens	Astoria Park: Great Lawn 02 Ditmars Blvd			1,	114,	9/21/2024	12:00:00	9/21/2024	19:00:00
796665	unga 79	Parks Department	Special Event	Manhattan	Ralph Bunche Park: Isaiah Wall Plaza			6,	17,	10/2/2024	0:00:00	10/2/2024	23:00:00
793750	csaa cross country- van co	Parks Department	Special Event	Bronx	Van Cortlandt Park: Cross Country Parade Ground Broadway 1			26,	50,	9/13/2024	16:00:00	9/13/2024	18:00:00
810298	yartung dhapa 2024	Parks Department	Special Event	Staten Island	Willowbrook Park: Archery Field			2,	122,	10/6/2024	9:00:00	10/6/2024	18:00:00
802967	cyclocross practice 2024	Parks Department	Special Event	Manhattan	Randall's Island Park: Wards Meadow Picnic Area Blue ,Randall's Island Park: Wards Me			11,	25,	11/6/2024	19:00:00	11/6/2024	21:00:00
756350	mothership rehearsals	Parks Department	Special Event	Manhattan	Marcus Garvey Park: Amphitheater			11,	25,	9/19/2024	11:00:00	9/19/2024	19:00:00
802755	ps186 school carnival	Parks Department	Special Event	Queens	Castlewood Playground: Castlewood Playground			13,	105,	9/13/2024	6:00:00	9/13/2024	15:00:00
806195	lego experiential activation	Street Activity Per Production Eve		Manhattan	EAST 42 STREET between LEXINGTON	North	Curb Lane Only	5,	14,	9/16/2024	8:00:00	9/17/2024	23:30:00
776305	striders of the caribbean br	Parks Department	Special Event	Queens	Brookville Park: Large Event Area			13,	105,	9/28/2024	11:00:00	9/28/2024	15:00:00
807885	jp urban gathering	Parks Department	Special Event	Brooklyn	Prospect Park: CONCERT GROVE PAVILION			55,	78,	10/3/2024	12:00:00	10/3/2024	18:00:00
786240	star track workshop for kid	Parks Department	Special Event	Queens	Kissena Park: Velodrome-Track-01			07,	109,	10/12/2024	11:00:00	10/12/2024	13:00:00
796106	miscellaneous	Parks Department	Special Event	Manhattan	Baruch Playground: Softball-01 ,Baruch Playground: Football-01 ,Baruch Playground: S			03,	7,	10/2/2024	10:00:00	10/2/2024	13:00:00
805705	sss cpw96 rudin lawn fall 2	Parks Department	Special Event	Manhattan	Central Park: Rudin Playground Lawn			64,	22,	9/30/2024	13:30:00	9/30/2024	17:00:00
756523	memorial 911 prayer	Parks Department	Special Event	Queens	Rockaway Beach and Boardwalk: Beach 94th Street			14,	100,	9/11/2024	11:00:00	9/11/2024	12:00:00
804117	cinema under the stars rair	Parks Department	Special Event	Brooklyn	Prospect Park: Breeze Hill Oval			55,	78,	9/20/2024	19:00:00	9/20/2024	22:00:00
801352	montefiore park summer ja	Parks Department	Special Event	Manhattan	Montefiore SquarePark: Montefiore SquarePark			9,	30,	9/28/2024	16:00:00	9/28/2024	19:30:00

5. Final_modeling_dataset_v2.csv

date	location_id	Latitude	Longitude	daily_event_c	lat_rad	lon_rad	past_7d_crime	night_crime_	month	monthly_pass	borough	day_of_wee	is_weeken	month_nu	season	is_holiday	future_ht_risk_1d
1/1/2021	40.78343_-73.96625	40.78343	-73.9663	0	0.711805	-1.29095	2	0.5	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.71427_-74.00597	40.71427	-74.006	0	0.710598	-1.29165	5	1	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.70789_-74.00857	40.70789	-74.0086	0	0.710487	-1.29169	4	1	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.82732_-73.92357	40.82732	-73.9236	0	0.712571	-1.29021	2	1	1/1/2021	0.1938281	BRONX	4	0	1	Winter	1	0
1/1/2021	40.6501_-73.94958	40.6501	-73.9496	0	0.709478	-1.29066	7	0.7142857	1/1/2021	0.1938281	BROOKLYN	4	0	1	Winter	1	0
1/1/2021	40.90056_-73.90639	40.90056	-73.9064	0	0.713849	-1.28991	0	0	1/1/2021	0.1938281	BRONX	4	0	1	Winter	1	0
1/1/2021	40.77205_-73.93014	40.77205	-73.9301	0	0.711607	-1.29032	5	0.8	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.66677_-73.88236	40.66677	-73.8824	0	0.709769	-1.28949	5	0.4	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.78788_-74.01431	40.78788	-74.0143	0	0.711883	-1.29179	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	41.15243_-74.02409	41.15243	-74.0241	0	0.718245	-1.29196	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.65749_-73.83875	40.65749	-73.8388	0	0.709607	-1.28873	0	0	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.76121_-73.86514	40.76121	-73.8651	0	0.711417	-1.28919	4	1	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.74705_-73.86014	40.74705	-73.8601	0	0.71117	-1.2891	4	1	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.68149_-73.83652	40.68149	-73.8365	0	0.710026	-1.28869	1	1	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.89788_-73.85236	40.89788	-73.8524	0	0.713803	-1.28897	4	1	1/1/2021	0.1938281	BRONX	4	0	1	Winter	1	0
1/1/2021	40.72677_-73.6343	40.72677	-73.6343	0	0.710816	-1.28516	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.71427_-73.95347	40.71427	-73.9535	0	0.710598	-1.29073	1	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.69983_-73.83125	40.69983	-73.8313	0	0.710346	-1.2886	1	1	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.63439_-73.95027	40.63439	-73.9503	0	0.709204	-1.29068	10	0.8	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.716_-73.9974	40.716	-73.9974	0	0.710628	-1.2915	8	0.875	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	41.11482_-74.14959	41.11482	-74.1496	0	0.717589	-1.29415	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.56233_-74.13986	40.56233	-74.1399	0	0.707946	-1.29398	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.80788_-73.94542	40.80788	-73.9454	0	0.712232	-1.29059	4	0.5	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.76844_-73.77708	40.76844	-73.7771	0	0.711544	-1.28765	0	0	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.69455_-73.73847	40.69455	-73.7385	0	0.710254	-1.28698	0	0	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	41.03399_-73.76291	41.03399	-73.7629	0	0.716178	-1.28741	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.65871_-73.64124	40.65871	-73.6412	0	0.709628	-1.28528	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0
1/1/2021	40.7001_-73.90569	40.7001	-73.9057	0	0.710351	-1.2899	4	0.75	1/1/2021	0.1938281	QUEENS	4	0	1	Winter	1	0
1/1/2021	40.60066_-74.19487	40.60066	-74.1949	0	0.708615	-1.29494	0	0	1/1/2021	0.1938281	MANHATTAI	4	0	1	Winter	1	0

6. Final_modeling_dataset_v1.csv

date	Latitude	Longitude	ht_related_crime	direct_reports	num_events	monthly_passen	borough	is_weekend	day_of_week	month	season	season_encode	is_holiday	is_near_holiday
2021-01-01	40.51203825	-74.24975495	1	0	0	611259	STATEN ISLANI	0	4	1	winter	0	1	0
2021-01-01	40.57428569	-74.10591441	3	0	0	611259	STATEN ISLANI	0	4	1	winter	0	1	0
2021-01-01	40.57664598	-73.9764804	2	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.59401906	-73.96085432	2	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.60221617	-74.0029508	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.607237	-73.956064	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.62318838	-74.14923769	3	0	0	611259	STATEN ISLANI	0	4	1	winter	0	1	0
2021-01-01	40.62816974	-73.94135878	4	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.63059985	-73.97370532	5	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.632316	-74.119802	1	0	0	611259	STATEN ISLANI	0	4	1	winter	0	1	0
2021-01-01	40.64472094	-74.07703272	3	0	0	611259	STATEN ISLANI	0	4	1	winter	0	1	0
2021-01-01	40.64885075	-73.95101651	6	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.65815775	-74.00044115	4	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.66412128	-73.94776484	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.67113	-73.91350206	4	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.67135982	-73.88181102	5	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.67449569	-73.93057133	5	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.676171	-73.951301	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.67998074	-73.77623391	2	0	0	611259	QUEENS	0	4	1	winter	0	1	0
2021-01-01	40.68078561	-73.97447512	2	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.68239828	-73.84007216	1	0	0	611259	QUEENS	0	4	1	winter	0	1	0
2021-01-01	40.6890014	-73.94502653	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.6894643	-73.92402909	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.6902236	-73.96027859	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.69543881	-73.98322538	1	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.69847382	-73.91776898	2	0	0	611259	BROOKLYN	0	4	1	winter	0	1	0
2021-01-01	40.69932416	-73.83157089	1	0	0	611259	QUEENS	0	4	1	winter	0	1	0
2021-01-01	40.70443503	-73.89378071	2	0	0	611259	QUEENS	0	4	1	winter	0	1	0
2021-01-01	40.70723982	-73.79272673	3	0	0	611259	QUEENS	0	4	1	winter	0	1	0
2021-01-01	40.71427	-74.00597	0	1	0	611259	MANHATTAN	0	4	1	winter	0	1	0
2021-01-01	40.71601201	-73.99733203	4	0	0	611259	MANHATTAN	0	4	1	winter	0	1	0
2021-01-01	40.71630999	-73.98316601	1	0	0	611259	MANHATTAN	0	4	1	winter	0	1	0
2021-01-01	40.7226266	-74.00811727	1	0	0	611259	MANHATTAN	0	4	1	winter	0	1	0
2021-01-01	40.72651564	-73.98829024	2	0	0	611259	MANHATTAN	0	4	1	winter	0	1	0