An Exploratory Study into NYC Motor Vehicle Collisions

Participant: David Wang
Research Supervisor: Claude Brathwaite

February 4, 2024

Contents

Abstract

New York City stands as a global hub for the financial world and a place for international cooperation with the Financial District and the United Nations. With a city holding such importance, safety is a major concern for city officials. One danger that people face in New York City is when they are on the many roadways that cross through the city. By exploring and analyzing the data that the NYPD publishes regarding motor vehicle collisions, we discover that collisions often occur every half-hour between 8 AM and 10 PM. In addition, a driver needs to be situationally aware at all times, especially at intersections, to decrease the likeliness of being in a collision. However, most collisions that occur in NYC often don't result in a driver being injured and instead, the vehicle would take the brunt of the damage.

Introduction

The concept of maintaining road safety encompasses a multifaceted challenge that governments worldwide grapple with. It involves a comprehensive approach to ensure the safety and security of all individuals using roadways. This endeavor demands the integration of various strategies, policies, and initiatives aimed at preventing accidents, reducing injuries, and ultimately saving lives.

Our work focuses on providing insight into the motor vehicle collisions that occur on NYC streets. By employing data visualizations and integrating diverse datasets sourced from New York City, alongside intricate data modeling utilizing Python libraries, we aim to offer readers more information on the complicated topic of implementing methods to prevent motor collisions in New York City.

Materials and Methods

Materials

To conduct our study, we utilized NYC Open Data's Motor Vehicle Collisions (Crashes) data.[1] The data tables contain information from all police-reported motor vehicle collisions in NYC in which the collision resulted in someone injured, or killed, or if there was at least $1000 worth of damage. We will be using Python to analyze our data by using the Pandas library. We will then visualize our data to answer questions using Python's Seaborn, Plotly, and Matplotlib.[2] [3] [4] We will then model our data using Scikit-Learn, an open source machine learning language library, to develop a model to predict whether or not a collision would result in an injury.[5]

In addition, we will also utilize the NYC Open Data's Motor Vehicle Collisions (Vehicle) data.[6] This dataset is part of the larger database that describes the hundreds of thousands of reports that the NYPD process. The Vehicle dataset contains data on the vehicle information associated with the collision. We will also utilize Python to analyze our data using the Pandas library, along with visualization tools with Seaborn, Plotly and Matplotlib.

To better understand the causes of a collision, we will be using NYC Planning zoning data[7] to investigate how much of a zoning district played a role in the collision. New York City is divided into three basic zoning districts Residence, Commercial, and Manufacturing.[8] Each of these districts is divided into different usage as well as special purpose districts. For our purpose, we will be generalizing our zoning district with our collision, but for future work, a collision can be specified specifically to what type of sub-division of a zoning district it is. By using Pandas and GeoPandas, we can concatenate our crash data tables with zoning information for each collision occurrence.

We will also be web-scraping NYC weather data from Weather Underground's BestForecast database. [9] By using Python's libraries of BeautifulSoup and Selenium, we will obtain general weather data for the NYC area for each of the collision entries found in our Motor Vehicle Collision data set. We will use the weather data to examine any correlations and observations we can make

to better understand why a collision occurred.

Methods

To begin a data exploration on any dataset, it is imperative to clean up our data from its raw form to perform analysis. To do so, we will:

- Understand our columns by looking at our data dictionaries
- Check for any null values in our columns
  - We will be removing rows with no geocoding as we are still left with a sizable amount of collision reports, however, it is possible to obtain the geocode of collisions by utilizing the data on the street in which the collision took place
- Remove redundant columns
  - For example, our crash data set contains all the required information in the Vehicle Type 1 column, Vehicle Type 3-5 was redundant so we dropped those columns to keep relevant columns which leaves about 1.4 million entries.
- Adjust and combine columns for merging of data.
  - Since we will be merging several datasets, we will need to ensure that we can combine them. To combine our collision and weather datasets, we will be utilizing the date and timestamps to merge. To merge our zoning and collision dataset, we will be merging based on the geocode of the collisions.

To develop a data model, we will be using Python's Scikit-Learn library, an open-source machine language learning library that provides a wide range of supervised and unsupervised learning algorithms.

- GridSearchCV

  - GridSearchCV does an exhaustive search with the available parameters of the model to search for the best hyper-parameters to obtain a high score for a scoring method. We will use the hyper-parameters to configure our model functions.

- Supervised Learning Algorithms - We will be using the following algorithms to model our data and then study their results to tune them to obtain the best results possible.
  - Decision Tree
  - Random Forest
  - K-Nearest Neighbors
  - Multinominal Naive Bayes

- Evaluation Methods [5]
  - Accuracy - evaluation of the fraction of correct predictions to total predictions $\texttt{accuracy}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} 1(\hat{y}_i = y_i)$
  - Precision - evaluation of the fraction of true positive to the sum of true positives $(P = \frac{TP}{TP+FP})$
  - Recall - evaluates the fraction of true positives to the sum of true positives and false negatives $(R = \frac{TP}{TP+FN})$
  - F1 Score - The F1 score can be interpreted as a harmonic mean of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. $(F_1 = 2\frac{P*R}{P+R})$
  - Confusion Matrix - A confusion matrix is such that it is equal to the number of observations known to be in group i and predicted to be in group j .Thus in binary classification, the count of true negatives is $C_{0,0}$, false negatives are $C_{1,0}$, true positives are $C_{1,1}$, and false positives are $C_{0,1}$.

Results

The results that we obtained and discussed from this research can be found in Appendix A and Appendix B.

Table 1 displays the datasets that we will be using to conduct our research on understanding the collisions that occurred in the NYC area. Notably, the NYC government entities provide a lot of data available for the public to conduct analysis such as this. By examining the data dictionaries found in Table 2 and 3, we can properly examine the trends we will explore.

To begin exploring our dataset, we're able to see the number of collisions reported that were made from June 2016 to the present. By taking advantage of that, we can examine if any trend was changed due to an event that occurred in the city. To approach this idea, Figure 1 depicts the number of reports made during the period between January 2018 and December 2021. Figure 2 visualizes a trend of the number of reports during each day in the year 2020.

With each collision report, officers have to evaluate the potential causes for the collisions to have occurred. There are many reasons that a collision could've occurred and sometimes it is unspecified at the time of the accident. However, when omitting that statistic to see the common causes for an accident, we obtain a visual in Figure 3 that showcases the Top 9 reasons for accidents from July 2012 to the present. Figure 4 and 5 visualizes the frequent impact point on the vehicles in which the accident cause was either the driver's inattention or their failure to yield the right-of-way. Since we explored those two specific accident causes, we also took the time to inspect the common area on a vehicle in which a point of impact often occurred.

Now, we can use the Vehicles portion of our collision dataset to investigate where our drivers are from, or at least, which state the vehicle is registered in. Given that this is a New York City-based database, most vehicles are from New York. We are more interested in what other states usually are involved in a collision so will be omitting the vehicles that are reported to be registered in New York. Figure 7 depicts the top ten states that have vehicles involved on the roads of New York City.

We see from Figure 7 that Florida is among one of the states in the top 10, since it isn't in the

Tri-state area, we can also look into the roadways in which a Floridian vehicle is involved in an accident. The top ten roadways can be seen in Figure 9.

When we combine our NYC Planning zoning dataset with our collision dataset, we can associate the zone that the collision occurred in. Table 4 shows the distribution of collisions for the entire dataset, while Table 5 shows the distribution in 2022.Figure 10 represents the temporal pattern of collisions reported in 2022. In 2022, we noticed that at midnight, an abnormal amount of collisions occurred. Table 6 represents the injuries that occurred across all 457 collision reports that were made at midnight in 2022. Figure 11 visualizes the causes of the collisions that occurred at midnight in 2022.

Table 7, 8, 9, 10 is the classification report for the data model techniques we utilized on our dataset. By using our performance metrics and confusion matrix, we can analyze how our models performed compared to each other.

Discussion

Our exploration of the NYC motor vehicle collision dataset begins against the backdrop of a recent global pandemic that led to shutdowns in cities across the world. In Figure 1, we can see a noticeable dip in the collisions reported. On March 20th, 2020, when Governor Cuomo issued a pause on non-essential business in New York. [10] With the federal and state governments issuing legislation and urging people to avoid going out and to practice social distancing, fewer people would be on the road to commute to work, school, and other activities. Since few people are going out, there are fewer vehicles on the road that could be involved in a collision. This can be further visualized in Figure 2 which illustrates a decrease in reported collisions in the days surrounding the government's recommendation to stay at home.

We then explored the common reasons for a collision to have occurred and can see that in Figure 3, the driver's inattention/distraction was a major cause across all five boroughs. By using the vehicle dataset, we can examine that the main impact point in collisions that involve a driver's inattention/distraction as the cause of the collision is the center front end of the vehicle. The next two common impact points are the right and left bumper. These two impact spots can be concluded when a vehicle is turning and ends up colliding with either another vehicle or pedestrian and fails to account for hazards on the road. A high number of collisions in which the center front end is impacted can be several reasons and scenarios that can be easily concluded without additional context.

The next collision factor we explore is the driver's failure to yield the right-of-way. Typically, this happens when a driver neglects to give priority to another vehicle or pedestrian, whether it's making a turn at an intersection, allowing pedestrians to cross, or approaching a STOP sign. This assertion finds support in Figure 5, which indicates that the Left Front Bumper is the most frequently impacted area in such scenarios. In New York City, vehicles often need to traverse lanes of oncoming traffic to execute left turns at intersections. The Right Front Bumper and the Center Front End are the next common impact points that further support the scenarios of a left turn in a large intersection having a higher chance of resulting in a collision.

Since we went through the vehicle impact points in Figure 4 and 5, we can explore the top ten impact locations across the entire dataset. From June 2012 to the present, many collisions that occurred involved the center front end of the vehicle. Followed by the Left Front Bumper and the Center Back End being the next common collision point. The Center Back End can be a scenario where a rear-ending of a vehicle occurred. Combined with Driver Inattention/Distraction being a major cause of collisions and the rise of smartphones being a distraction on the road, it is safe to correlate those scenarios. [11]

Given that this dataset is New York City-based, it makes sense that the vehicles registered with New York are involved in the most collisions on the roads of New York City. However, when we omit New York, we get Figure 7. Given that people commute to work from New Jersey, and Pennsylvania, it isn't too alarming that those two states are the next highest counts. However, the fact that Florida is among the most is quite peculiar. The one explanation for the high number of Floridians leaving NYC or traveling to NYC for holidays. This is supported by Figure 8 in which the peaks of collision during the period between Jan 2018 and Nov 2021 correspond to June 2018, June 2019, Dec 2019, and Sept 2020. These months are usually when people travel for vacation or for the holidays. It is worth noting that the number of Floridian vehicles involved in a collision dipped when COVID-19 quarantine measures first started. Another supporting detail to the fact that people are traveling for vacation or holidays can be found in Figure 9. The roadway with the most collisions that involved Floridian vehicles is the Belt Parkway, followed by the Brooklyn Queens Expressway, Long Island Expressway, and Cross Bronx Expressway. These highways some of New York's highways and highways are one of the main roadways to cross state lines.

Upon combining our NYC Planning zoning dataset and our collision dataset, it is immediately noticeable that collisions often occur in residential districts. The zoning district of 'Park' indicates the collision occurred near a park district. It is notable to mention that zoning districts do overlap but for our observation, we can find the temporal pattern of collisions that occur in residential districts in Figure 10. It is immediately notable that there is a high concentration of collisions that occur between 8 AM and 10 PM, with a spike at midnight. Another notable thing is that collision

reported a spike every half hour in that time interval. In the case of 8 AM, that is a time when people are commuting to work.

The most abnormal time in our visualization is midnight, one might think that is the time when people are driving while under the influence. However, that is not the case as we can see in Figure 11 where alcohol involvement is barely in the top 9 causes. The main cause is driver distractions, which can also include a driver's fatigue due to the late hour. Furthermore, out of the 457 reports made at midnight in 2022, about 60% of collisions result in no injury occurring at the time of the collision.

To the data model a dataset, tuning, testing, and experimenting is needed. Before we incorporate our zoning and weather data, we will develop a model to predict whether or not a collision will cause an injury. We will begin by utilizing our borough, period of day, and contributing factors to predict whether or not a collision will occur. In one of our approaches, we utilized a Random Forest model. Table 7 showcases our classification report and confusion matrix. One of the main reasons why there is an extremely poor calculation of whether or not a collision results in an injury is because of an imbalanced classification. Only 21% of our training data indicated an accident caused an injury. This makes the results more understandable as to why our results are skewed to our model being extremely accurate in predicting that a collision won't result in an injury. Our initial Random Forest parameters were based on one iteration using GridSearchCV from Scikit-Learn and aimed for a high f1 score.

While using the same features as our model in Table 7, we perform a random under-sampling on the data we use to train our model. Random under-sampling involves randomly selecting examples from the majority class to delete from the training dataset.[12] We see an instant improvement in predicting a collision will result in an injury, while losing our prediction on a collision not resulting in an injury. This is quite desirable as our model is balanced in predicting if a collision will result in an injury or not. We aim to get a more accurate model while incorporating more features to be more accurate and usable.

We utilize the same training data to determine a baseline for our K-Neighbor model. It shows

promise as it performs relatively similarly but with a lower recall score for predicting a collision, which will result in an injury, and a lower precision score for both classifications. We also utilize a Multinomial Naive Bayes model to continue to compare our models and note that it has a slightly lower F1 Score when compared to our Random Forest model. These models and data processing continue to emphasize that data needs to be balanced and prepared to work with.

Conclusion

We started our project by collecting data from the Motor Vehicle Collision dataset that is published on NYC Open Data, obtained zoning data from NYC Planning, and scrapped weather information from Weather Underground

We discovered several insights from visualizing our dataset:

- COVID-19 Quarantine protocols had a drastic impact on the decreasing number of collisions that occurred on NYC roadways.
- The fewer people on the road, the less chance someone will be involved in a collision.
- Driver distractions were the leading cause of collisions that occurred. Those types of collisions often had the point of impact at the center of the front end of a vehicle.
- Driver's inability to yield the right of way was another lead cause of collisions. Those types of collisions often had the point of impact at the left front bumper.
- New Jersey registered vehicles are the most common out-of-state registration that is mostly involved in accidents on NYC roadways.
- A majority of collisions occur every half hour between 8 AM EST and 10 PM EST.

We trained supervised learning models with our training data. These models included Random Forest, K-Nearest Neighbors, and Multinomial Naive Bayes. Our models obtain an accuracy of around 80% with our test data. However, our model was only accurate in predicting if a collision would not result in an injury due to most collisions recorded didn't result in a death or injury. After downsampling our data to train our data without bias, we got an accuracy of around 55% with our models.

Appendix A: Tables

Table 1: Material Used

Data	Data Origin
Motor Vehicle Collisions Information	NYC Open Data - NYPD
Zoning Districts	NYC Planning
NYC Historical Weather	Weather Underground's BestForecast

Table 2: NYC Motor Vehicle (Crashes) Column Information

Column Name	Column Description
COLLISION_ID	Unique record code generated by system
ACCIDENT_DATE	Occurrence date of collision
ACCIDENT_TIME	Occurrence time of collision
BOROUGH	Borough where collision occurred
ZIP CODE	Postal code of incident occurrence
LATITUDE	Latitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
LONGITUDE	Longitude coordinate for Global Coordinate System, WGS 1984, decimal degrees (EPSG 4326)
LOCATION	Latitude , Longitude pair
ON STREET NAME	Street on which the collision occurred
CROSS STREET NAME	Nearest cross street to the collision
OFF STREET NAME	Street address if known
NUMBER OF PERSONS INJURED	Number of persons injured
NUMBER OF PERSONS KILLED	Number of persons killed
NUMBER OF PEDESTRIANS INJURED	Number of pedestrians injured
NUMBER OF PEDESTRIANS KILLED	Number of pedestrians killed
NUMBER OF CYCLIST INJURED	Number of cyclists injured
NUMBER OF CYCLIST KILLED	Number of cyclists killed
NUMBER OF MOTORIST INJURED	Number of vehicle occupants injured
NUMBER OF MOTORIST KILLED	Number of vehicle occupants killed
CONTRIBUTING FACTOR VEHICLE 1	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 2	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 3	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 4	Factors contributing to the collision for designated vehicle
CONTRIBUTING FACTOR VEHICLE 5	Factors contributing to the collision for designated vehicle
VEHICLE TYPE CODE 1	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)
VEHICLE TYPE CODE 2	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)
VEHICLE TYPE CODE 3	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)
VEHICLE TYPE CODE 4	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)
VEHICLE TYPE CODE 5	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)

Table 3: NYC Motor Vehicle (Vehicle) Column Information

UNIQUE_ID	Unique record code generated by system
COLLISION_ID	Unique crash identification code
ACCIDENT_DATE	Occurrence date of collision
ACCIDENT_TIME	Occurrence time of collision
VEHICLE_ID	Vehicle identification code assigned by system
STATE_REGISTRATION	State where driver license was issued
VEHICLE_TYPE	Type of vehicle based on the selected vehicle category (ATV, bicycle, car/suv, ebike, escooter, truck/bus, motorcycle, other)
VEHICLE_MAKE	Vehicle make
VEHICLE_MODEL	Vehicle model
VEHICLE_YEAR	Year the vehicle was manufactured
TRAVEL_DIRECTION	Direction vehicle was traveling
VEHICLE_OCCUPANTS	Number of vehicle occupants
DRIVER_SEX	Gender of driver
DRIVER_LICENSE_STATUS	License, permit, unlicensed
DRIVER_LICENSE_JURISDICTION	NYPD, Port Authority, TBTA, MTA, etc.
PRE_ACDNT_ACTION	Going straight, making right turn, passing, backing, etc.
POINT_OF_IMPACT	Location on the vehicle of the initial point of impact (i.e. driver side, passenger side rear, etc.)
VEHICLE_DAMAGE	Location on the vehicle where most of the damage occurred
VEHICLE_DAMAGE_1	Additional damage locations on the vehicle
VEHICLE_DAMAGE_2	Additional damage locations on the vehicle
VEHICLE_DAMAGE_3	Additional damage locations on the vehicle
PUBLIC_PROPERTY_DAMAGE	Public property damaged (Yes or No)
PUBLIC_PROPERTY_DAMAGE_TYPE	Type of public property damaged (ex. Sign, fence, light post, etc.)
CONTRIBUTING_FACTOR_1	Factors contributing to the collision for designated vehicle
CONTRIBUTING_FACTOR_2	Factors contributing to the collision for designated vehicle

Table 4: Collision Distribution Across NYC Zoning Districts (June 1st, 2012 - Dec 2nd, 2023)

Zone	Number of Collisions Reported
Residential	489115
Commercial	252229
Park	181189
Mixed Manufacturing/Residential	119237

Table 5: Collision Distribution Across NYC Zoning Districts in 2022

Zone	Number of Collisions Reported
Residential	15303
Commercial	6162
Park	5742
Mixed Manufacturing/Residential	3738

Table 6: Injuries Reported at Midnight in 2022

Collision Reported Injury	Count
No Injuries	271
Injury	186

Table 7: (Classification Report: Baseline Random Forest ($criterion =' gini', max_depth = 15, n_estimators = 50, n_jobs = -1$)

	precision	recall	f1-score	support
No	0.79	1.00	0.88	213499
Yes	0.75	0.02	0.03	58847
accuracy			0.79	272346
macro avg	0.77	0.51	0.46	272346
weighted avg	0.78	0.79	0.70	272346

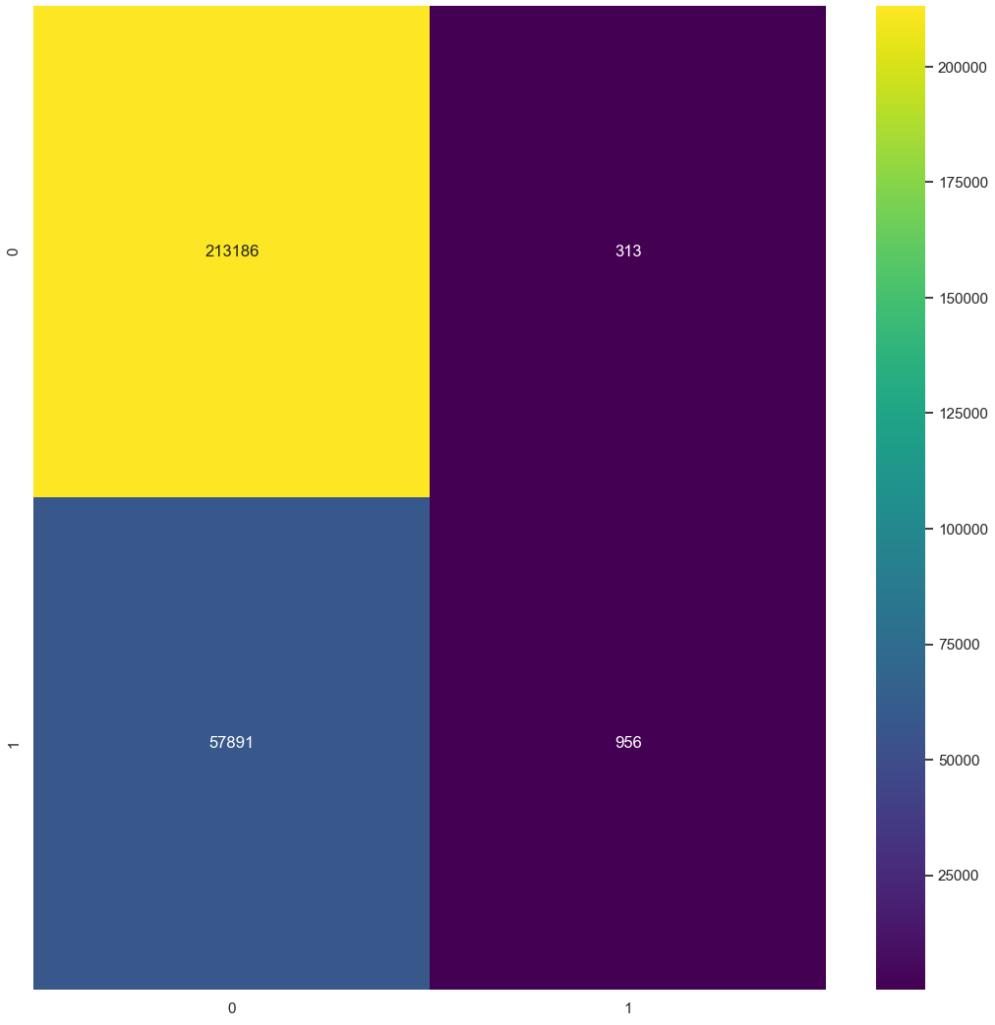


Table 8: (Classification Report: RandomUnderSampler Random Forest, criterion = 'gini', $\max_depth = 15, n_estimators = 50, n_jobs = -1$)

	precision	recall	f1-score	support
No	0.57	0.51	0.54	47242
Yes	0.56	0.61	0.58	47287
accuracy			0.56	94529
macro avg	0.56	0.56	0.56	94529
weighted avg	0.56	0.56	0.56	94529

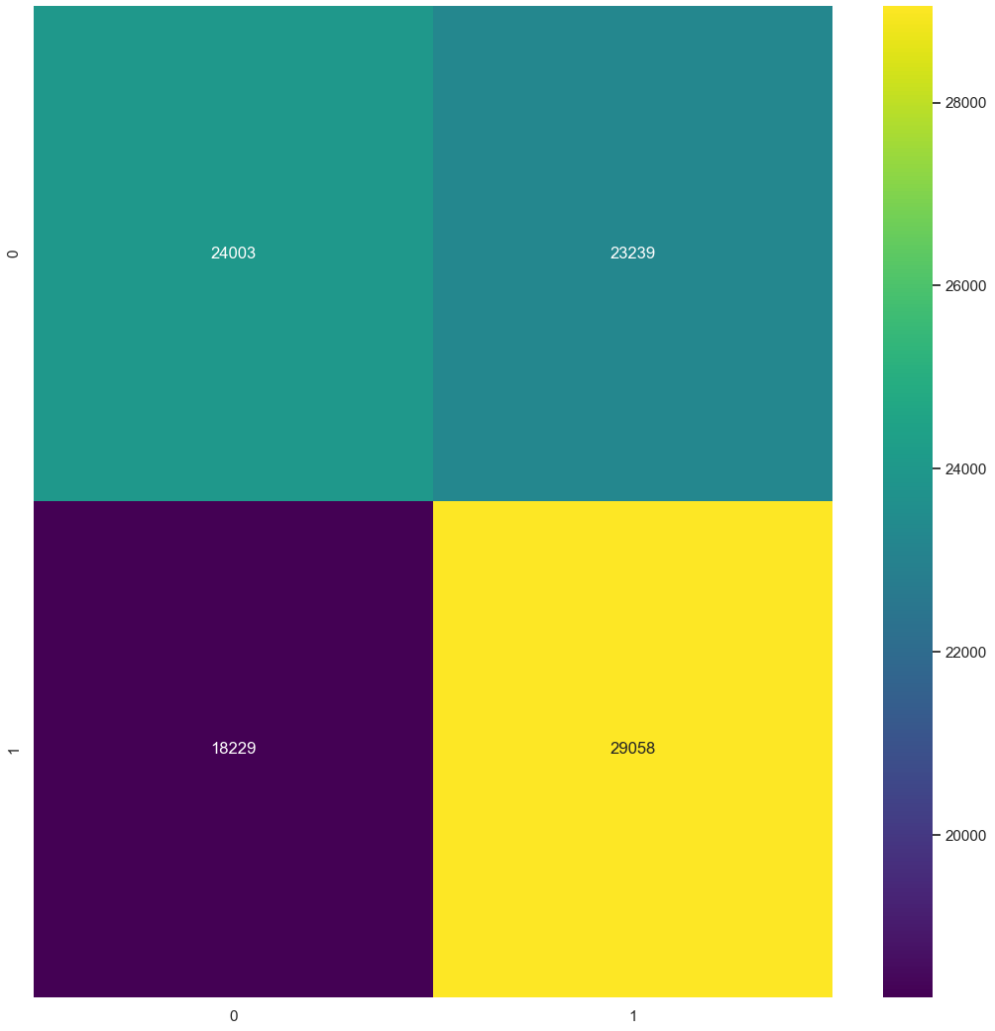


Table 9: (Classification Report: RandomUnderSampler KNeighbors)

	precision	recall	f1-score	support
No	0.52	0.57	0.55	47242
Yes	0.53	0.47	0.50	47287
accuracy			0.52	94529
macro avg	0.52	0.52	0.52	94529
weighted avg	0.52	0.52	0.52	94529

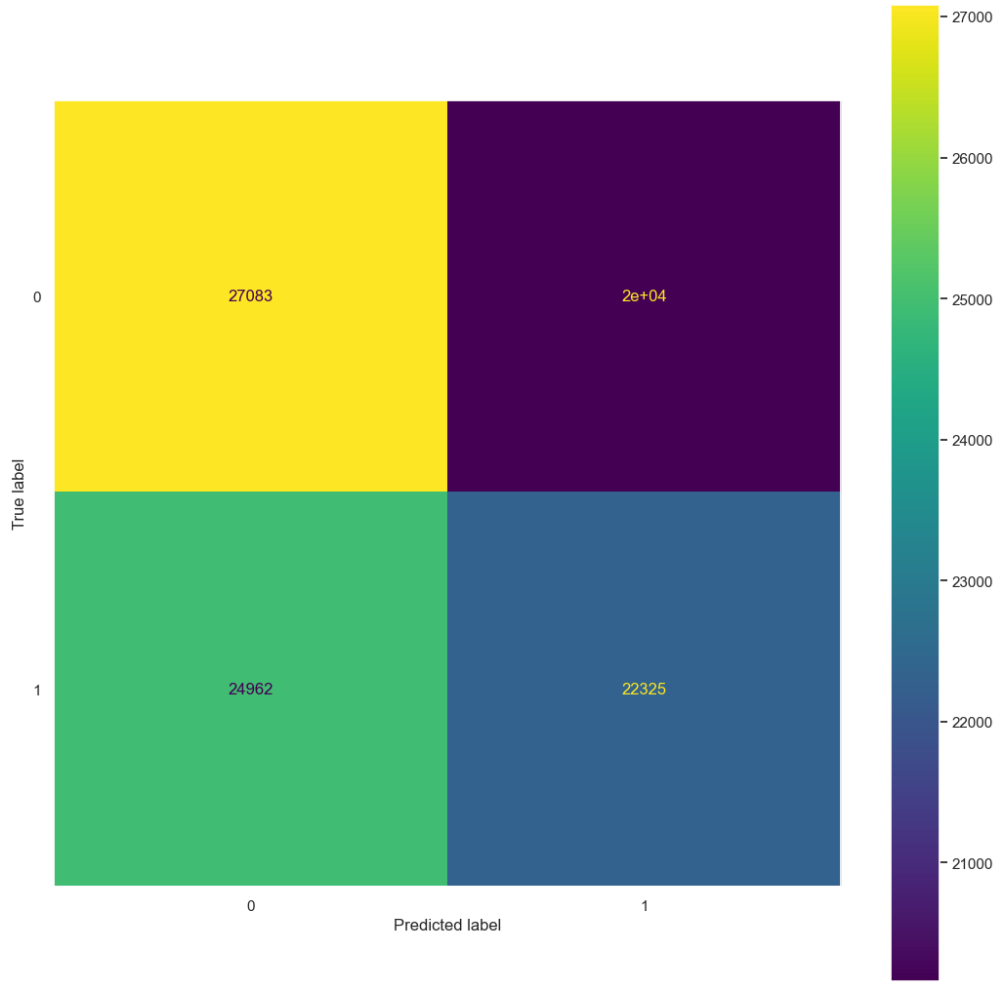
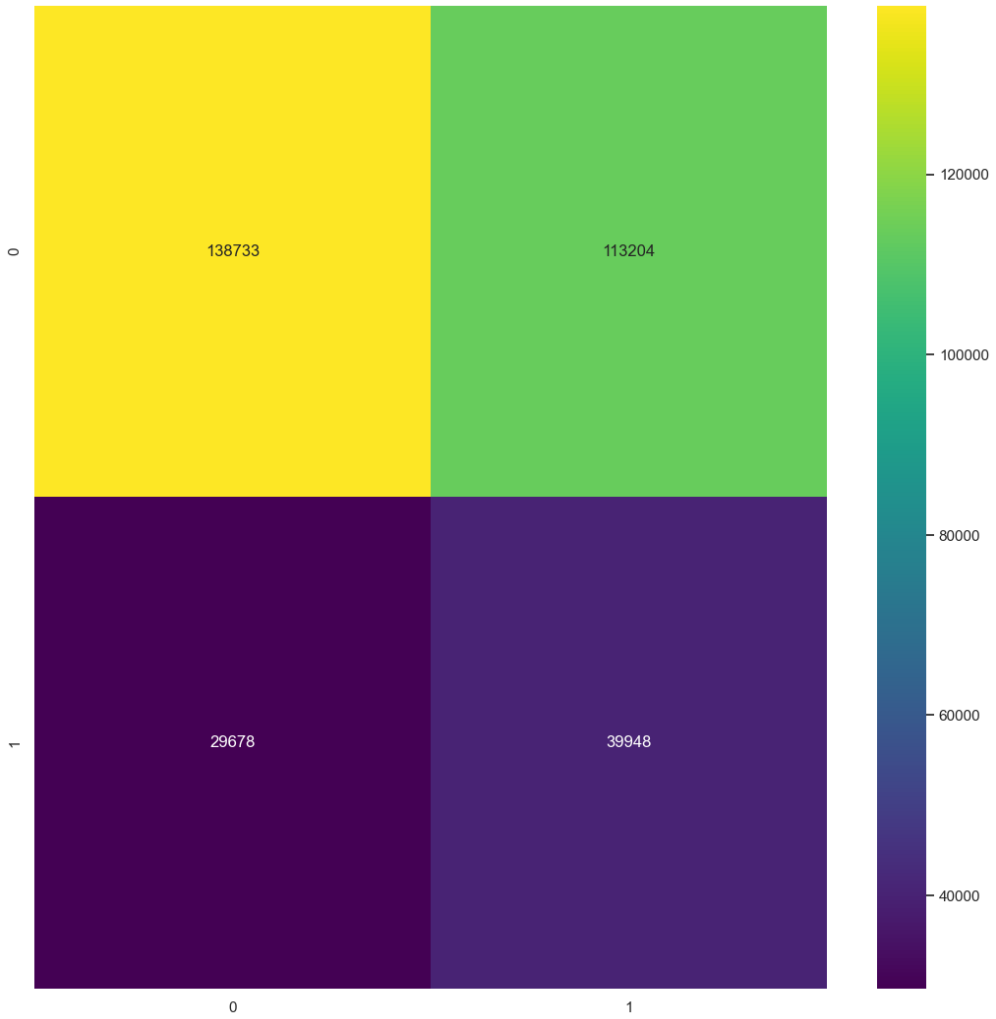


Table 10: (Classification Report: RandomUnderSampler Multinomial Naive Bayes)

	precision	recall	f1-score	support
No	0.82	0.55	0.66	251937
Yes	0.26	0.57	0.36	69626
accuracy			0.56	321563
macro avg	0.54	0.56	0.51	321563
weighted avg	0.70	0.56	0.59	321563



Appendix B: Figures

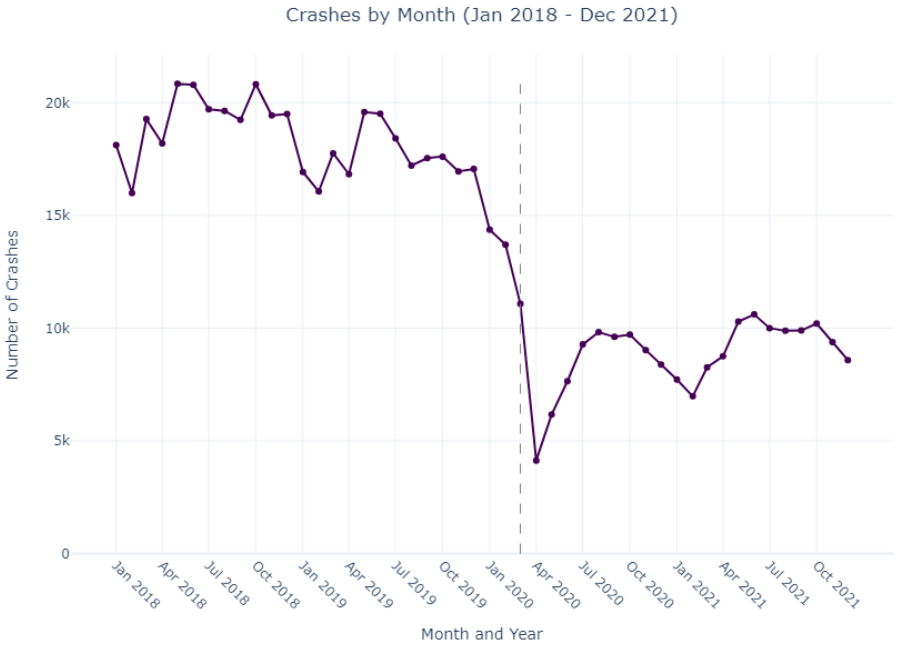


Figure 1: Trend of collisions reported between Jan 2018 and Dec 2021

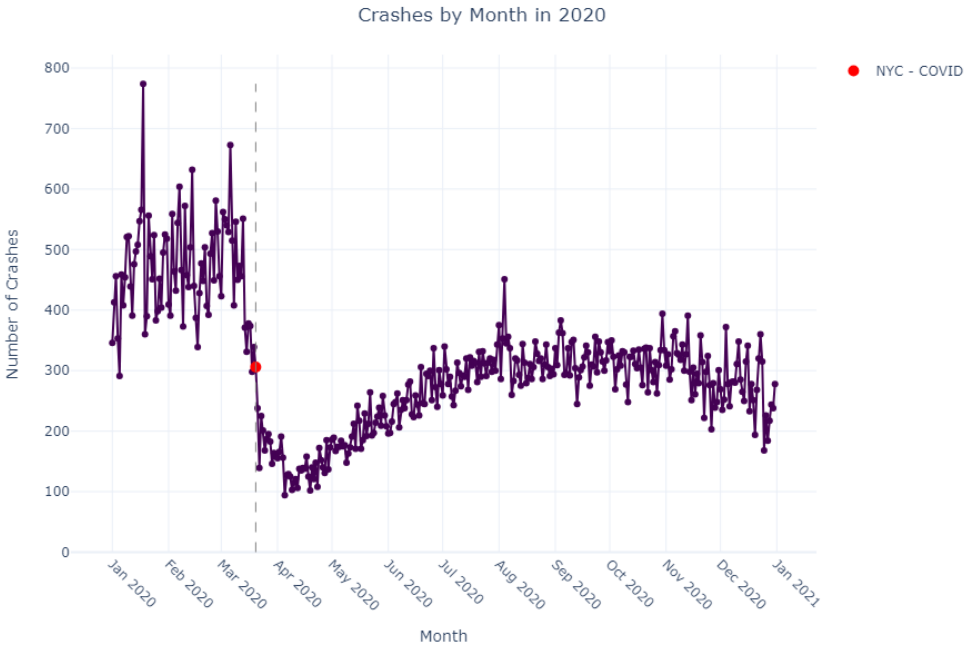


Figure 2: Collision reported during 2020 in NYC

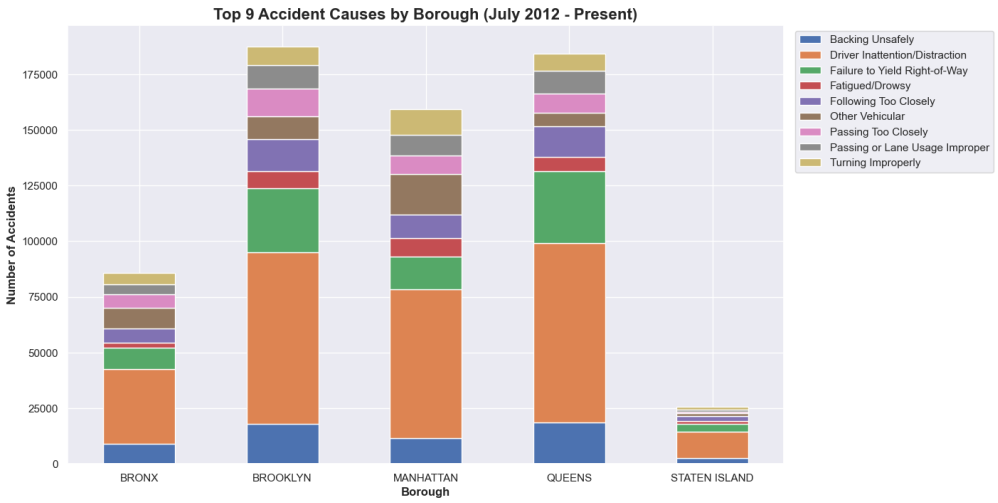


Figure 3: Top 9 Causes to Collisions in NYC

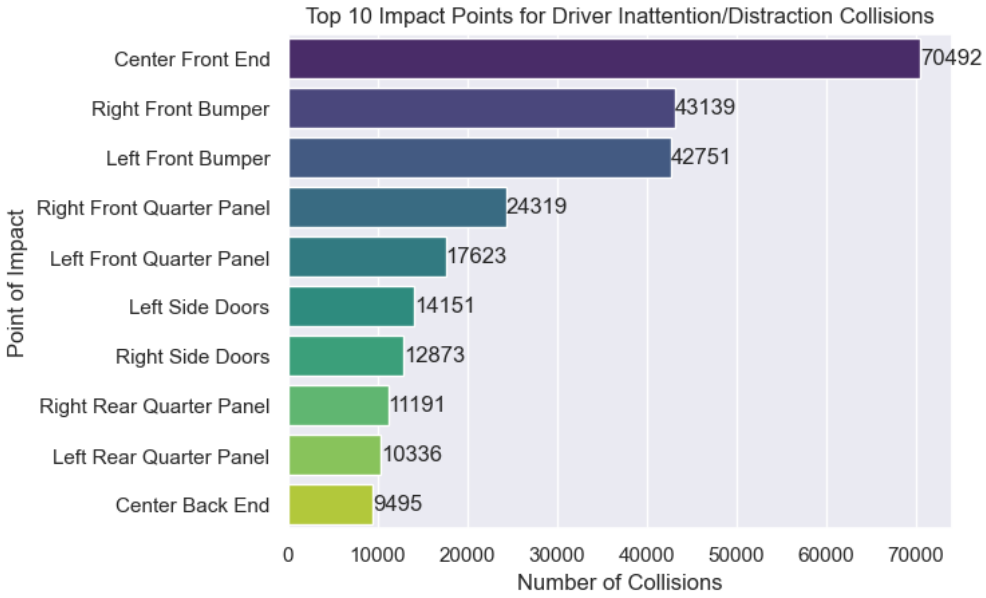


Figure 4: Top 10 Vehicle Impact Points for Collisions where Driver Inattention/Distraction was a Cause of the Collision

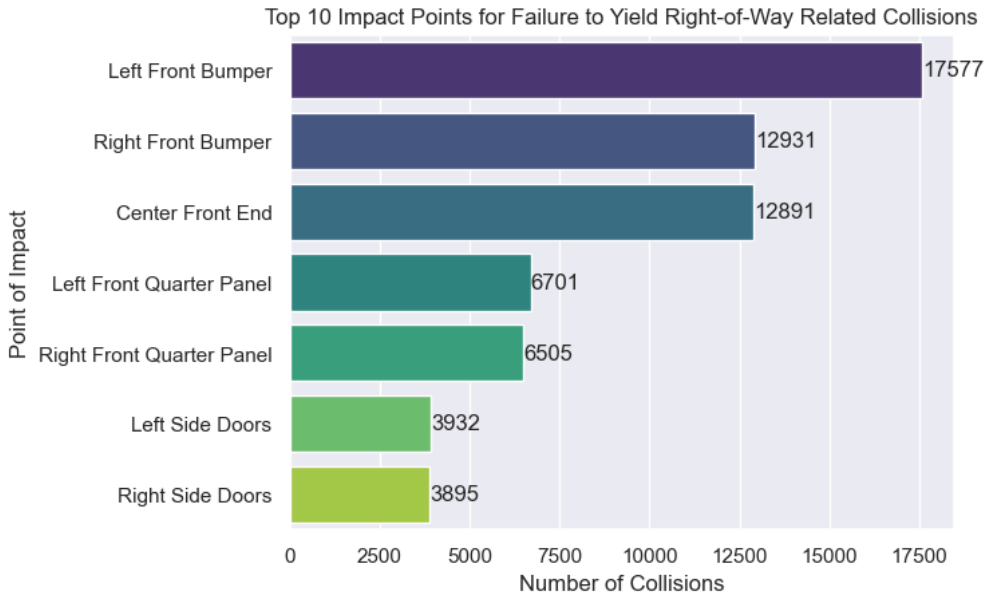


Figure 5: Top 10 Vehicle Impact Points for Collisions where Failure to Yield to Right Of Way was a Cause of the Collisions

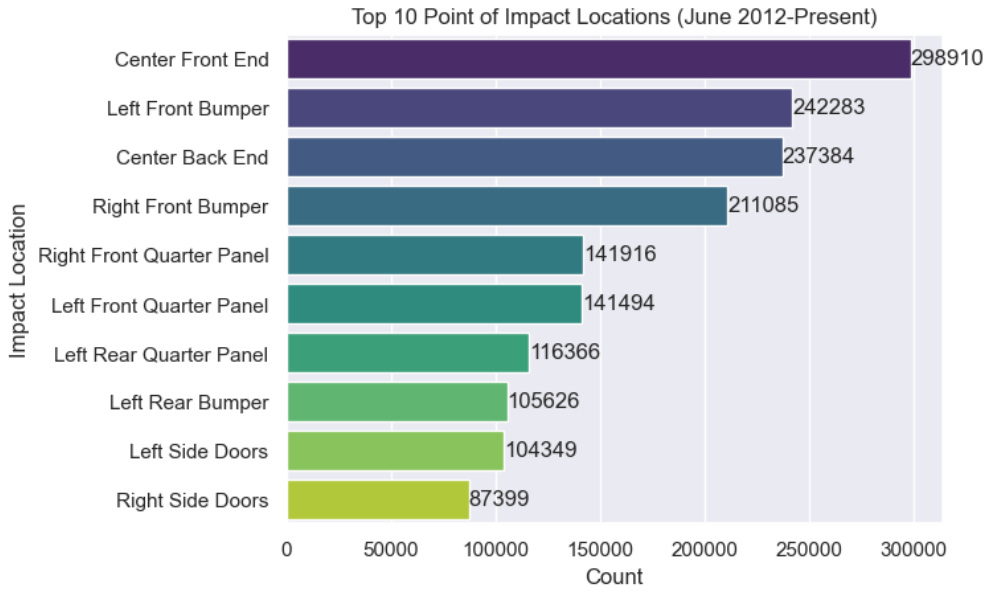


Figure 6: Top 10 Vehicle Impact Points for Collisions in NYC

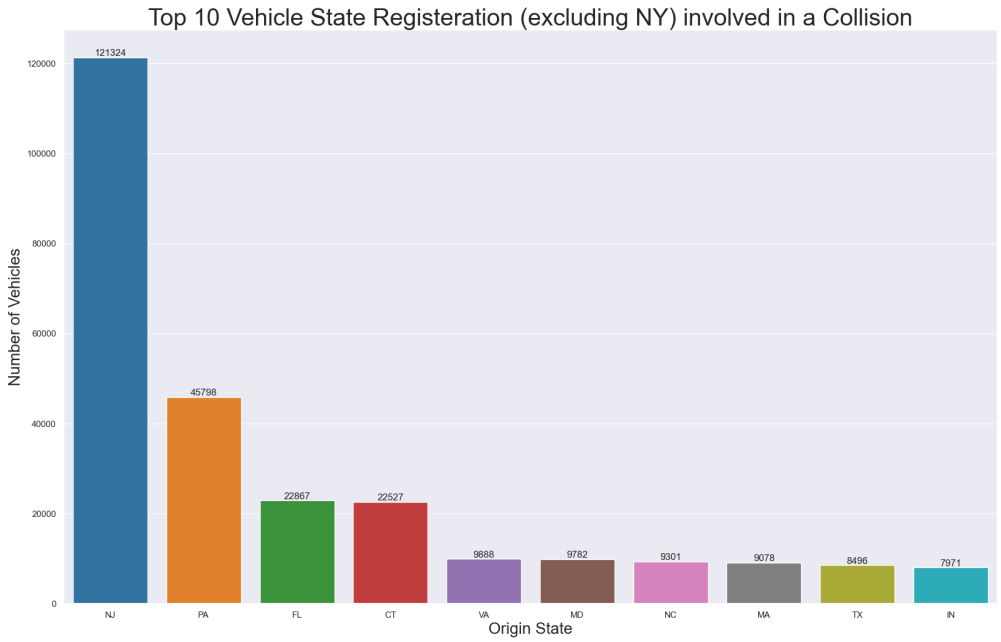


Figure 7: Top 10 States (excluding NY) where vehicles involved in collisions are registered to

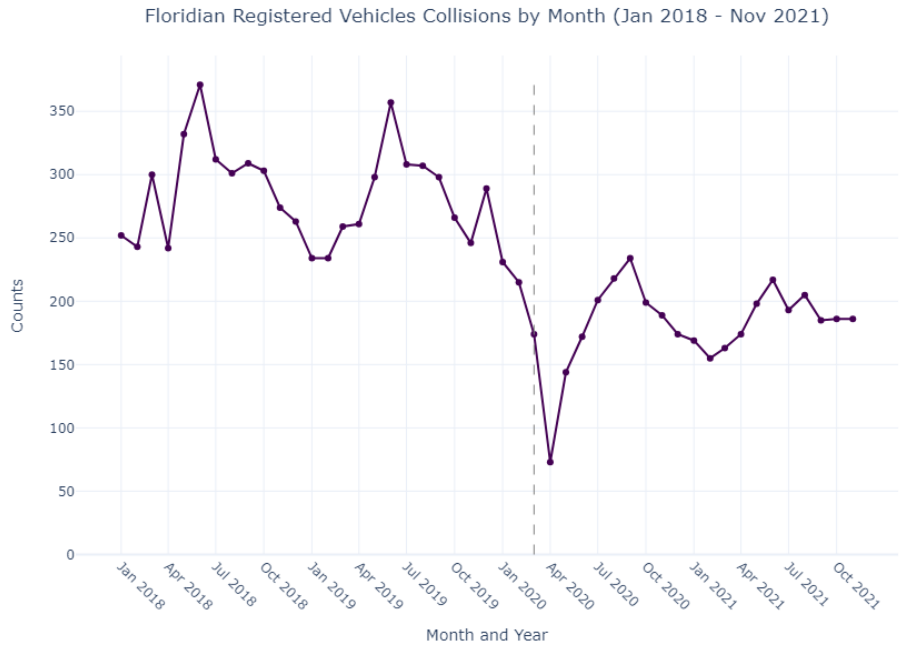


Figure 8: Trend of collisions with Floridian vehicles reported between Jan 2018 and Nov 2021

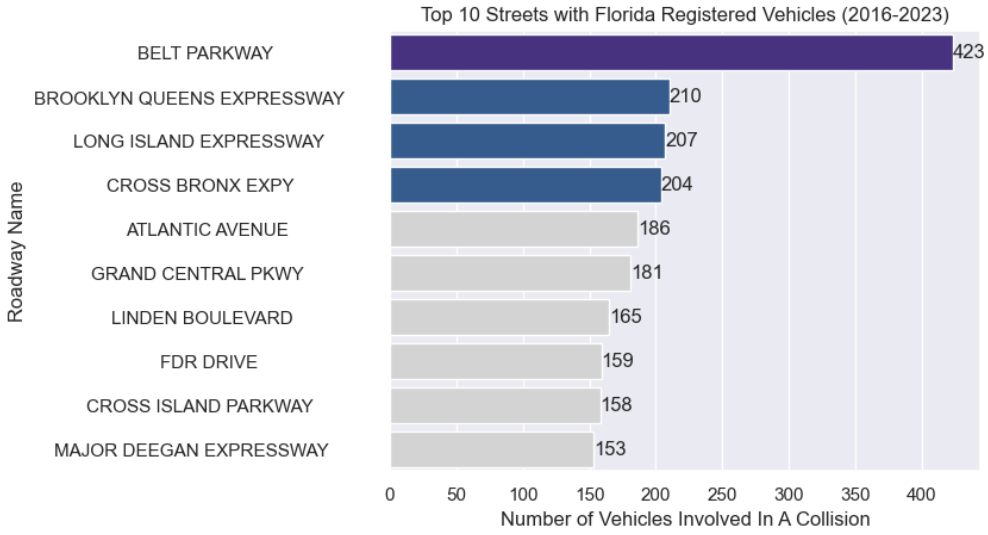


Figure 9: Top 10 Roadways where Floridian registered vehicles were involved in a collision.

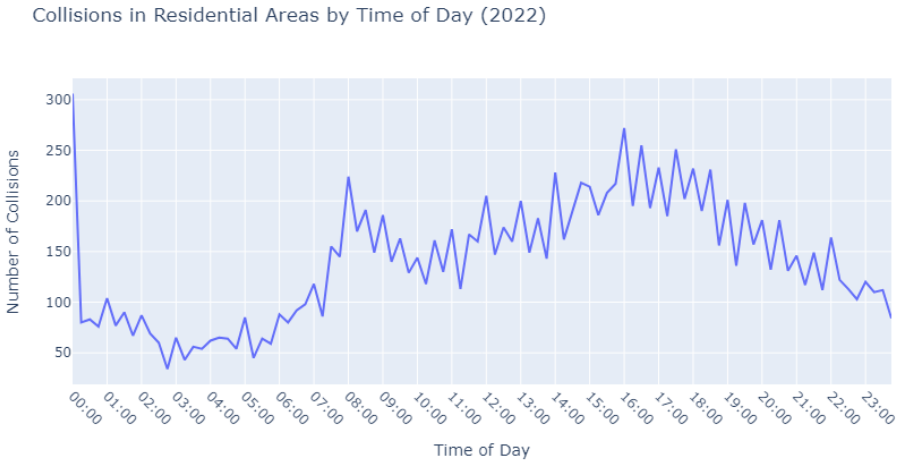


Figure 10: Temporal Patterns of Reported Collisions in 2022.

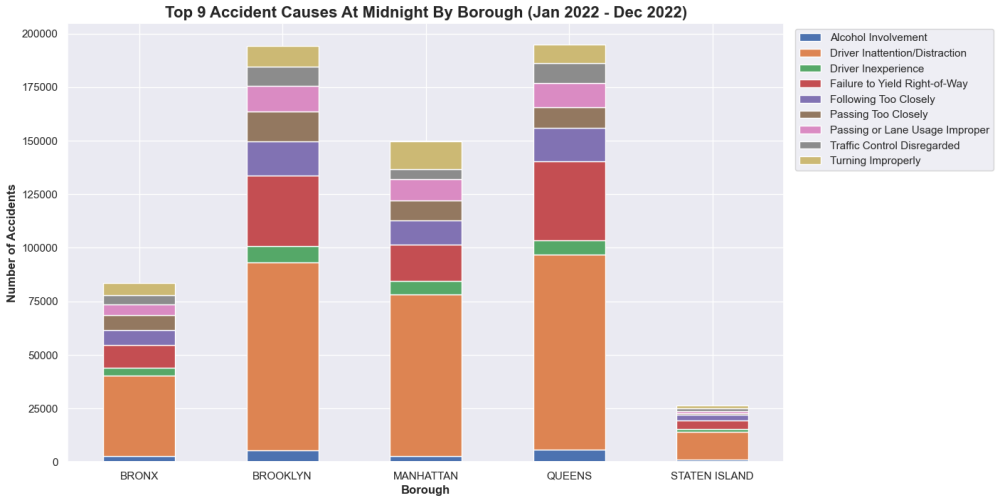


Figure 11: Top 9 Accident Causes for Collisions at Midnight in 2022.

Appendix C: Future Plans

The future idea is to develop a web interface where people can explore the dataset, explore the graphs used in this report, test out the model, and contribute or provide ideas. As of the writing of this report, the project has been uploaded to a GitHub repository[13]

The following is a list of ideas for how to take this project further:

- Deploy the project to a web interface for researchers and other users to use.
- Incorporate NYC Traffic Volume data from the NYC Department of Transportation.
- Incorporate data from NYC MTA datasets to incorporate bus routes.
- Develop a map interface for users to explore the data to draw their conclusions.

References

[1] NYC Open Data: NYPD, "Motor Vehicle Collisions - Crashes," https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95?source=post_page-----9f06c94140f2----------------------------, 2023.

[2] M. L. Waskom, "seaborn: statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. [Online]. Available: https://doi.org/10.21105/joss.03021

[3] P. T. Inc. (2015) Collaborative data science. Montreal, QC. [Online]. Available: https://plot.ly

[4] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] NYC Open Data: NYPD, "Motor Vehicle Collisions - Vehicles," https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Vehicles/bm4k-52h4, 2023.

[7] NYC Planning, "NYC GIS Zoning Features," https://www.nyc.gov/site/planning/data-maps/open-data/dwn-gis-zoning.page#metadata, 2023.

[8] ——, "About Zoning Districts," https://www.nyc.gov/site/planning/zoning/districts-tools.page, 2023.

[9] Weather Underground, "NYC Weather History," https://www.wunderground.com/history/daily/us/ny/new-york-city/KLGA, 2023.

[10] A. M. Ivan Pereira. (2021) A year of COVID-19: What was going on in the US in March 2020. [Online]. Available: https://abcnews.go.com/Health/year-covid-19-us-march-2020/story?id=76204691

[11] C. A. C. Ortega, M. A. Mariscal, W. Boulagouas, S. Herrera, J. M. Espinosa, and S. García-Herrero, "Effects of mobile phone use on driving performance: An experimental study of workload and traffic violations," *International Journal of Environmental Research and Public Health*, vol. 18, no. 13, p. 7101, 2021. [Online]. Available: https://doi.org/10.3390/ijerph18137101

[12] J. Brownlee, "Random oversampling and undersampling for imbalanced classification," 2021.

[13] D. Wang, "NYC Collision Study," https://github.com/dwang312/NYC-Collision-Study, 2024.