An Empirical Analysis of Injurious Traffic Collisions in New York City

Jennifer Ganeles, Afzal Hossain, Abigail Strick

Over the past four years, injuries due to motor vehicle collisions in New York City have increased by a whopping 18% (Furfaro & Moore, 2019). This startling statistic necessitates a deeper analysis into the causes and contributions of New York City's injurious traffic collisions. Past research surrounding motor vehicle collisions has often been framed around the dangerous behaviors that impair one's judgement or ability to drive. Driving recklessly while intoxicated, high, tired, or on a cell phone are all ways in which individual actions have been found to increase the probability of injury or even death on the road (Stubig et al. 2012; et al. 2013; Bener 2017). However, for every person that crashes as a result of such risky or distracted behavior, there is oftentimes another party that is simply the victim of poor circumstances. For the average person who is not at fault, what other factors might contribute to a higher probability of danger? Fewer studies have focused on answering this question by exploring the circumstantial (rather than behavioral) predictors of traffic injury.

For this reason, the present study aims to look at the contributing factors of harmful vehicle collisions in New York City that go beyond the actions of individual drivers. More specifically, this paper seeks to determine how larger environmental patterns, such as time and location, correlate with the number of traffic injuries in all five boroughs of New York City. Our empirical analysis of motor vehicle collisions in New York City will explore trends and predictors associated with the total number of traffic injuries incurred by both motorists and non-motorists alike. We will then build a logistic regression model to predict the likelihood of a vehicle collision resulting in injuries based on multiple predictor variables.

Looking at the larger temporal and geographical patterns surrounding vehicle collisions will not only provide more insight into the rising number of traffic injuries in New York City; it will also better inform traffic policy meant to increase road safety. As the third most congested city in the world, New York City has already attempted to implement certain policies in order to reduce the number of motor vehicles on the road. These include added toll fares (Manskar, 2018), as well as other initiatives meant to clear lanes, curbs, intersections, and highways (Congestion Action Plan). Despite these efforts, however, decongesting city streets has proven to be a difficult and perhaps unfeasible endeavor (Furfaro et al., 2018). As traffic injuries continue to rise, policy must be aimed at improving, rather than decongesting, city streets. Our analysis of motor vehicle injuries in NYC will prove useful in identifying traffic trends in need of such improvement, as well as the specific boroughs and city streets most affected. Furthermore, injuries from motor vehicle collisions continue to rise despite bans on texting and handheld cell phone use (Marsh, 2017). Given the ineffectiveness of such bans, it can be argued that traffic policy must not be guided by driver tendencies alone, but by the environmental factors surrounding motor vehicle injuries as well.

In a recent literature review conducted by Masuri et al. (2017), environmental and temporal factors such as population growth, poor road conditions, weekends, and festive seasons were found to increase traffic collisions in developing countries. However, more research is needed to confirm whether these

findings can be applicable to more developed geographic areas. In exploring the temporal predictors of traffic injuries in New York City, we hypothesize that motor vehicle injuries will occur most often in the winter months due to both holidays and icy conditions. We also predict a higher number of traffic injuries during rush hour on weekdays due to congestion and during the evening on weekends, perhaps as a result of higher alcohol consumption. As for geographical predictors, we hypothesize that the highest number of injuries will occur on the busiest streets, especially in the borough of Brooklyn where it is the most populated.

In order to study NYC motor vehicle injury, the New York City Police Department (*NYPD*) is the most reliable and accurate source available. The data we used was collected by *NYPD* under 'Local Law 11' which was passed in 2011 when New York City's "Facade Inspection Safety Program" (FISP) was adopted not only to make building facades safe, but also to protect pedestrians on the road. The dataset is reviewed and updated every month by a *TrafficStat* Unit that was developed to improve targeted enforcement, step up internal accountability, and increase public transparency. This dataset is available at *NYC OpenData* as a public resource. The dataset contains almost a million and a half records and has twenty-nine variables that contain all motor collisions from the last seven years. Each record in the dataset contains collision data in *NYC* by city, borough, cross street, type of vehicle, date, contribution factor, and injury/fatality record. Therefore, this dataset was extremely efficient to measure the frequency of *NYC* motor vehicle collisions as well as the major environmental factors that may contribute to these collisions.

Time Series Analysis:

We conducted a time series analysis of traffic collisions from 2013 through 2018, exploring patterns of injury by year, month, and hour. As expected, we found an overall increase in traffic collision injuries as of 2013, and a cyclical pattern showing drastic monthly variation (see *Figure 1*).

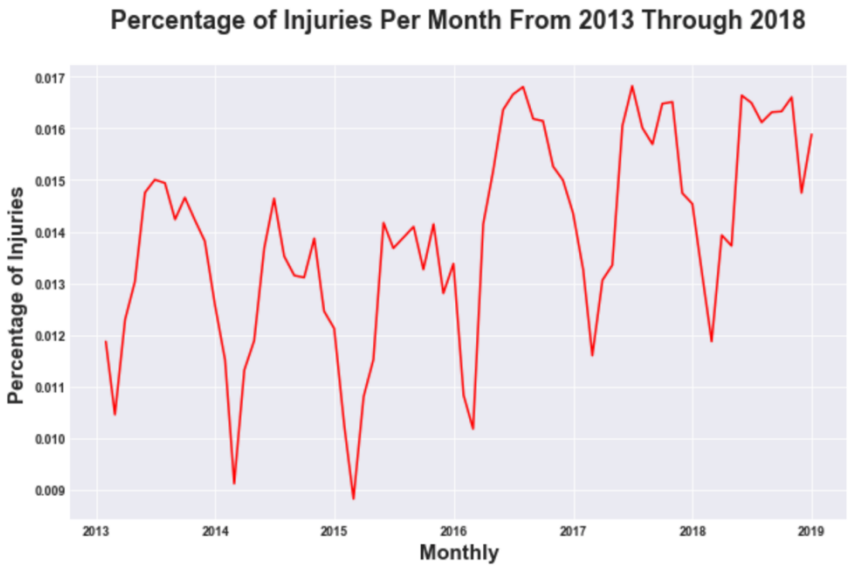


Figure 1. Percentage of injuries by month from 2013 through 2018

A closer look at this monthly variation is shown in *Figure* 2, where all traffic injuries from 2013 to 2018 are aggregated by month. Our analysis shows a major dip in traffic injury during the month of February with a steady rise into the spring and summer seasons. Interestingly, we see a peak in traffic injury in the month of July, in which there are no icy conditions. With a minor decline in November, injuries begin to rise again in December as the winter holiday season picks up.

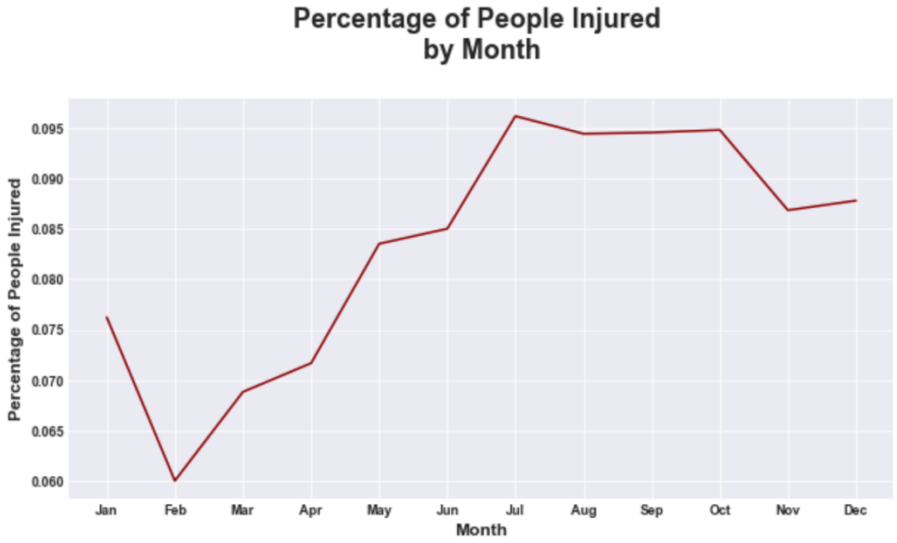


Figure 2. Percentage of all traffic injuries from 2013 through 2018 aggregated by month (non-sampled)

Our final time series analysis compares weekdays to weekends by exploring the hourly differences in traffic injury aggregated from July 2012 to February 2019. We found that during the weekdays, traffic injuries are lowest at 2 AM and highest around 8 AM and 4 PM, suggesting a spike in injury during morning and evening rush hours. For the weekends, traffic injuries are lowest at 6 AM and highest at 4 PM.

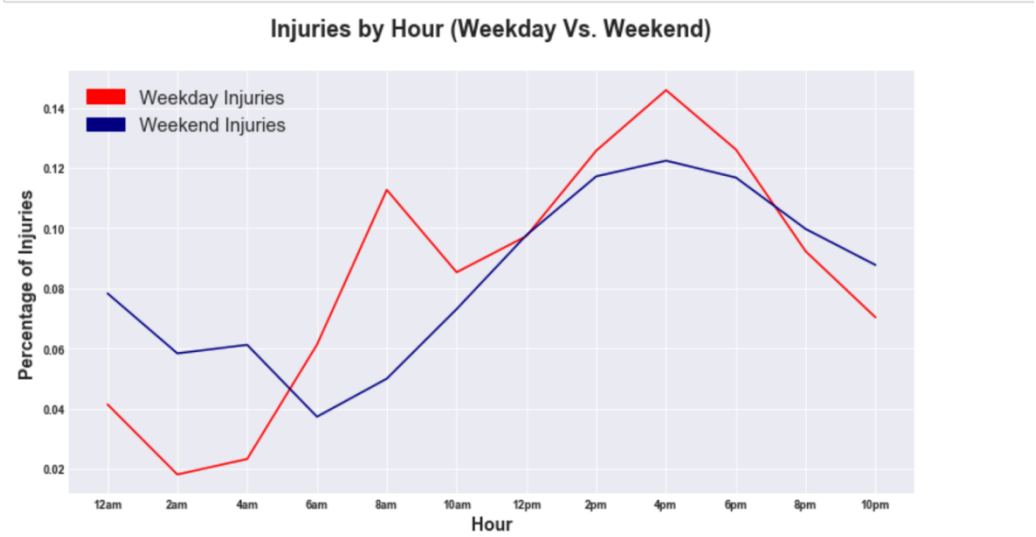


Figure 3. Aggregated percentage of traffic injuries by hour from July 2012 to February 2019 for both weekdays and weekends (non-sampled)

Geographical Analysis:

In addition to studying the temporal factors surrounding traffic injury in NYC, we also analyzed the geographical predictors of such injuries. Our first geographical analysis explores the rate of traffic collision injuries by borough (see *Figure 4*). After adjusting the traffic injury rate by square mile for each borough, we found the highest percentage of injuries to occur in Manhattan (38%), followed by Brooklyn (26%), Bronx (19%), Queens (13%) and Staten Island (4%).

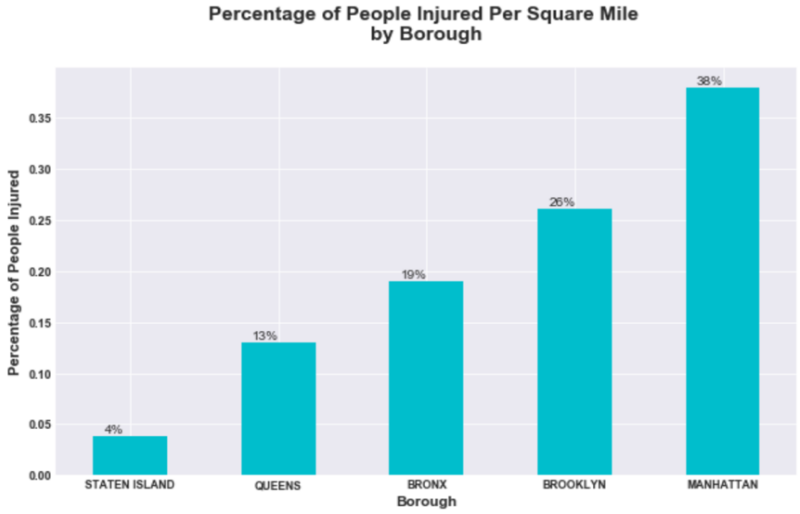


Figure 4. Number of traffic injuries per square mile by borough in New York City (2012-2019)

Our analysis also explored which roads in each borough were found to have the highest number of traffic injuries. We found that each borough has at least one highly dangerous road, with the exception of Staten Island. The most dangerous streets are mostly main roads (avenues and boulevards) and include Atlantic Avenue, Northern Boulevard, Broadway, Flatbush Avenue, Queens Boulevard, 3rd Avenue, Jamaica Avenue, Eastern Parkway, Linden Boulevard, and Bedford Avenue (see *Figure 5*).

Figure 5. Most dangerous streets in NYC in terms of traffic collision injuries (2012-2019)

Other Exploratory Data Analyses:

In addition to looking at the temporal and geographical trends surrounding motor vehicle injuries, we explored the correlations between the primary and secondary contributing factors of injuries, as defined by our dataset. We found that collisions resulting from traffic violations are positively correlated with alcohol/drugs, though negatively correlated with technical faults. Collisions resulting from human ignorance, on the other hand, are negatively correlated with alcohol/drugs and positively correlated with technical faults.

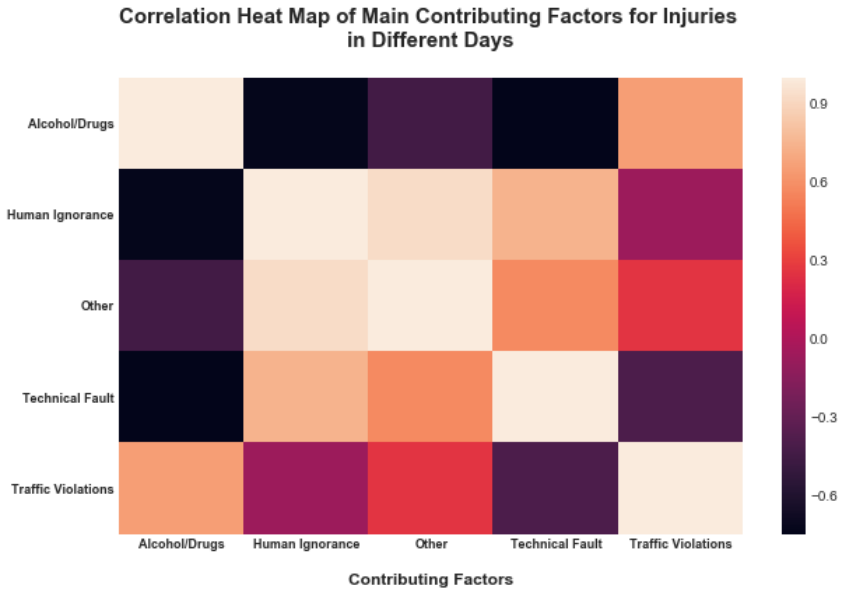


Figure 6. Correlational heat map of main contributing factors for traffic collision injuries between 2012 and 2019

Regression Analysis:

We used a logistic regression model to predict the likelihood of a vehicle collision resulting in injuries based on predictor variables, based off the variables we used for the exploratory data analysis. Using logistic regression classification, our goal is to predict whether the people will be injured or not. Therefore, we want to capture the importance (i.e., predictive power) of reasons for motor vehicle injuries and put weights on them. In order to have more accurate results we used the SMOTE algorithm, which works by creating synthetic samples from the minor class (no-subscription) rather than creating copies. It randomly chooses one of the k-nearest-neighbors and uses it to create similar, but randomly tweaked, new observations.

<u>Logistic Regression Model:</u>

$$Pr(Injured \mid Vehicle\ Collision) = \text{logit}^{-1}(\beta_0 + \beta_1 \mathbb{1}_{Borough\ Brooklyn} + \ldots + \beta_{46} \mathbb{1}_{Primary\ Reason\ Traffic\ Law\ Violations})$$

Dependent Variable:

- **Injured** (Yes=1 or No=0)

Independent variables:

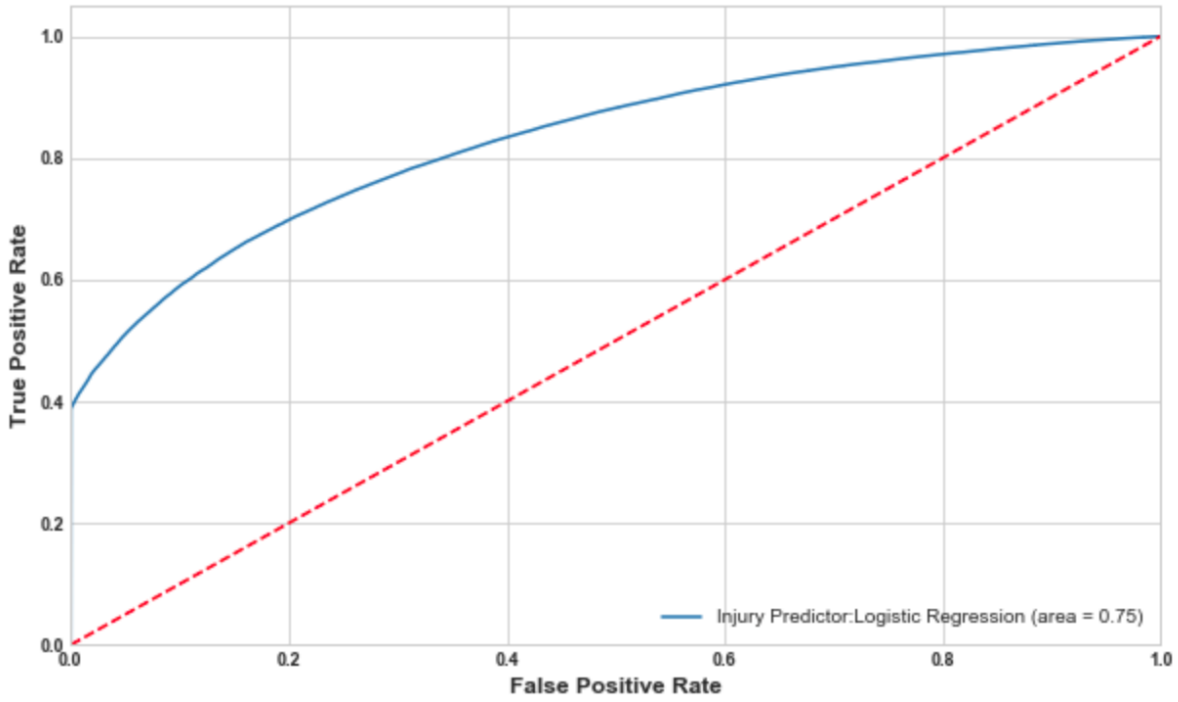
1. **BOROUGH:** Factor variable that contains the name of each NYC borough
2. **Hour**: The dataset uses a 24 hour clock
3. **WeekDay**: Day of the week (ex: Saturday, Friday…)
4. **Weekend**: Whether it is a weekend or weekday (Weekend = 1, Else = 0)
5. **Street_Type_1**: "ON Street Name" where it took place (Avenue, Street, Road, Boulevard, Other)
6. **Street_Type_2**: "Cross Street Name" where it took place (Avenue, Street, Road, Boulevard, Other)
7. **Street_Type_3**: "OFF Street Name" Where it took place (Avenue, Street, Road, Boulevard, Other)
8. **Primary_Reason:** Primary Contributing factor (Other, Traffic Law Violations, Alcohol/Drugs ,Technical Fault)
9. **Primary_Reason_2:** Secondary contributing factor (Other, Traffic Law Violations, Alcohol/Drugs ,Technical Fault)
10. **Vehicle_Type_1**: Primary vehicle type involved (Private Vehicle, Commercial Vehicle, Other)
11. **Vehicle_Type_2**: Second vehicle type involved (Private Vehicle, Commercial Vehicle, Other)
12. **Vehicle_Type_3**: Third vehicle type involved (Private Vehicle, Commercial Vehicle, Other)

Confusion Matrix:

N= 490187	Predicted : NO	Predicted: YES	
Actual: NO	TN=204059	FP=40942	245001
Actual: YES	FN=81443	TP=163743	245186
	285502	204685	

- **Accuracy:** Overall, how often is the classifier correct?
    - *(TP+TN)/N = (163743+204059)/490187* = **0.7503**
- **Misclassification Rate (Error Rate):** Overall, how often is it wrong?
    - *(FP+FN)/N = (40942+81443)/ 490187* = **0.2496**

The dotted line represents the ROC curve of a purely random classifier; a good classifier stays as far away from that line as possible.



Overall, our model performed well with 75% accuracy. Even though predicting vehicle collision injuries is extremely difficult due to the unpredictability of human behavior, we were able to find patterns and produce a decent accuracy level, thanks to our large sample.

Discussion:

After analyzing the graphs above, and our key takeaways, it becomes clear that the growing traffic injuries in New York City vary by both time and location. Overall, we see that vehicle injuries have increased from 2013 to 2018. From our time series analysis, we also see that vehicle injuries are cyclical and are almost predictable based on the season, day, and time of day. Regarding our hypotheses, our results did not support them completely. The spike in percentage of people injured in July and the drop in February show that weather conditions cannot entirely be blamed for motor vehicle injuries. This points to other conclusions that can be made regarding injury spikes and drops throughout the year. We see that the holiday seasons seem to be better predictors than bad weather. Perhaps drivers are more cautious when driving in icy conditions and are less careful when the weather is better.

Our prediction of injuries on weekdays and weekends were fairly accurate. The rush hours in the morning and evening show spikes in injuries. There are more injuries in the early morning hours of weekends than on weekdays. This may be attributed to people staying out later and drinking on weekends than weekdays, but further research is needed to test the interaction effect between temporal trends and drinking on the rate of traffic injury.

The percentage of injuries varies by borough as well as by street. The largest percentage of injuries per square mile occur in Manhattan, perhaps since Manhattan is a bustling borough (i.e. Manhattan has a theater district, club scene, businesses, flagship stores, etc). Manhattan is also a drive-through borough to get to and from New Jersey. However, Brooklyn has the most injury-prone streets, where the most injury-prone streets tend to be the longest streets that lead to bridges and tunnels.

It is important to note that our logistic regression model contained mostly environmental factors, with few behavioral variables. Our regression model performed well with 75% accuracy, providing further evidence that circumstantial factors play a key role in predicting traffic injuries. One limitation of our analysis was that our dataset contained many missing data and values. There is also a lack of unknown factors in the variables, including items labeled as "unspecified" which we took out of our model. Overall, our analyses showed that there are several circumstantial predictors of traffic injury that are worth exploring further. A deeper analysis into these patterns will no doubt inform traffic policy, as well as provide invaluable information to drivers, cyclists, and pedestrians who wish to lower their chance of injury.

Conclusion:

Vehicle collisions and their resulting injuries are a serious problem for New York City. The Mayor's Office spends a lot of its resources and taxpayers' money to reduce the number of collisions. However, a data driven solution is ultimately the best way forward in order to manage this problem. Our research report not only highlighted the growing frequency of injuries resulting from traffic collisions in NYC, but more importantly, it provided insight into the major contributing factors of this problem by looking at variables such as time and location. Using the variables from our exploratory analysis, we then successfully created a logistic regression model to help predict the likelihood of motor vehicle collisions resulting in injury. This will help NYC administrations effectively use their resources and aid towards reducing traffic related injuries. Though the generalizability of our findings are perhaps limited to New York City, this kind of work can be reproduced for the entire New York state, in order to tackle statewide vehicle collision problems beyond the five boroughs of NYC. In future research, it might be worthwhile to use other classification models like KNN or Random Forests in order to improve our prediction. Furthermore, ride share data can be used to better understand the impact of those on vehicle collisions.

References:

Asbridge, M., Brubacher, J. R., & Chan, H. (2012). Cell phone use and traffic crash risk: A culpability analysis. *International Journal of Epidemiology*, 42(1), 259-267.

Bener, A., Yildirim, E., Özkan, T., & Lajunen, T. (2017). Driver sleepiness, fatigue, careless behavior and risk of motor vehicle crash and injury: Population based case and control study. *Journal of Traffic and Transportation Engineering* (English Edition), 4(5), 496-502.

Gordon, A. (2018, October 08). Is NYC doing enough to address the looming climate crisis? Retrieved from https://ny.curbed.com/2018/10/8/17952564/un-climate-change-report-new-york-transportation

Furfaro, D., Moore, T. (2019, January 03). NYC traffic injuries are up despite drop in fatalities. Retrieved f rom https://nypost.com/2019/01/02/nyc-traffic-injuries-are-up-despite-drop-in-fatalities/

Furfaro, D., Rosner, E., & Brown, R. (2018, June 16). Why driving in NYC has somehow gotten even slower. Retrieved February 25, 2019, from https://nypost.com/2018/06/15/why-driving-in-nyc-has-somehow-gotten-even-slower

Lau-Barraco, C., Braitman, A. L., Linden-Carmichael, A. N., & Stamates, A. L. (2016). Differences in weekday versus weekend drinking among nonstudent emerging adults. Experimental and Clinical Psychopharmacology, 24(2), 100-109.

Manskar, N. (2018, January 19). It Might Soon Cost $11 To Enter Manhattan: Report. Retrieved February 25, 2019, from https://patch.com/new-york/new-york-city/congestion-pricing-plan-charges-11-enter-manhattan-reports

Marsh, A. (2017, February 6). Distracted driving ...overblown? Retrieved April 1, 2019, from h ttps://www1.nyc.gov/assets/dcas/downloads/pdf/fleet/Distracted_driving_article_Fleet_Owner _2-6-2017.pdf

Masuri, M. G., Isa, K. A., & Tahir, M. P. (2017). Children, Youth and Road Environment: Road traffic accident. *Asian Journal of Environment-Behaviour Studies*, 2(4), 13.

New York City Department of Transportation. (n.d.). Congestion Action Plan. Retrieved February 25, 2019, from https://www1.nyc.gov/html/dot/html/motorist/congestion-plan.shtml

Stübig, T., Petri, M., Zeckey, C., Brand, S., Müller, C., Otte, D., . . . Haasper, C. (2012). Alcohol intoxication in road traffic accidents leads to higher impact speed difference, higher ISS and MAIS, and higher preclinical mortality. *Alcohol*, 46(7), 681-686.