# Mini-Project 2

## 1 Introduction

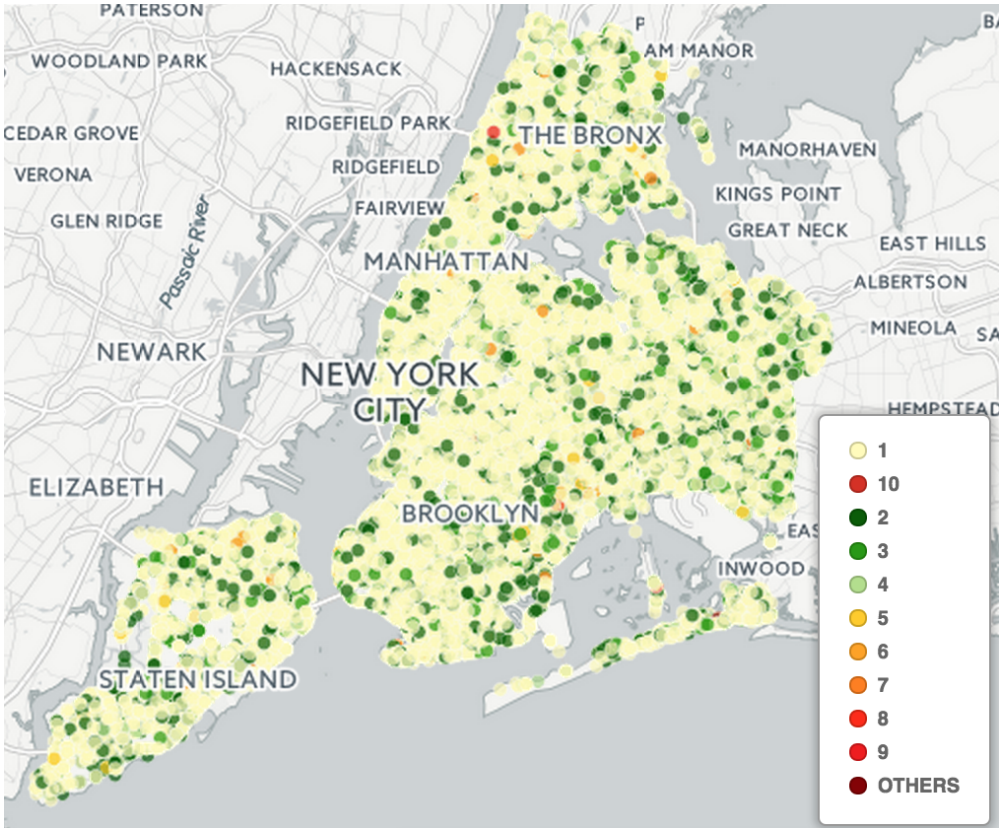


Figure 1: Intensity of injuries from accidents in NYC [1]

Quite a lot of accidents happen on the streets of New York City almost everyday. Some of these accidents injuries, from figure 1, some of them were even fatal, from figure 2 and some just ended up with property damage, be it private or public property. This simply means accidents are bad. The data [2] that is being analyzed here contains information about such accidents, courtesy NYPD. This data contains some details about the collisions that occurred from July, 2012 to September, 2015.

The following are some columns from the data set (not exhaustive):

1. Date
2. Time
3. Borough
4. Latitude
5. Longitude
6. Zip Code
7. On Street Name
8. Cross Street Name
9. Contributing factors
10. Types of vehicles involved
11. Number of persons injured
12. Number of persons killed

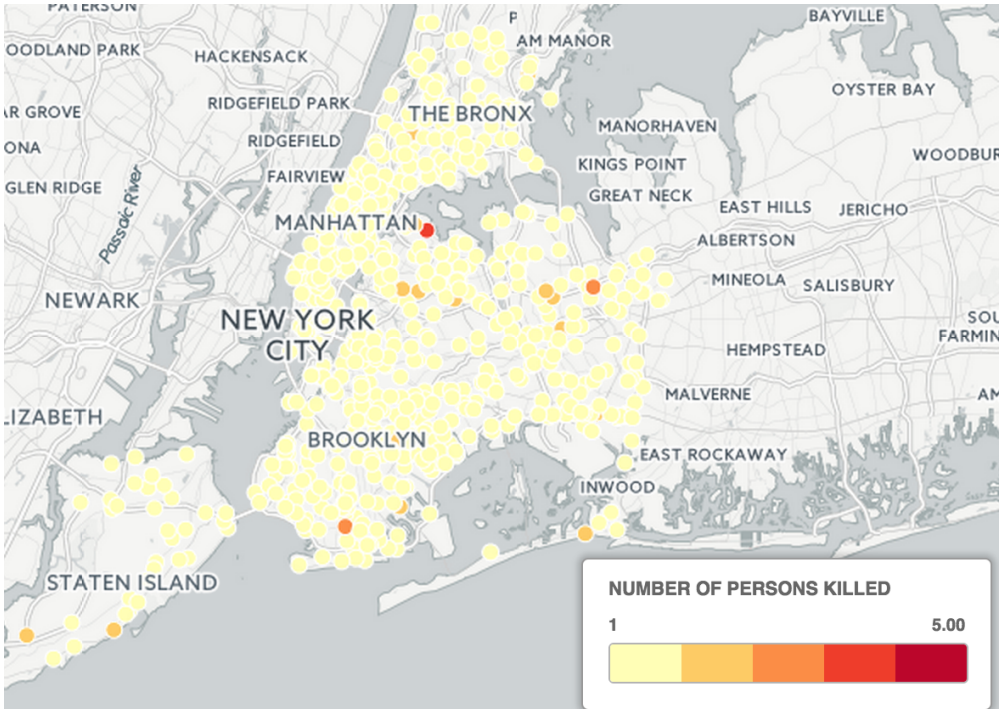


Figure 2: Intensity of fatalies from accidents in NYC [1]

# 2 Problem Statement

The objective is to make New York City a better place by predicting the chance of getting injured or injuring any citizen of New York City based on one's regular living patterns.

Based on the given data [2], I try to classify the information to answer injury-like or injury-unlike based on some features obtained from the data set. By constructing a model using Naive Bayesian classifier method, I was able to predict with about 70% accuracy, the chance of injury to either party involved in the accident.

# 3 Observations

By Exploratory Data Analysis and Visualization, I tried to convert the data to categorical data. I tried to observe the accident location on map but it didn't show any conclusive results as in almost every street or intersection there was a case of accident. So my first thought was to check if accidents depend on time of the day or have some relation with day of the week. For example, number of people traveling on the streets during weekends may be less or there could be comparatively more traffic during the peak hours as people commute to work.

## 3.1 Accidents based on time of the day

From figure 3, percentage of accidents on weekends are lesser than weekdays which can be attributed to less traffic patterns on weekends. Additionally, on Friday, percentage of accidents are high, leading us to probably attribute this observation due to impatience and excitement towards the weekend.

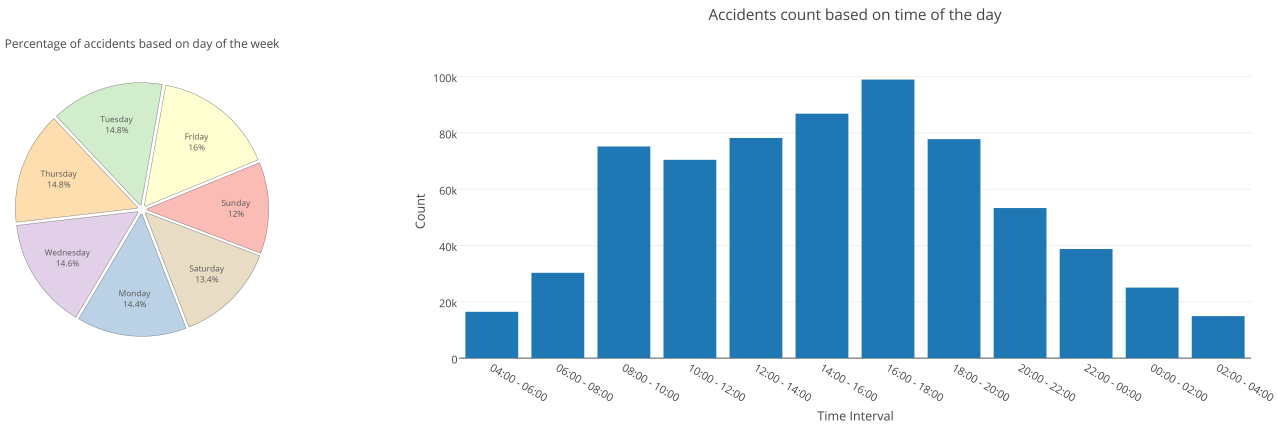


Figure 3: Accidents based on date and time

Also, based on the time of day, we can observe that most number of accidents happen from **04:00PM to 06:00PM** which is typically the departing time from offices and second most is from **08:00AM to 10:00AM** again typically the start time from home.

## 3.2 Accidents based on some factors

From figure 4, most of the accidents are due to driver inattention/distraction which can be majorly due to technology such as radio, cell phones, navigators or failure to follow traffic rule or fatigue. From the above observations based on date and time, we can infer that fatigue and failure to follow traffic may be due to tiring day at office.

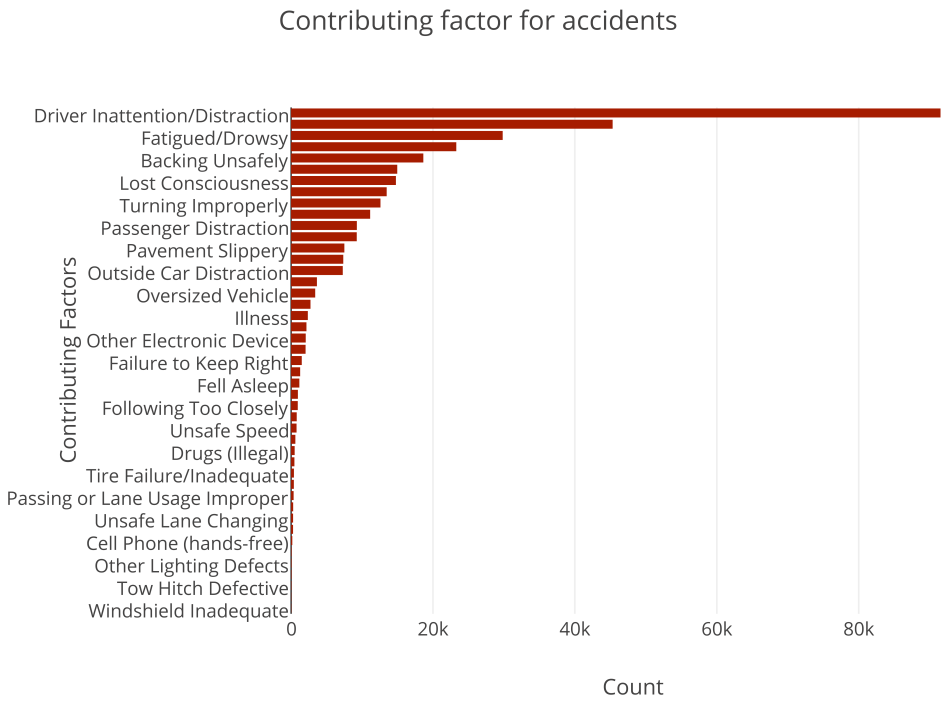


Figure 4: Contributing factors for accidents

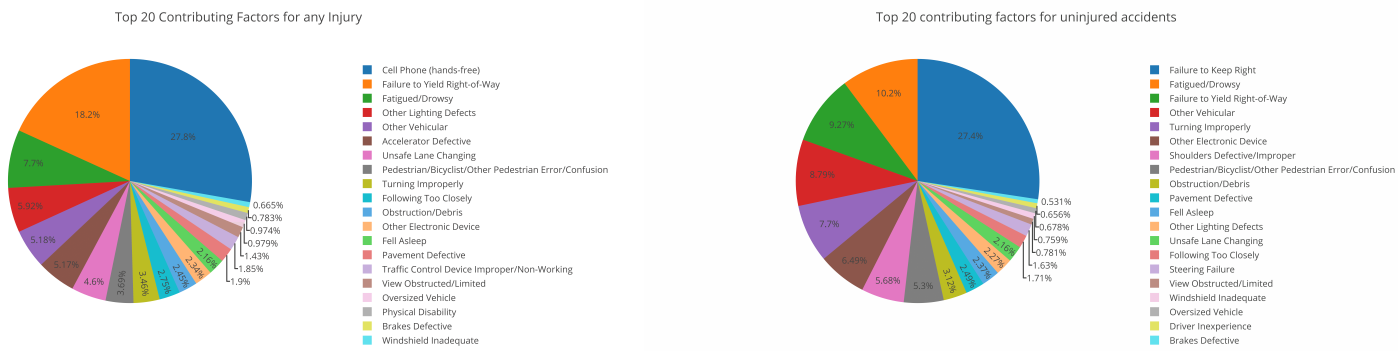


Figure 5: Percentage distribution of contributing factors

From figure 5, we can observe that 'cell phones' usage is top reason for injuries. This can be attributed to use of cell phones in crucial turn or during heavy traffic areas. Second most reason is due to "Failure to Yield Right-of-Way" which could be attributed to injury of pedestrians trying to cross roads or due to collision with other parked vehicle on the road.

Whereas, top factors for non-injury collisions are "Failure to keep right" and "Fatigued/Drowsy". This could mean that even though the mistake might be from the contributing vehicle of the collision, the other party might have been careful to avoid injuries or the collision might have been with some public/private property in the city.

## 3.3 Accidents based on type of vehicles

From figure 6, "Passenger Vehicle" is the major cause for accidents, which I'm assuming to be regular cars used for commute by the citizens of NYC. This is complimenting with previous observations where we observed that 'Driver Inattention/Distraction' is the top most factor for motor collisions. From figure 7, collisions from passenger vehicles lead to injuries in most of the cases and collisions with SUVs looks to be more secure as it tops the chart of no injury. We might infer that SUVs are safer than other passenger vehicle class and one suggestion to NYC citizens might to be consider going for SUVs.

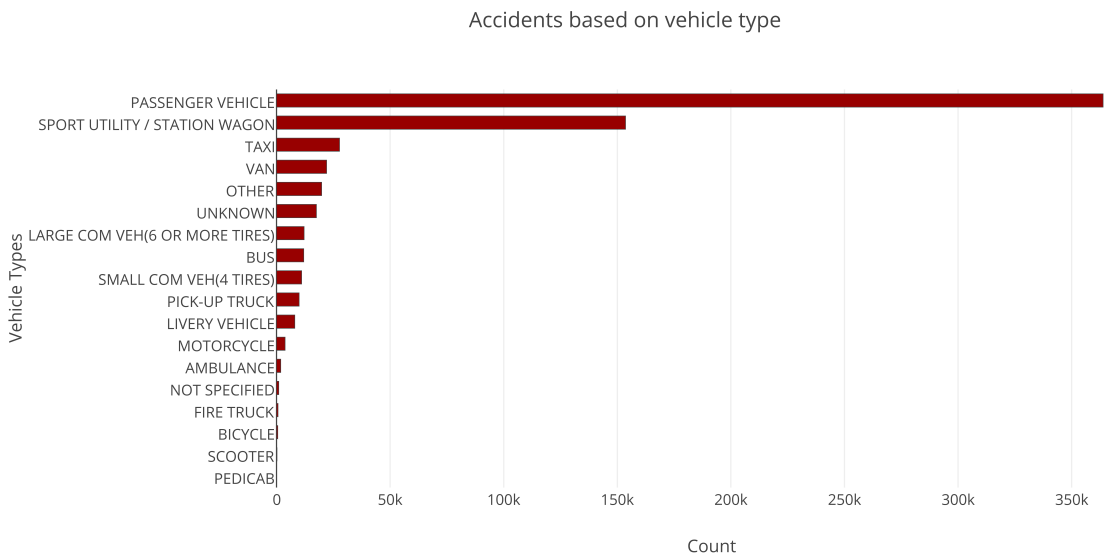


Figure 6: Accidents based on vehicle types

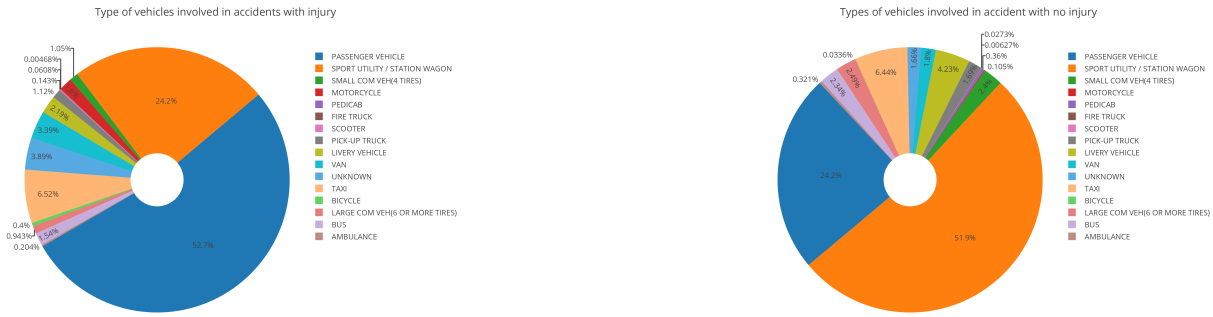


Figure 7: Percentage of vehicle types

## 3.4 Accidents based on type of road

Street names denote some properties about the street. For example, avenues are generally broader than streets and expressway or highways are broader than avenues. Since there are not much highways or expressways through the cities, from chart 8, most of the accidents have taken place on avenues and streets.

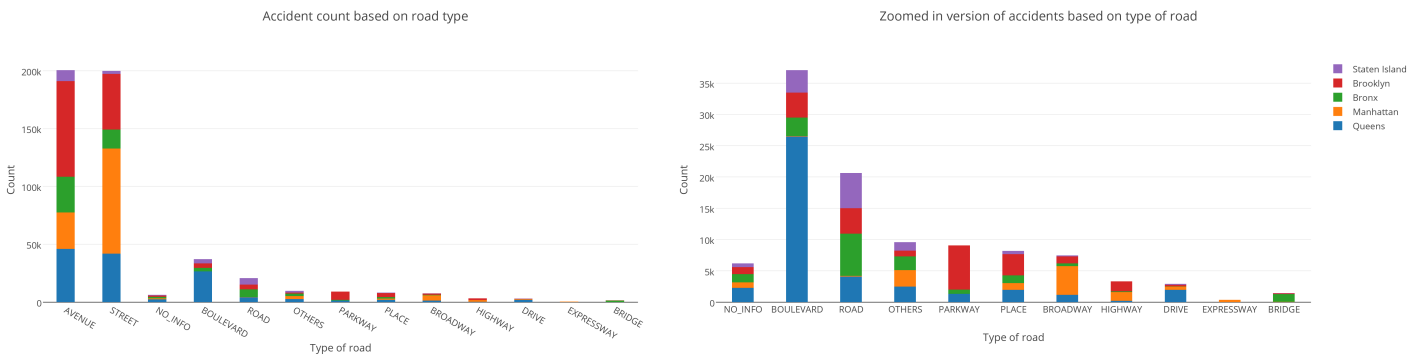


Figure 8: Accidents based on road type grouped by borough

# 4 Fitting Data to a Model

Based on the visualization of various parameters seen in section 3, I have categorized the data for producing a Naive-Bayesian model to predict the chance of injuring a New York citizen.

As the percentage of injuries/kills (about 20%) is far less than non-injuries, I'm **over-sampling** positive data to bring an equal spilt.

As an input data to NB, I pass the data with 41.8% of positive value and remaining with negative values with default probability of 0.001.
Following are the features used by NB model:

1. cfv1 - Contributing factors to the accident 1 (categorical)

2. cfv2 - Contributing factors to the accident 2 (categorical)

3. date_time - Cross product of number of days and different time periods (categorical)
4. roadtype - Cross product of onstreet and crsstreet (intersection information) (categorical)
5. vehicletype1 - Vehicles involved in the accident (categorical)
6. vehicletype2 - Vehicles involved in the accident (categorical)
7. injury_status - 0/1 value representing happening of injury to either party or not

The cleaned up data set, without **null** values and not so useful values in features such as "Unspecified", "Unknown", "others", contains about **81,487** rows [3] of data.

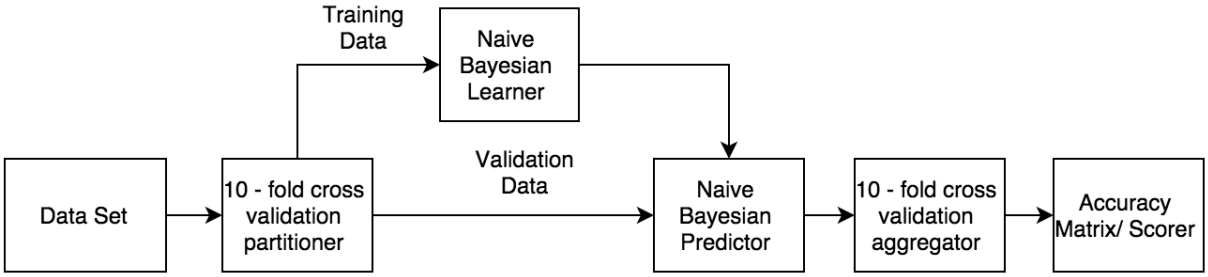


Figure 9: Naive - Bayesian Model Implementation[4]

# 5 Results

By using a tool named "Knime" [5] for applying NB technique, following are the results obtained. With 10-fold Cross Validation, the model's average accuracy was 69.8%. The table below enumerates percentage of error, size of the test set, error count and accuracy.

| **Fold Num** | **Error in %** | **Size of test set** | **Error count** | **Accuracy in %** |
|---|---|---|---|---|
| Fold 1 | 30.58043932 | 8149 | 2492 | 69.41956068 |
| Fold 2 | 31.00993987 | 8149 | 2527 | 68.99006013 |
| Fold 3 | 30.5190821 | 8149 | 2487 | 69.4809179 |
| Fold 4 | 30.19145803 | 8148 | 2460 | 69.80854197 |
| Fold 5 | 30.34728188 | 8149 | 2473 | 69.65271812 |
| Fold 6 | 29.43919499 | 8149 | 2399 | 70.56080501 |
| Fold 7 | 28.75552283 | 8148 | 2343 | 71.24447717 |
| Fold 8 | 30.10185299 | 8149 | 2453 | 69.89814701 |
| Fold 9 | 30.06503866 | 8149 | 2450 | 69.93496134 |
| Fold 10 | 31.29602356 | 8148 | 2550 | 68.70397644 |
| | | | **Total** | 697.6941658 |
| | | | **Average Accuracy** | **69.76941658 ≈ 69.8%** |

| | M1 | M2 | M3 | M4 |
|---|---|---|---|---|
| U1 | 5 | 3 | - | 1 |
| U2 | 4 | - | - | 1 |
| U3 | 1 | 1 | - | 5 |
| U4 | 1 | - | - | 4 |
| U5 | - | 1 | 5 | 4 |

Below is the confusion matrix obtained which contains TrueNegative, FalseNegative, TruePositive and FalsePositive values.

**Confusion Matrix** [5]

| | | **Prediction** | | |
|---|---|---|---|---|
| | | False | True | Total |
| **Actual** | False | 38703 | 16182 | 54885 |
| | True | 8452 | 18150 | 26602 |
| | Total | 47155 | 34332 | - |

# 6 Conclusion

From the visualizations, we can conclude that majority of accidents were due to passenger vehicles and occurred during the peak office hours of the day (08:00 AM to 10:00 AM and 04:00 PM to 06:00 PM). Also, there were considerably less number of accidents over the weekends as the traffic during weekends is less. Additionally, major contributing factor for accidents was 'Driver Inattention/Distraction' which was again due to either work tension or fatigue. Number of accidents may be reduced if we increase the awareness among the citizens to drive carefully especially after work which in turn decreases number of injuries.

We can use the above model to predict the chance of getting injured or injuring some one with a probability of about 70% given the values of the features used. Here, the percentage of accuracy is low, as the amount of data leading to injuries were significantly less and further due to the fact that we oversampled those data to construct an unbiased model. Also, the features attributing to classification were almost uniformly distributed which led to poor prediction of classification.

The accuracy can be further improved if we try to fit the model using Neural Networks or Support Vector Machines technique.

# References

[1] CartoDB `https://cartodb.com/`

[2] NYPD Motor Vehicle Collisions Data Set `https://data.cityofnewyork.us/Public-Safety/NYPD-Motor-Vehicle-Collisions/h9gi-nx95`

[3] PostgreSQL DB `http://www.postgresql.org/`

[4] Flow Charts Online `https://www.draw.io`

[5] Knime `https://www.knime.org/`

[6] Plot.ly - Plot charts online `https://plot.ly`