Motor Vehicle Collisions

Team 5 - Jiaxi Liu, Alice Pascalev, Abhimanyu Barun, Pranav Sai

Contents

- Topic Introduction
- Data Description
    - Data Source
    - Data Cleaning
    - Dataset Information
- Exploratory Data Analysis
- SMART Questions
    1. How does other variables affect the number of motorist injured?
    2. Effect of time of the day and time of year on the number of vehicles involved
    3. Can we predict number of vehicles involved in the accident?
    4. Modeling Type of Car - 1 involved in a two car accident
    5. Modeling Type of Car - 2 involved in a two car accident
- Issues Faced

Topic Introduction

Topic Introduction

More than 46,000 people die in car crashes each year, according to Annual United States Road Crash Statistics (ASIRT). Furthermore, the latest evidence has found that post-pandemic, reckless driving in the US has surged, with more cases of speeding, unbuckled seatbelts, and impaired driving[1]. Being struck by a vehicle is the leading cause of injury-related death of children in the United States, according to the CDC.[1]

Therefore, the topic of our project is Motor Vehicle Collisions, focusing on motor vehicle collisions, particularly those that require police reports.

1. Baumgaertner, E. (2021, December 8). *Car crash deaths have surged during COVID-19 pandemic. Here's why*. Los Angeles Times. Retrieved December 13, 2022, from https://www.latimes.com/world-nation/story/2021-12-08/traffic-deaths-surged-during-covid-19-pandemic-heres-why
2. https://www.cdc.gov/injury/features/child-injury/index.html

Data Description

Data Source

We are utilizing the data set ‘**Motor Vehicle Collisions - Crashes**’, which we accessed at Data.gov. This data is gathered from NYPD and includes data from all motor vehicle collisions in New York City that require a police report. This data set has 1,000,000 rows and 18 variables.

This data was made public in 2014, the year NYC’s Vision Zero program was implemented, with the goal of safer streets and zero fatalities. The data is from 2014 until Nov 18, 2022.

“Police report (MV104-AN) is required to be filled out for collisions where someone is injured or killed, or where there is at least $1000 worth of damage.” [1]

VISION ZERO
Building a Safer City

[1] https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95

Motor Vehicle Collision - Crashes

CRASH TIME
NUMBER OF PERSONS INJURED
NUMBER OF PERSONS KILLED
NUMBER OF PEDESTRIANS INJURED
NUMBER OF PEDESTRIANS KILLED
NUMBER OF CYCLIST INJURED
NUMBER OF CYCLIST KILLED
NUMBER OF MOTORIST INJURED
NUMBER OF MOTORIST KILLED

CONTRIBUTING FACTOR VEHICLE 1, CONTRIBUTING FACTOR VEHICLE 2, CONTRIBUTING FACTOR VEHICLE 3, CONTRIBUTING FACTOR VEHICLE 4, CONTRIBUTING FACTOR VEHICLE 5

VEHICLE TYPE CODE 1, VEHICLE TYPE CODE 2, VEHICLE TYPE CODE 3, VEHICLE TYPE CODE 4, VEHICLE TYPE CODE 5

Data Cleaning

- Dropped NA values
- Convert Crash Date and Crash Time columns, to datetime format, split into Year, Month, Day and Hour, Minute columns
- Convert Zip Code from object -> integer
- Convert month number to name
- Create Vehicle Count column
- Due to seven-figure number of observations, only kept top 20 Vehicle Types for each Vehicle Type Code
    - Consolidated categories with misspellings, extra words, case differences (ie. SEDAN, Sedan, 3-door sedan)

Dataset info

Time Indicator:

Year, Month, Day, Hour, Minute

Location Indicator:

Borough, Zip code, Latitude, Longitude

# of injured/killed:

Persons, Pedestrians, Cyclist, Motorist

Vehicle info:

Contributing factor vehicle 1-5, Vehicle type code 1-5, Vehicle count

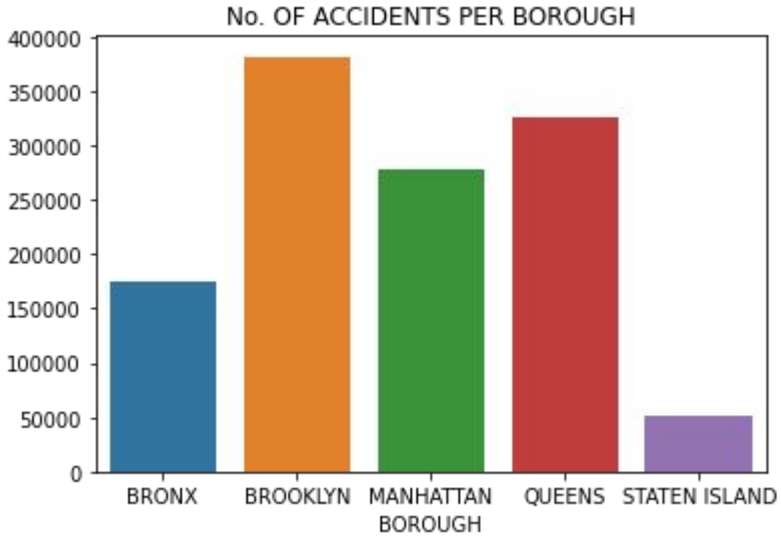
Exploratory Data Analysis

Accidents per Borough

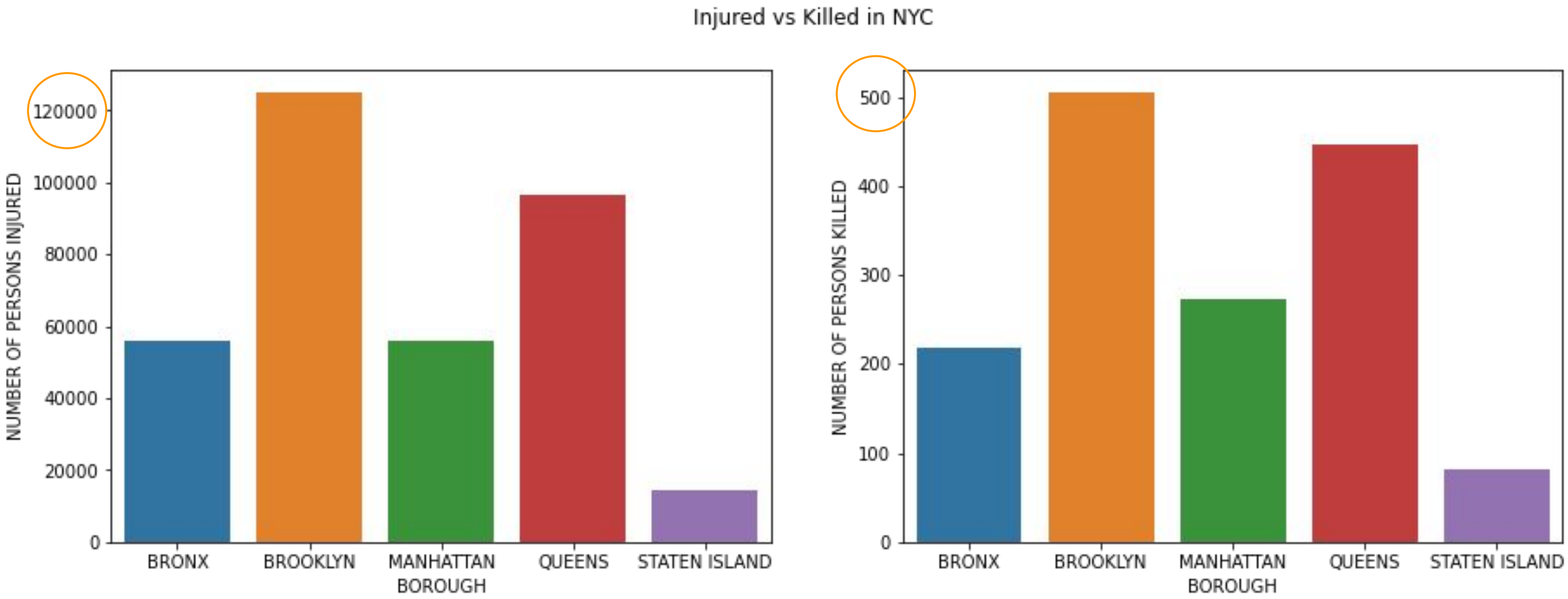
- Road accidents are more frequent in Brooklyn, Queens and Manhattan.

- Staten Island has the lowest number of accidents

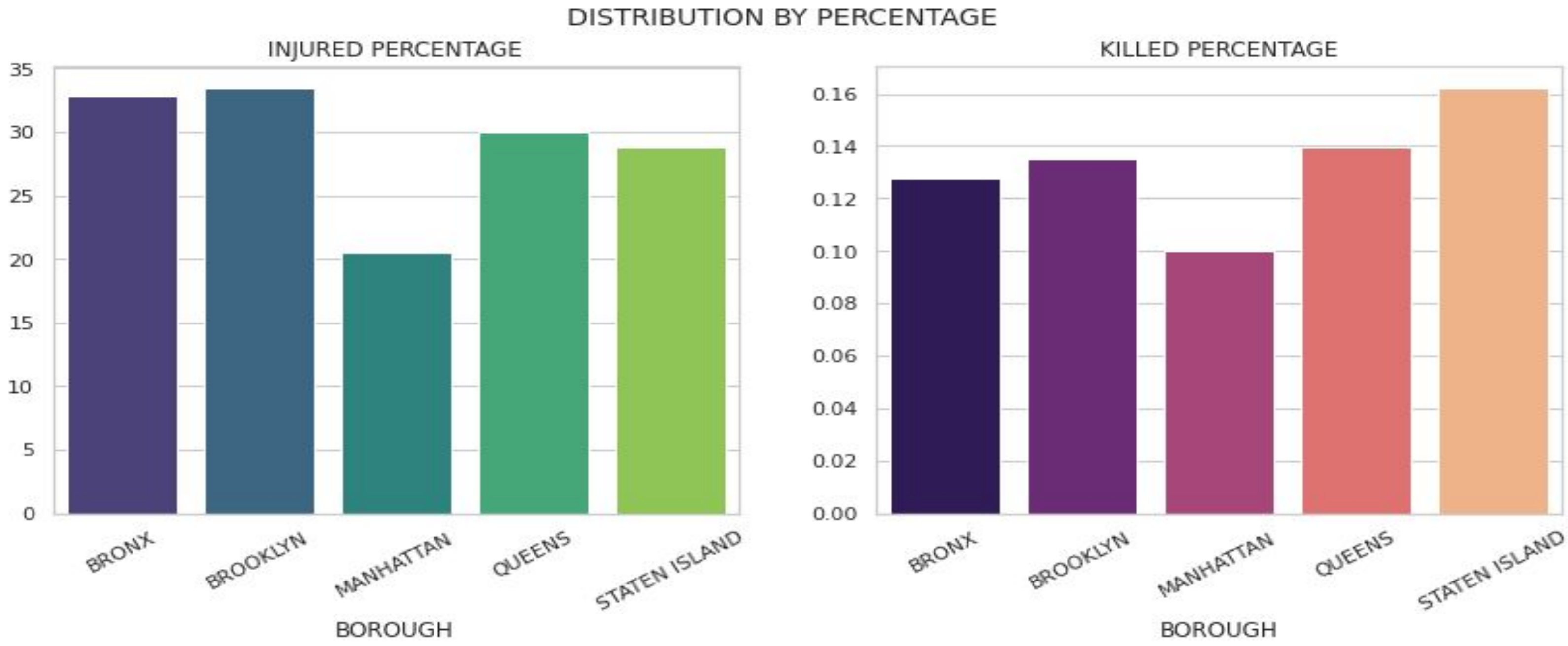
- Brooklyn has the highest number of accidents



Distribution of People Injured and Killed in NYC

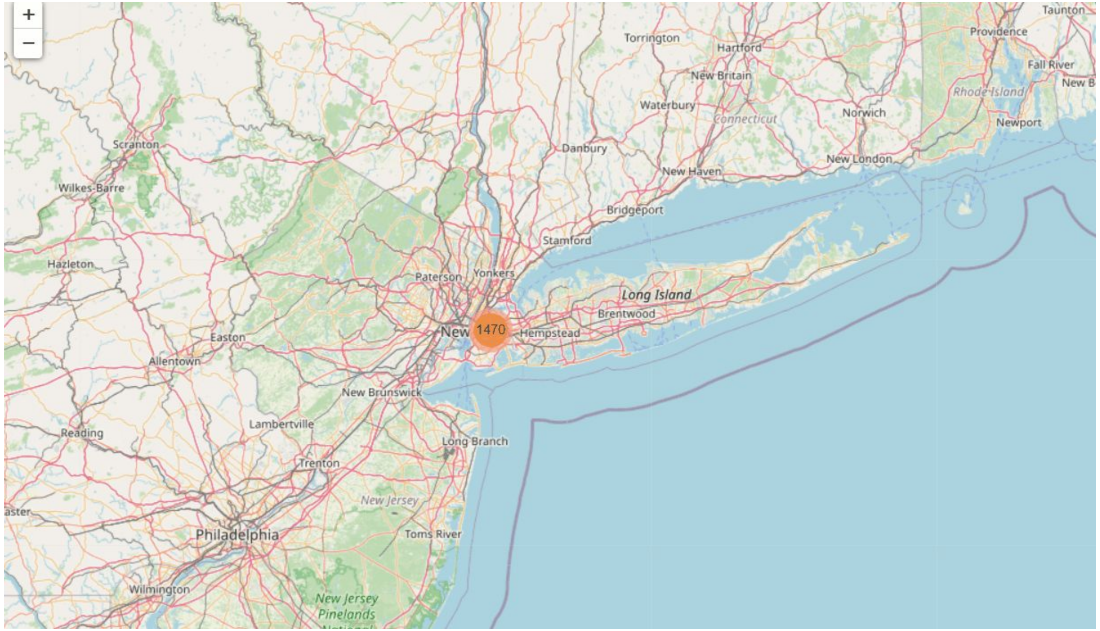


Distribution by Percentage

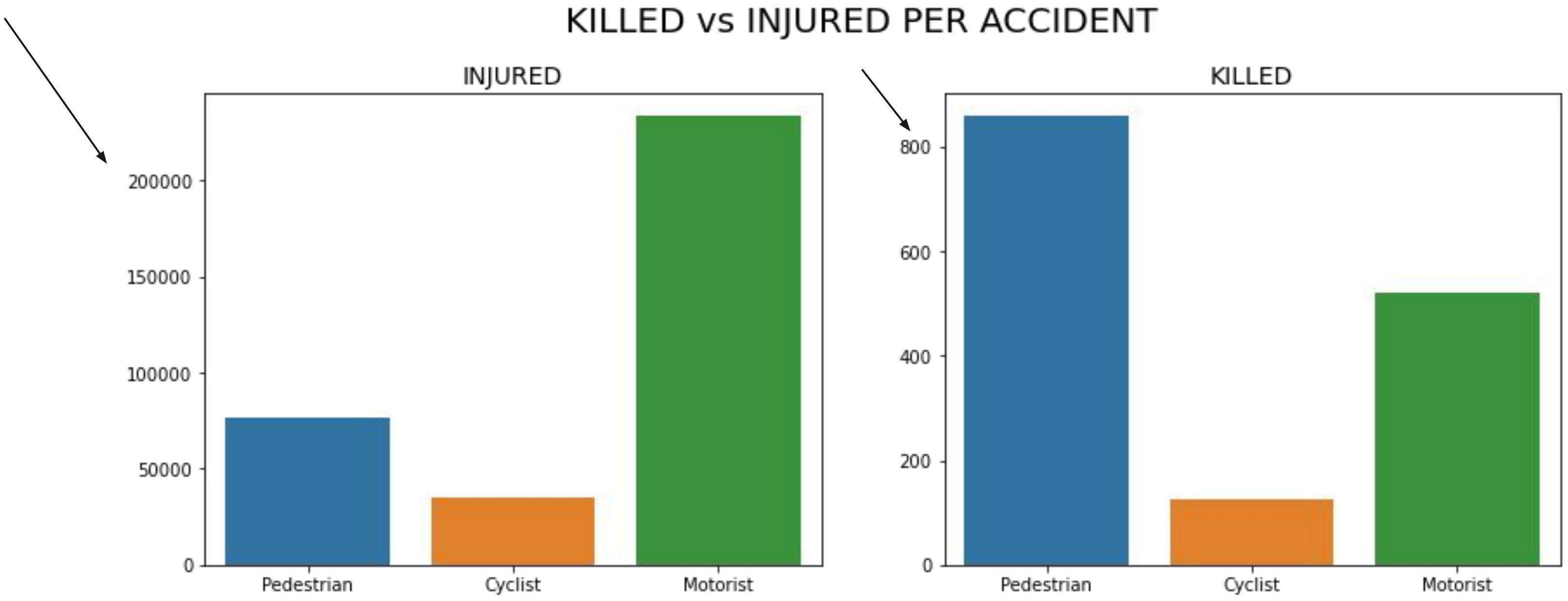


Impact of location on number of fatal accidents

Location wise lethal accidentes

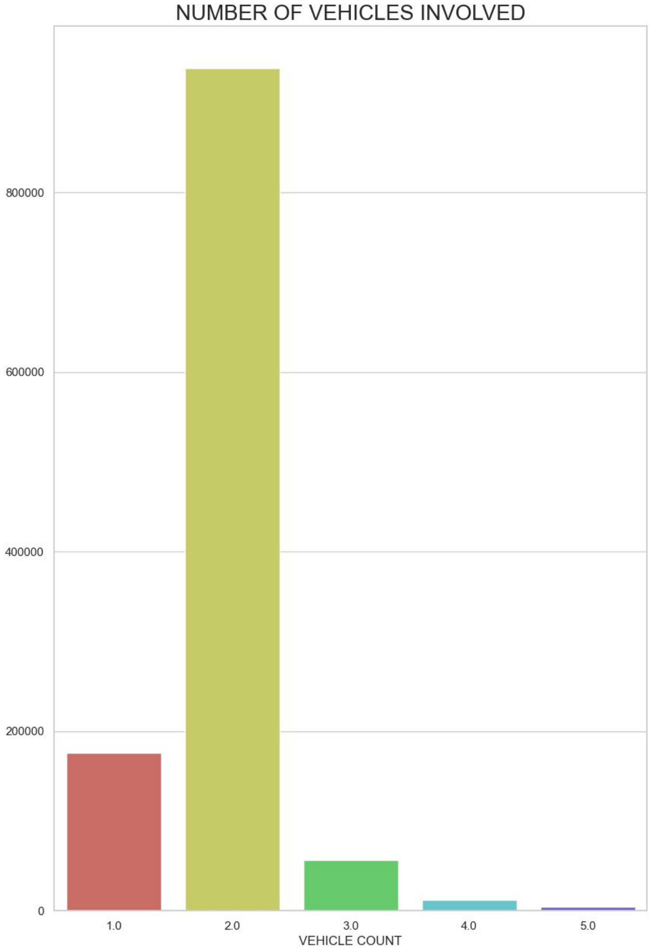


Killed and Injured per Accident



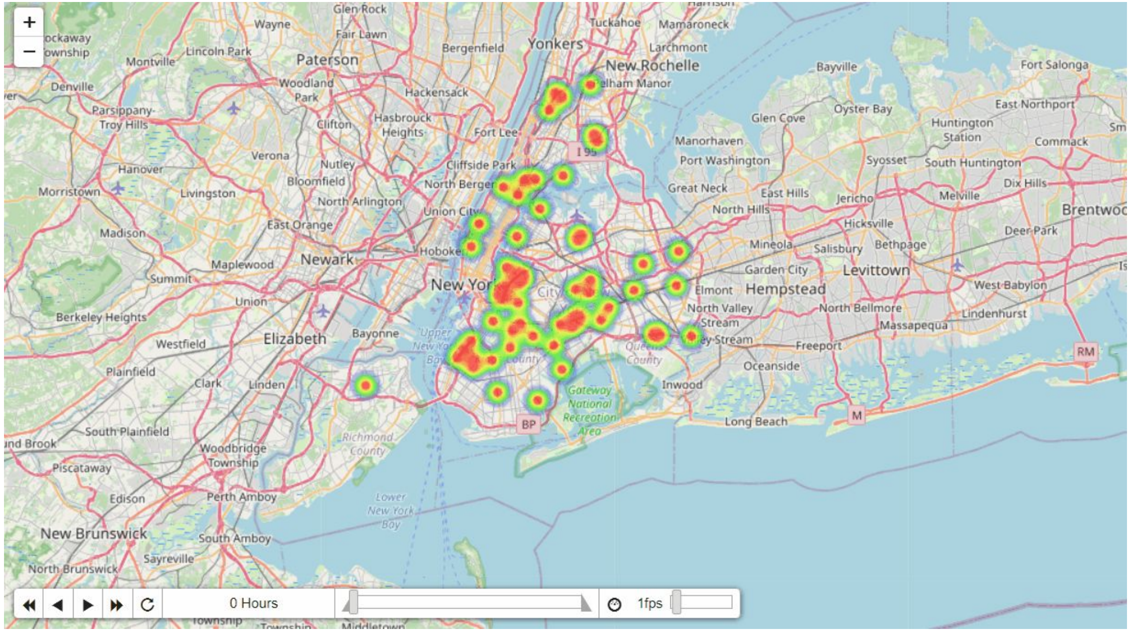
Vehicles Involved

We see that the vast majority of collisions involved two vehicles

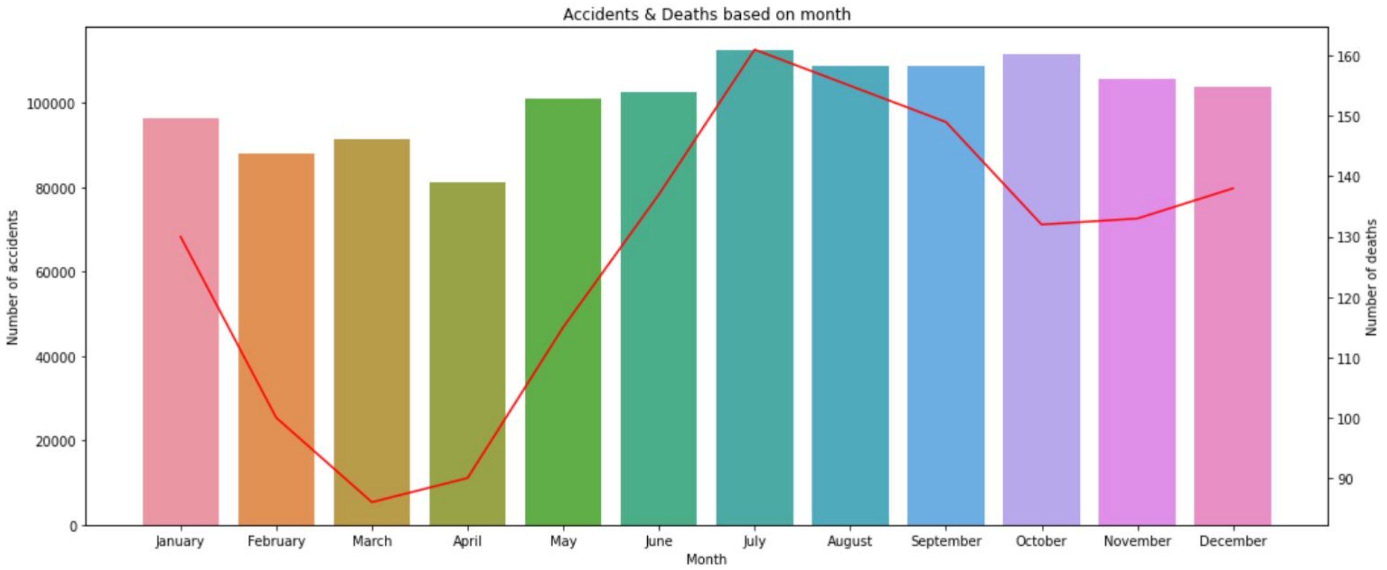


Impact of time of day on location of fatal crashes

Location and time wise lethal accidents



Accidents and Deaths based on Month



Please note: Does not include December 2022 statistics

Highest Frequency (Accidents)

1. **July** 112,401
2. **October** 111,716

Lowest Frequency (Accidents)

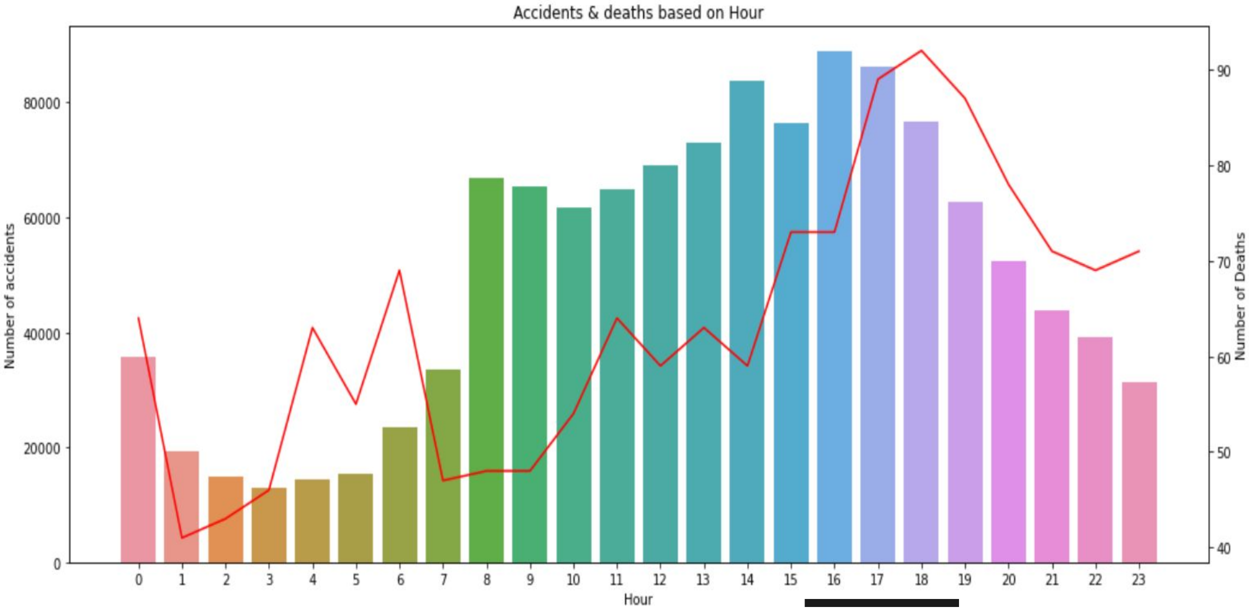
1. **April** 81,082
2. **February** 87,912

# deaths increase by 88% from month of march to July

Accidents and Deaths based on Hours

Most accidents are observed during the day

And the fatalities are highest at the end of the day from 6PM to 8PM in the evening



Rush Hour

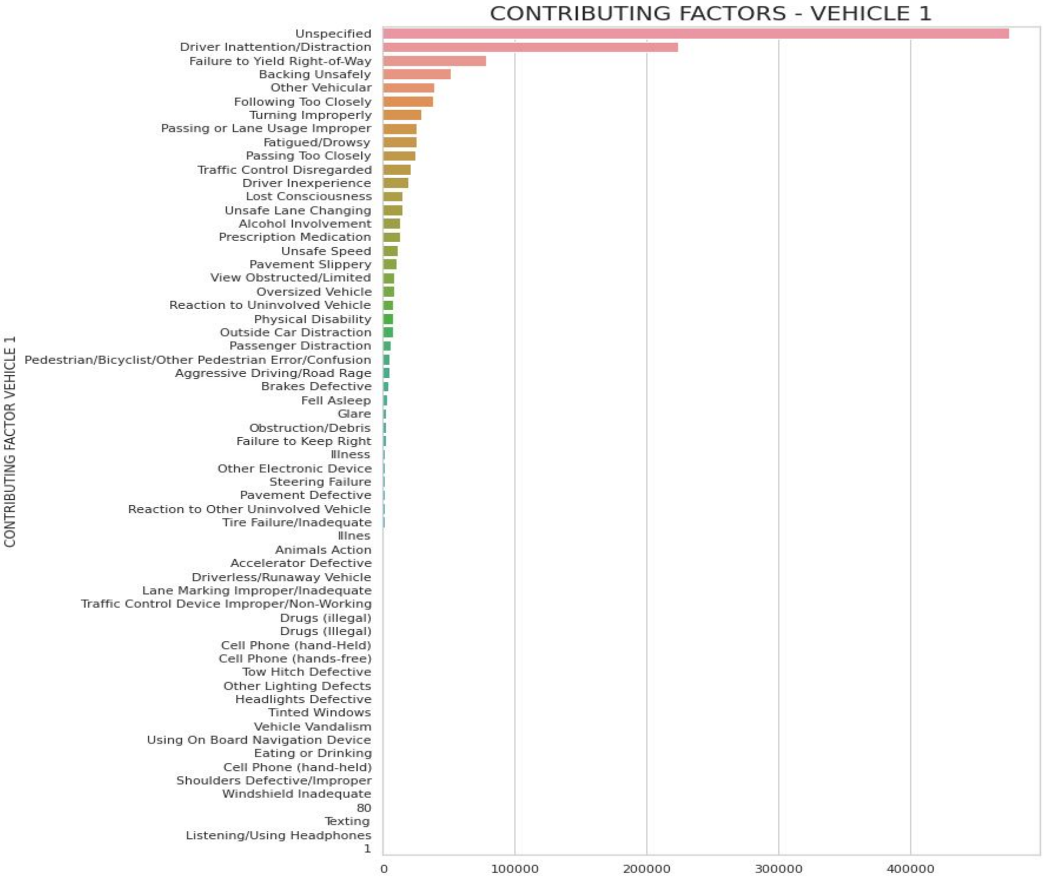
Contributing Factors of Vehicle 1

Over 20000 accidents were due to Driver Inattention or some other distraction.

Failure to Yield Right-of-Way and Driving Unsafely are the two next most common contributing factors to an accident.

Similar amount of Accidents were caused due to errors of another vehicle on road and Following too closely.

Other common contributing factors include - Turning Improperly, Improper lane usage, fatigue and passing too closely.

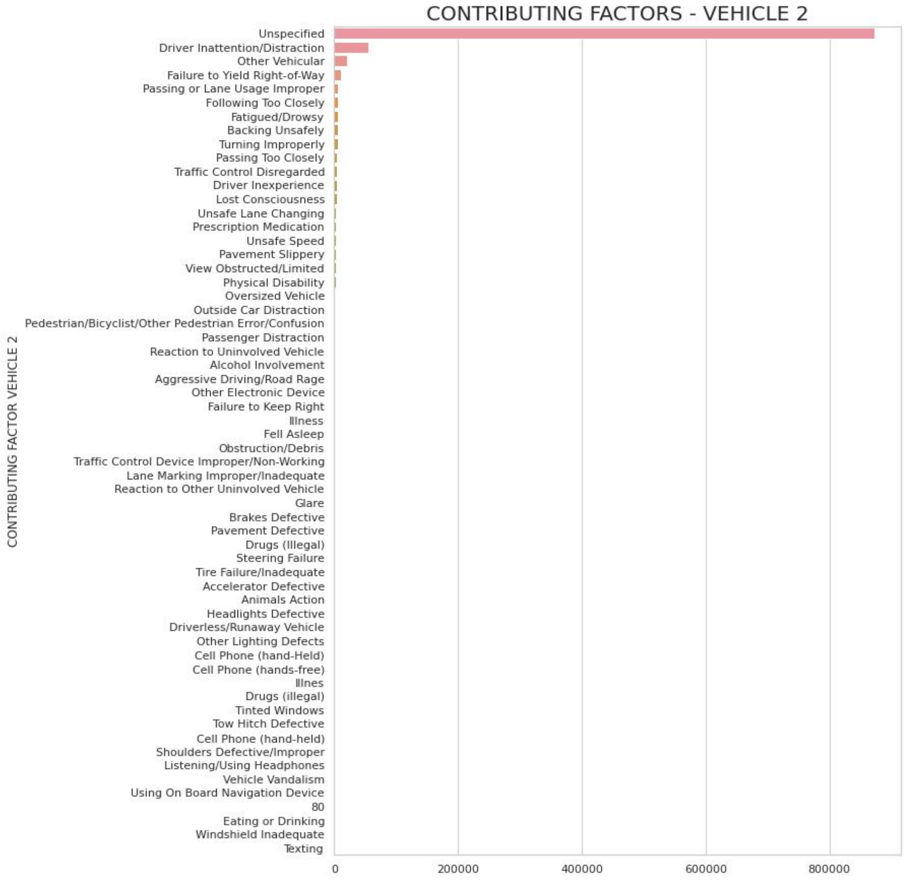


Contributing Factors of Vehicle 2

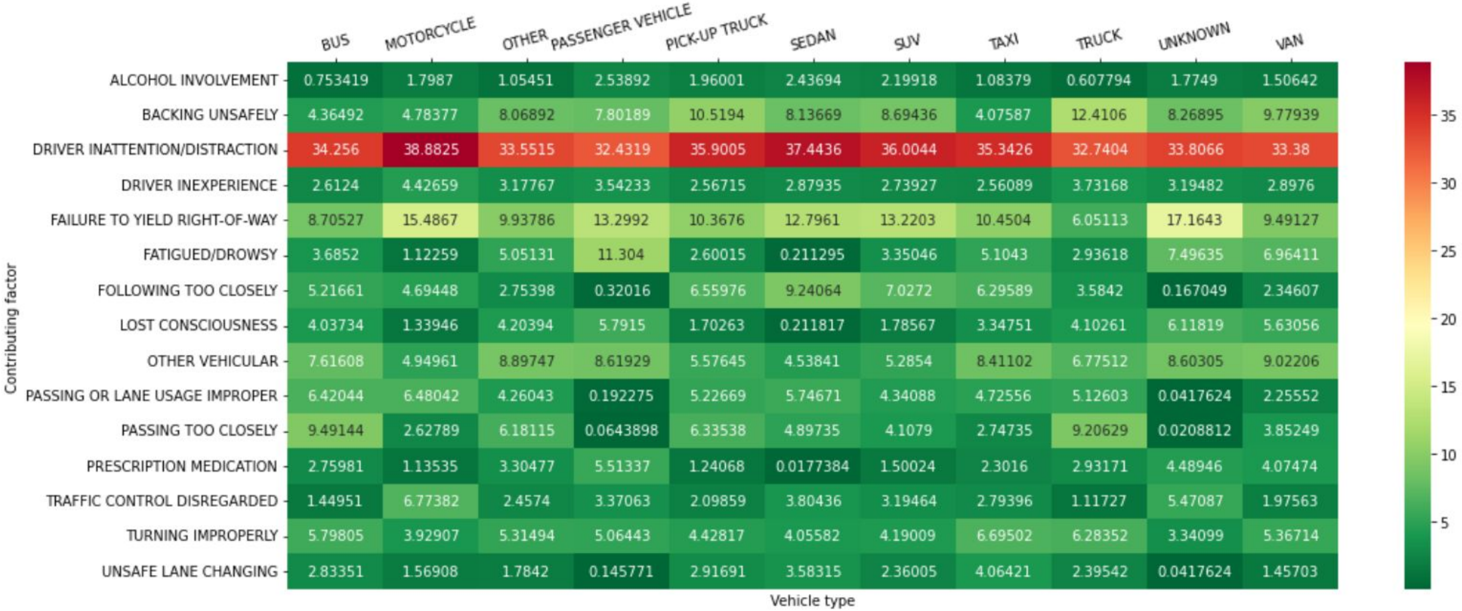
Most accidents were due to Driver Inattention or some other distraction.

Errors of another vehicle on road and Failure to Yield Right-of-Way are the two next most common contributing factors to an accident.

Similar amount of Accidents were caused Improper Lane usage, following too closely, Fatigue, backing unsafely and Turning improperly.

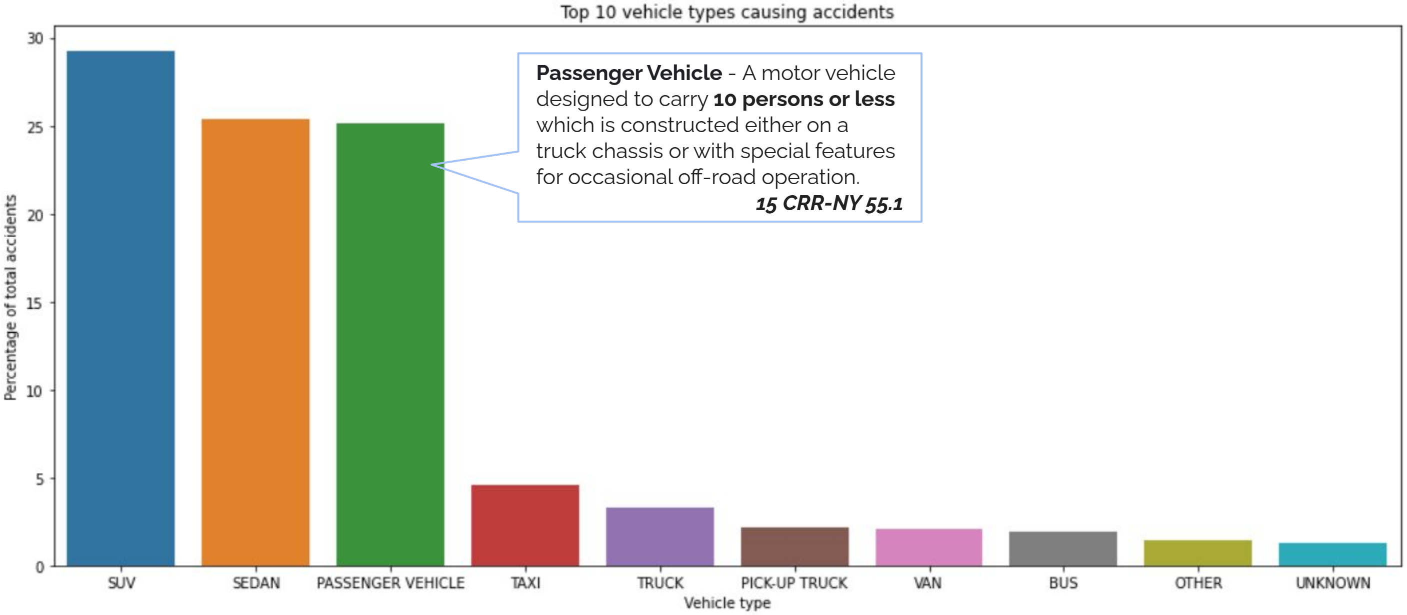


Data Visualization



	BUS	MOTORCYCLE	OTHER	PASSENGER VEHICLE	PICK-UP TRUCK	SEDAN	SUV	TAXI	TRUCK	UNKNOWN	VAN
ALCOHOL INVOLVEMENT	0.753419	1.7987	1.05451	2.53892	1.96001	2.43694	2.19918	1.08379	0.607794	1.7749	1.50642
BACKING UNSAFELY	4.36492	4.78377	8.06892	7.80189	10.5194	8.13669	8.69436	4.07587	12.4106	8.26895	9.77939
DRIVER INATTENTION/DISTRACTION	34.256	38.8825	33.5515	32.4319	35.9005	37.4436	36.0044	35.3426	32.7404	33.8066	33.38
DRIVER INEXPERIENCE	2.6124	4.42659	3.17767	3.54233	2.56715	2.87935	2.73927	2.56089	3.73168	3.19482	2.8976
FAILURE TO YIELD RIGHT-OF-WAY	8.70527	15.4867	9.93786	13.2992	10.3676	12.7961	13.2203	10.4504	6.05113	17.1643	9.49127
FATIGUED/DROWSY	3.6852	1.12259	5.05131	11.304	2.60015	0.211295	3.35046	5.1043	2.93618	7.49635	6.96411
FOLLOWING TOO CLOSELY	5.21661	4.69448	2.75398	0.32016	6.55976	9.24064	7.0272	6.29589	3.5842	0.167049	2.34607
LOST CONSCIOUSNESS	4.03734	1.33946	4.20394	5.7915	1.70263	0.211817	1.78567	3.34751	4.10261	6.11819	5.63056
OTHER VEHICULAR	7.61608	4.94961	8.89747	8.61929	5.57645	4.53841	5.2854	8.41102	6.77512	8.60305	9.02206
PASSING OR LANE USAGE IMPROPER	6.42044	6.48042	4.26043	0.192275	5.22669	5.74671	4.34088	4.72556	5.12603	0.0417624	2.25552
PASSING TOO CLOSELY	9.49144	2.62789	6.18115	0.0643898	6.33538	4.89735	4.1079	2.74735	9.20629	0.0208812	3.85249
PRESCRIPTION MEDICATION	2.75981	1.13535	3.30477	5.51337	1.24068	0.0177384	1.50024	2.3016	2.93171	4.48946	4.07474
TRAFFIC CONTROL DISREGARDED	1.44951	6.77382	2.4574	3.37063	2.09859	3.80436	3.19464	2.79396	1.11727	5.47087	1.97563
TURNING IMPROPERLY	5.79805	3.92907	5.31494	5.06443	4.42817	4.05582	4.19009	6.69502	6.28352	3.34099	5.36714
UNSAFE LANE CHANGING	2.83351	1.56908	1.7842	0.145771	2.91691	3.58315	2.36005	4.06421	2.39542	0.0417624	1.45703

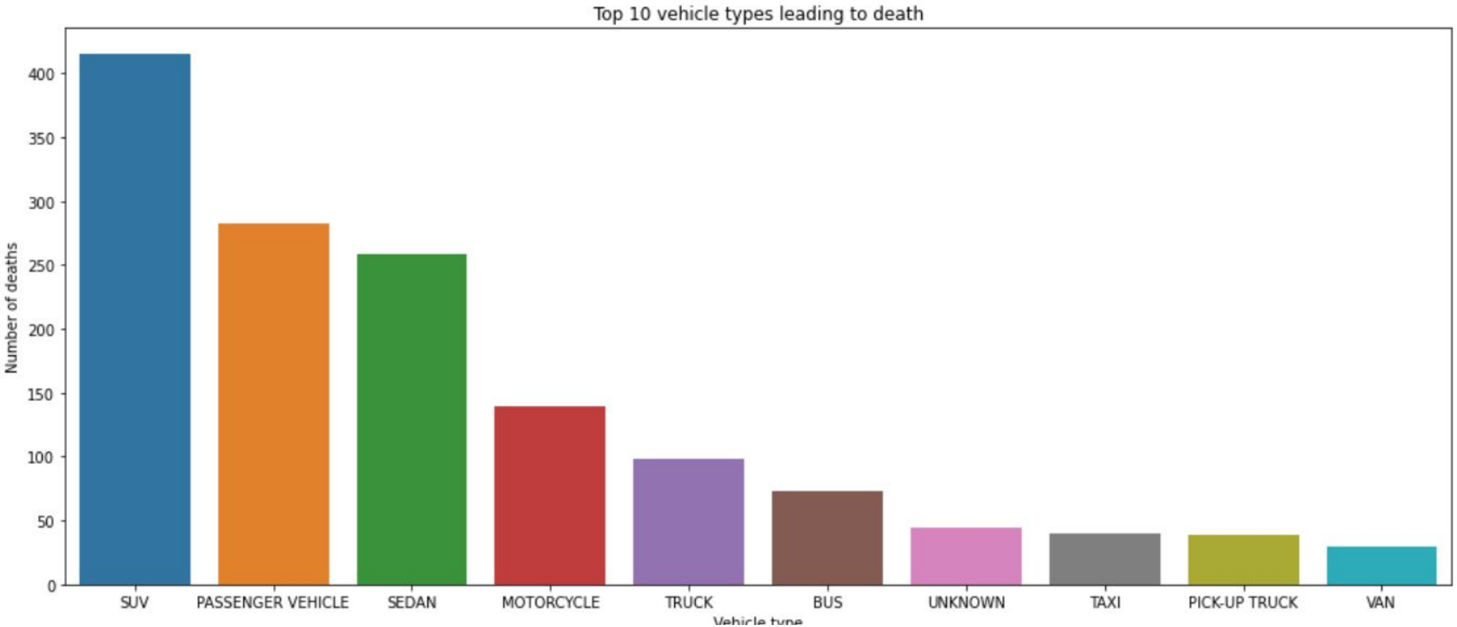
Top 10 Vehicle types involved in accidents



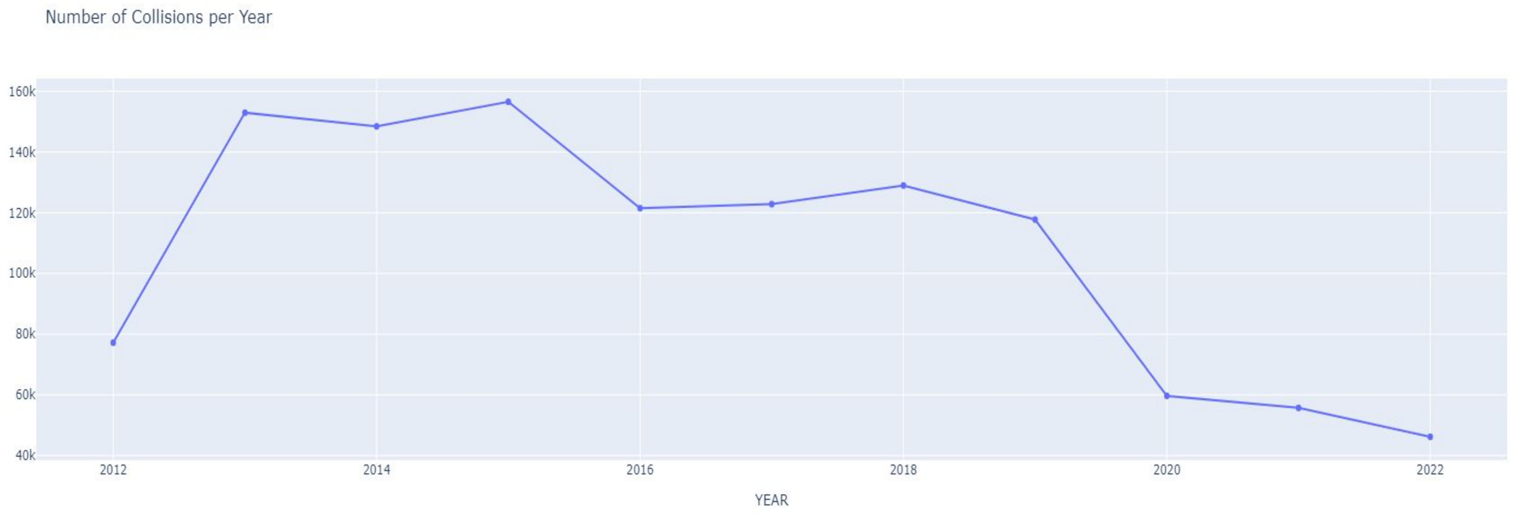
Passenger Vehicle - A motor vehicle designed to carry **10 persons or less** which is constructed either on a truck chassis or with special features for occasional off-road operation.

15 CRR-NY 55.1

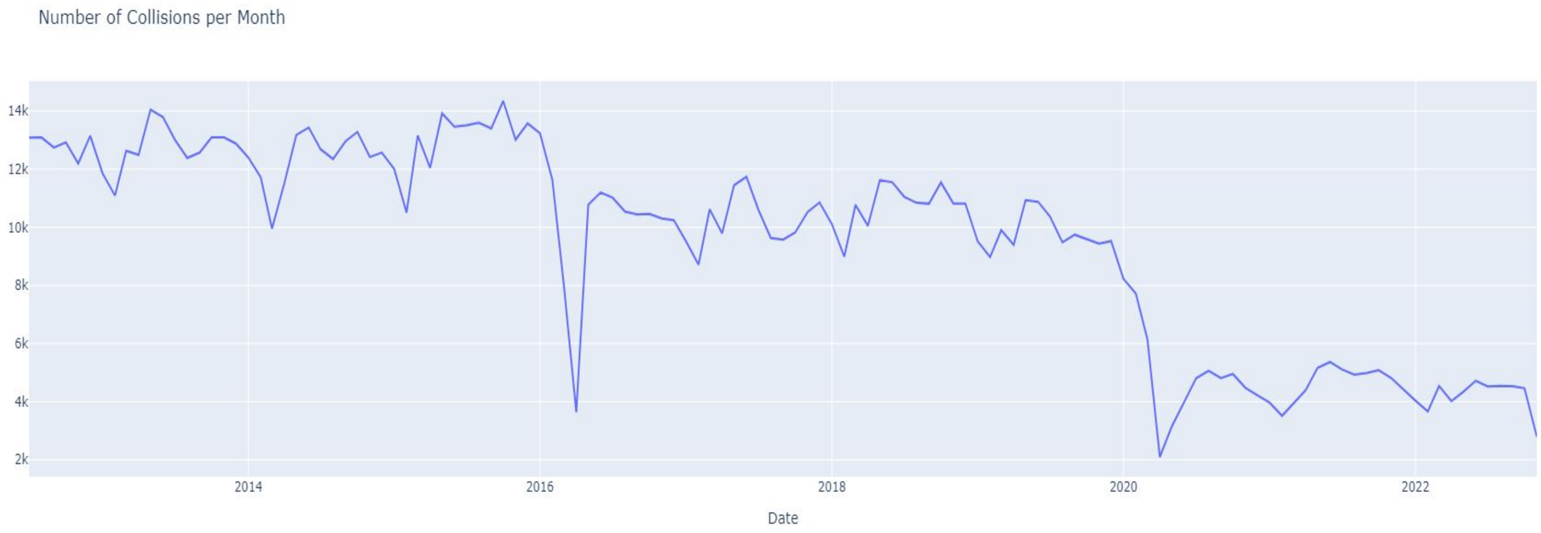
Top 10 Vehicle type involvement leading to Death



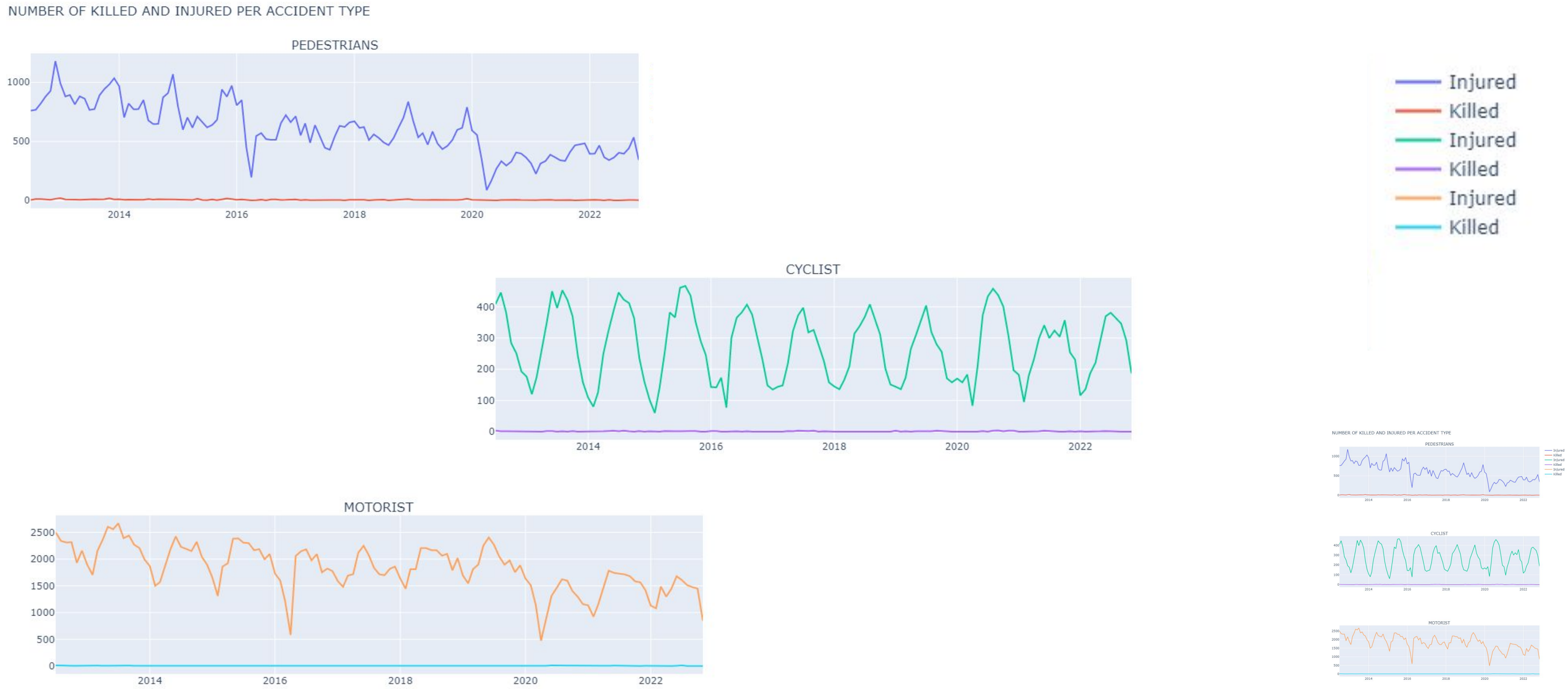
Number of Collisions per year



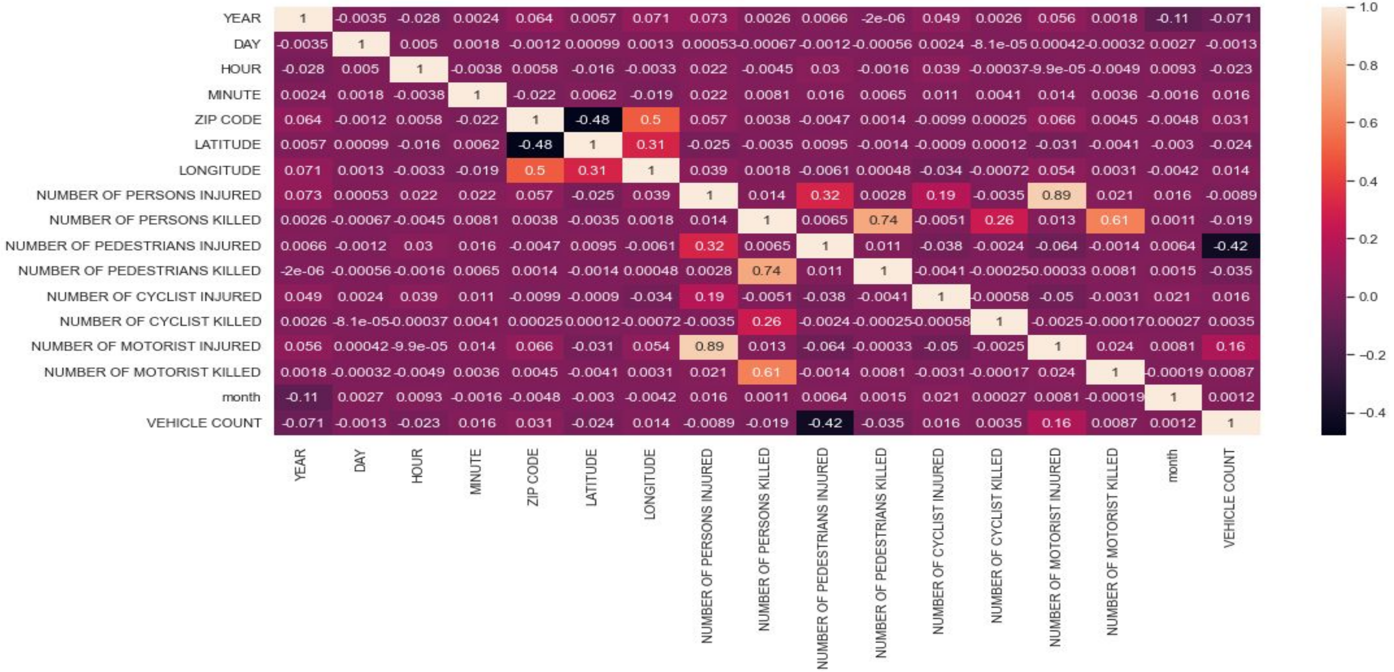
Number of Collisions per month



Number of killed and injured per accident type



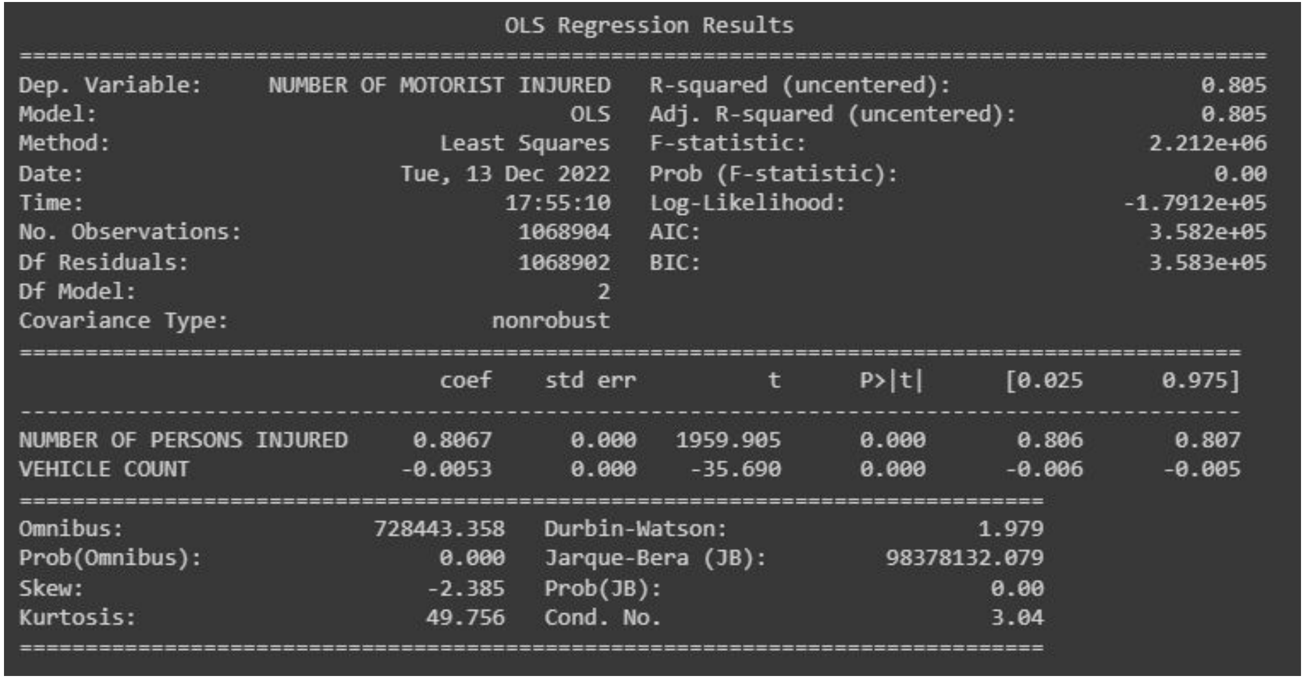
Data Visualization



	YEAR	DAY	HOUR	MINUTE	ZIP CODE	LATITUDE	LONGITUDE	NUMBER OF PERSONS INJURED	NUMBER OF PERSONS KILLED	NUMBER OF PEDESTRIANS INJURED	NUMBER OF PEDESTRIANS KILLED	NUMBER OF CYCLIST INJURED	NUMBER OF CYCLIST KILLED	NUMBER OF MOTORIST INJURED	NUMBER OF MOTORIST KILLED	month	VEHICLE COUNT
YEAR	1	-0.0035	-0.028	0.0024	0.064	0.0057	0.071	0.073	0.0026	0.0066	-2e-06	0.049	0.0026	0.056	0.0018	-0.11	-0.071
DAY	-0.0035	1	0.005	0.0018	-0.0012	0.00099	0.0013	0.00053	-0.00067	-0.0012	-0.00056	0.0024	-8.1e-05	0.00042	-0.00032	0.0027	-0.0013
HOUR	-0.028	0.005	1	-0.0038	0.0058	-0.016	-0.0033	0.022	-0.0045	0.03	-0.0016	0.039	-0.00037	-9.9e-05	-0.0049	0.0093	-0.023
MINUTE	0.0024	0.0018	-0.0038	1	-0.022	0.0062	-0.019	0.022	0.0081	0.016	0.0065	0.011	0.0041	0.014	0.0036	-0.0016	0.016
ZIP CODE	0.064	-0.0012	0.0058	-0.022	1	-0.48	0.5	0.057	0.0038	-0.0047	0.0014	-0.0099	0.00025	0.066	0.0045	-0.0048	0.031
LATITUDE	0.0057	0.00099	-0.016	0.0062	-0.48	1	0.31	-0.025	-0.0035	0.0095	-0.0014	-0.0009	0.00012	-0.031	-0.0041	-0.003	-0.024
LONGITUDE	0.071	0.0013	-0.0033	-0.019	0.5	0.31	1	0.039	0.0018	-0.0061	0.00048	-0.034	-0.00072	0.054	0.0031	-0.0042	0.014
NUMBER OF PERSONS INJURED	0.073	0.00053	0.022	0.022	0.057	-0.025	0.039	1	0.014	0.32	0.0028	0.19	-0.0035	0.89	0.021	0.016	-0.0089
NUMBER OF PERSONS KILLED	0.0026	-0.00067	-0.0045	0.0081	0.0038	-0.0035	0.0018	0.014	1	0.0065	0.74	-0.0051	0.26	0.013	0.61	0.0011	-0.019
NUMBER OF PEDESTRIANS INJURED	0.0066	-0.0012	0.03	0.016	-0.0047	0.0095	-0.0061	0.32	0.0065	1	0.011	-0.038	-0.0024	-0.064	-0.0014	0.0064	-0.42
NUMBER OF PEDESTRIANS KILLED	-2e-06	-0.00056	-0.0016	0.0065	0.0014	-0.0014	0.00048	0.0028	0.74	0.011	1	-0.0041	-0.00025	-0.00033	0.0081	0.0015	-0.035
NUMBER OF CYCLIST INJURED	0.049	0.0024	0.039	0.011	-0.0099	-0.0009	-0.034	0.19	-0.0051	-0.038	-0.0041	1	-0.00058	-0.05	-0.0031	0.021	0.016
NUMBER OF CYCLIST KILLED	0.0026	-8.1e-05	-0.00037	0.0041	0.00025	0.00012	-0.00072	-0.0035	0.26	-0.0024	-0.00025	-0.00058	1	-0.0025	-0.00017	0.00027	0.0035
NUMBER OF MOTORIST INJURED	0.056	0.00042	-9.9e-05	0.014	0.066	-0.031	0.054	0.89	0.013	-0.064	-0.00033	-0.05	-0.0025	1	0.024	0.0081	0.16
NUMBER OF MOTORIST KILLED	0.0018	-0.00032	-0.0049	0.0036	0.0045	-0.0041	0.0031	0.021	0.61	-0.0014	0.0081	-0.0031	-0.00017	0.024	1	-0.00019	0.0087
month	-0.11	0.0027	0.0093	-0.0016	-0.0048	-0.003	-0.0042	0.016	0.0011	0.0064	0.0015	0.021	0.00027	0.0081	-0.00019	1	0.0012
VEHICLE COUNT	-0.071	-0.0013	-0.023	0.016	0.031	-0.024	0.014	-0.0089	-0.019	-0.42	-0.035	0.016	0.0035	0.16	0.0087	0.0012	1

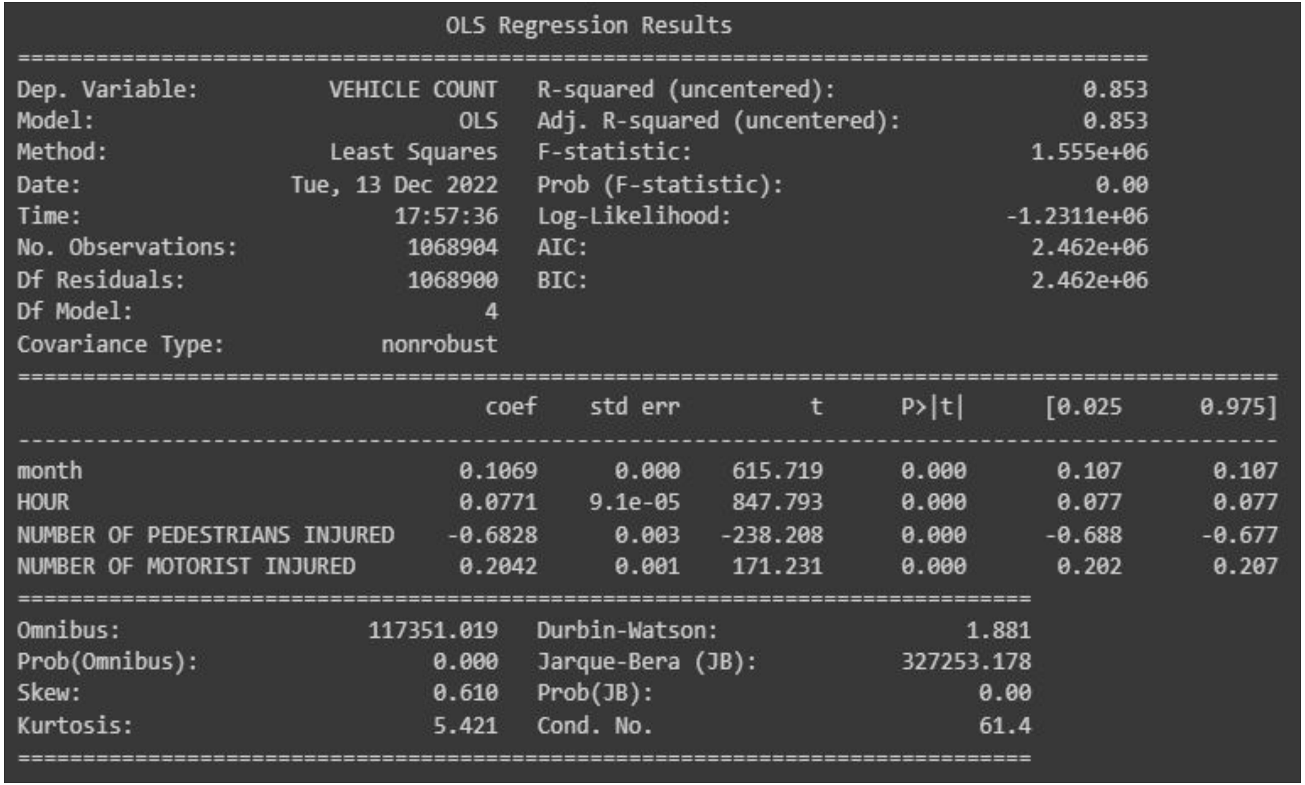
SMART QUESTIONS

1. How does other variables affect the number of motorist injured?



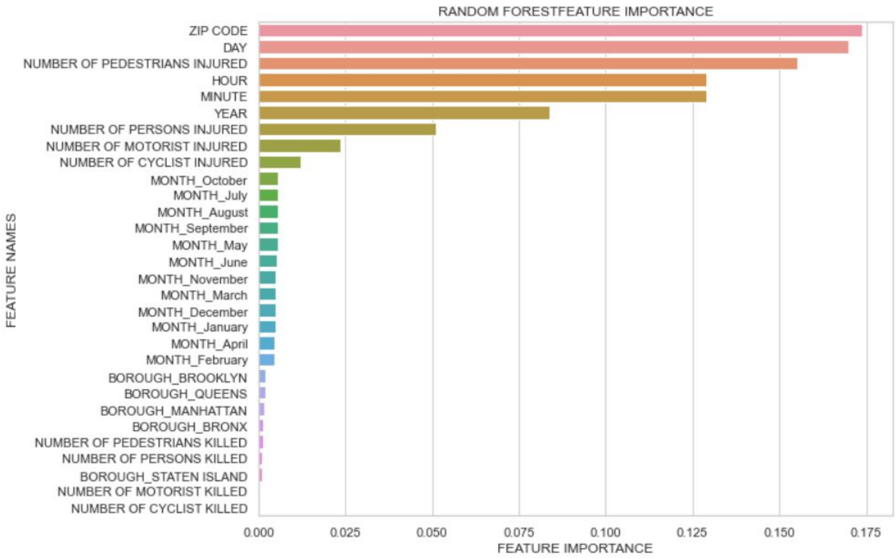
```
                                    OLS Regression Results
================================================================================================
Dep. Variable:     NUMBER OF MOTORIST INJURED   R-squared (uncentered):                   0.805
Model:                                    OLS   Adj. R-squared (uncentered):              0.805
Method:                         Least Squares   F-statistic:                          2.212e+06
Date:                        Tue, 13 Dec 2022   Prob (F-statistic):                        0.00
Time:                                17:55:10   Log-Likelihood:                     -1.7912e+05
No. Observations:                     1068904   AIC:                                  3.582e+05
Df Residuals:                         1068902   BIC:                                  3.583e+05
Df Model:                                   2
Covariance Type:                    nonrobust
==============================================================================================
                               coef    std err          t      P>|t|      [0.025      0.975]
----------------------------------------------------------------------------------------------
NUMBER OF PERSONS INJURED    0.8067      0.000   1959.905      0.000       0.806       0.807
VEHICLE COUNT               -0.0053      0.000    -35.690      0.000      -0.006      -0.005
==============================================================================
Omnibus:                   728443.358   Durbin-Watson:                   1.979
Prob(Omnibus):                  0.000   Jarque-Bera (JB):         98378132.079
Skew:                          -2.385   Prob(JB):                         0.00
Kurtosis:                      49.756   Cond. No.                         3.04
==============================================================================
```

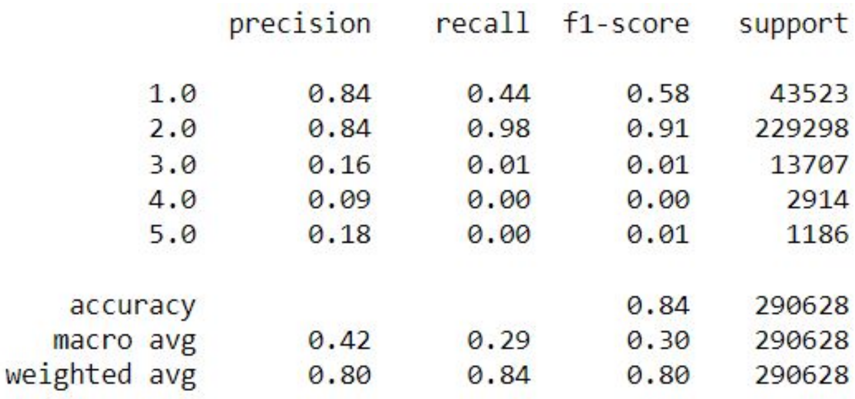
2. How does the time of the day and time of year affect the number of vehicles involved?



```
                                 OLS Regression Results
========================================================================================
Dep. Variable:          VEHICLE COUNT   R-squared (uncentered):                    0.853
Model:                            OLS   Adj. R-squared (uncentered):               0.853
Method:                 Least Squares   F-statistic:                           1.555e+06
Date:                Tue, 13 Dec 2022   Prob (F-statistic):                         0.00
Time:                        17:57:36   Log-Likelihood:                      -1.2311e+06
No. Observations:             1068904   AIC:                                   2.462e+06
Df Residuals:                 1068900   BIC:                                   2.462e+06
Df Model:                           4
Covariance Type:            nonrobust
==================================================================================================
                                     coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------------
month                              0.1069      0.000    615.719      0.000       0.107       0.107
HOUR                               0.0771    9.1e-05    847.793      0.000       0.077       0.077
NUMBER OF PEDESTRIANS INJURED     -0.6828      0.003   -238.208      0.000      -0.688      -0.677
NUMBER OF MOTORIST INJURED         0.2042      0.001    171.231      0.000       0.202       0.207
==============================================================================
Omnibus:                   117351.019   Durbin-Watson:                   1.881
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           327253.178
Skew:                           0.610   Prob(JB):                         0.00
Kurtosis:                       5.421   Cond. No.                         61.4
==============================================================================
```

3. Can we predict number of vehicles involved in the accident?



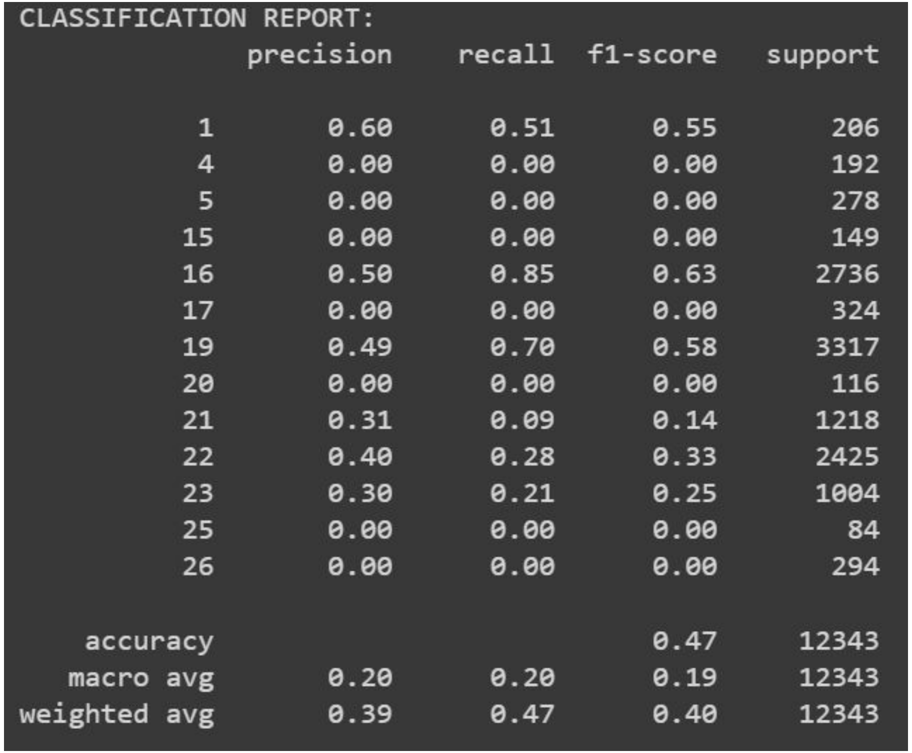


```
              precision    recall  f1-score   support

         1.0       0.84      0.44      0.58     43523
         2.0       0.84      0.98      0.91    229298
         3.0       0.16      0.01      0.01     13707
         4.0       0.09      0.00      0.00      2914
         5.0       0.18      0.00      0.01      1186

    accuracy                           0.84    290628
   macro avg       0.42      0.29      0.30    290628
weighted avg       0.80      0.84      0.80    290628
```

4. What type of Car - 1 was involved in a two car accident?

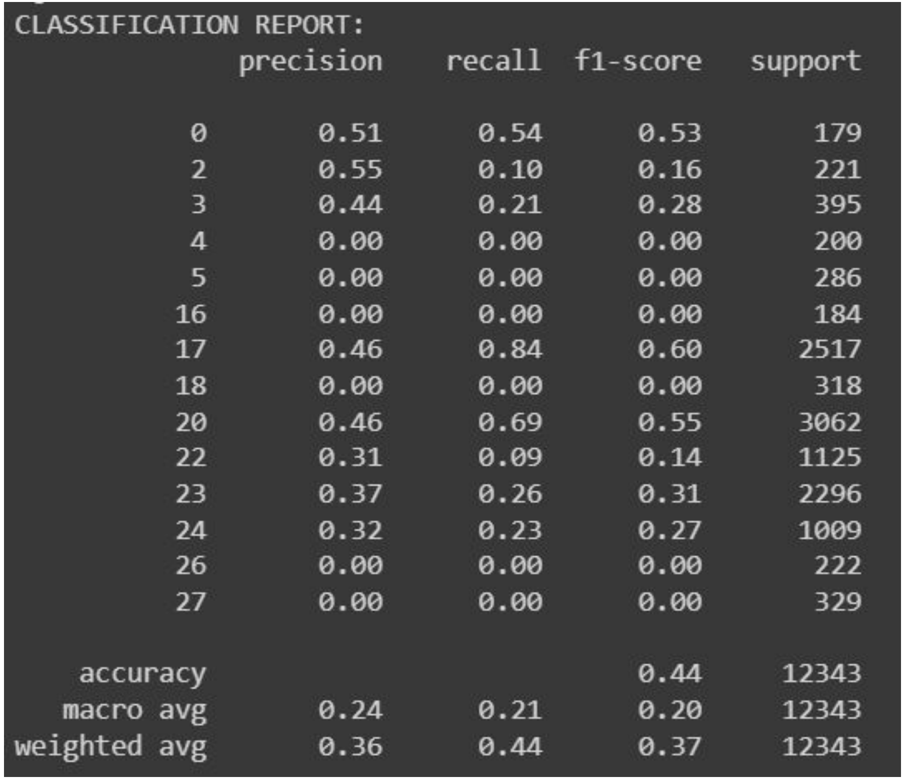


```
CLASSIFICATION REPORT:
              precision    recall  f1-score   support

           1       0.60      0.51      0.55       206
           4       0.00      0.00      0.00       192
           5       0.00      0.00      0.00       278
          15       0.00      0.00      0.00       149
          16       0.50      0.85      0.63      2736
          17       0.00      0.00      0.00       324
          19       0.49      0.70      0.58      3317
          20       0.00      0.00      0.00       116
          21       0.31      0.09      0.14      1218
          22       0.40      0.28      0.33      2425
          23       0.30      0.21      0.25      1004
          25       0.00      0.00      0.00        84
          26       0.00      0.00      0.00       294

    accuracy                           0.47     12343
   macro avg       0.20      0.20      0.19     12343
weighted avg       0.39      0.47      0.40     12343
```

5. What type of Car - 2 was involved in a two car accident?



```
CLASSIFICATION REPORT:
              precision    recall  f1-score   support

           0       0.51      0.54      0.53       179
           2       0.55      0.10      0.16       221
           3       0.44      0.21      0.28       395
           4       0.00      0.00      0.00       200
           5       0.00      0.00      0.00       286
          16       0.00      0.00      0.00       184
          17       0.46      0.84      0.60      2517
          18       0.00      0.00      0.00       318
          20       0.46      0.69      0.55      3062
          22       0.31      0.09      0.14      1125
          23       0.37      0.26      0.31      2296
          24       0.32      0.23      0.27      1009
          26       0.00      0.00      0.00       222
          27       0.00      0.00      0.00       329

    accuracy                           0.44     12343
   macro avg       0.24      0.21      0.20     12343
weighted avg       0.36      0.44      0.37     12343
```

Issues Faced with the Dataset

- The **large size** of the dataset **restricted** the type of models we could use, or how we used them:
  - **KNN -** long run time, could not get highly accurate results
  - **Random Forest -** had to adjust n_jobs to -1, and n_estimators to 100, in order to get an output in a more efficient, timely manner
    - **n_jobs=-1 :** use all available CPUs to handle the 1,000,000+ observations
    - **n_estimators=100:** large number of trees for large number of variables, observations
- The data in the survey is added manually i.e. there is not a fixed format.
  - This forced us to remove many rows from the dataset.
  - Same type of vehicles were labelled differently in many rows for ex. Ambulance was mentioned as AMBULANCE, Ambulance and AMBU or Taxi was mentioned as TAXI and Taxi.
  - Unclear categories with only 1 occurrence (ie. 994 , UHUAL, TCN)

Mercé
Xièxie
Sagolun
Obrigado
ありがとう
Takk
Tak
Hvala
Grazie
Sagolun
Matondo
Obrigado
Efharisto
Kiitos
Mahalo
Mauruuru
Dankjewel
Köszönöm
Thank You
Danke
Gracias
شكرا
ありがとう
Grazie
Gracias
ありがとう
Merci
Toda
Misaotra
Mahalo
Köszönöm
Merci
Спасибо
Mahalo
Mercé
Dankie
Hvala
Xièxie
Toda
Faleminderit
Efharisto
Kiitos
Mercé
Takk
Tak
Thank You
Rahmat
Terima Kasih
Thank You
Dankie
Kiitos
Спасибо
Hvala
Dankie
Faleminderit
Dankjewel
Sagolun
شكرا
Terima Kasih
Tak
Faleminderit
Grazie
Terima Kasih
Xièxie
Gracias
Rahmat
Dankjewel
Mauruuru
Merci
Matondo
Toda
Efharisto
Misaotra
Obrigado
Takk
Danke
Köszönöm
Danke
شكرا
Matondo
Спасибо