Reducing Cycling Deaths in NYC: A Causal Analysis and Model for a Real-Time Dashboard

Table of Contents

EXECUTIVE SUMMARY **2**

1. INTRODUCTION **3**

2. CONTEXT: CYCLING IN NEW YORK **3**

3. LITERATURE REVIEW **6**

4. METHODOLOGY **9**

5. MODEL BUILDING **16**

5.1 LINEAR REGRESSION – BOROUGH LEVEL 16

5.2 LOGISTIC REGRESSION 17

5.3 LINEAR REGRESSION MODEL – PREDICTIVE ANALYSIS 18

6. CONCLUSIONS AND NEXT STEPS **21**

Reducing Cycling Deaths in NYC: A Causal Analysis and Model for a Real-Time Dashboard

Civic Analytics Project – Team 11

Executive Summary

New York City has an ambitious agenda to reduce traffic accidents which kill or injure more than 4,500 citizens each year. In support of the Mayor's 'Vision Zero' initiative, this project develops a predictive tool to identify expected numbers of traffic accidents by census tracts at different times of day. The tool specifically addresses the risk of cycle accidents.

Two models were constructed. A logistic regression model was developed to predict whether accidents are likely to result in injury or death. Making use of weather, time, and location data, the model shows a 79% accuracy rate. Extending the analysis to develop a tool designed for real-time predictions based on API data, a multiple linear regression model for expected number of injuries per census tract was developed. The model demonstrates strong effects of hour-of-day on predicted injury levels: holding other variables constant, with accident risk 25% higher at 5pm-6pm holding other variables constant. Positive linear relationships are also identified between (i) wet weather, and (ii) low visibility weather and accident rates.

The use case for the model is as follows: the NYPD and NY Department of Transport can employ the model through a web interface to identify predicted accident rates by census tract at the current time of day, month of year, weather conditions, and recent 311 complaints history of the census tract. Based on the predicted accident volume, the agencies could direct additional traffic police to the highest risk areas. The predicted accident levels could also be used to prioritize investment in better street furniture or safer road layouts where predicted accident levels are high.

Several limitations are noted, including the need to fully integrate 311 data based on more thorough feature selection, and to conduct census-level analysis for a larger number of census tracts. The model also needs to be retrained periodically based on up-to-date accident data to incorporate changes to the built environment – eg. construction of cycle lanes – that may make the areas more or less dangerous. The tool nevertheless represents a useful model of predictive analysis for the city agencies based on data that can be streamed from APIs in real-time.

1. Introduction

Each year in New York, more than 250 citizens and killed and 4,000 injured in traffic accidents. The city government is pursuing a 'Vision Zero' initiative that aims to drastically reduce this number across all accident categories – including those where cyclists are involved. Currently 5 agencies are involved in more than 60 actions to bring down road deaths[1]. However, despite there being plentiful data on road accidents, the city lacks predictive tools to help allocate resources to where they can make the greatest difference.

To contribute to the Vision Zero initiative, this project will address two questions:

1. What factors contribute to deaths or serious injuries among cyclists?

2. How can New York's enforcement resources, specifically the deployment of traffic policemen, be optimized to reduce cyclist deaths?

The project will proceed in two stages. First, a multi-factor regression analysis will be conducted to determine which variables most strongly influenced the likelihood of cyclist-involved accidents. Secondly, the team will design the model for a real-time dashboard for the Department of Transport (DOT) and NYPD. The purpose of the dashboard would be to predict likely levels of cyclist injuries and fatalities at high-risk intersections, allowing the agencies to deploy more officers when cyclists are likely to be injured and conserve their resources when injuries are unlikely.

2. Context: Cycling in New York

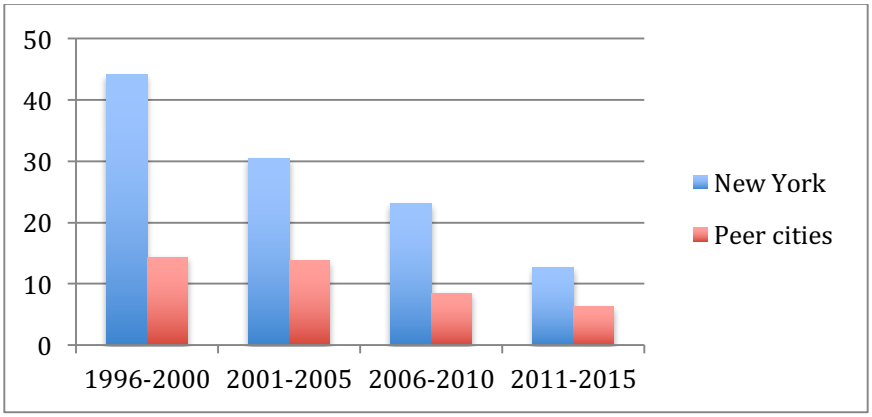
Regarding bicycles, New York has done some powerful advances in the last decade, firstly building cycle lanes, and secondly the 'Citibike' system. These advances have augmented the number of cyclists in the city: the DOT notes that daily cycle trips reached 450,000 in 2015 compared with 240,000 in 2009. In the same period, the number of people using a bicycle rose from 521,000 to 778,000.

The number of cycling deaths has remained flat since 2000, at about 12 to 24 per year, despite the large growth in trip numbers. This represents an impressive 71% decline in the cycle accident rate, measured by casualties per million rides[2]. Nevertheless, the cycle fatality rate in New York is double that of a group of peer cities (Fig 1). Compared with this peer group of other large and dense US cities including Boston, Philadephia and Seattle, it remains a dangerous place to ride.

---

[1] http://www.nyc.gov/html/dot/html/bicyclists/bike-ridership-safety.shtml
[2] NYC DOT (2016). Safer Cycling: Bicycle Ridership and Safety in New York City. link

Figure 1: Cycling Deaths per 1 million journeys



Source: NYC DOT, Safer Cycling (ibid.)

Factors affecting cycle accidents

Much is known about the factors that cause bike crashes, and New York City has major investments and policy initiatives underway to address these (Box 1). The challenge for this project is to find practical ways to improve our understanding of bike crashes, and ultimately to reduce them, by using urban informatics.

In describing the problem of bike fatalities in New York, several points stand out from the city and academic literature:

1. Most fatalities happen at intersections. Accident records suggest that 65% of New York's cycling fatalities and 89% of serious injuries take place at intersections[3]. As cycle safety improved in the past decade, the percentage of fatalities taking place at intersections actually rose.

2. Cycle accidents follow seasonal/temporal patterns. Crash frequency differs systematically by time of day, with 6pm-9pm seeing most crashes. The temporal distribution differs by season and on weekdays vs. weekends (Fig 2).

3. Traffic and weather affect accident risk. Studies from California and Italy identify statistically significant relationships between environmental variables - such as slippery roads after a rain storm - and accident risk. Underlying traffic volumes also strongly affect the number of accidents[4].

4. Cycle lanes reduce accident risk but are not ubiquitous. The city has built bike lanes and improved street layouts to cut fatalities at identified high-risk locations. A

[3] The Transportation Research Board notes that 'gridiron' street layouts, common in New York, are prone to much higher accident rates than city geographies with more arcs and curves.

[4] Prati G, De Angelis, M, Marín (2017). Characteristics of cyclist crashes in Italy using latent class analysis and association rule mining. PLoS ONE 12(2). (link)

DOT study found the fatality/injury rate fell by 35% or more on 6 out of 8 corridors where separated bike lanes were introduced. However, improvements have not been made everywhere due to funding constraints or difficult road layouts, leaving some locations at higher risk due to physical and layout factors.

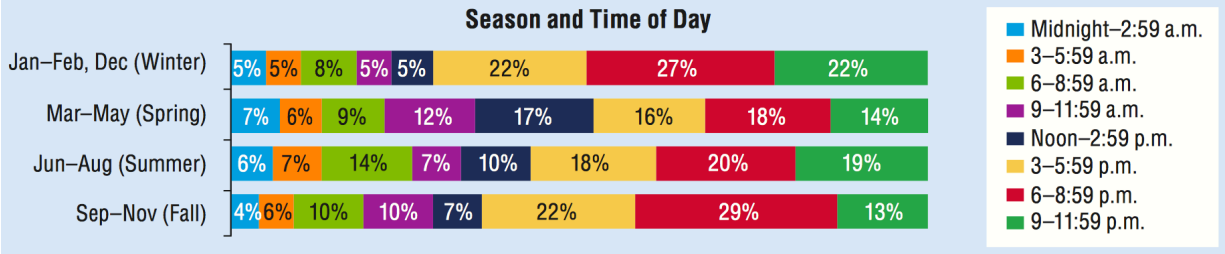
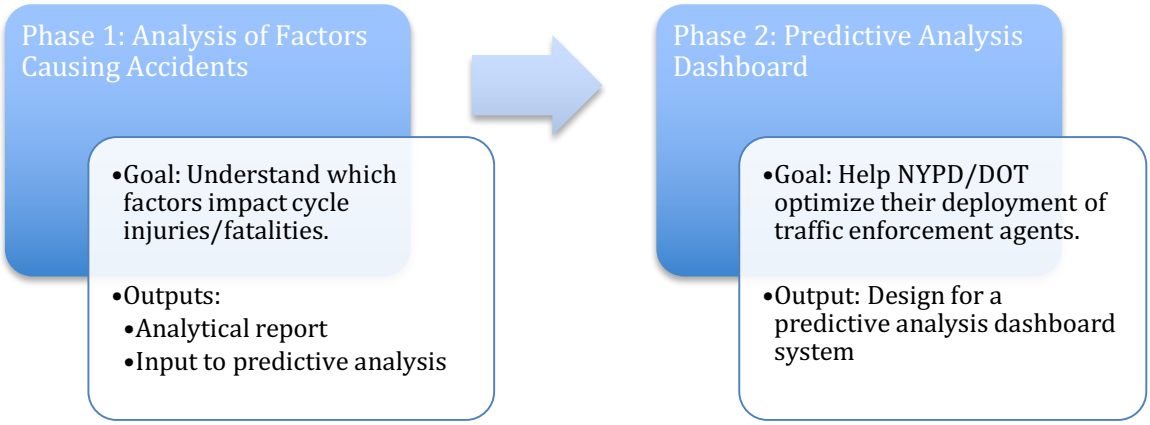


Figure 2: Distribution of US cycle accident fatalities in 2015 by time/season/day. Source: Highway Traffic Safety Administration (link).

Study approach

The project aims to support the city government's Vision Zero initiative in reducing fatalities and serious injuries affecting cyclists.

Figure 3: Goals and outputs of the project's two phases



A model for the dashboard was designed as follows:

- The dashboard model should be usable by NYPD and DOT in a web browser.
- It should be able to pull weather, traffic and incident data (as well as time of day and year) in real-time.
- The tool should be trained on the 2012-2016 data. Based on the relationship between dependent variables and the independent variable (cyclist serious injuries/fatalities), the dashboard would indicate predicted injuries/fatalities for the current one-hour period (incident per million cycle journeys).

Box 1: **What is NYC currently doing in this area? Summary of latest Vision Zero commitments:**

- Continue building at least 50 miles of bike lanes per year
- Roll out new approaches to intersection design and evaluate their effectiveness
- Focus and deploy enforcement resources to intersections with high rates of cyclist deaths/fatalities
- Tailor enforcement to address dangerous driver and cyclist behavior
- Improve education and outreach
- Expand bicycle count data

Source: Safer Cycling: Bicycle Ridership and Safety in New York City (link)

3. Literature Review

The need to reduce road traffic accidents has motivated an extensive literature that comes from several academic and policy communities: transport engineers, public health specialists, public administration and urban planning specialists, and government agencies such as state or city Departments of Transport. Given our research interest around using big data for dynamic prediction of likely cycle accidents, we searched for and identified relevant studies that examined several questions, in particular: (i) what causes road traffic accidents; (ii) factors affecting their severity; (iii) methods to identify accident hotspots; and (iv) effectiveness of counter-measures.

Physical factors promoting accidents

Studies of cycling accidents are a subset of a larger literature on the causes of road traffic accidents between motor vehicles and other road users. Scholars examining varieties of urban street layout have concluded that certain common types, such as the 'gridiron' form found in Manhattan and the roundabout form used in European cities, are associated with higher accident frequency (Rifaat et al, 2010). Collisions involving cyclists and pedestrians commonly take place at intersections, whereas street layouts with warped parallels and loops create fewer opportunities for accidents of this type.

Specific causal analysis of accidents is best developed in the transport engineering literature. In the United States, the Federal Highway Administration (FHWA) categorizes collisions into 38 categories based on their physical configuration (FHWA, 2014). Transport engineers model accident avoidance in terms of:

- **perception reaction time**, the time taken to realize a collision is imminent and take corrective action; and
- **maneuver time**, the time required to brake or take other evasive action.

Calculations of the necessary reaction times are embedded in standards for road layout and signage implemented by bodies such as the FHWA, the New York Department of Transport (DOT) and the American Association of State Highway and Transportation Officials (AASHTO), as well as in design standards for cycle lanes. This approach is important for our study as it highlights the causal mechanism of accidents which can be exacerbated by environmental conditions. Where visibility at intersections is low or signage is misleading, both reaction times may increase, while they may also vary systematically with weather conditions (eg. slippery roads caused by rain) and light (eg. increased perception time at dusk).

An important observation from the literature is the non-linearity of cycling accidents with respect to number of cycle journeys in a city. As the number of cyclists increases, the accident rate per journey is observed (across large numbers of cities examined) to decrease. This is ascribed to behavior modification among motorists when they expect or experience [more] people walking or cycling" (Jacobsen 2003). This is good news to those who wish to encourage cycling but felt held back by concerns about high injury rate (Elvik 2009).

City-specific studies of accident risk

Since the 1990s, heightened policy interest in promoting cycling and walking as non-polluting and health-enhancing forms of urban mobility has prompted a large number of city-specific studies. Using police-reported bike accident data from 2003 to 2010, a study in Berkeley shows that bike routes with lower motor vehicle traffic had statistically significant lower bike collision rates than the parallel main traffic roads (Minikel et al, 2010).

In Berkeley, the researchers have studies that the bicycle boulevards can provide cyclists a safer choose to ride. It has implemented a network of 7 bicycle boulevards spanning the city. The boulevards are heavily traffic calmed with a combination of chicanes, traffic circles, traffic diverters and barriers, and speed humps. Berkeley's citywide speed limit of 25 mph applies to all of the bicycle boulevards and arterials. The researchers used average daily traffic (AADT) of 5440 for Milvia St. compared to 27,599 on Shattuck Ave. and 23,502 on Martin Luther King Jr. Blvd.

Besides street design, street condition is also a contributing factor to bike accidents. According to a study in Umea, Sweden, half of the single bike accidents were caused by physical road surface defects and winter road maintenance significantly reduced bike related injuries (Nyberg et al. 1996). Street visibility is very important for biking safety as well. Wanvik used a before and after comparison to show that street lighting reduced bike accident injury risks by 60% and this effect on injury reduction increases in darkness (Wanvik P, 2009).

There is extensive evidence of cycle collisions following regular patterns. The City of Denver mapped out the time of day of crashes. 85% of collisions occurred at

intersections and 15 at mid-block locations. Most collisions between 5pm-6pm (City of Denver, 2015). New York has found that crashes rise after daylight savings time comes in, since cyclists are not accustomed to suddenly earlier dusk and poor light conditions. The city's traffic analysis showed that between 6 p.m. and 7 p.m. during the week — the prime commuting time — the weekday hourly average rate of severe injuries and fatalities involving pedestrians rose to 2.44 in mid-December, or nearly triple the average rate of 0.84 in August (NYC DOT 2016).

Study methodologies

In order to get better understanding of our data and justify our choice of methodology, data to be collected, instruments to be used, we need to review the different methodological approaches that has been used in this field to investigate similar questions.

Studies have included multivariate linear regression, logistic regression investigating categories variables such as severe crashes and fatalities (Gabriele et al 2013, Cai et al 2016), decision trees, and machine learning techniques such as K-means (Mohamed et al 2013). The literature employs a variety of techniques depending on whether they are investigating absolute crash numbers, likelihood of categorical variables such as crash severity, or of ordered injury severity levels.

We also learn from a study about characteristics of multimodal conflicts in Manhattan and Brooklyn boroughs of how to identify expected high-conflict locations through geographic analysis in ArcGIS(Conway et al., 2013). This study used bivariate correlation analyses to select useful variables to predict multimodal conflict frequencies, and conducted statistical tests to evaluate the specific characteristics of high-conflict areas.

Some of the most sophisticated studies categorize crashes into different clusters and use techniques such as K-means and latent class analysis to trace likelihood of specific typologies to underlying factors (see box).

Based on our review, we found many studies employing police accident datasets and using GIS together with multi-variate analytical techniques to identify key influencing factors. As yet, there are few examples of applying datasets on these regressors in real-time to provide predictive insights on future accidents. Nevertheless, the City of Boston provides one such example. The Citywide Analytics Team, which reports to the Mayor, has made available map visualizations of accident location together with crowdsourced information on site-specific factors that citizens believe make specific locations unsafe (City of Boston 2017).

In conclusion, we find an extensive literature from urban planning and transport engineering scholars, city agencies such as New York, Boston and Denver, and community groups aiming to improve safety. Studies consistently find statistically significant predictors of crash likelihood including time and weather, and presence

or absence of safety features. The literature confirms the viability of our project, suggests additional factors to consider such as lighting and weather, and highlights the future scope for big data approaches such as our own to help operationalize safety research techniques in a helpful real-time form that can support city government decisions.

4. Methodology

4.1. Outline

As outlined in the introduction and literature review, our project aims to help New York City officials achieve the Vision Zero goal of reduced cycle deaths through predictive analytics. We approach this by building a tool that gives predicted levels of accidents for specific geographic areas of the city based upon live data streams. Our prototype model will use a limited number of variables that are nonetheless expected to contain significant information about the distribution of cycle accidents.

We initially include four groups of variables:

- time, at the level of month, day of week and hour of day;
- weather, including precipitation and temperature;
- 311 data complaints relevant to road conditions; and
- traffic safety features such as presence of bike lanes.

We also adopt an approach to normalizing cycle accidents by estimated number of cycle journeys, making use of Citibike data and bike counts to get estimates at a suitable spatial and temporal scale. The sections below outline the data used, give descriptive statistics, and set out the proposed methodology.

4.2. Data

The project uses the bicycle collision, weather, census tract, safety features / 311 complaints, and Citibike data described below.

Cycle accidents

We utilize the Motor Vehicle Collision dataset. This is compiled by the NYPD and maintained by the Department of Transport (DOT). It contains reported motor vehicle collision location, time, number of injuries or fatalities, and causal factors as reported by the officers. In spatial terms, the incidents are mapped to the nearest intersection. The complete dataset contains 1,149,041 entries from 2012 to 2017. We carried out several data-cleaning steps:

- Because of the geospatial nature of our analysis, we drop all rows with geospatial locational data missing.
- A dummy variable for cyclist involvement was created, and incidents without bicycle involvement are dropped. November 21, 2017

- The cleaned dataset contains 25,725 entries (2.23% of the original data points). The following points are observed from Figures 1-4:
- The peak of bicycle accidents takes place during the summer months of June to October.
- There is also a pronounced daily pattern of accident distributions, which peak between 7pm-9pm.
- Accident volumes are higher in Manhattan and Brooklyn, in line with higher cycle traffic density in these Boroughs.
- Most frequent contributory factors reported by NYPD officers are physical vehicle condition, behavioral factors, and other.

Weather

Weather data for New York was downloaded from open source site Weather Underground (www.wunderground.com).

The raw data contains a column for 'condition' and columns with readings on temperature, precipitation, wind speed and direction, amounts of precipitation, air pressure, and other items. For our purposes, the condition column is particularly relevant. It contains categorical data with the following values:

We carry out several data cleaning steps:

- The data is merged into a single large dataframe, and columns converted into the necessary types (strings to datetime, and numeric to float).
- Dummy variables are created for the 15 weather conditions as shown in Figure 4.

```
df1231_dummy = pd.get_dummies(df1231.Cond,prefix = '',prefix_sep='')
df1231_dummy.head()
```

	Clear	Fog	Haze	Heavy Rain	Heavy Snow	Light Freezing Fog	Light Freezing Rain	Light Rain	Light Snow	Mist	Mostly Cloudy	Overcast	Partly Cloudy	Rain	Scattered Clouds	Snow	Unknown
0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Figure 4: Construction of Dummy Varies for Weather Condition 1

Weather Condition: Categories Overcast', 'Partly Cloudy', 'Mostly Cloudy', 'Scattered Clouds', 'Clear', 'Haze', 'Light Rain', 'Rain', 'Fog', 'Unknown', 'Light Snow', 'Snow', 'Heavy Rain', 'Heavy Snow', 'Mist', 'Light Freezing Rain', 'Light Freezing Fog'

The weather data has seasonal patterns as shown in Figure 5 for four variables of potential interest: temperature, humidity, visibility and wind speed. We note that the literature review showed potential links between weather and personal safety behavior such as wearing helmets. Also, weather is likely to influence the perception and reaction time that plays a key role in individual cycle accidents, including

through difficulties seeing vehicles when visibility is low, or longer maneuver times by both motor vehicles and cyclists when roads are slippery. As such, the variation and seasonal regularities revealed by the weather data below are highly likely to contain valuable predictive insight into accident likelihood.

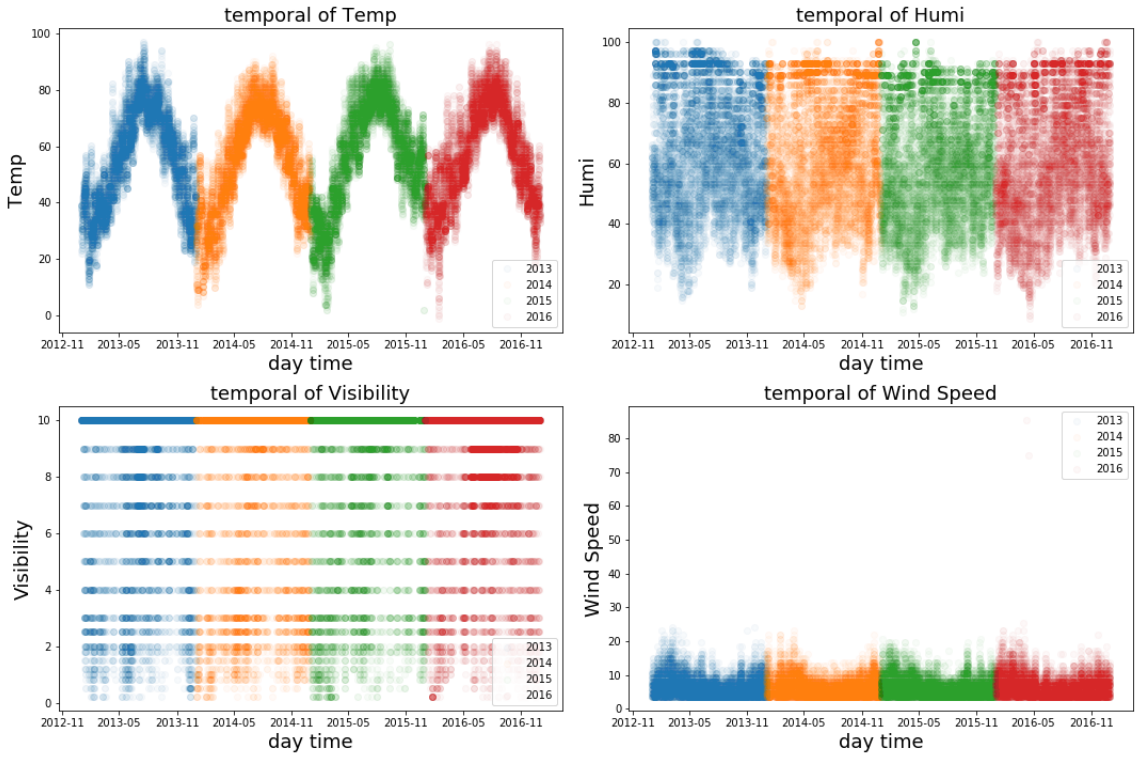


Figure 5: Seasonal Change in Key Weather Variables, 2012-2016 November 21, 2017

Census tract

Census tract data for New York City was downloaded from Open Data. We merged the census data with collision data and used Carto to visualize it. In Figure 6, the census tracts with purple color stands for the low-low clusters, which means in this area, collisions counts are low, however in the dark blue area, mostly the Manhattan midtown and downtown, shows high-high clusters, means larger number of collisions happens in those area.

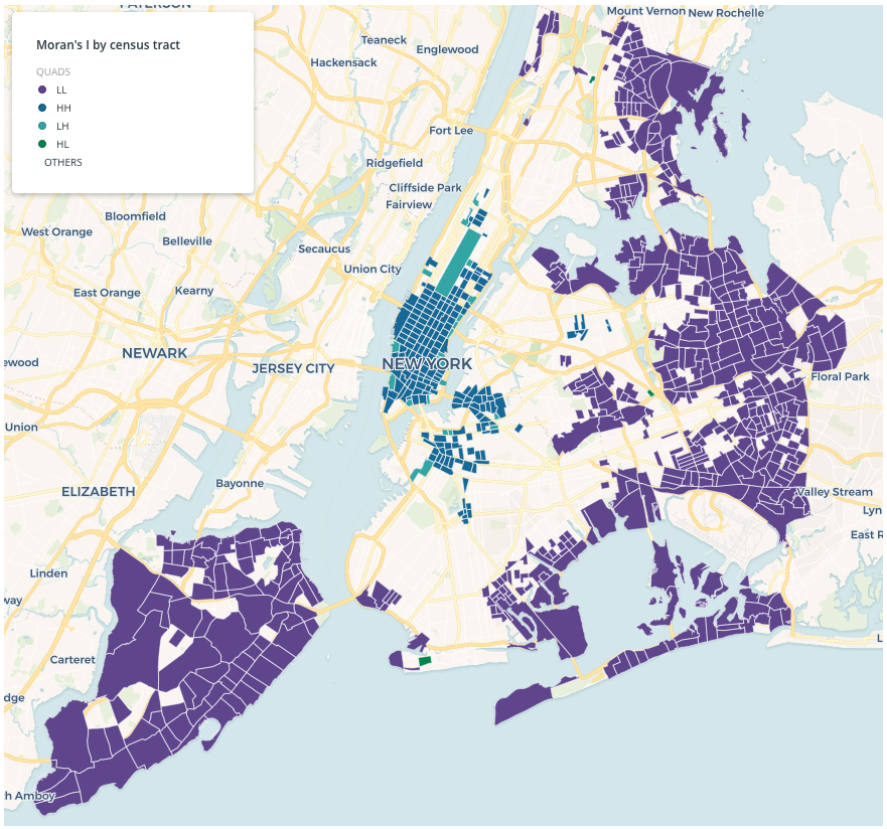


Figure 6: Collision counts clusters map

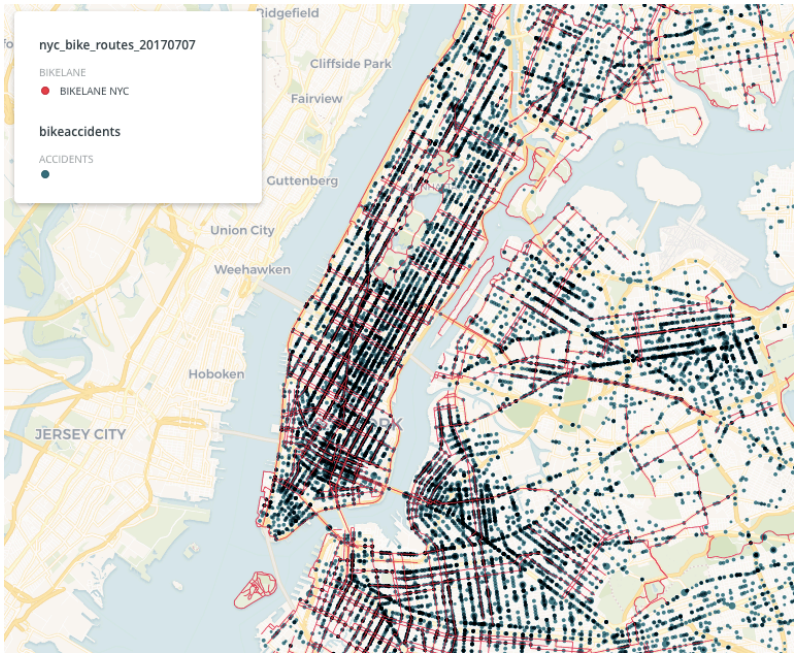


Figure 7: Bike Lane Distribution Data on the presence of bike lanes is available from DOT shape files.

This was mapped as shown in Figure 7. Red lines represent bike lanes and black dots represent the presence of one of more accidents. We note that many accidents

occur on road where bike lanes are present, as expected given that these are heavily trafficked routes, but many accidents also occur where bike lanes are not present. The figure suggests that exploring the influence of bike lanes on accident rates per cycle trip will be valuable. Thus our use of a normalization index will be important to identify accident hotspots on a per trip basis as well as an absolute basis. 311 Complaint types to be incorporated include damaged highway signs, traffic complaints including blocked streets, trash not picked up, and sanitation complaints. The mixture of 311 complaint types to be incorporated will be optimized based on the regression analysis.

Normalization factor: Citibike journeys

We use Citibike data for the purpose of normalizing accident volumes with regard to overall cycle usage, as described below. Citibike journey starts times are hypothesized to follow similar patterns to overall bike ridership across areas covered by the Citibike network. To explore this, a dataset was constructed of Citibike rides from 2014-2016.

Given the similar distribution of Citibike ridership with traffic volumes, and the clear daily / monthly / yearly variations, we believe Citibike ridership may function effectively as a normalization index. Citibike ridership is observed to be undergoing a continued growth trend, at the same time that its volumes are proportional with cycle traffic volumes. We therefore use three-yearly average ride volumes (Table 1).

Month	Average trip counts
January	365144
February	327514
March	566955
April	778774
May	1013462
June	1112806
July	1144877
August	1233399
September	1297481
October	1204954
November	904459
December	671796

Table 1: Average Citibike ride volumes by month (three-year average, 2014-2016)

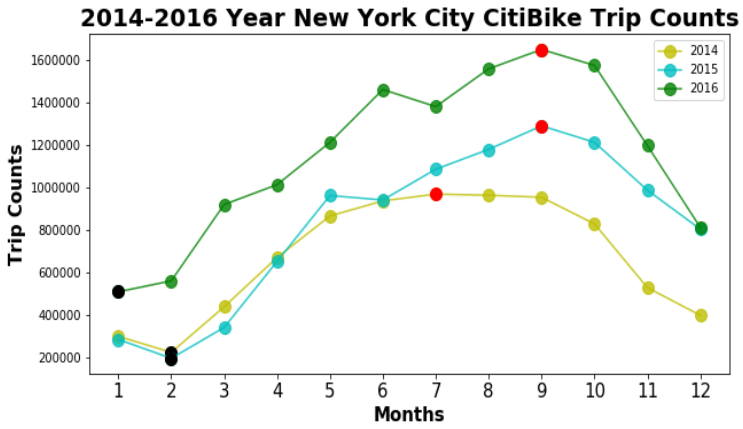


Figure 8: Citibike Trip Volumes by Month November 21, 2017

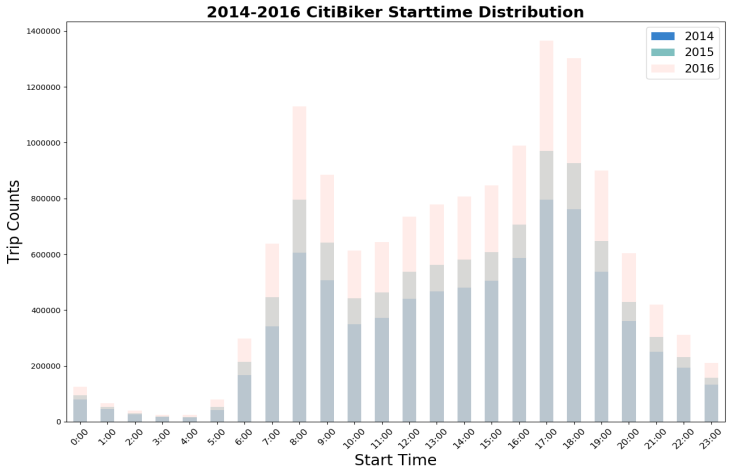


Figure 9: Citibike Trip Volumes By Time of Day

4.3 Data Visualization

The accident dataset was visualized on a map of New York using Carto. Figure 10 uses a choropleth scale from blue to yellow to show accident intensity by spatial unit, with bike lanes overlaid in red. We see lower accident volumes in outer areas where cycle traffic density is known to be low, and high accident volumes in much of Manhattan, and parts of Brooklyn and Queens. We also observe areas of dense parts of the city, such as the Upper West Side and Lower East Side, where accident volumes appear to be relatively lower.

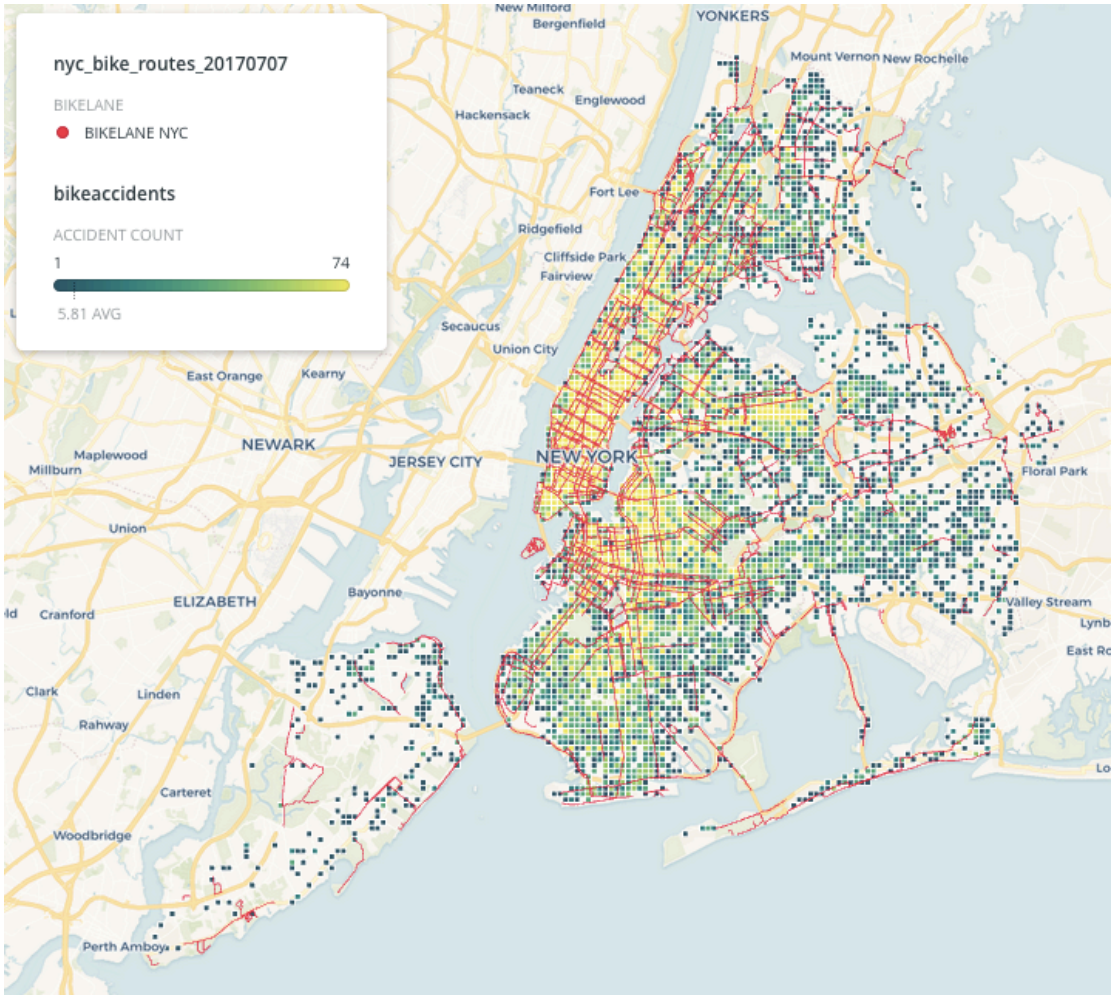


Figure 10: Accident Intensity By Spatial Unit, 2014-2016

We further examined the dataset to explore the average number of injuries per accident. This helps produce a 'danger map' as shown in Figure 11. The map visualizes the amount of injuries from accidents in every square divided by the total number of accidents. The lighter the color, the more dangerous a place is. In particular, we see a number of locations in Brooklyn, Queens and Bronx where accidents produce higher numbers of injuries. The map highlights different areas compared with Figure 10. This may indicate that people actually ride their bikes more in the less dangerous places.

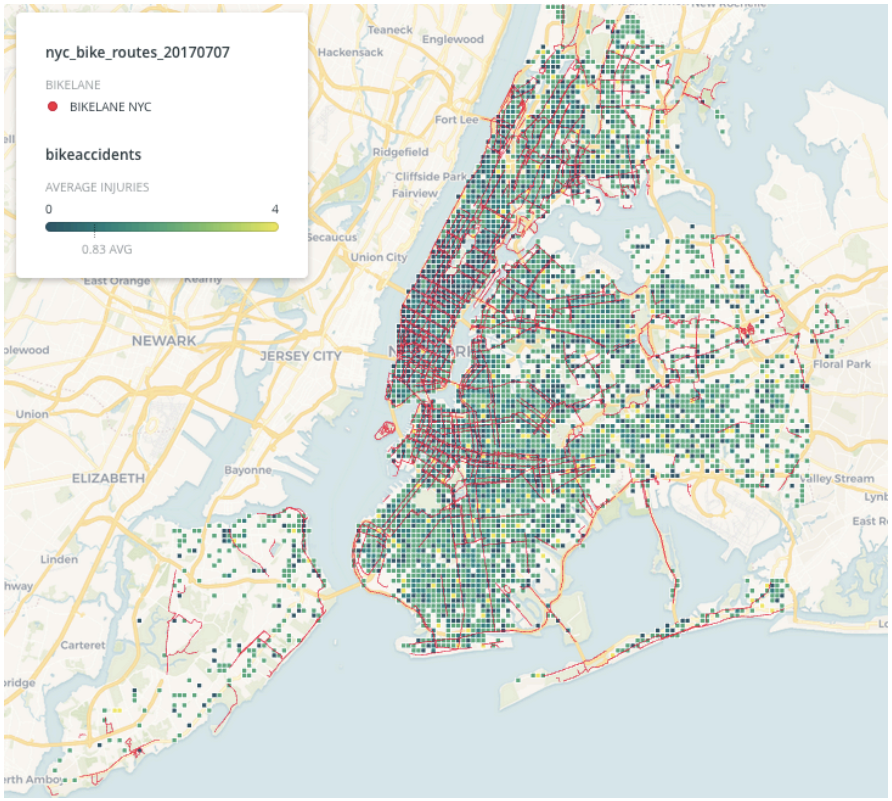


Figure 11: Cycle Traffic Accident Injury Intensity (2014-2016) November 21, 2017

5. Model building

5.1 Linear regression – Borough level

After aggregating the raw dataframe, we arrived at a new dataframe with collision counts in each borough on each day with their corresponding weather conditions. To understand the relation between bike crashes and bike collision caused injuries, we will regress the number of collisions and bike accident related death and injury in each borough on each day on temperature, humidity, wind speed and dummy variables we created to indicate whether the biker would encounter wet road condition or difficult visibility. We will also use the boroughs as dummy variables to see the locational relations.

The regression summaries are listed below, note that Bronx is omitted automatically to avoid multicollinearity:

```
============================================================
                            death_injury  borough_count
------------------------------------------------------------
Intercept                   -0.4304**      -0.5324***
                            (0.1851)       (0.2048)
C(BOROUGH)[T.BROOKLYN]      3.0354***      3.6832***
                            (0.0671)       (0.0742)
C(BOROUGH)[T.MANHATTAN]     2.0211***      3.1958***
```

```
                              (0.0665)        (0.0736)
C(BOROUGH)[T.QUEENS]          0.9346***       1.1754***
                              (0.0705)        (0.0780)
C(BOROUGH)[T.STATEN ISLAND]  -0.9029***      -1.1471***
                              (0.1468)        (0.1624)
Temp                          0.0549***       0.0663***
                              (0.0015)        (0.0017)
Humi                         -0.0166***      -0.0200***
                              (0.0016)        (0.0018)
Wind_Speed                   -0.0504***      -0.0546***
                              (0.0146)        (0.0161)
wet                           0.0765         -0.0985
                              (0.0851)        (0.0941)
Difficult_visibility         -0.0503          0.0209
                              (0.1551)        (0.1715)
R-squared                     0.34            0.40
No. observations              8816            8816
=============================================================
Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01
```

From above, we can observe that weather have statistically significant relations with bike collisions. Each unit increase of temperature will increase death or injury by 5.5% and bike collisions by 6.6%. However, when there is higher humidity or higher wind speed, both death injury counts, and bike collision counts drops. On the dummy variables that indicates whether bikers would encounter wet road conditions or difficult visibilities, we observed statistically significant coefficients that share different signs.

To explain the coefficients for the Borough dummies, normalizing the Bronx, we observe that after controlling for weather conditions, both collisions and death and injuries are more likely to take place in Brooklyn instead of Manhattan. Biking on Staten Island is less likely to result in bike collisions and bike collision related injury or death.

5.2 Logistic regression

One of the models we tried to run, was a logistic model, designed to understand if the variables we have (weather, time of the day, moth and location) have an impact on the outcome of an accident. This means that we tried to train a model to predict with a given input if an observation (accident) resulted in injury (or death) = 1 or not = 0.

We used a 70-30% split for the training and test data, and the following are the results we had:

Model Accuracy: 0.788620373203

Model Precision: 0.782585579518

Confusion Matrix: Where 1 means injury or death and 0 means No injury or death

		Predicted value	
		0	**1**
True value	**0**	0	1382
	1	0	5156

From the results we can see a very high accuracy and precision rates, but only because our dataset has mostly accidents that resulted in injury and what the model is doing is predicting, with the input variables, that every accident will have an injury/lethal outcome. This last affirmation can be confirmed by looking at the confusion matrix.

This is a model that tells us something very interesting, and it is that given the input variables we had, it will always predict that every reported accident resulted in injury. It is very likely that the reason for this happening is that the main motivation for people to report an accident involving a bicycle is because the collision resulted in injury or death to anyone involved. We can draw another conclusion from this model and it is that probably the variables we are considering have an impact on weather an accident will occur or not, but not on the outcome of that accident.

5.3 Linear regression model – predictive analysis

This model uses linear regression through ordinary least squares to model the dependent variable of number of cycle injuries per hour for census tract $_{CT}$.

Model statement

Number of predicted accidents for census tract 'c' and hourly time period 't' is given by the following function:

$$Y_{ct} = a + \beta_1 \text{time}_{ct} + \beta_2 \text{weather}_{ct} + \beta_3 \text{complaints}_{ct} + \beta_4 \text{temporal lag}_{ct}$$

Variable specification

The following variables will be used for our base model:

Time:
• Time of day • Day of week • Week of year

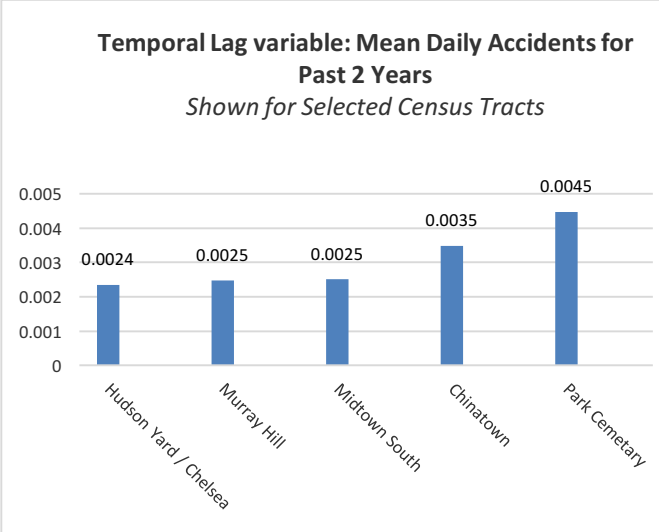
Weather
• Temperature • Visibility • Wind speed • Precipitation (from condition dummies)

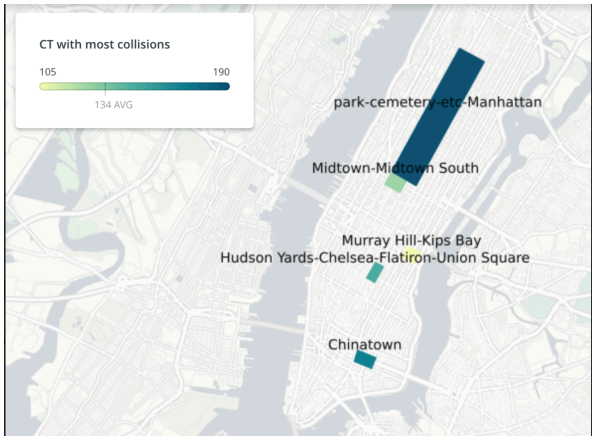
Complaints
• Composite of traffic-related 311 complaints

• Composite of trash-related 311 complaints

Temporal lag

• Average daily accident rate for census tract in past 2 years.





Feature selection:

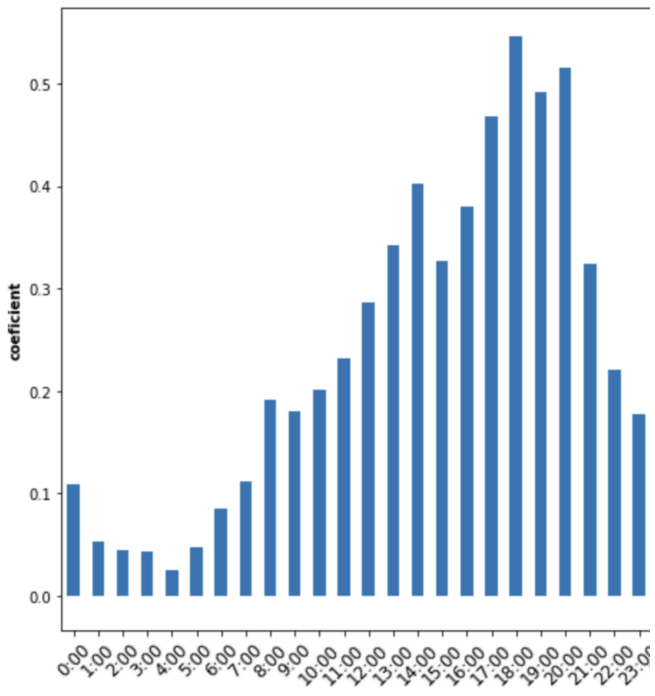
Weather variables were added in a step-wise fashion to determine which contribute most to explaining variance in accidents. Based upon this procedure, two dummies were constructed, bringing together the significant features around wet weather ('wet' dummy) and low visibility ('low_vis' dummy).

Regression output:

For the dataset covering the whole city, the regression produced the output below:

Dep. Variable:	y	**R-squared:**	0.109		
Model:	OLS	**Adj. R-squared:**	0.086		
Method:	Least Squares	**F-statistic:**	4.756		
Date:	Fri, 01 Dec 2017	**Prob (F-statistic):**	3.07E-13		
Time:	15:04:44	**Log-Likelihood:**	-759.28		
No. Observations:	1000	**AIC:**	1571		
Df Residuals:	974	**BIC:**	1698		
	coef	**std err**		**coef**	**std err**
Intercept	0.2188	0.016	**h13**	0.1714	0.084
h0	-0.2124	0.079	**h14**	0.083	0.083
h1	-0.2114	0.081	**h15**	0.0074	0.08
h2	-0.2145	0.077	**h16**	0.1863	0.083
h3	-0.1111	0.075	**h17**	0.1757	0.08
h4	-0.1987	0.075	**h18**	0.486	0.081
h5	-0.0385	0.08	**h19**	0.2271	0.076
h6	-0.2389	0.078	**h20**	0.2391	0.079
h7	-0.1176	0.08	**h21**	0.1699	0.078
h8	0.0273	0.079	**h22**	-0.0956	0.078
h9	-0.0903	0.08	**h23**	0.0662	0.078
h10	-0.1307	0.083	**wet**	0.0362	0.058
h11	0.044	0.083	**low_vis**	0.0407	0.081
h12	-0.0051	0.082	**lag**	0.0532	0.004

Figure: Regression Coefficients for Hour of Day Dummy Variables



Discussion:

The regression coefficients indicate that, holding other variables constant, accidents are likely to be 30% if it is 8pm-9pm. We note a statistically significant linear relationship between the wet weather and low visibility variables, and particularly strong linear relationships with several time-of-day dummies such as 4pm-5pm. The regression can be expanded before the project is finalized.

6. Conclusions and next steps

The project has demonstrated strong patterning in the distribution of cycle accidents in New York City based on time, weather and spatial variables. We identified linear relationships that are statistically significant and can be used for predictive analysis. The r^2 from the predictive linear model is currently low, but can be increased with improved modelling of the temporal and spatial variables, as well as full incorporation of the 311 data.

The next step, beyond the scope of the current project, is to deploy the model for each census tract in NYC. Based on applying the data cleaning and regression analysis steps to separate accident datasets for each individual census tract, different coefficients will be derived for each – reflecting local temporal patterns of cycling, as well as a different lag variable reflecting each census tract's historical accident rates. The normalization model as developed conceptually in this paper can also be applied. The model can then be applied through a Javascript interface and Geopandas to show a real-time map of predicted accident levels per hour. The project has therefore shown proof-of-concept for this tool, which would be a valuable urban data science contribution for the NYPD, NY DOT, and local residents.

References

Amoh-Gyimah, R., Saberi, M., & Sarvi, M. (2016). Macroscopic modeling of pedestrian and bicycle crashes: A cross-comparison of estimation methods. *Accident Analysis and Prevention*, *93*, 147–159. https://doi.org/10.1016/j.aap.2016.05.001

Cai, Q., Lee, J., Eluru, N., & Abdel-Aty, M. (2016). Macro-level pedestrian and bicycle crash analysis: Incorporating spatial spillover effects in dual state count models. *Accident Analysis & Prevention*, *93*, 14–22. https://doi.org/10.1016/j.aap.2016.04.018

Conway, A., Cheng, J., Peters, D., & Lownes, N. (2013). Characteristics of Multimodal Conflicts in Urban On-Street Bicycle Lanes. *Transportation Research Record: Journal of the Transportation Research Board*, *2387*, 93–101. https://doi.org/10.3141/2387-11

Eric Minikel. (2012). Cyclist safety on bicycle boulevards and parallel arterial routes in Berkeley, California. *Accident Analysis & Prevention*, *45*, 241–247. https://doi.org/10.1016/J.AAP.2011.07.009

Harris, M. A., Reynolds, C. C. O., Winters, M., Cripton, P. A., Shen, H., Chipman, M. L., … Harris, A. (2013). Comparing the effects of infrastructure on bicycling injury at intersections and non-intersections using a case–crossover design. *Inj Prev*, *19*, 303–310. https://doi.org/10.1136/injuryprev-2012-040561

Mohamed, M. G., Saunier, N., Miranda-Moreno, L. F., & Ukkusuri, S. V. (2013). A clustering regression approach: A comprehensive injury severity analysis of pedestrian-vehicle crashes in New York, US and Montreal, Canada. *Safety Science*, *54*, 27–37. https://doi.org/10.1016/j.ssci.2012.11.001

Nicaj, L., Stayton, C., Mandel-Ricci, J., Mccarthy, P., Grasso, K., Woloch, D., & Kerker, B. (2009). Bicyclist fatalities in New York City: 1996-2005. *Traffic Injury Prevention*, *10*(2), 157–161. https://doi.org/10.1080/15389580802641761

Nyberg ', P., Bjbrnstig ', U., & Bygren, L.-0. (n.d.). Road characteristics and bicycle accidents. *Scand J SOC Med*, *24*(4). Retrieved from http://journals.sagepub.com/doi/pdf/10.1177/140349489602400410

NYC Department of Transport (2016). Safer Cycling: Bicycle Ridership and Safety in New York City. Retrieved from http://www.nyc.gov/html/dot/downloads/pdf/bike-safety-study-frlandscapeformat2017.pdf

Prati, G., Pietrantoni, L., & Fraboni, F. (2017). Using data mining techniques to predict the severity of bicycle crashes. *Accident Analysis and Prevention*, *101*, 44–54. https://doi.org/10.1016/j.aap.2017.01.008

Pucher, J., & Dijkstra, L. (2000). Making walking and cycling safer: lessons from Europe. *Transportation Quarterly*, *54*(3), 25–50. https://doi.org/10.1258/0007142001903184

Rifaat, S., Tay, R., & de Barros, A. (n.d.). Effect of Street Pattern on Road Safety. *Transportation Research Record: Journal of the Transportation Research Board*. https://doi.org/10.3141/2147-08

Wanvik, P. O. (2009). Effects of road lighting: An analysis based on Dutch accident statistics 1987-2006. *Accident Analysis and Prevention*. https://doi.org/10.1016/j.aap.2008.10.003