Spatiotemporal and Weather Trends in New York City Cycling Accidents

Nolan Betts

***Abstract*—Since the inception of New York's bike share service in 2013, the city has experienced a notable increase in cycling participation and cycling accidents. This study aims to explore the geographical, temporal, and environmental trends with respect to these accidents. Focusing on factors such as weather, location, and time of day, this work explores two different machine learning approaches to uncover patterns. An unsupervised clustering approach analyzes five distinct accident clusters that are characterized by different feature patterns. In order to better understand how traffic volume may impact cycling safety, the second approach leverages traffic normalization to create a predictive model for accident severity. While the predictive model showed limited success, the analysis of its results uncovered specific high-risk locations for cycling safety. The findings provide valuable insights for policymakers and urban planners to create a safer cycling environment in New York City.**

***Index Terms*—Cycling Safety, Bicycle Crashes, New York City Transportation, Traffic Data Analysis, Bike Share Safety**

I. Introduction

CYCLING in urban areas has become a prominent mode of transportation across the world. Research has shown that an increase in cycling has provided many benefits ranging from personal health, environment and emissions, to reduced traffic congestion [1]. While the United States has lagged behind many other developed nations, New York City has seen a rise in daily cycling between 2011 and 2021 of 104%[2]. This rise in daily ridership can be attributed to the advent of Citibike, New York's bikeshare program. Since the inception of Citibike, New York has added 1456 miles of bike lanes, including 590 protected bike lanes[2]. In 2023 alone, the program facilitated over 35 Million trips, demonstrating the impact this program has on promoting cycling as a viable transit method[3].

Alongside this increase in ridership, we have been presented with an increase in concern for rider safety. In 2023 alone, there were 5179 cyclist injuries and 29 fatalities reported in New York City[3]. Globally, roughly 40,000 cyclists lose their lives each year to road traffic related incidents [4]. Cyclists, particularly in North America, are disproportionately exposed to heightened risk of death as compared to car occupants [5], [6]. Per kilometer traveled, studies report that cyclists are 12 times more likely to be killed in an accident than car passengers [7]. Although New York City has made efforts to improve cyclist safety through constructing cyclist only infrastructure, it is clear that there are remaining safety concerns that need to be addressed [8].

One existing study showed that the introduction of 32 km/h zones in London decreased cycling fatalities by 49.6% [4]. Another study highlights that factors such as population density and trip length are correlated with safer cycling [9]. Additionally, some existing work suggests that policies that increase walking and cycling appear to simultaneously improve safety [10]. While many of these existing studies are able to highlight important factors with respect to cycling incidents, they fail to cohesively integrate spatiotemporal trends, traffic normalization and other dynamic environment factors like weather and time of day. Many of these studies also fail to represent or study New York City specifically which may present a misrepresentation.

This study aims to address these limitations by leveraging a variety of data sources provided by New York City including vehicle collisions, Citibike trip data, and historical weather data to reveal trends in cycling accidents. By leveraging geospatial tools such as H3 to create hexagonal mappings and integrating traffic normalization methods, this paper aims to identify high risk zones, times, and weather patterns. This study is split into two parts. First, it studies collision data involving cyclists from 2012 to 2024 (Fig 1) using an unsupervised clustering approach. The second part of the study implements a traffic normalization methodology and attempts to solve a regression problem to predict incident severity. Highlighting these patterns can allow for evaluation of the impact of environmental and infrastructural factors on cycling safety. Ultimately, it seeks to provide insight for how urban planners and policy makers can improve bike safety in New York City.

The paper is structured as follows. Section II reviews relevant research and identifies gaps in this research. Section III provides a description of the data, technologies, data preparation activities, and methodologies. Section IV details the results of the unsupervised clustering analysis as well as a traffic normalization risk analysis. Section V discusses implications of these findings and limitations of the study. Section VI concludes with other potential uses for discoveries and recommendations for future direction.

II. Research Background

Cycling has proven to be a critical focus for analyzing transportation safety. Evidence consistently exposes the role of infrastructure and urban planning in reducing accidents and severity of accidents. Most notably, research shows that protected bike lanes, traffic calming measures, and dedicated cycling zones positively contribute to the safety of cyclists[11], [12], [13]. In one Canadian study, researchers found that protected bike lanes reduced cyclist injuries in children by 50% [12]. Another European study found a positive correlation

N. Betts Master's Student of Artificial Intelligence, University of Texas, Austin, TX, 78712 USA e-mail: nolanfbetts@utexas.edu.

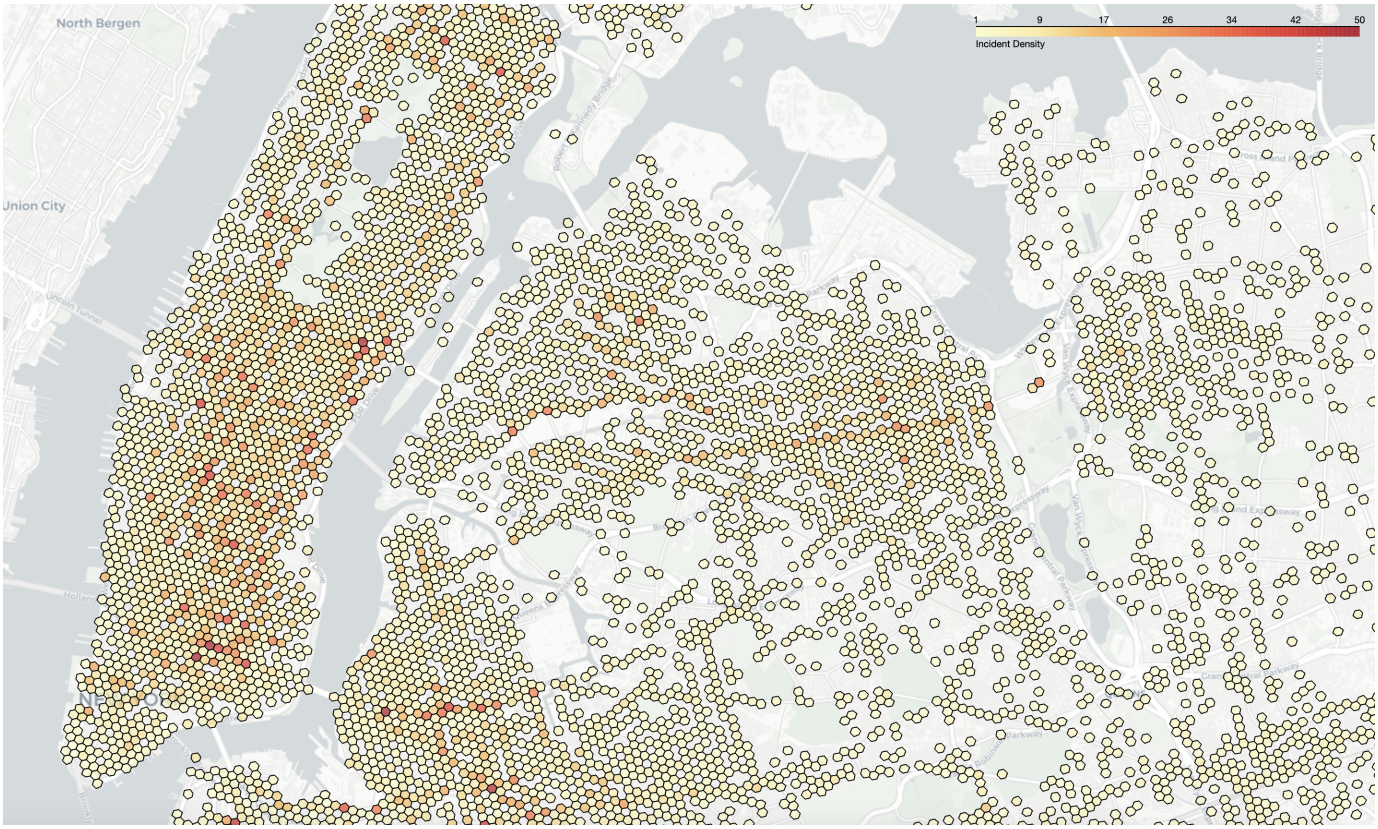


Fig. 1: Hexagonal Heatmap of Cycling Incidents (June 2012 - October 2024)

between cycling infrastructure investment and a reduction in cyclist fatalities [7]. New York City has shown a commitment to grow its number of cyclists while simultaneously improving its cycling infrastructure, efforts that are showing increased cycling activity and reduced cycling incidents [1].

Despite these advancements, walking and cycling is still much more dangerous in US cities than many others around the world [7]. Cyclists in New York City are still faced with a much more dangerous cycling environment. Cyclists in North America are 8 to 30 times more likely to be involved in a traffic accident than European cyclists [5]. These disparities can be credited to inconsistent infrastructure coverage, sharing space with high speed vehicles (i.e. cars), and insufficient traffic enforcement [7]. Among the less understood risk factors includes the rise of eBike and other motorized micro mobility vehicles occupying cycling lanes [2]

The use of geospatial tools such as H3 to analyze New York City specific collision data offers the potential to uncover new trends in cycling collisions specifically for New York City. H3 grids allow for the segmentation of geographical space into uniform, indexable, hexagonal cells [14]. These capabilities allow for the discretization of things like incident and traffic data allowing for easier identification of hot spots. This study attempts a traffic normalization procedure leveraging Citibike trip data and NetworkX pathing algorithms. Traffic normalization can be difficult to obtain and often omitted in similar cyclist safety studies.

Environmental factors also prove to play a pivotal role in cyclist participation and safety. Existing research shows that weather conditions along with temporal variables are significant factors in cyclist collisions and outcomes [15], [4], [16]. Commuting times in the morning and evening have been identified for having high accident volumes and increased traffic volumes [4], [16].

In addressing some of the gaps mentioned above, this study builds on an infrastructure based approach for analyzing cycling safety by focusing on spatial, temporal and environmental factors. By taking a detailed look at New York City specifically this study is able to provide cycling safety incidents for a community that is attempting to grow it's adoption of cycling safely. This study also attempts to include traffic normalization to not only highlight areas where accidents frequently occur, but areas where accident frequency and severity outweighs its normalized traffic volume. This approach positions this research as a step forward in developing targeted safety interventions that prioritize cyclist safety in the urban environment of New York City.

III. Methodology

A. Datasets

This work leveraged several different data sources to complete this analysis. From NYC Open Data I used the "Motor Vehicle Collisions - Crashes"[17]. This dataset includes data about all of the reported vehicle collisions in New York City. I filtered this data to collisions involving only bicycles (Table I). I generated weather data from Visual Crossing Weather in order to gather information about the weather

conditions at the time each incident occurred (Table II) [18]. From citibikenyc/Lyft, I used the Citi BikeTrip Histories data [3]. Citi Bike is New York City's bike sharing system and this data tracks all of the rides that users take [19]. I used this data to generate a traffic normalization for the collision data (Table III).

TABLE I: Motor Vehicle Incident Data Fields

Field Number	Field Name
1	CRASH DATE
2	CRASH TIME
3	BOROUGH
4	ZIP CODE
5	LATITUDE
6	LONGITUDE
7	LOCATION
8	ON STREET NAME
9	CROSS STREET NAME
10	OFF STREET NAME
11	NUMBER OF PERSONS INJURED
12	NUMBER OF PERSONS KILLED
13	NUMBER OF PEDESTRIANS INJURED
14	NUMBER OF PEDESTRIANS KILLED
15	NUMBER OF CYCLIST INJURED
16	NUMBER OF CYCLIST KILLED
17	NUMBER OF MOTORIST INJURED
18	NUMBER OF MOTORIST KILLED
19	CONTRIBUTING FACTOR VEHICLE 1
20	CONTRIBUTING FACTOR VEHICLE 2
21	CONTRIBUTING FACTOR VEHICLE 3
22	CONTRIBUTING FACTOR VEHICLE 4
23	CONTRIBUTING FACTOR VEHICLE 5
24	COLLISION_ID
25	VEHICLE TYPE CODE 1
26	VEHICLE TYPE CODE 2
27	VEHICLE TYPE CODE 3
28	VEHICLE TYPE CODE 4
29	VEHICLE TYPE CODE 5

TABLE II: Weather Data Fields

Field Number	Field Name
1	name
2	datetime
3	temp
4	feelslike
5	dew
6	humidity
7	precip
8	precipprob
9	preciptype
10	snow
11	snowdepth
12	windgust
13	windspeed
14	winddir
15	sealevelpressure
16	cloudcover
17	visibility
18	solarradiation
19	solarenergy
20	uvindex
21	severerisk
22	conditions
23	icon
24	stations

TABLE III: Citibike Trip Data Fields

Field Number	Field Name
1	ride_id
2	rideable_type
3	started_at
4	ended_at
5	start_station_name
6	start_station_id
7	end_station_name
8	end_station_id
9	start_lat
10	start_lng
11	end_lat
12	end_lng
13	member_casual

B. Technologies

H3 is an open source hexagonal geospatial grid system that was developed by Uber [14]. I leveraged H3 in order to create unique and indexable geospatial units for my analysis. When creating a hexagonal grid using H3, you specify a resolution. In this case I chose a resolution of 10 which has an average hexagon area of .0150475km^2. I felt that this resolution provided an appropriate level of granularity for the streets and intersections of New York City. Each hexagon in H3 has a unique key which allows for easy looks up and tracking. I used H3 in conjunction with folium to generate the hexagonal figures that are displayed in this analysis. Folium is a popular python package that makes it easy to visualize data on an interactive leaflet map [20]. Along with using H3 to create visualizations, it was also instrumental in my analysis for creating mappings between traffic, routing, and incident information.

Sklearn or Scikit-Learn is a widely used open source python framework built for machine learning. This package library includes many easy to use algorithms out of the box. This work made use of many of the supervised and unsupervised learning algorithms to perform work such as unsupervised clustering and regression for predictive modeling [21].

NetworkX is a python library that is used for analyzing and processing various types of graphs and networks [22]. In this study, NetworkX is used to find shortest path navigation instructions between two geographical points. The nodes and geographical information used to perform this task comes from OpenStreetMap, often referred to as OSM. OSM is a widely used open source library that provides a multitude of rich geographical data [23]. For this work, OSM was able to provide the geographical coordinates of bike accessible routes within New York City.

In order to visualize the results of this study, several visualization methods were implemented. Along with folium as mentioned above, the study also made use of Matplotlib and Seaborn. Matplotlib is a widely used visualization library that allows users to build complex visualizations from basic building blocks [24]. Seaborn is an abstracted visualization library that was built on top of Matplotlib to enable more complex visualizations with a simplified user process [25]. Many of the more basic visualizations used Matplotlib while the more complex figures leverages Seaborn.

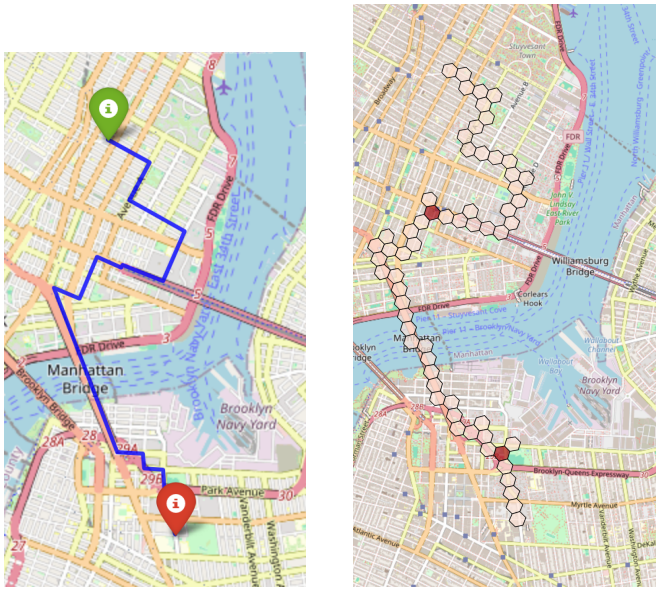
C. Data Preprocessing

In order to perform this analysis, there were several data preprocessing steps that I had to take. In order to create a

mapping for weather data, I rounded the time of the vehicle incidents to the nearest hour. This allows us to still see trends in the data with granularity up to the hour while removing some noise in the date time provided by more specific minutes and seconds. Another preprocessing activity involved leveraging H3 to find the unique hexagon for the location of each incident. I was able to use the latitude and longitude information provided in the incident data to return a valid H3 hexagon identifier.

Data preprocessing for the unsupervised clustering involved combining the weather data and vehicle collision data as well as performing encoding on categorical variables. This stage of the study considers incidents from June 2012 to November 2024. To start on this effort I crafted the set of features that were to be used for clustering. These features included different vehicle incident specific data points such as location, number of cyclists injured or killed, and location as well as weather specific data such as wind speed, temperature and precipitation.The data for incident and weather data was mapped by a rounded time value. This time value provides fields such as Month, Hour, and Day of the Week. Many of these features needed to be expanded using one hot encoding as they presented as categorical variables. One hot encoding is a process used to convert categorical values into numerical or binary fields [26]. In totality, the feature selection process yielded 80 different features.

One of the biggest challenges came in the second part of the analysis when creating a representation for traffic normalization. While the incident data clearly tells us where and when an accident involving a bicycle occurred, it does not tell us any information about the traffic volume at the particular location. In order to create a traffic normalization I leverage the Citibike Trip Histories data. For 2023, this data contained over 35 million records [3]. I rounded this data to the nearest hour, to match our standardization for incidents. I then extracted the start station and endstation information. I used this to determine how many times each unique trip occurred in each hour. For example, I could then say at timestamp 2023-01-01 06:00:00 from station 4762.05 to station 4724.03 there were 3 Citibike trips taken. Using this information I was also able to identify about 2 million unique station pair trips taken in the year. For this step it was necessary to consider bidirectional routes as one way streets can create discrepancies in navigation. Now that I had more information about Citibike volume and locations, I needed to see how these trips routed through the city. I created an offline instance of OSM as my usage exceeded API limits. OSM contains data about roads, buildings, addresses, shops, points of interest, railways, trails, transit, land use, and natural features [23]. My offline instance of OSM was filtered to only include navigational nodes that are accessible by bicycle. I used the latitude and longitudes of the start and end stations along with NetworkX's Astar [22] search algorithm to determine the shortest routes between all of the station pairs found. The output from Networkx provided a list of latitude and longitude pairs as navigational directions. I converted these points to H3 and used the H3 grid_path_cells function to find all hexagons connecting two points (Fig 2) [14].



(a) Route visualization using line segments

(b) Route visualization using H3 hexagonal binning

Fig. 2: Comparison of route visualization methods

The result of this data processing allowed me to make an inference about potential bike traffic data at a particular location at a different time. I used the following process to generate a projected traffic count at the time and location of a particular incident. Using the timestamp of the incident, I gathered all of the CitiBike trips taken in that hour. I then see how many times the incident H3 index occurs leveraging the start and end station routing information scaled by the frequency count. This provides me with a cycling traffic metric of how many bicycled presumably passed through the incident H3 area at a given time (Fig 3).

D. Methods

This study was split into two different approaches. The first approach was an unsupervised clustering approach that aimed to find clusters with similar characteristics to help to explain why some accidents occur. This part of the study analyzes all reported traffic accidents in New York City from 2012 to current. The second part of this study focuses specifically on 2023 traffic data and leverages Citibike trip data to create a normalized traffic score. This stage of the study attempts to predict incident severity and identify unique hotspot locations throughout the city.

1) Unsupervised Clustering: In the first stage of the study we are focused on unsupervised clustering. The general goal in this stage is to uncover any underlying patterns or structure in the data without providing labels. To complete this process, I started first with feature selection to provide the clustering algorithm with a set of data to perform clustering on. This feature set proved to be rather large so I performed Principal Component Analysis (PCA) to reduce the dimensionality while retaining the most important information from the top principal components [27]. In order to determine the number of principal components I considered the cumulative variance

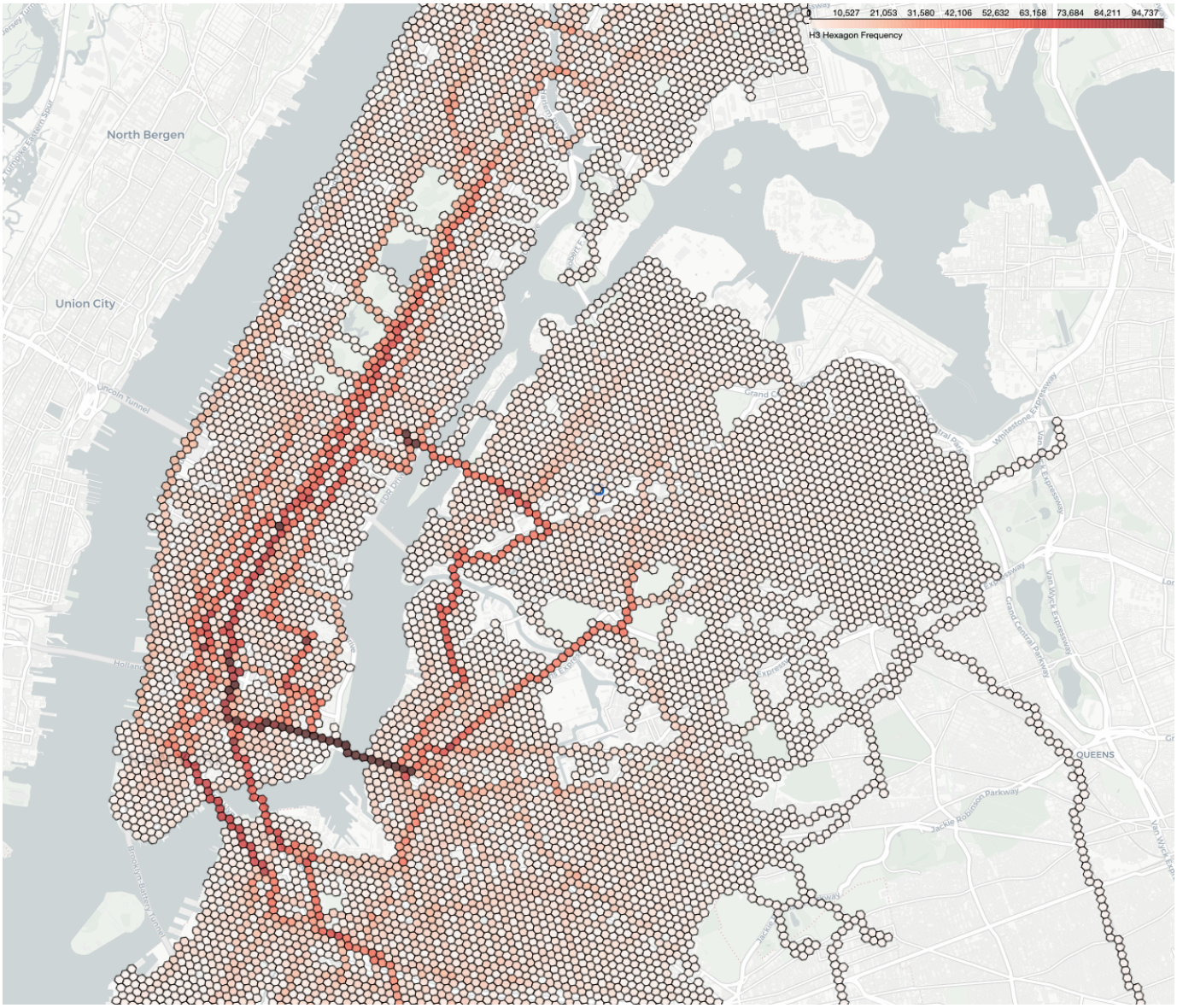


Fig. 3: Overlay of all possible routes between stations with hexagonal counts

vs. the number of principal components. Now that the number of components was decided, I used the elbow method to help determine the optimal number of clusters. With the number of clusters decided I ran KMeans clustering on my dimensionality reduced dataset. In the analysis focused on cluster comparison, I leverage various different charting techniques to make observations about the results.

To start on this effort I crafted the set of features that was used for clustering. These features included different vehicle incident specific data points such as location, number of cyclists injured or killed, and location as well as weather specific data such as wind speed, temperature and precipitation. As mentioned in the preprocessing section, the data for incident and weather data was mapped by a rounded time value. This time value provides fields such as Month, Hour, and Day of the Week. Many of these features needed to be expanded using one hot encoding as they presented as categorical variables. In totality, the feature selection process yielded 80 different features (Table IV).

With these 80 features I then used PCA to reduce the dimensionality of the data. I decided that dimensionality reduction was an appropriate step because many of the weather features were highly correlated, the data was noisy, and a lower dimensionality would help with KMeans performance and visualizations. In this step I used PCA functionality provided by sklearn. I compared cumulative explained variance to the number of PCA components to determine the number of components required. I found that 70% of variance could be preserved when using the top 20 principal components. I felt that this preserved enough of the initial data to perform the unsupervised exploratory analysis.

TABLE IV: Model Feature Fields

Time Features	
1-24	Hour_0 through Hour_23
Weather Measurements	
25	temp
26	feelslike
27	dew
28	humidity
29	precip
30	precipprob
31	snow
32	snowdepth
33	windgust
34	windspeed
35	winddir
36	sealevelpressure
37	cloudcover
38	visibility
39	solarradiation
40	solarenergy
41	uvindex
42	severerisk
Precipitation Types	
43	precip_Clear
44	precip_Rain
45	precip_Snow
46	precip_Freezing_Rain
47	precip_Ice
Weather Conditions	
48	Clear
49	Overcast
50	Partially_cloudy
51	Rain
52	Snow
53	Ice
54	Freezing_Drizzle
Day of Week	
55-61	Monday through Sunday
Month	
62-73	January through December
Geospatial	
74	h3_center_lat
75	h3_center_long
Borough	
76	BRONX
77	BROOKLYN
78	MANHATTAN
79	QUEENS
80	STATEN ISLAND

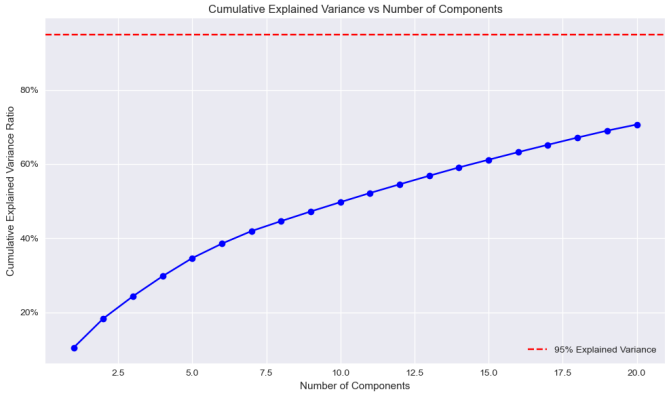


Fig. 4: Cumulative Variance vs Number of Components

With dimensionality reduction complete, I then moved on to clustering. I used the elbow method to determine the optimal number of clusters for my KMeans algorithm. Using Kmeans and the _intertia property offered by sklearn I charted the sum of squared distances between the centroids. I found the elbow point to be where k was equal to 5 which indicates that 5 is the optimal number of clusters for this problem. Finally, I was able to run the KMeans clustering algorithm provided by Sklearn to generate 5 clusters on the top 20 principal components from my dataset.

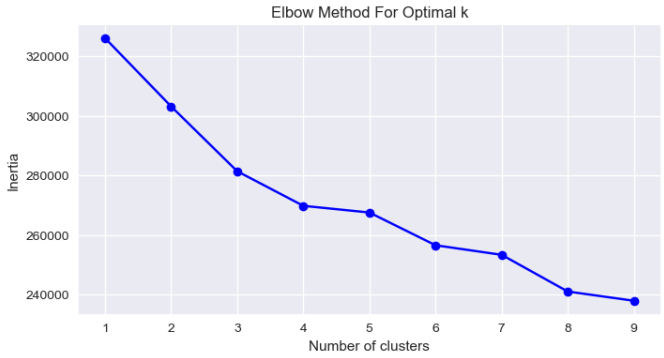


Fig. 5: Elbow Method for Optimal k

After the completion of Kmeans I performed analysis on the clusters. I leveraged a scatter plot from matplotlib to visually analyze the top 2 principal components of the clusters. I grouped the initial dataset by cluster and took the mean of each feature for each cluster to look for dominating features by cluster. I again used matplotlib but this time to chart the average temperature and precipitation by cluster. I then used seaborn to create visualizations such as a heatmap showing the distribution of accidents throughout the year by cluster, and a chart analyzing the hourly distribution by cluster. These visualizations and their spatial, temporal, and environmental significance will be further explored in the Results and Discussion sections.

2) Traffic Normalized Regression: For the second application in this study, I used the processed traffic volume data outlined in the data processing section of this report. This section of the study only focuses on incidents and traffic from 2023. The goal for this portion of the study was to better understand incidents with respect to the proposed traffic volume of the area. I created a process to assess a normalized risk score for each incident and used a linear regression model to attempt to predict the risk score for a given accident. Along with this, a similar feature selection with one hot encoding of categorical fields was employed. To further study temporal and weather trends, this study specifies fields such as is_weekend, and is_rush_hour (Table V)

In order to create the traffic normalized risk score, I created a simple heuristic. This heuristic counts cyclist injuries as 1 point and a cyclist death as 10 points. The score is then normalized by the traffic volume defined by preprocessing. For example, an incident that involved 2 injured cyclists and a cyclist death would be recorded as 12. In this example we will say that our traffic volume at the time for that hex location is 100. Our normalized score would then be calculated as 12/100.

I trained the model using a RandomForestRegressor and KFOLD Cross Validation for modeling accident severity. The

dataset was split 80/20 and performance was evaluated using R-squared scoring. Additionally, the study calculates feature importance to describe which selected features are most influential. Predictions are made on the test set using the trained model. To analyze the model's accuracy, this work uses Root Mean Squared Error (RMSE).

For analyzing this portion of the study, several different visualization methods were used. A bar plot was used for feature importance. A spatial heatmap was created to analyze the normalized high severity geolocations. Several plots were included to analyze severity with respect to temporal patterns. Finally, this study leverages a scatter plot and a histogram to measure the performance of the model in predicting severity.

TABLE V: Model Input Features

Temporal Features	
Hour Indicators	hour_0 through hour_23
Day Indicators	day_0 through day_6
Month Indicators	month_1 through month_12
Derived Time	is_weekend, is_rush_hour
Weather Features	
Continuous	temp
	feelslike
	humidity
	windspeed
	cloudcover
	visibility
	solarradiation
Precipitation Types	
Categories	preciptype_Clear
	preciptype_rain
	preciptype_rain,snow
	preciptype_snow
Geospatial Features	
Coordinates	h3_center_lat
	h3_center_long
Network	neighbor_count

Variable	C0	C1	C2	C3	C4
Weather					
Temp	14.67	20.44	15.29	19.82	20.03
Clear	0.62	0.98	1.00	1.00	0.98
Rain	0.37	0.02	0.00	0.00	0.02
Snow	0.03	0.00	0.00	0.00	0.00
Weekday					
Mon	0.13	0.14	0.13	0.14	0.14
Tue	0.15	0.14	0.14	0.16	0.15
Wed	0.15	0.16	0.14	0.15	0.16
Thu	0.16	0.15	0.14	0.16	0.15
Fri	0.16	0.16	0.17	0.16	0.15
Sat	0.13	0.13	0.14	0.13	0.14
Sun	0.12	0.12	0.14	0.09	0.13
Month					
Jan	0.07	0.03	0.05	0.02	0.03
Feb	0.05	0.03	0.05	0.03	0.03
Mar	0.07	0.04	0.05	0.06	0.04
Apr	0.08	0.05	0.05	0.07	0.06
May	0.10	0.09	0.06	0.10	0.10
Jun	0.09	0.14	0.08	0.11	0.13
Jul	0.08	0.16	0.09	0.13	0.15
Aug	0.08	0.16	0.11	0.12	0.15
Sep	0.10	0.11	0.14	0.14	0.12
Oct	0.13	0.08	0.13	0.10	0.08
Nov	0.08	0.06	0.12	0.08	0.06
Dec	0.09	0.04	0.06	0.03	0.05
Hour					
00-06	0.10	0.05	0.22	0.00	0.08
06-12	0.19	0.13	0.09	0.39	0.15
12-18	0.31	0.44	0.04	0.47	0.41
18-24	0.39	0.36	0.63	0.14	0.35
Location					
Lat	40.73	40.66	40.73	40.72	40.75
Long	-73.94	-73.96	-73.94	-73.94	-73.93
Borough					
BX	0.08	0.00	0.09	0.08	0.13
BK	0.30	1.00	0.32	0.33	0.00
MN	0.26	0.00	0.24	0.23	0.35
QN	0.16	0.00	0.16	0.16	0.24
SI	0.01	0.00	0.01	0.01	0.02

TABLE VI: Cluster Analysis of Traffic Accident Patterns

IV. Results

A. Unsupervised Clustering

When considering Cluster 0, I found that this cluster had the lowest mean temperature of 14.67°C and generally occurred during times that experienced precipitation. This cluster saw the lowest amount of clear day weather and had the highest average values for rain and snow. We can also see that this cluster is generally evenly distributed throughout the year with a slight lean towards the Fall months. At a more granular level, it can be observed that this cluster falls in line with a common trend of incidents beginning around morning commute time, rising throughout the day, peaking at evening commute times, and falling off into the night.

Cluster 1 follows similar daily temporal patterns as Cluster 0. However, this Cluster follows dramatically different temporal patterns throughout the year. This cluster favors the Summer months, particularly with June, July, and August occupying 47% of the sample space. It can also be shown that this cluster most significantly favors warmer weather with an average temperature of 20.44°C, the highest among the 5 clusters. This cluster experiences clear days and very little precipitation with 98% of incidents occurring during clear weather.

When considering more granular temporal trends, Cluster 2 exhibits some distinguishable properties. While 50% of the incidents occur between August and November and temperatures are lower (15.29°C), there are more differences available to observe at weekly and daily intervals. It can be observed that a majority of these incidents occur overnight between 6pm and 8am with 96% of incidents taking place during this time. This Cluster has a distribution that favors weekend incidents showing that 45% of these incidents take place between Friday and Sunday.

Cluster 3 appears to have an inverse of the daily temporal trends seen in Cluster 2. In this Cluster, 93% of the incident mass resides in the time between 8am and 6pm. This Cluster experiences entirely clear weather and favors warmer temperatures. Most notably, this Cluster has the highest weekly proportion of incidents with 77% of incidents occurring during the weekday between Monday through Friday.

In the final grouping, Cluster 4, the hourly temporal pattern falls in line with Cluster 0, and Cluster 1. This general temporal trend is the most common among the clusters. This cluster heavily favors the summer months, another trend that occurs most commonly among the clusters.

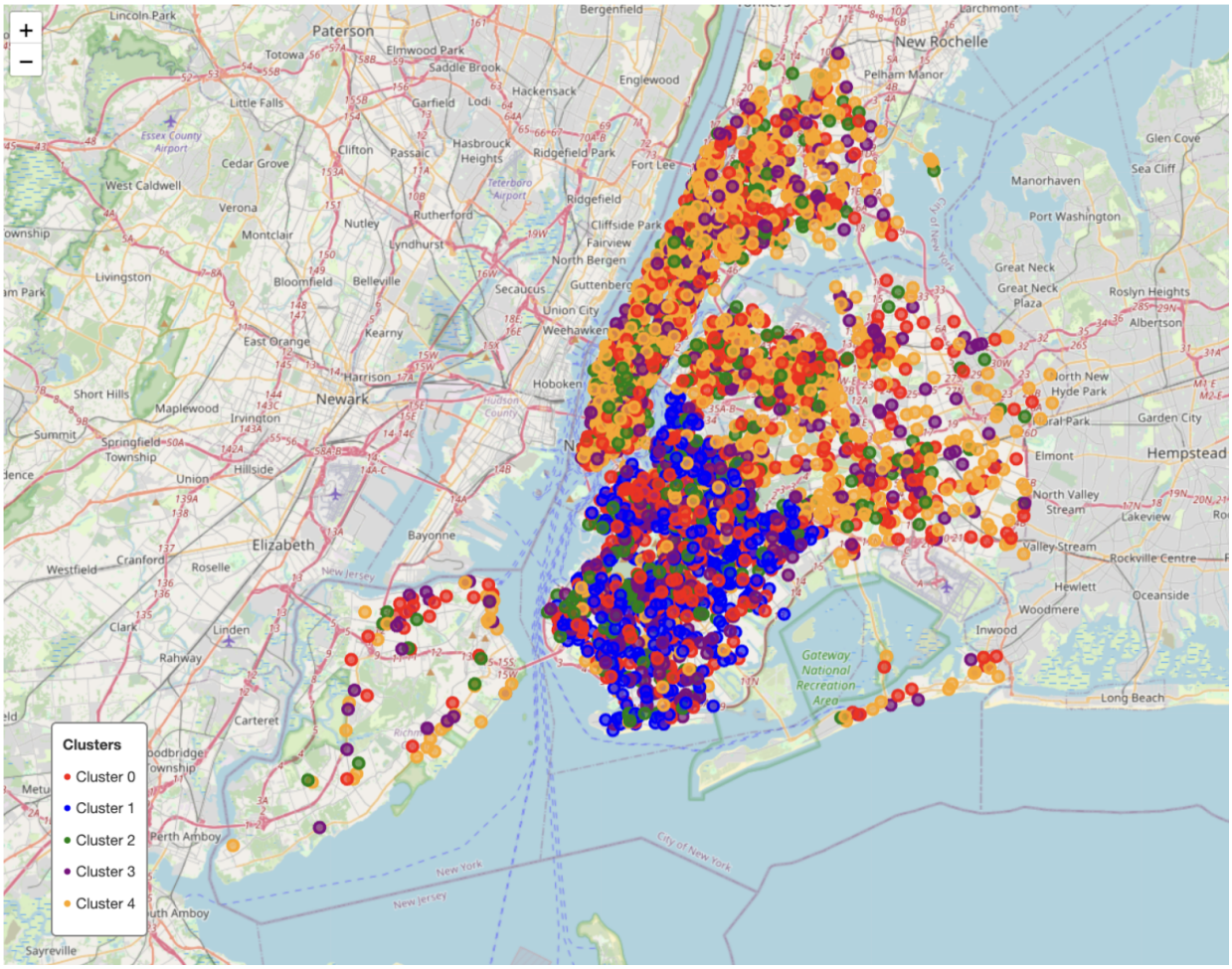


Fig. 6: Spatial Clustering.

The entirety of Cluster 1 is located within the Brooklyn borough. Cluster 0, Cluster 2, and Cluster 3 are spread out among the boroughs. Cluster 4 is spread out among the boroughs with an exceedingly limited presence in Brooklyn.

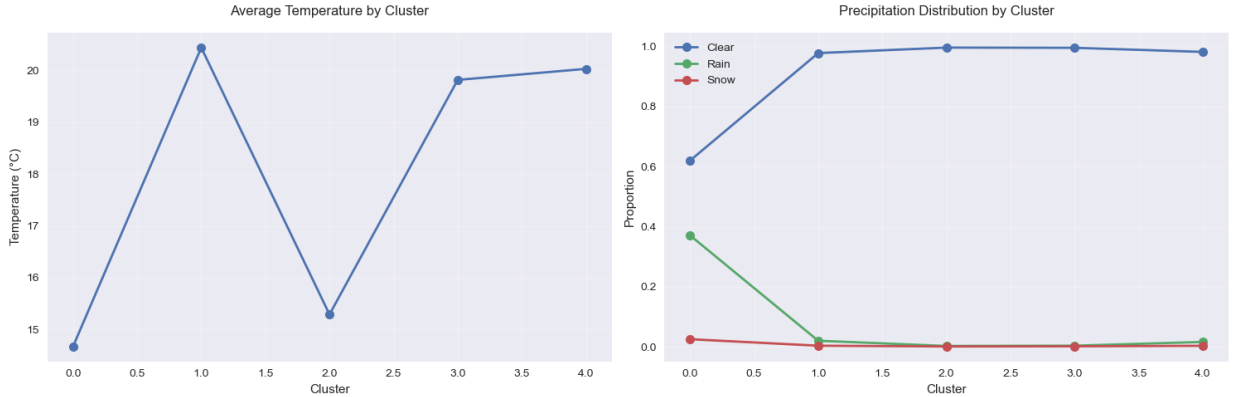


Fig. 7: Weather Clustering.

Cluster 0 has the lowest mean temperature. Clusters 1 has the highest mean temperature. Cluster 2 has a mean temperature similar to Cluster 0. Clusters 3 and 4 trend higher and are similar to Cluster 1's mean temperature. Precipitation is a significant factor for Cluster 0, with 40% of incidents involving some form of precipitation. Clusters 1 and 4 experience mostly clear weather with only 2% of incidents involving precipitation. Clusters 2 and 3 experience 100% clear weather.

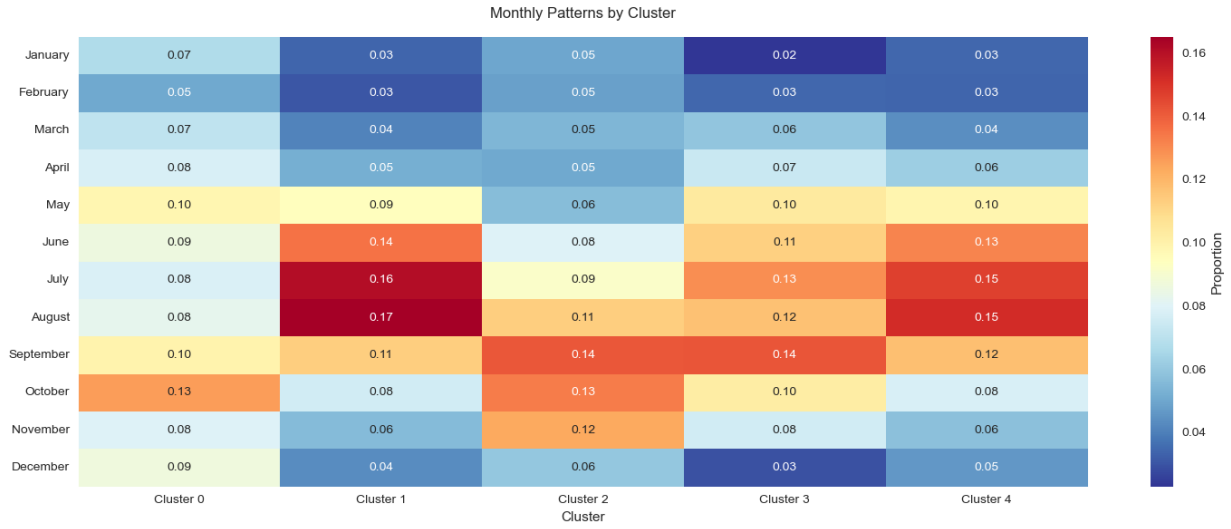


Fig. 8: Temporal Monthly.

Cluster 0 is the most evenly distributed temporally among the clusters. Clusters 1, 3, and 4 generally favor the summer months from May to September. Cluster 2 focuses on Fall incidents with a majority of the distributed mass occurring between August and November.

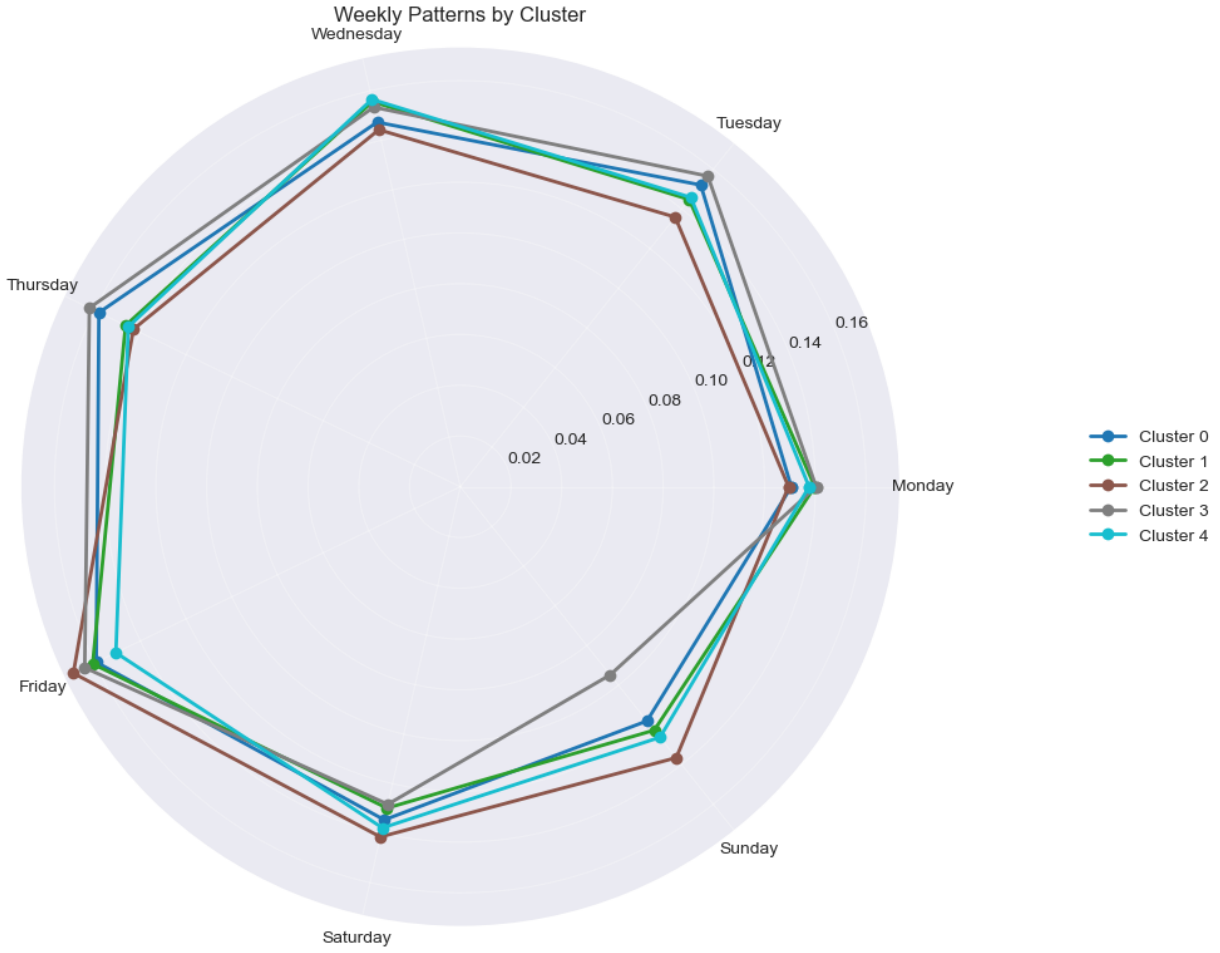


Fig. 9: Temporal Weekly.

All Clusters are generally evenly distributed among the days of the week. Cluster 2 has the largest weekend presence out of the 5 clusters with Friday, Saturday, and Sunday occupying 45% of the clusters distribution. in contrast, Cluster 3 has the least amount of distribution occupying the weekend, 37%.

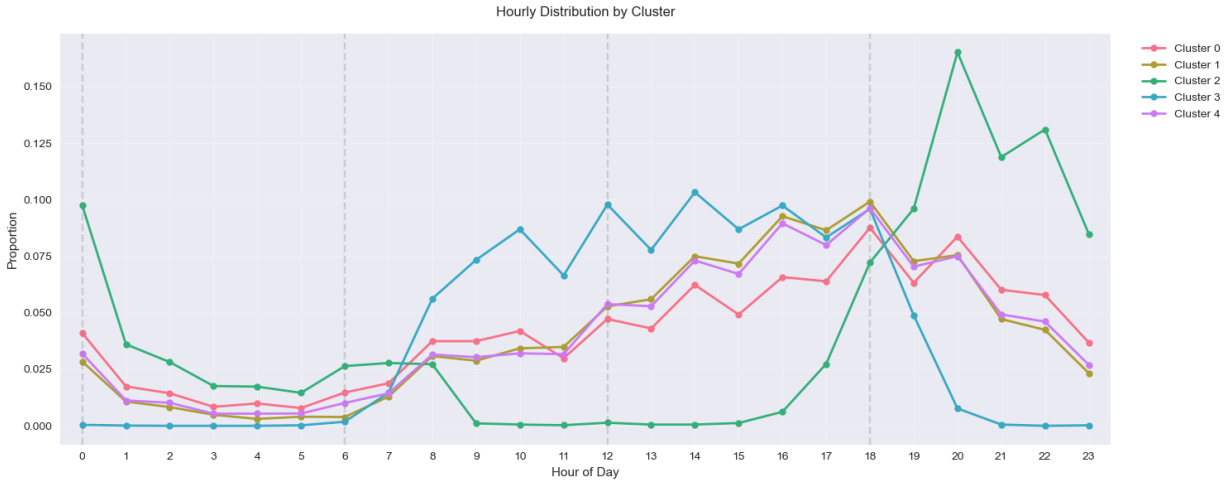


Fig. 10: Temporal Hourly.

Clusters 0, 1, and 4 follow similar trends. Each have a majority of events occurring in the middle 2 quartiles of the day, between 6:00 and and 18:00, with some residual distribution lying outside. Cluster 2 incidents occur mostly during the 1st and 4th quartiles of the day. Cluster 3 incidents occur mostly during the 2nd and 3rd quartiles of the day with a small amount of incidents occurring outside of this time.

Fig. 11: Top Two Principal Components.

This figure shows the top two principal components. It is shown that the clustering algorithm was able to differentiate these clusters as they are generally occupying distinct space within the plot. It can be observed that Cluster 1 and Cluster 4 have more overlapping space among these top 2 principal components than other Clusters.

B. Traffic Normalized Regression

The results of this portion of the study can be analyzed starting with the R^2 scores for the training set and the test set. For the training set, the observed R^2 score was *.643*. When calculating the R^2 score for the test set, the observed score was *.286*. The calculated RMSE was *.321*. These values and their significance are further analyzed in the results and limitation sections.

TABLE VII: H3 Hotspots by Severity

H3_INDEX	Severity	Incidents	Traffic Rate
8a2a100f048ffff	10.0	1	1.0
8a2a100a9a17fff	10.0	1	1.0
8a2a100daae7fff	5.0	1	1.0
8a2a100da35ffff	3.0	1	1.0
8a2a100d9437fff	3.0	1	1.0
8a2a100f3daffff	2.5	1	4.0
8a2a100f2ac7fff	2.0	1	1.0
8a2a100d2a8ffff	2.0	1	1.0
8a2a100aa8d7fff	2.0	1	1.0
8a2a10770c87fff	2.0	1	1.0

The most severe hotspots, given by the proposed severity algorithm, have low traffic rates and low incident rates. These severity score results indicate locations with death or multiple cyclist involvement. Geographically, these areas are spread out across the sample space.

TABLE VIII: H3 Hotspots by Traffic Volume

H3_INDEX	Severity	Incidents	Traffic Rate
8a2a100d20f7fff	0.001681	1	595.0
8a2a1072c8f7fff	0.002825	1	354.0
8a2a100d2c6ffff	0.003448	1	290.0
8a2a10725a77fff	0.003509	1	285.0
8a2a100d262ffff	0.007099	2	274.0
8a2a10725a5ffff	0.004000	1	250.0
8a2a100d2097fff	0.006315	2	231.0
8a2a100d26d7fff	0.004608	1	217.0
8a2a100d2717fff	0.004808	1	208.0
8a2a100d21affff	0.030684	2	205.0

The hotspots with the highest traffic, given by the traffic normalization algorithm, show the areas that received the most traffic during a given incident time interval. We see high traffic rates, but still relatively low incident counts, leading to a low severity score.

TABLE IX: H3 Hotspots by Incident Volume

H3_INDEX	Severity	Incidents	Traffic Rate
8a2a1008d567fff	0.310523	7	10.714286
8a2a100d2677fff	0.050644	7	54.428571
8a2a100d6d97fff	0.034806	6	34.000000
8a2a100da557fff	0.137692	6	10.500000
8a2a100d3797fff	0.202474	6	25.166667
8a2a1072cb5ffff	0.093795	6	12.666667
8a2a1072cad7fff	0.019809	6	81.000000
8a2a100d66b7fff	0.021732	6	87.000000
8a2a100decc7fff	0.021705	5	60.200000
8a2a1072d377fff	0.022878	5	52.000000

The highest incident hotspots show H3 indices that contain up to 7 incidents in the year 2023. The traffic rates for these areas vary from 11 cyclists per hour to 90 cyclists per hour. The higher incident counts show a trend towards a higher severity score.

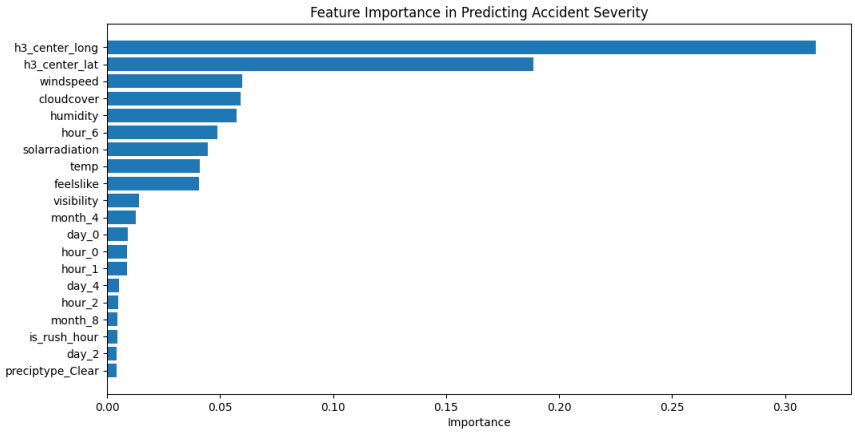


Fig. 12: Feature Importance.

This figures shows the top 20 features by importance. At the top of the chart we see that location (latitude, and longitude) have the highest impact. Weather features follow and begin to mix with temporal patterns. Among weather features windspeed, cloudcover, humidty, and temperature have the largest impact. Among temporal features, early morning, the day Monday, and the month April, have the highest feature impact.

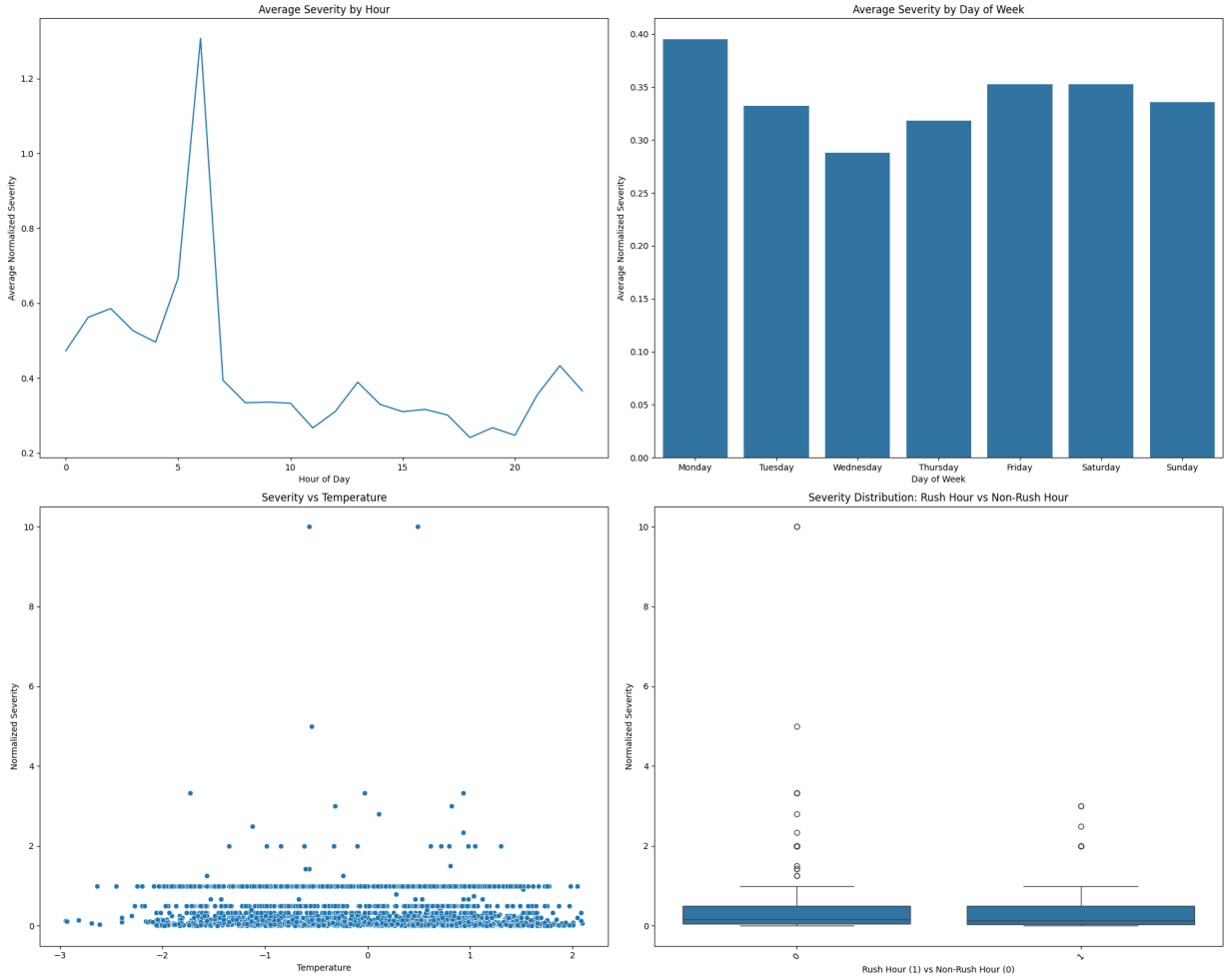


Fig. 13: Temporal Hotspot Patterns.

The study shows that normalized severity raises during the evening and morning hours peaking at 6:00. Throughout the week, the study shows a generally even distribution of severity with Monday having the highest average severity score. Severity and temperature seem to be fairly evenly distributed. We see slightly more incident weight with lower temperatures. Severity during Rush-Hour *hours: [7, 8, 9, 16, 17, 18]* shows increased incident severity.

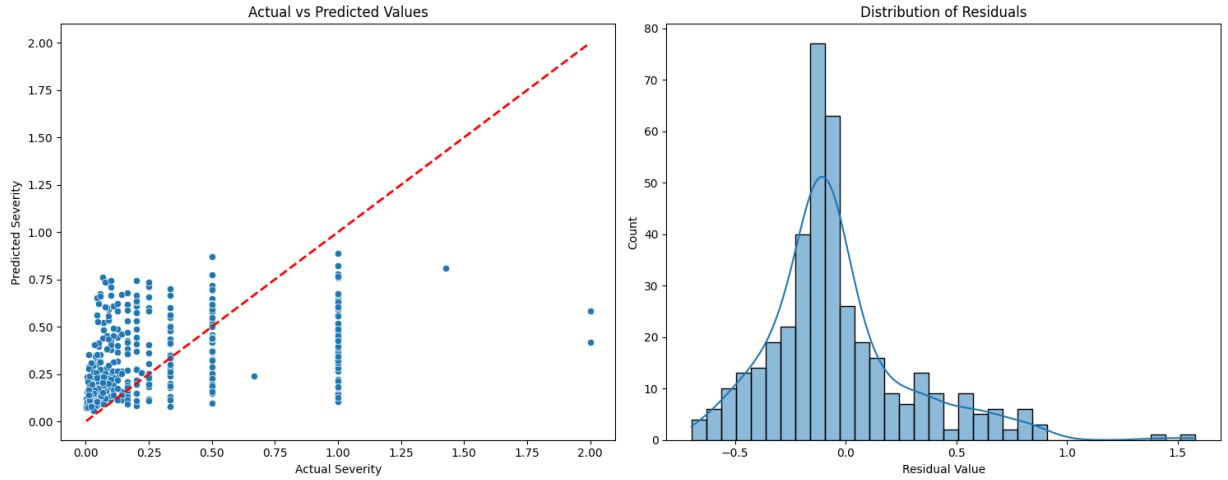


Fig. 14: Model Performance.

The actual vs. predicted scatter points are significantly far from the line. The model does not accurately predict high severity cases. While the distribution is roughly normal, the model has a slight positive bias. There are also some severe mispredictions, shown by the far right outliers. The model performs poorly at predicting severity. This is further analyzed in the discussion.

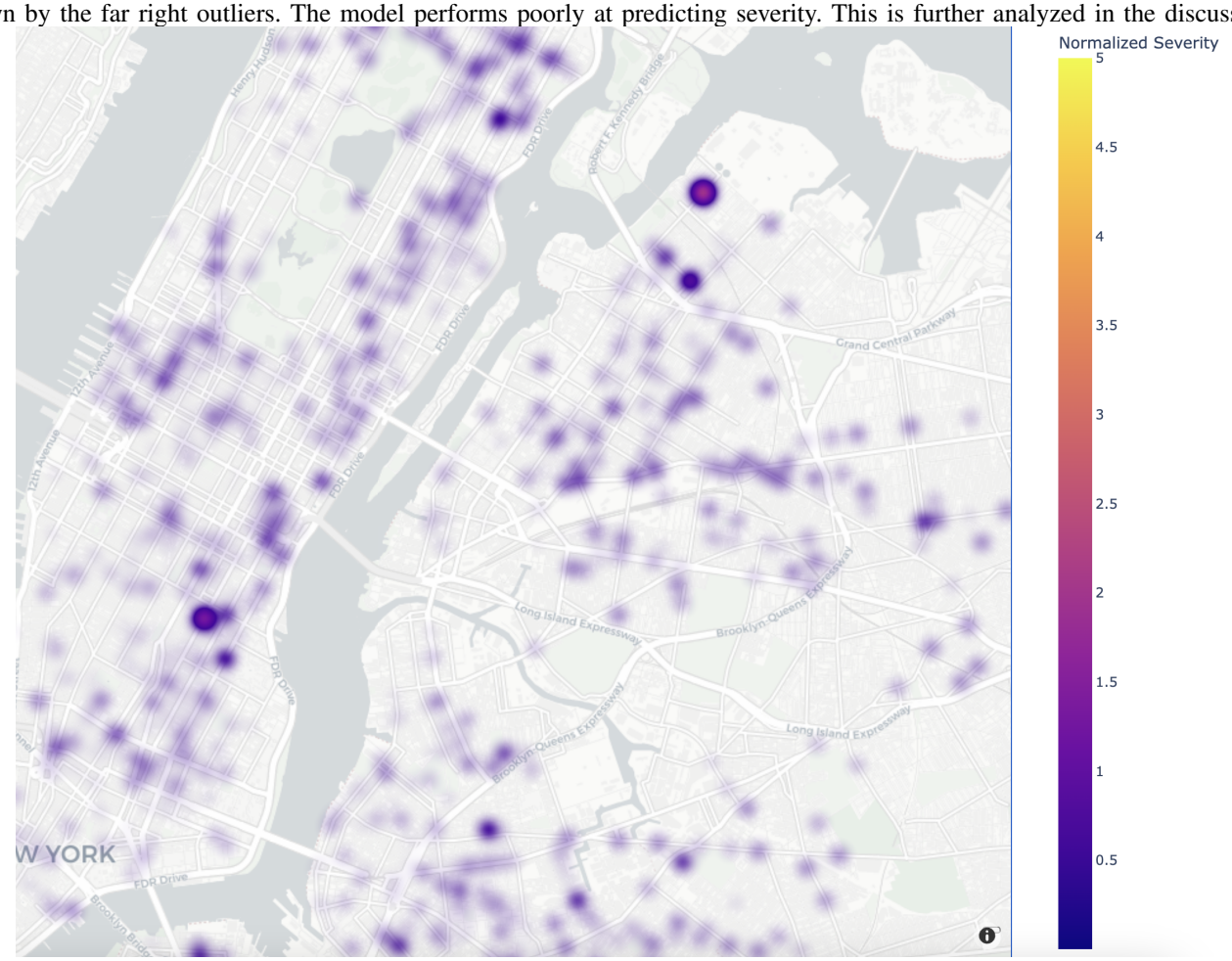


Fig. 15: Geographical Hotspots.

This figure visualizes the geographical areas of interest where the normalized severity was high. This represents areas where a disproportionate amount of cycling incident severity occurs. The significance of these results is further analyzed in the discussion.

V. Discussion

A. Significance of Findings

Using unsupervised clustering and traffic normalization, the study revealed several significant patterns associated with cycling accidents in New York City.

Temporally, the clustering exercise showed that Clusters 0, 1, and 4 have peak accident rates during evening commuting times. Additionally, Cluster 2 presented with a majority of accidents occurring overnight. This suggests that there are a different set of risk factors attributing to overnight accidents. We also see that Clusters 1, 3, and 4 have a high density of incidents occurring during summer months. This suggests that there may be a unique set of features associated with summer ridership that results in this type of accident.

The study shows that the weather during an incident is a significant risk factor suggesting that there are important environmental features associated with cycling incidents. Cluster 0 showed the highest correlation with precipitation having 40% of incidents occur during rain or snow. Conversely, all other groupings in the study had clear weather. Mean temperature varied drastically among the Clusters ranging from 14.67°C to 20.44°C. These results indicate that environmental features present patterns in incident occurrence.

Cluster 1 in this work is entirely contained within the borough Brooklyn. Other clusters in this study were distributed across the 5 different boroughs. This suggests that there may be geographically significant risk factors local to Brooklyn. For example, local infrastructure or traffic patterns unique to Brooklyn, could be related to accidents in this area.

In the traffic normalization portion of the study, this work reveals that the areas with the highest number of incidents are not necessarily the most risky cycling locations. When adjusting for traffic volume, it is shown that there are areas with a disproportionately high level of risk. These areas should be further examined and targeted for cycling safety improvements.

B. Comparison with Previous Research

This study confirms and extends some findings from previous research. This study provides a unique approach to traffic normalization for New York City that could be extending to provide further insights.

This study confirms the findings that there is an increased risk in cycling during peak or commuting hours. The same study promotes the idea that cyclists are at greater risk because of increased traffic volume and aggressive driving [16]. While this study is unable to comment on driver behavior, it is able to attribute traffic volume with an increased risk for cyclists. On top of these findings, this work extends previous findings by showing that a unique set of features exists for off peak incidents. In the unsupervised approach it is revealed that Cluster 2 occurs almost exclusively off peak. These incidents should be further examined to better understand why this type of accidents occurs.

Existing work reveals that environmental factors such as precipitation have an impact on cycling participation due to perceived dangers [15]. This study confirms that the perceived dangers are a valid concern. Cluster 0 from the unsurprised approach can be largely categorized by the precipitation feature. This suggest that rain or snow is significant in causing cycling accidents. City planners should consider educating or providing other safety measures to combat this type of accident from occurring.

This work extends the understanding of previous environmental and temporal findings. It is shown that warmer or Summer months are associated with a higher incident rate. It can be observed that there is a correlation temporally and environmentally. While the clusters in this study ranged in temperature, it was shown that a majority occur during warmer temperature.

This study creates a new avenue of study unique to New York City by created a traffic normalization method leveraging the Citibike trip data. This allows researchers to better understand the importance of traffic volume at a given time and location. While the study was unable to accurately predict severity, it was able to identify hotspots and their relation to traffic volume (Tables:VII VIII IX).

C. Limitations

When considering the accuracy and incident rates of cycling accidents in New York City, there are some limitations. One issue that arises is that cycling accidents are grossly under reported . In the case of minor incidents or accidents not requiring medical care, these incidents are often not recorded. We also understand very little about accident severity [5]. While many studies propose that personal safety equipment and speed are the biggest influences for cyclist fatalities, we do not have information about these features in our dataset. Fatalities also occupy an exceedingly small portion of our dataset making their occurrence appear more like noise preventing us from uncovering any meaningful trends. Traffic volume proved to be computationally expensive to produce and provides further limitation as mentioned below.

Analyzing trends such when performing unsupervised clustering without traffic normalization presents issues. For example, many of our clusters favored warmer weather. However, it can be shown with Citibike data that cycling participation trends higher during times of warmer weather. This correlation between ridership and warmer weather as well as the correlation between ridership and incidents can skew results by over weighting or under weighting the significance that weather has on an accident occurring.

There are also some obvious shortcomings involved with the traffic normalization methods used in this study. For one, the study only considers point to point shortest path traffic patterns between Citibike stations. Many cyclists in New York are not Citibike users and our incident data is not contained to these users. This limitation also limited the geographical area that I was able to study. While the incident data covers all of the 5 boroughs in their entirety, the Citibike traffic normalization is only able to cover the area where the Citibike network exists. This area includes all of Manhattan, parts of Queens, Brooklyn, and the Bronx, and none of Staten Island. This limits the amount of incident data that I was able to

study. Not only is this geographical area limited, but I also was not able to find a way to account for non-Citibike cyclists nor Citibike users not following shortest path routing. While traffic normalization is often omitted and I feel like this was a good step for future research to build upon, these limitations should not be ignored.

The model for predicting severity performed poorly in this study. When analyzing the R^2 scores for the training set and test set, it can be observed that the model is likely overfit. The training set does not retain enough of the variance and the test set is even worse. Several attempts were made to improve this model through feature engineering, model choice, and hyper parameter tuning. However, I was unable to achieve desirable results for model performance. I believe that this is because there are not that many fatal accidents which makes severity difficult to predict. Another cause could be that there is unpredictability within the dataset. For example, the data does not tell whether or not a cyclist is wearing protective gear which could have a unforeseen impact on accident severity. This data also does not include information about traffic speed which has been shown to be a major factor for incident severity in other studies [28]

D. Implications and Use Cases

This data could be used to help New York City be more mindful in their planning. By providing information about the discovered clusters, New York City could attempt to implement additional safety features that have proven successful in other studies.

As shown in the results, commuting times are proven as a significant risk factor for cyclists. New York could implement temporal based traffic calming features to limit the risk for cyclists. In our analysis we have highlighted some specific problem areas from 2023. These are areas where the risk for cyclists outweighs the traffic volume for the area. This means that these areas have a disproportionate amount of cycling incidents with respect to traffic volume and should be further studied to understand the root cause. City planners could also use the analysis here to alert Citibike users directly if they are participating in high risk cycling.

New York City could implement rules that require cyclists to increase their viability during weather or nighttime, a method that has proven successful in Australia. "Australia requires riders at night and in hazardous weather that may restrict visibility to have a flashing or steady white light visible for 200 metres from the front and rear of the bicycle, and a red reflector visible for at least 50 metres on the rear of the bicycle." [4]

Alerting users to significant time, weather, or geographical risks may help to create a safer environment for New York City cyclists.

VI. Conclusion

A. Future Expansions

Building upon traffic normalization to build a better incident severity and hotspot classifiers could yield a pathway for compelling use cases. If this research was expanded to be able to determine hotspots in real time with passable accuracy, it could be used in navigation applications to create navigation pathways for users that balance safety and efficiency. Many navigation applications today do not consider aspect of safety in their route calculations [29]. To explore this topic further, I would focus on creating a model to determine hotspots in real time, a process that has seen significant progress through the wide spread use of GPS devices in recent years [30]. From here I would explore different search based algorithms. For example, I may create an extension of an A* algorithm with a heuristic based search that weights a pathway that collides with a hotspot higher. In return, our cost to go down a troubled pathway would be higher and thus likely not taken by the algorithm. We could also have severity thresholds that create blockers in our navigation paths if the risk exceeds a threshold thus forcing the search based algorithm to avoid the area entirely. This algorithm could be tested by comparing incident data against a snapshot of data at that time. The study could then compare a regular shortest path algorithm with the extended heuristic based search to see which algorithm does a better job of avoiding areas with known collisions. The goal of this research extension would be to further enhance cycling safety by creating a navigation path that balances safety and efficiency.

B. Final Remarks

Overall, this research advances the study of cycling safety in New York City. This studies combination of several different data sources and machine learning approaches this work was able to uncover new areas of concern with respect to cycling safety. In the unsupervised clustering approach it was shown that there are various types of accidents that occur. City and urban planners can use this information to target the unique underlying causing. This study also presented a novel approach to traffic normalization in New York City. While predicting severity proved to be difficult, the results ultimately revealed specific hotspot locations that should be further analyzed to improve cycling safety.

Acknowledgment

I would like to thank Dr. Jiao, Professor Choi, and all of the Teaching Assistants for Case Studies in Machine Learning. This paper and these findings would not be possible without all of the hard work and dedication put into this course by the teaching staff. Thank you for all of the great learning opportunities.

REFERENCES

[1] W. Yu, C. Chen, B. Jiao, Z. Zafari, and P. Muennig, "The cost-effectiveness of bike share expansion to low-income communities in new york city," *Journal of Urban Health*, vol. 95, no. 6, pp. 888–898, Nov. 2018.

[2] C. H. Basch, D. Ethan, J. Fera, B. Kollia, and C. E. Basch, "Micromobility vehicles, obstructions, and rider safety behaviors in new york city bike lanes," *Journal of Community Health*, vol. 48, pp. 522–527, Feb. 2023.

[3] C. B. NYC, "Citi bike nyc system data," 2024, accessed: 2024-11-25. [Online]. Available: https://citibikenyc.com/system-data

[4] World Health Organization, "Cyclist safety: an information resource for decision-makers and practitioners," Geneva, 2020, licence: CC BY-NC-SA 3.0 IGO.

[5] C. C. Reynolds, M. A. Harris, K. Teschke, P. A. Cripton, and M. Winters, "The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature," *Environmental Health*, vol. 8, no. 1, Oct. 2019.

[6] D. Castells-Graells, C. Salahub, and E. Pournaras, "On cycling risk and discomfort: urban safety mapping and bike route recommendations," *Computing*, vol. 102, pp. 1259–1274, Dec. 2019.

[7] J. Pucher and L. Dijkstra, "Promoting safe walking and cycling to improve public health: Lessons from the netherlands and germany," *American Journal of Public Health*, vol. 93, no. 9, pp. 1509–1516, Sep. 2003.

[8] N. Y. C. D. of Transportation (NYC DOT), "Nyc dot press release: Pr15-096," 2015, accessed: 2024-11-25. [Online]. Available: https://www.nyc.gov/html/dot/html/pr2015/pr15-096.shtml

[9] S. D. S. Fraser and K. Lock, "Cycling for transport and public health: a systematic review of the effect of the environment on cycling," *European Journal of Public Health*, vol. 21, no. 6, pp. 738–743, Oct. 2010.

[10] P. L. Jacobsen, "Safety in numbers: more walkers and bicyclists, safer walking and bicycling," *Injury Prevention*, vol. 9, no. 3, pp. 205–209, Sep. 2003.

[11] T. Ahmed, A. Pirdavani, G. Wets, and D. Janssens, "Bicycle infrastructure design principles in urban bikeability indices: A systematic review," *Sustainability*, vol. 16, no. 6, p. 2545, Jan. 2024.

[12] D. Rosenfield, P. Fuselli, and S. Beno, "Improving cycling safety for children and youth," *Paediatrics & Child Health*, vol. 29, no. 5, pp. 324–328, Aug. 2024.

[13] S. Daraei, K. Pelechrinis, and D. Quercia, "A data-driven approach for assessing biking safety in cities," *EPJ Data Science*, vol. 10, no. 1, Mar. 2021.

[14] H3 Development Team, *H3 Documentation*, 2024, accessed: 2024-11-26. [Online]. Available: https://h3geo.org/docs

[15] H. Wu *et al.*, "Can infrastructure, built environment, and geographic factors negate weather impact on cycling?" *Journal of Transport and Land Use*, vol. 17, 2024, accessed: 2024-11-25. [Online]. Available: https://www.jtlu.org/index.php/jtlu/article/view/2318

[16] L. Ayad, H. Imine, C. Lantieri, and F. De Crescenzio, "Pedal towards safety: The development and evaluation of a risk index for cyclists," *Infrastructures*, vol. 9, no. 1, p. 14, 2024. [Online]. Available: https://www.mdpi.com/2412-3811/9/1/14

[17] New York City Open Data, "Motor vehicle collisions - crashes," 2024.

[18] Visual Crossing, "Weather data," 2024, accessed: 2024-11-26. [Online]. Available: https://www.visualcrossing.com/weather-data

[19] Citi Bike NYC, "Citi bike nyc," 2024, accessed: 2024-11-26. [Online]. Available: https://citibikenyc.com/

[20] Folium Development Team, "Folium documentation," 2024, accessed: 2024-11-26. [Online]. Available: https://python-visualization.github.io/folium/latest/

[21] S. learn Developers, "Scikit-learn: Machine learning in python," 2024, accessed: 2024-11-26. [Online]. Available: https://scikit-learn.org/

[22] D. A. S. Aric A. Hagberg and P. J. Swart, "Networkx: Network analysis in python," https://networkx.org/, 2008. [Online]. Available: https://networkx.org/

[23] O. contributors, "Openstreetmap: The free wiki world map," https://www.openstreetmap.org, 2024. [Online]. Available: https://www.openstreetmap.org

[24] M. Developers, "Matplotlib: Visualization with python," https://matplotlib.org, 2024.

[25] S. Developers, "Seaborn: statistical data visualization," https://seaborn.pydata.org/, 2024.

[26] D. Becker, "Using categorical data with one-hot encoding," https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding, n.d.

[27] I. T. Jolliffe and J. Cadima, "Principal component analysis: a review and recent developments," *Adaptive data analysis: theory and application*, vol. 374, no. 2065, 2016. [Online]. Available: https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202

[28] World Health Organization, "Save lives: A road safety technical package," 2017, accessed: 2024-12-01. [Online]. Available: https://www.who.int/publications/i/item/save-lives-a-road-safety-technical-package

[29] K. Hübner, B. Schünemann, T. Schilling, and I. Radusch, "On assessing road safety aspects of a cycling router application," in *2017 15th International Conference on ITS Telecommunications (ITST)*, 2017, pp. 1–8.

[30] M. Bíl, R. Andrášik, and Z. Janoška, "Identification of hazardous road locations of traffic accidents by means of kernel density estimation and cluster significance evaluation," *Accident Analysis & Prevention*, vol. 55, pp. 265–273, Jun. 2013.