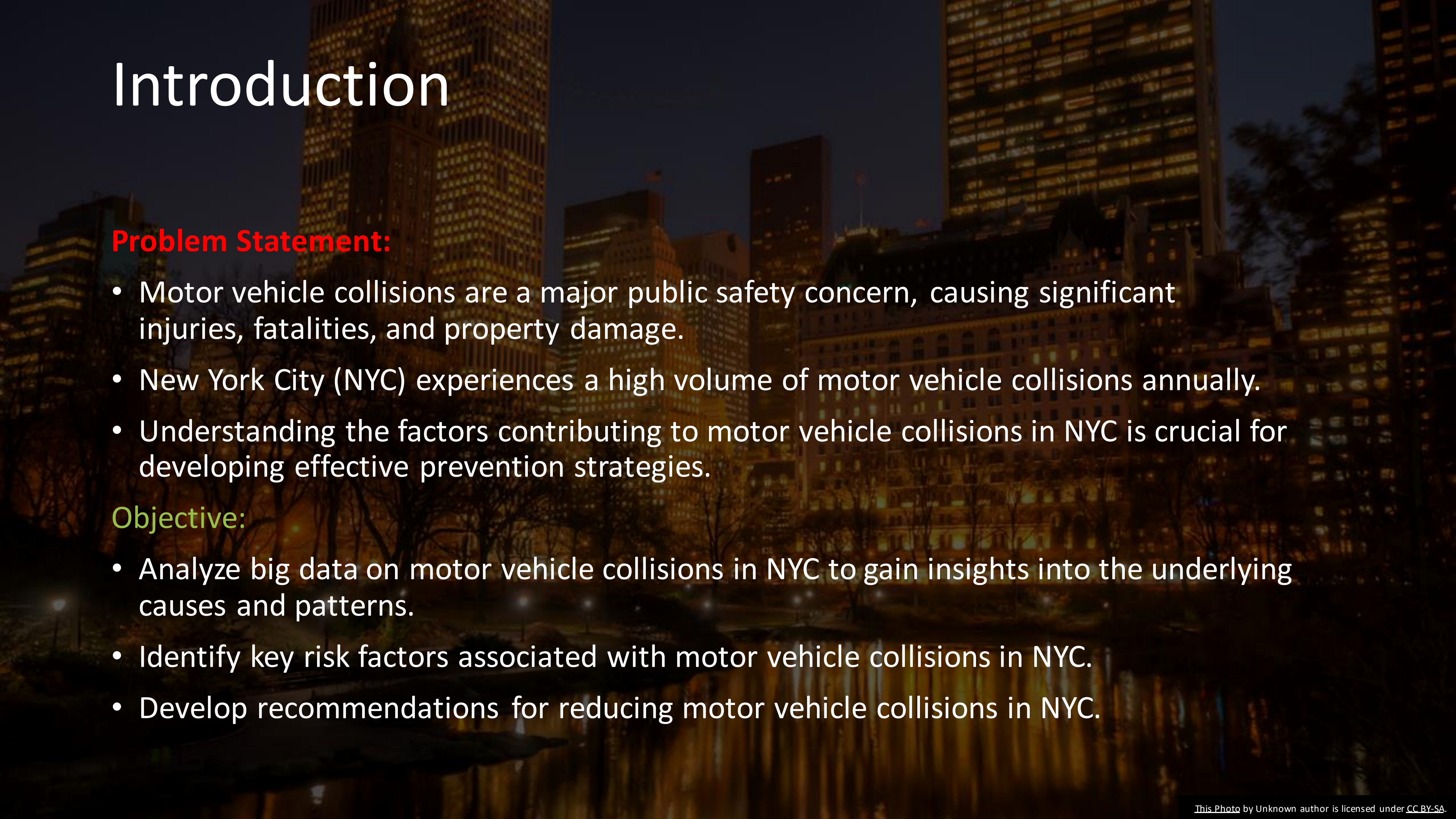
Title: Big Data Project on Motor Vehicle Collisions in NYC: Insights and Reasons

Presenter: Garima Goyal and Sachin Garg

Date: December 8th, 2023



Introduction

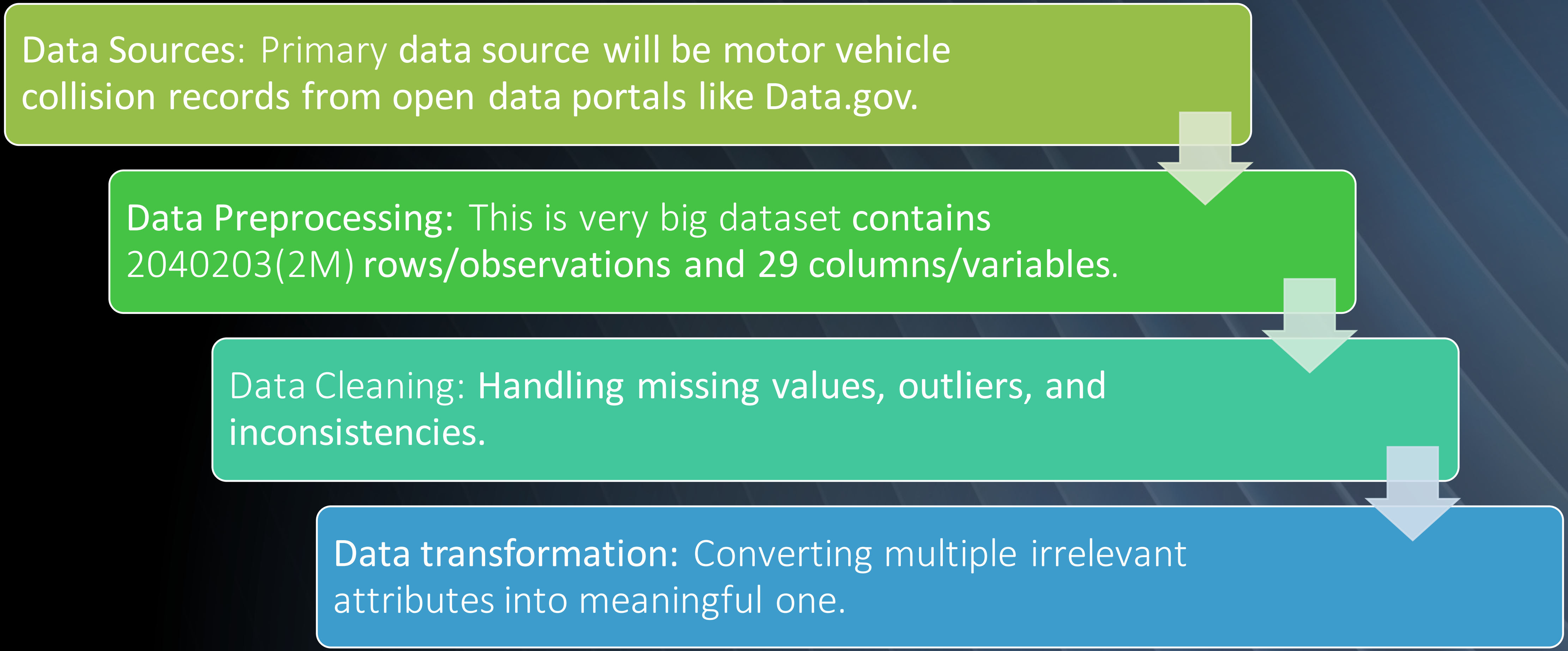
Problem Statement:

- Motor vehicle collisions are a major public safety concern, causing significant injuries, fatalities, and property damage.
- New York City (NYC) experiences a high volume of motor vehicle collisions annually.
- Understanding the factors contributing to motor vehicle collisions in NYC is crucial for developing effective prevention strategies.

Objective:

- Analyze big data on motor vehicle collisions in NYC to gain insights into the underlying causes and patterns.
- Identify key risk factors associated with motor vehicle collisions in NYC.
- Develop recommendations for reducing motor vehicle collisions in NYC.

Data Collection and Preprocessing



Data Sources: Primary data source will be motor vehicle collision records from open data portals like Data.gov.

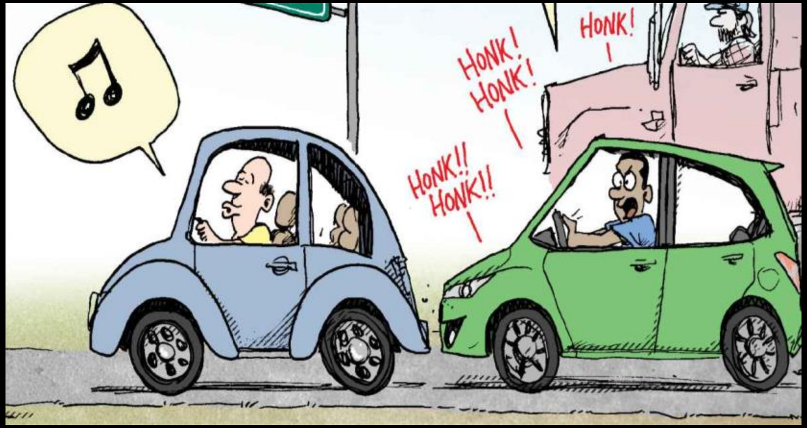
Data Preprocessing: This is very big dataset contains 2040203(2M) rows/observations and 29 columns/variables.

Data Cleaning: Handling missing values, outliers, and inconsistencies.

Data transformation: Converting multiple irrelevant attributes into meaningful one.

MISSING VALUE COLUMNS

Columns	Values
CRASH DATE	0
CRASH TIME	0
BOROUGH	634709
ZIP CODE	634951
LATITUDE	231352
LONGITUDE	231352
LOCATION	231352
ON STREET NAME	430849
CROSS STREET NAME	766605
OFF STREET NAME	1701727
NUMBER OF PERSONS INJURED	18
NUMBER OF PERSONS KILLED	31
NUMBER OF PEDESTRIANS INJURED	0
NUMBER OF PEDESTRIANS KILLED	0
NUMBER OF CYCLIST INJURED	0
NUMBER OF CYCLIST KILLED	0
NUMBER OF MOTORIST INJURED	0
NUMBER OF MOTORIST KILLED	0
CONTRIBUTING FACTOR VEHICLE 1	6504
CONTRIBUTING FACTOR VEHICLE 2	312925
CONTRIBUTING FACTOR VEHICLE 3	1895102
CONTRIBUTING FACTOR VEHICLE 4	2007586
CONTRIBUTING FACTOR VEHICLE 5	2031372
COLLISION_ID	0
VEHICLE TYPE CODE 1	13042
VEHICLE TYPE CODE 2	384146
VEHICLE TYPE CODE 3	1900233
VEHICLE TYPE CODE 4	2008690
VEHICLE TYPE CODE 5	2031640



Factors Contributing to Accidents

Data cleaning (Outliers and Inconsistencies)

We fix the some typos/misspelling ('Illnes', 'Illness'), ('Cell Phone (hand-Held)', 'Cell Phone (hand-held)'), data inconsistency ('Fell Asleep', 'Drowsy', Lost Consciousness') and invalidation ('80', '1') for all the CONTRIBUTING FACTOR VEHICLE involved.

Data cleaning (Outliers and Inconsistencies)

Also, there were multiple attributes for contributing factor vehicle which are similar(in a way) to each other, we specified them in few different category to better understanding of insights.

For example:

Over speeding

Following Too Closely

Driver Inattention/Distraction

Pavement Slippery or Defective

Fatigued/Drowsy/Sleep/Unconscious

Outside Car Distraction

Drugs or Alcohol Involvement

Driver Inexperience

Obstruction/Debris

Unsafe Speed

Failure to Yield Right-of-Way

Traffic Control Disregarded

2) For the vehicle type code, lot of data inconsistencies, misspelling values and invalid data. So, to encounter that we are performing 'fuzzy-wuzzy' method (a python library, used to compare strings and determine the similarity between them, which can be useful in big data analytics).
Similarly, we worked on vehicle type code which are similar(in a way) to each other, we specified them in few different category to better understanding of insights

- SUV /Sedan cars
- Taxi
- Pick-Up/Tow-Trucks
- Bike/Scooter/2-Wheel Vehicles
- Van
- Bus
- All Other Type Vehicles 30798
- Other Commercial/Utility Vehicles
- Heavy Truck/Trailor
- City/Government Vehicles
- Construction/Dump Vehicles
- Other Mini Vehicles

TAXI
RAM
MTA
AMBULANCE
MENU
FOOD TRUCK
SCHOOL BUS
UNITED STATES POSTAL SERVICE

Standardizing Formats:

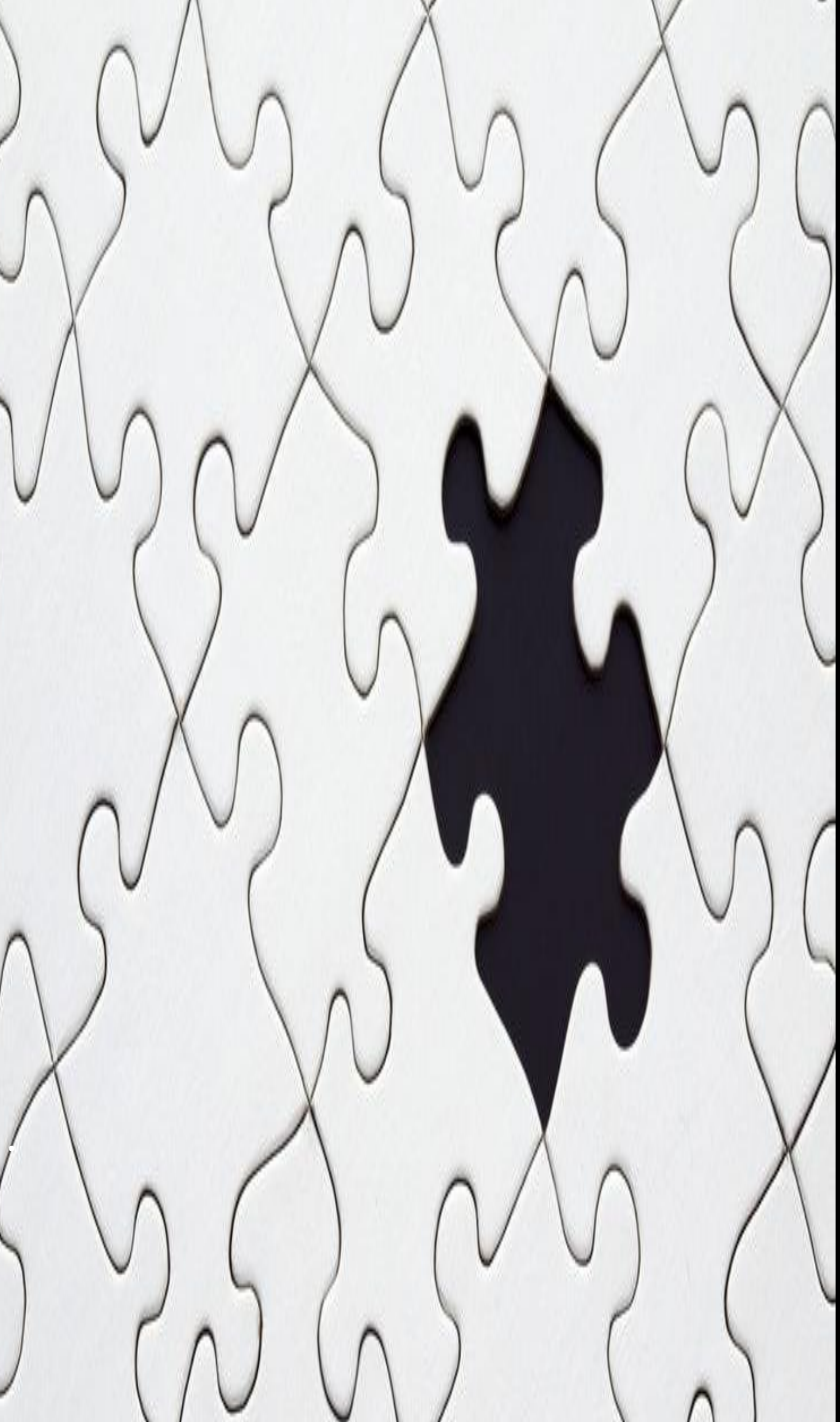
Standardize the format of street names (e.g., convert to uppercase, remove leading/trailing spaces).

Handling Abbreviations:

Expand abbreviations to make the data more consistent. (e.g., convert St to Street, Ave to Avenue, etc.)

Removing Special Characters:

Remove any special non-alphanumeric values or symbols that might cause issues.



Imputing Missing Data

For the columns 'NUMBER OF PERSONS INJURED' and 'NUMBER OF PERSONS KILLED' contains only 18 and 31 missing values respectively, which are very low values related to the whole dataset. So, we use the simple statistical 'mean' strategy to impute the missing values in those columns.

For the columns 'CONTRIBUTING FACTOR VEHICLE 1' and 'VEHICLE TYPE CODE 1' contains 6452 and 12940 missing values respectively, which are only 0.0031% and 0.0063% of the total values. So, we just use the simple 'mode' strategy to impute missing values.

In the other contributing factor vehicle columns, some of the missing values are not true. For example if vehicle 1 and 2 are involved in an accident, vehicle 3, 4 and 5 are not applicable and similar for all cases. So, to counter those missing values, we assume and label them as 'Not Applicable'. Same we performed with off street, on street and across street.

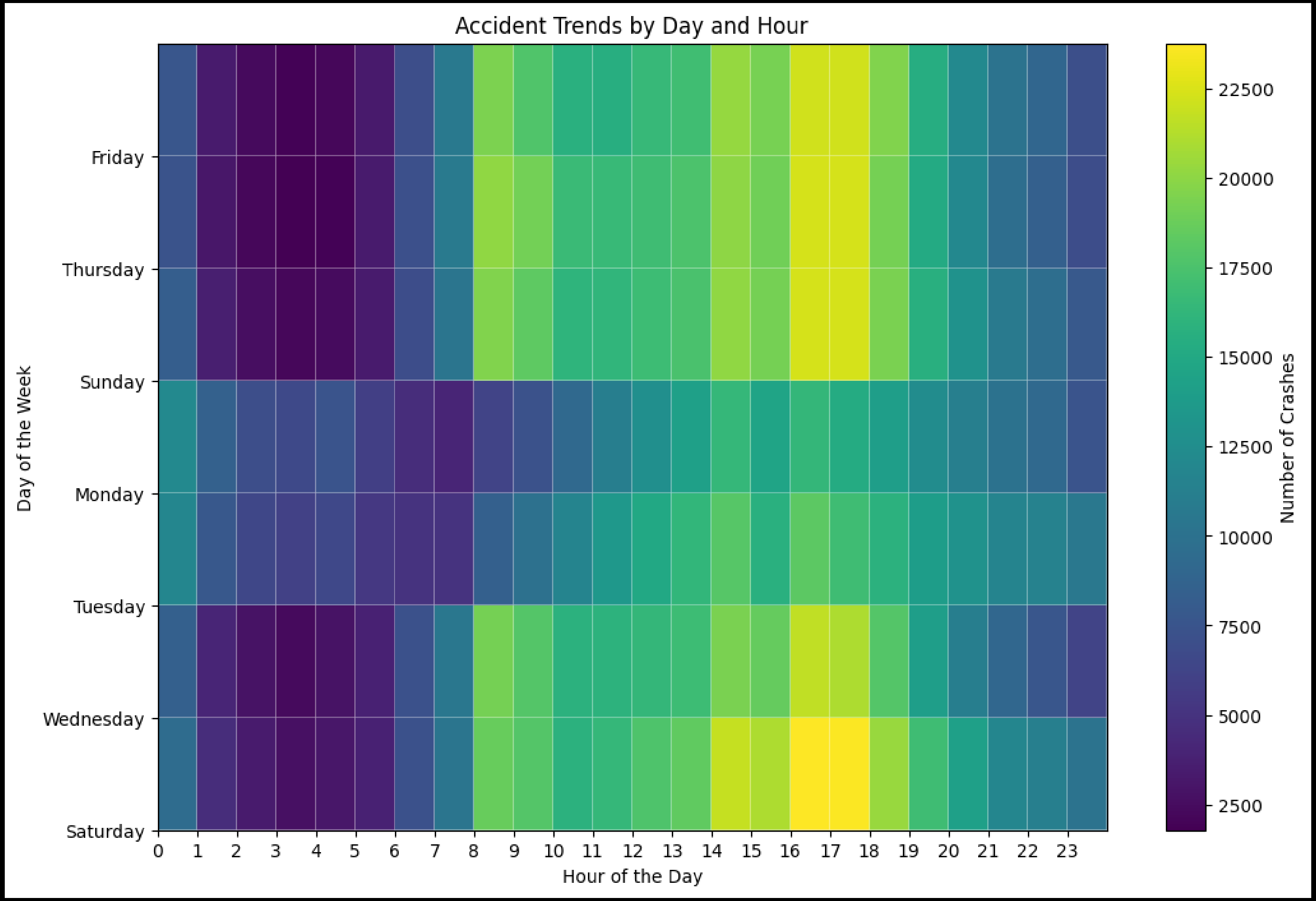
We have imputing the locations using Geo pandas, the approach we have usedis the geometry of nearby features to estimate the missing values.

For missing zip-codes, we have used Machine learning algorithms- Random Forest Regressor along with Simple Imputer, One Hot Encoder, Label Encoder, mean absolute error.

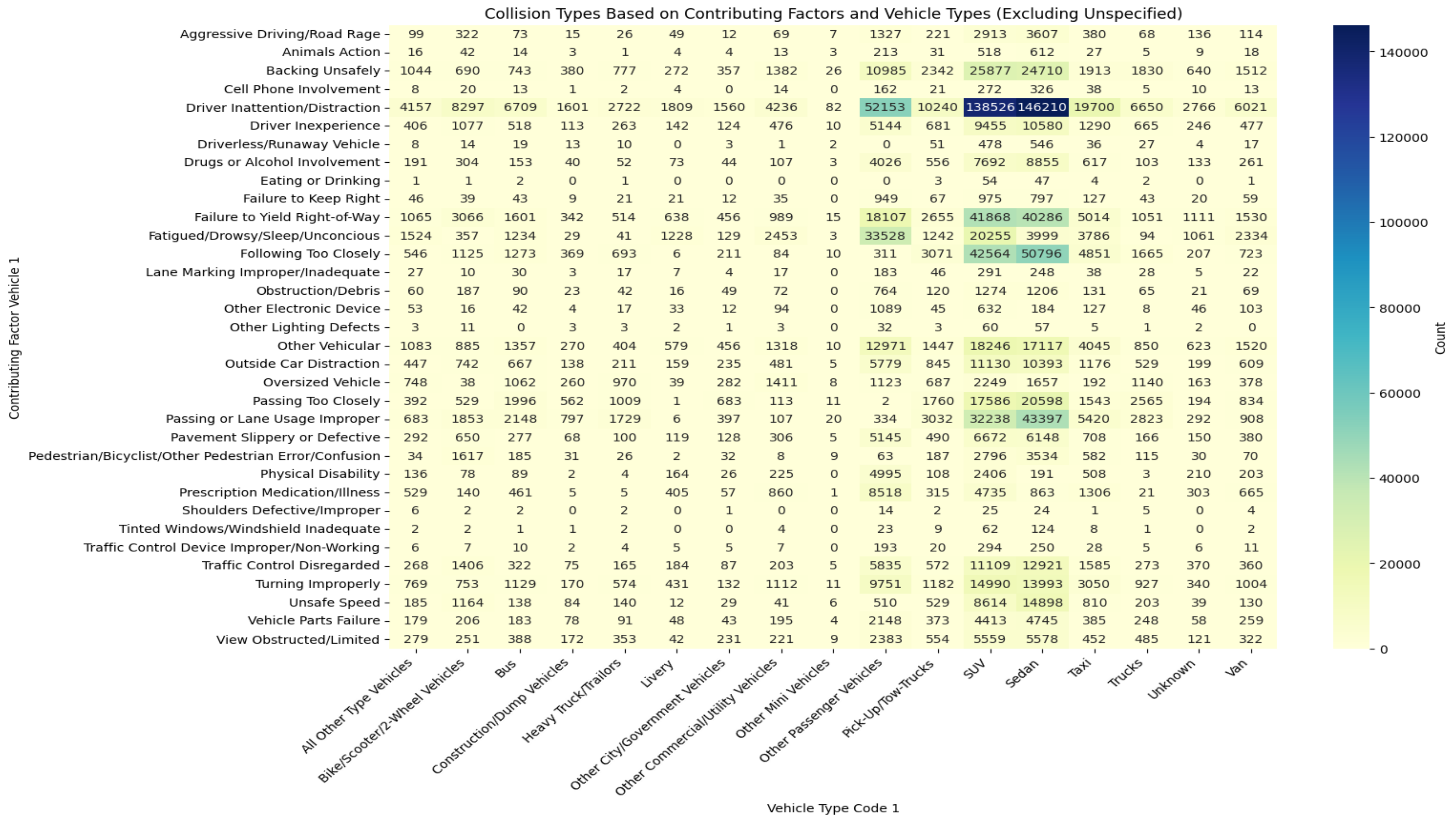
Descriptive Statistics:

- Number of collisions - The total number of collisions in the dataset over a certain timeframe (e.g. per year). Gives an overall sense of frequency.
- Collision severity - The number or percentage of collisions by severity level (e.g. property damage only, injury, fatality). Shows the harm caused.
- Collision type - The number or percentage of collisions by type (e.g. rear-end, angle, pedestrian-involved). Describes the common configurations.
- Primary collision factor - The main factor that contributed to collisions, like speeding, impairment, distraction. Helps understand causes.
- Temporal statistics - Number or share of collisions over timeframes like month, day of week, time of day. Reveals temporal patterns.
- Environmental statistics - Number or proportion of collisions related to factors like weather, lighting, road surface condition. Shows environmental influences.
- Vehicle statistics - The number/share of different vehicle types involved - passenger cars, trucks, motorcycles etc.
- Demographic statistics - Age and gender distributions of drivers involved in collisions. Allows demographic analysis.
- Location statistics - Collisions summarized by geographic units - state, county, city, intersection etc. Pinpoints high frequency areas.

Accident Trends by Day and Hour



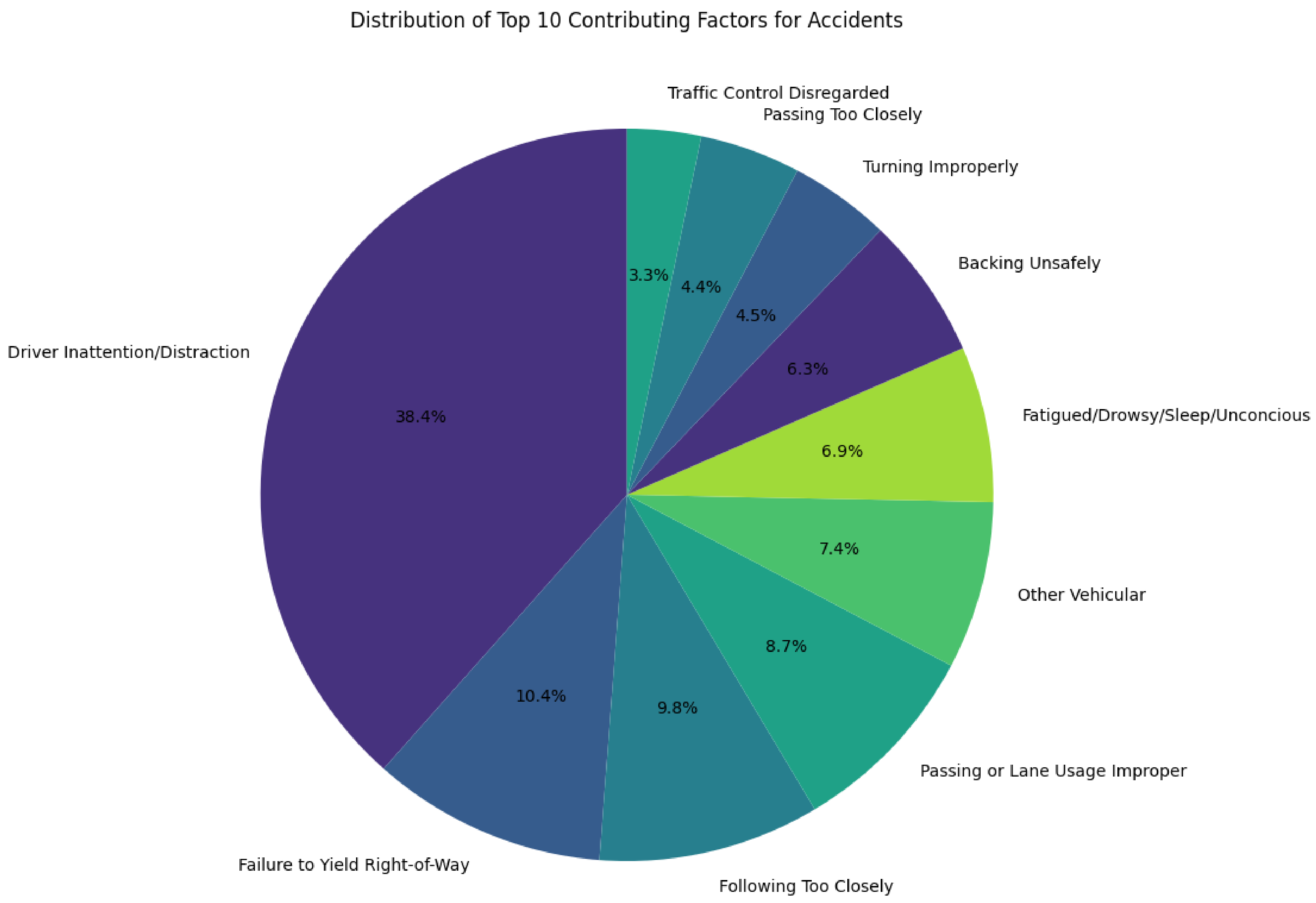
Collision Types Based on Contributing Factors and Vehicle Types



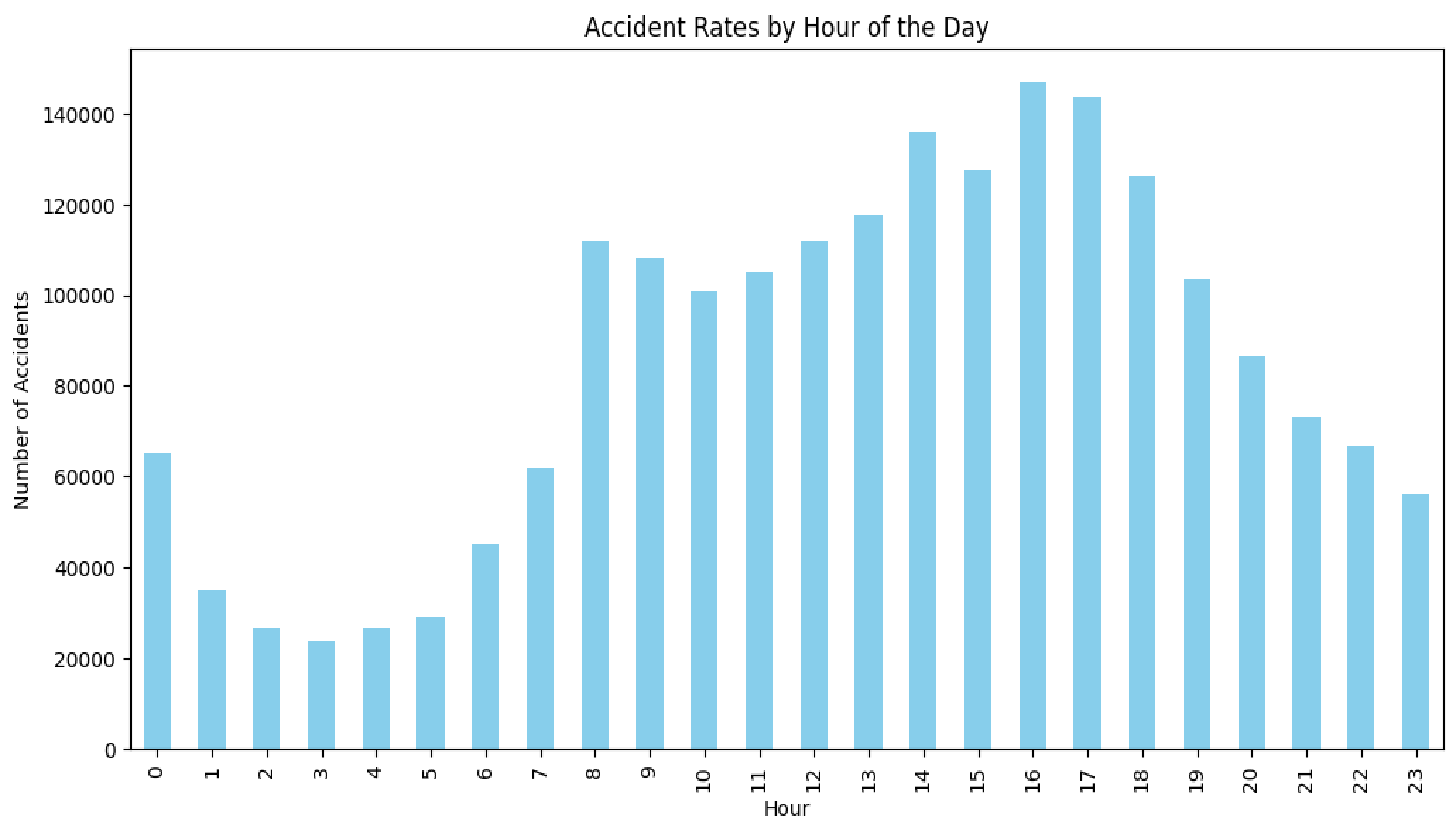
Collision Types Based on Contributing Factors and Vehicle Types (Excluding Unspecified)

Contributing Factor Vehicle 1	All Other Type Vehicles	Bike/Scooter/2-Wheel Vehicles	Bus	Construction/Dump Vehicles	Heavy Truck/Trailors	Livery	Other City/Government Vehicles	Other Commercial/Utility Vehicles	Other Mini Vehicles	Other Passenger Vehicles	Pick-Up/Tow-Trucks	SUV	Sedan	Taxi	Trucks	Unknown	Van
Aggressive Driving/Road Rage	99	322	73	15	26	49	12	69	7	1327	221	2913	3607	380	68	136	114
Animals Action	16	42	14	3	1	4	4	13	3	213	31	518	612	27	5	9	18
Backing Unsafely	1044	690	743	380	777	272	357	1382	26	10985	2342	25877	24710	1913	1830	640	1512
Cell Phone Involvement	8	20	13	1	2	4	0	14	0	162	21	272	326	38	5	10	13
Driver Inattention/Distraction	4157	8297	6709	1601	2722	1809	1560	4236	82	52153	10240	138526	146210	19700	6650	2766	6021
Driver Inexperience	406	1077	518	113	263	142	124	476	10	5144	681	9455	10580	1290	665	246	477
Driverless/Runaway Vehicle	8	14	19	13	10	0	3	1	2	0	51	478	546	36	27	4	17
Drugs or Alcohol Involvement	191	304	153	40	52	73	44	107	3	4026	556	7692	8855	617	103	133	261
Eating or Drinking	1	1	2	0	1	0	0	0	0	0	3	54	47	4	2	0	1
Failure to Keep Right	46	39	43	9	21	21	12	35	0	949	67	975	797	127	43	20	59
Failure to Yield Right-of-Way	1065	3066	1601	342	514	638	456	989	15	18107	2655	41868	40286	5014	1051	1111	1530
Fatigued/Drowsy/Sleep/Unconcious	1524	357	1234	29	41	1228	129	2453	3	33528	1242	20255	3999	3786	94	1061	2334
Following Too Closely	546	1125	1273	369	693	6	211	84	10	311	3071	42564	50796	4851	1665	207	723
Lane Marking Improper/Inadequate	27	10	30	3	17	7	4	17	0	183	46	291	248	38	28	5	22
Obstruction/Debris	60	187	90	23	42	16	49	72	0	764	120	1274	1206	131	65	21	69
Other Electronic Device	53	16	42	4	17	33	12	94	0	1089	45	632	184	127	8	46	103
Other Lighting Defects	3	11	0	3	3	2	1	3	0	32	3	60	57	5	1	2	0
Other Vehicular	1083	885	1357	270	404	579	456	1318	10	12971	1447	18246	17117	4045	850	623	1520
Outside Car Distraction	447	742	667	138	211	159	235	481	5	5779	845	11130	10393	1176	529	199	609
Oversized Vehicle	748	38	1062	260	970	39	282	1411	8	1123	687	2249	1657	192	1140	163	378
Passing Too Closely	392	529	1996	562	1009	1	683	113	11	2	1760	17586	20598	1543	2565	194	834
Passing or Lane Usage Improper	683	1853	2148	797	1729	6	397	107	20	334	3032	32238	43397	5420	2823	292	908
Pavement Slippery or Defective	292	650	277	68	100	119	128	306	5	5145	490	6672	6148	708	166	150	380
Pedestrian/Bicyclist/Other Pedestrian Error/Confusion	34	1617	185	31	26	2	32	8	9	63	187	2796	3534	582	115	30	70
Physical Disability	136	78	89	2	4	164	26	225	0	4995	108	2406	191	508	3	210	203
Prescription Medication/Illness	529	140	461	5	5	405	57	860	1	8518	315	4735	863	1306	21	303	665
Shoulders Defective/Improper	6	2	2	0	2	0	1	0	0	14	2	25	24	1	5	0	4
Tinted Windows/Windshield Inadequate	2	2	1	1	2	0	0	4	0	23	9	62	124	8	1	0	2
Traffic Control Device Improper/Non-Working	6	7	10	2	4	5	5	7	0	193	20	294	250	28	5	6	11
Traffic Control Disregarded	268	1406	322	75	165	184	87	203	5	5835	572	11109	12921	1585	273	370	360
Turning Improperly	769	753	1129	170	574	431	132	1112	11	9751	1182	14990	13993	3050	927	340	1004
Unsafe Speed	185	1164	138	84	140	12	29	41	6	510	529	8614	14898	810	203	39	130
Vehicle Parts Failure	179	206	183	78	91	48	43	195	4	2148	373	4413	4745	385	248	58	259
View Obstructed/Limited	279	251	388	172	353	42	231	221	9	2383	554	5559	5578	452	485	121	322

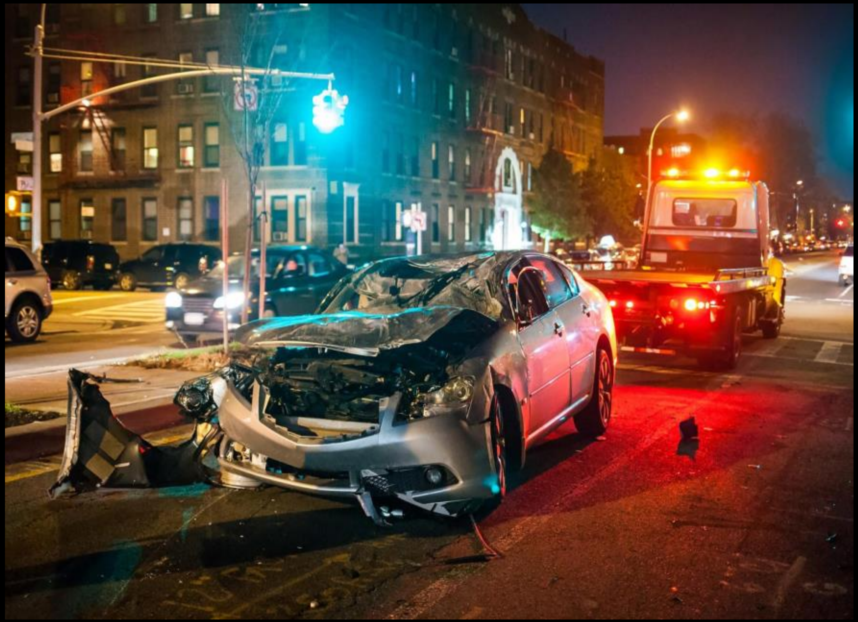
10 Most Common Contributing Factors for Accidents

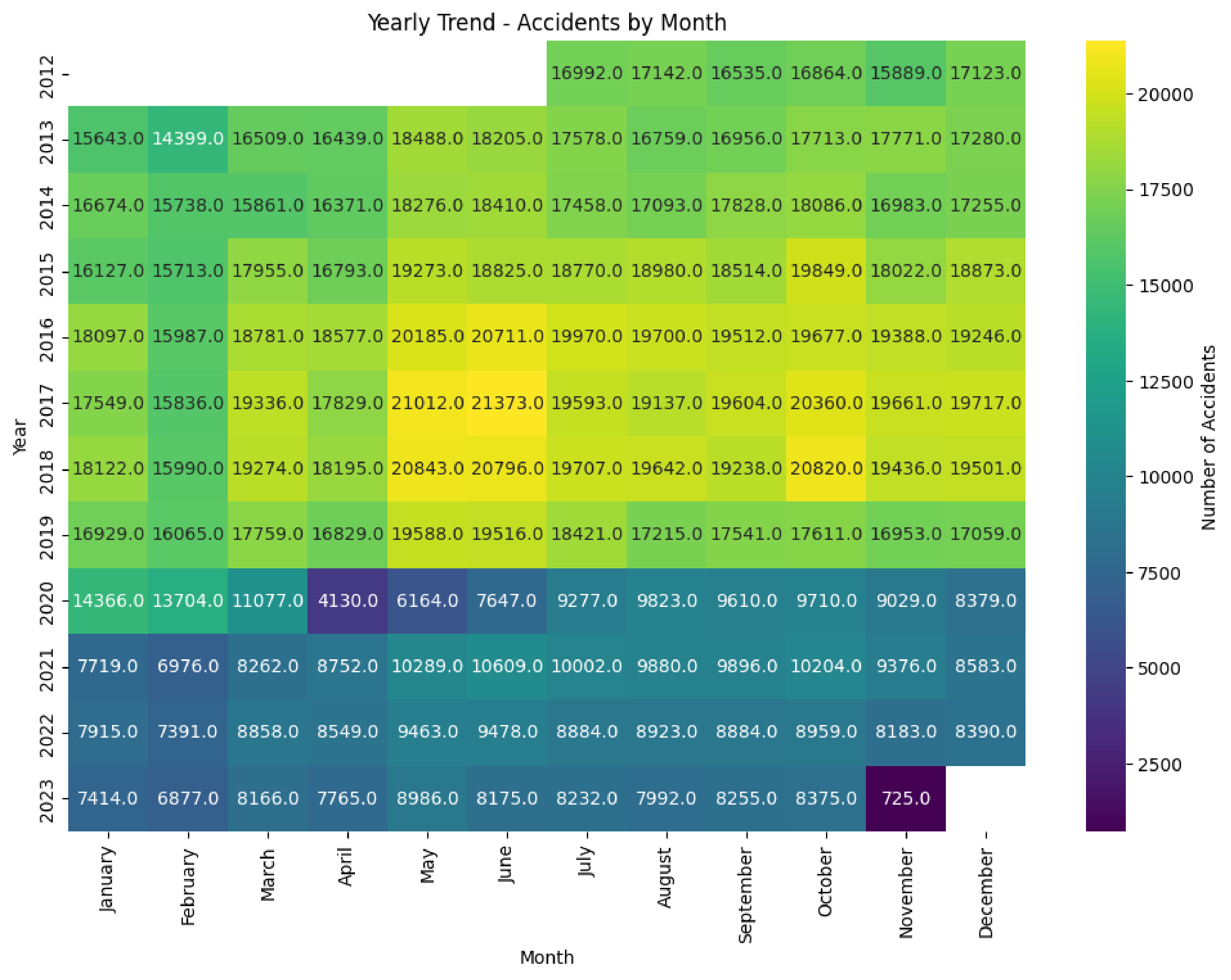


Accident Rates by Hour of the Day



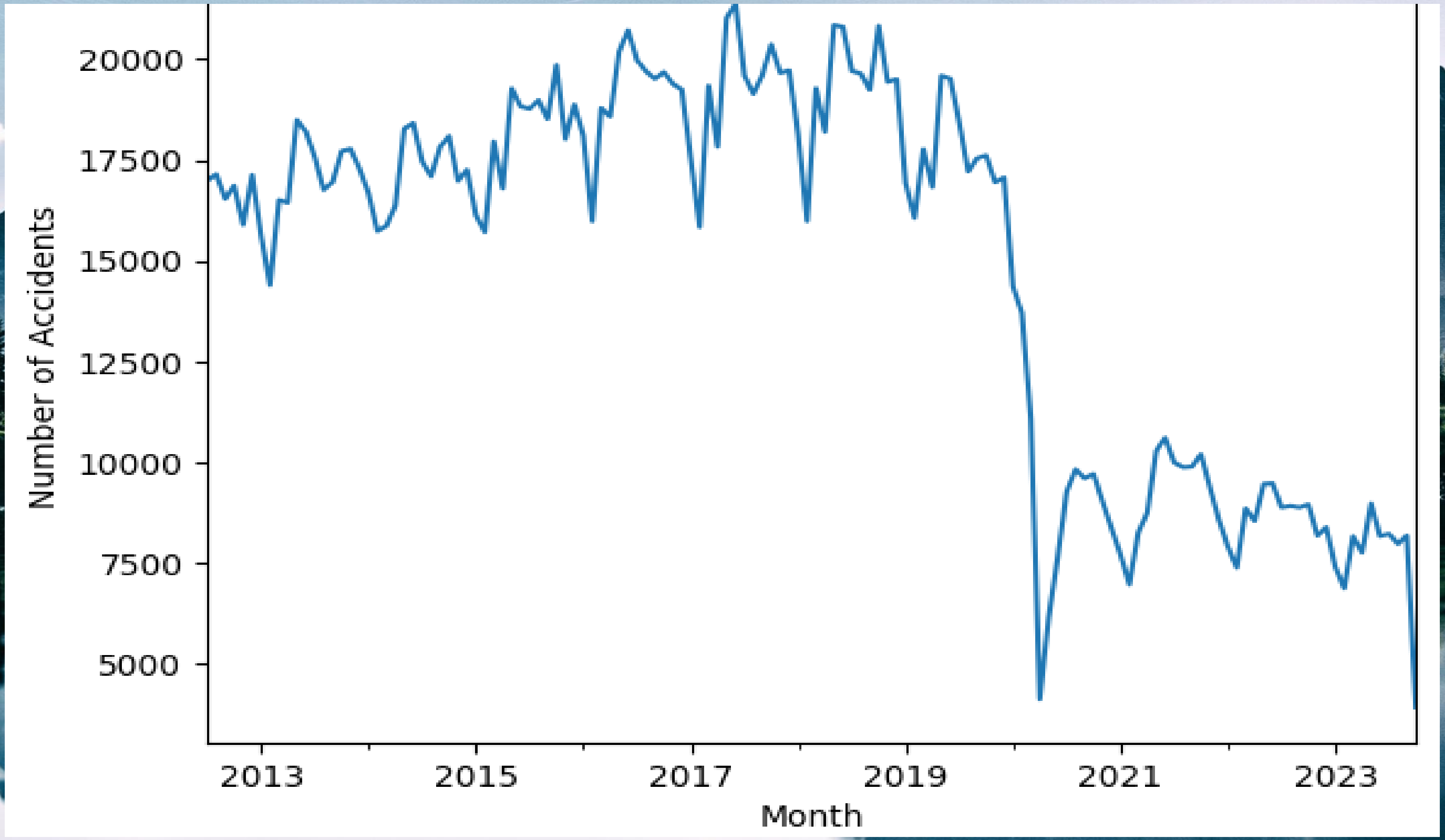
Yearly and Monthly Trend





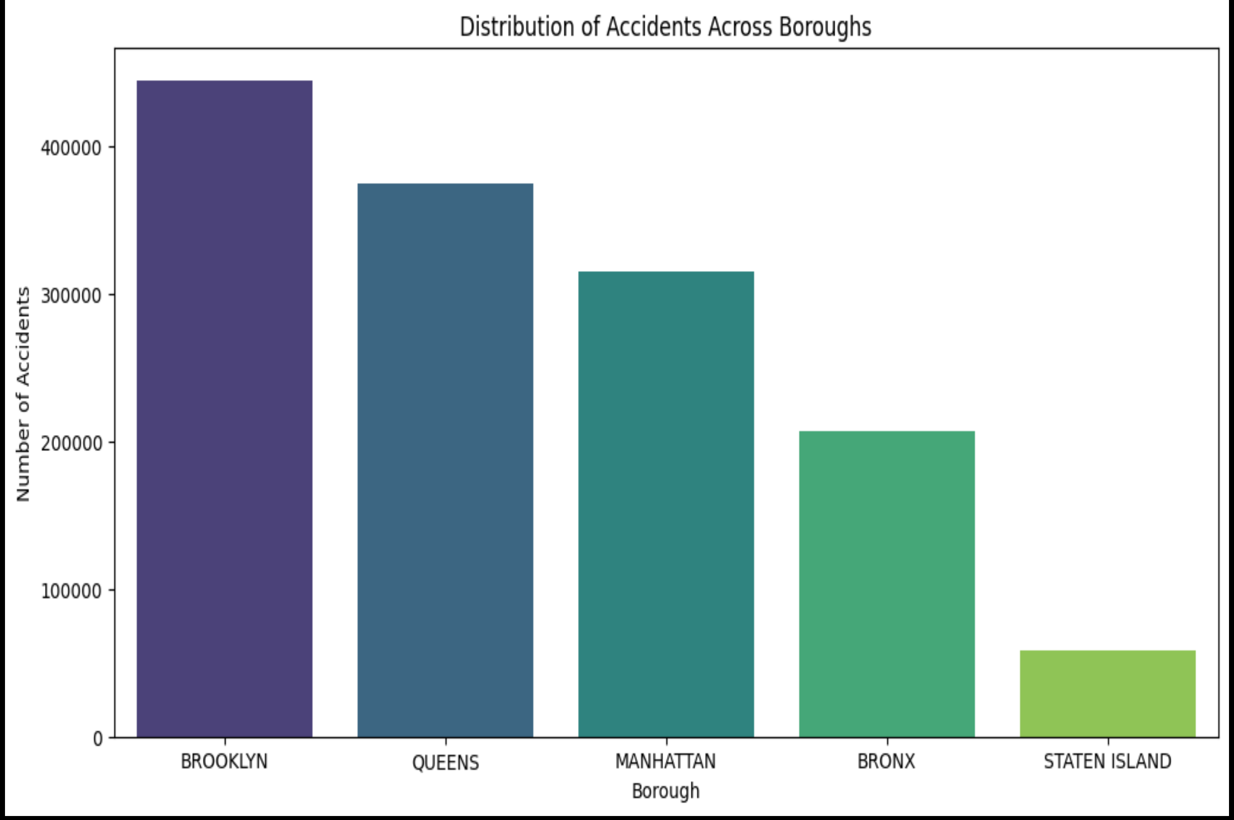
Yearly Trend - Accidents by Month

Year	January	February	March	April	May	June	July	August	September	October	November	December
2012							16992.0	17142.0	16535.0	16864.0	15889.0	17123.0
2013	15643.0	14399.0	16509.0	16439.0	18488.0	18205.0	17578.0	16759.0	16956.0	17713.0	17771.0	17280.0
2014	16674.0	15738.0	15861.0	16371.0	18276.0	18410.0	17458.0	17093.0	17828.0	18086.0	16983.0	17255.0
2015	16127.0	15713.0	17955.0	16793.0	19273.0	18825.0	18770.0	18980.0	18514.0	19849.0	18022.0	18873.0
2016	18097.0	15987.0	18781.0	18577.0	20185.0	20711.0	19970.0	19700.0	19512.0	19677.0	19388.0	19246.0
2017	17549.0	15836.0	19336.0	17829.0	21012.0	21373.0	19593.0	19137.0	19604.0	20360.0	19661.0	19717.0
2018	18122.0	15990.0	19274.0	18195.0	20843.0	20796.0	19707.0	19642.0	19238.0	20820.0	19436.0	19501.0
2019	16929.0	16065.0	17759.0	16829.0	19588.0	19516.0	18421.0	17215.0	17541.0	17611.0	16953.0	17059.0
2020	14366.0	13704.0	11077.0	4130.0	6164.0	7647.0	9277.0	9823.0	9610.0	9710.0	9029.0	8379.0
2021	7719.0	6976.0	8262.0	8752.0	10289.0	10609.0	10002.0	9880.0	9896.0	10204.0	9376.0	8583.0
2022	7915.0	7391.0	8858.0	8549.0	9463.0	9478.0	8884.0	8923.0	8884.0	8959.0	8183.0	8390.0
2023	7414.0	6877.0	8166.0	7765.0	8986.0	8175.0	8232.0	7992.0	8255.0	8375.0	725.0	

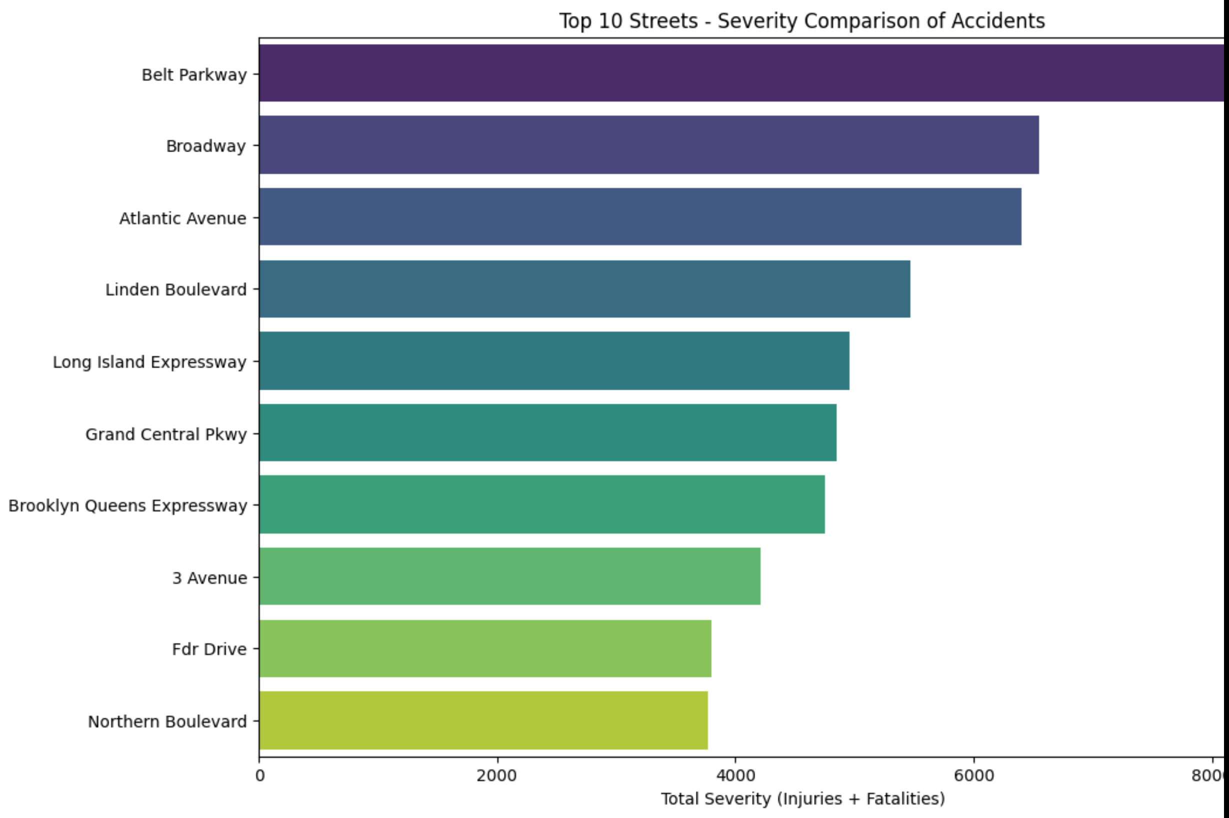


20000
17500
15000
12500
10000
7500
5000
Number of Accidents
2013
2015
2017
2019
2021
2023
Month

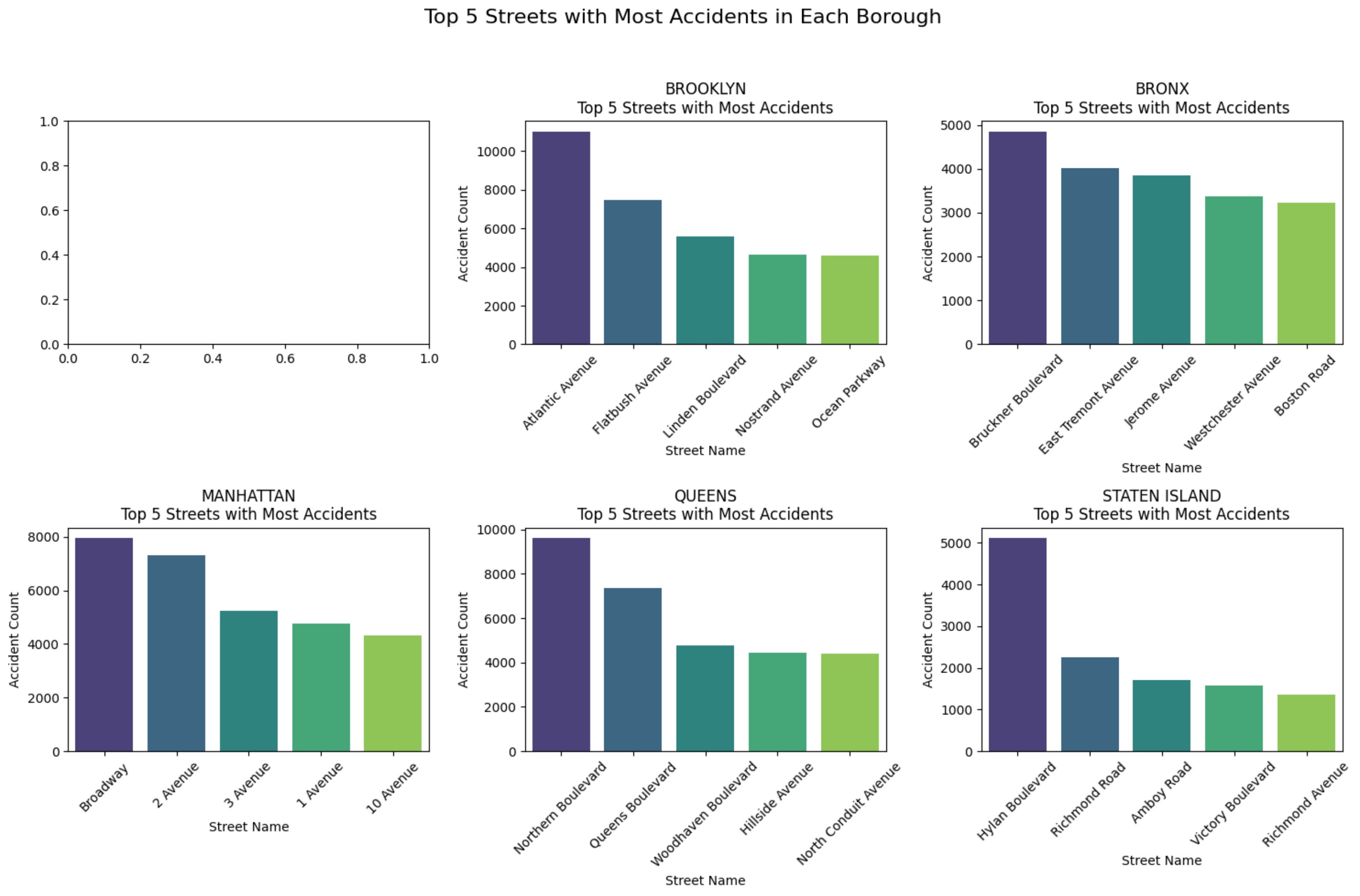
Which Borough has Most Accidents?



Which Streets have Most Dangerous Accidents Occur?



Top 5 Streetes with Most Accidents in Each Borough



Conclusion

By summarizing the distribution of these key variables, we can gain valuable insights into the overall patterns and trends of motor vehicle collisions. This information can be used to:

- Develop targeted interventions: Focus resources and efforts on areas, times, vehicle types, and contributing factors posing the highest risk.
- Prioritize safety initiatives: Allocate funding and implement programs based on the most prevalent risks and vulnerabilities identified.
- Design effective prevention strategies: Develop campaigns and policies tailored to address specific risk factors and conditions.
- Track progress and evaluate effectiveness: Monitor changes in collision patterns over time to assess the impact of interventions and safety initiatives.

Thank You So Much!