CAR CRASHES DATA ANALYSIS: CINCINNATI

report

Andrea Nappi andi.nappi@gmail.com
24-06-2024

SCOPE OF THE PROJECT:

The aim of this project is to analyze data on individuals involved in car crashes in Cincinnati from February 23, 2010, to January 14, 2021. The goal is to identify predictors of crash severity and potential injuries. Additionally, the project will seek to extract insights that can inform effective precautionary measures.

The data:

The dataset used is "Cincinnati Car Crash Data Since 2010" from Kaggle, the data has been obtained from data.cincinnati-oh.gov. In compliance with privacy laws, all Public Safety datasets are anonymized. Each row of the dataset corresponds to a person that has been involved in a car crash.

DATA EXPLORATION AND CLEANING:

The data has been explored and cleaned, checking for potential duplicates, removing irrelevant and redundant features, and editing some modalities for easier interpretation or to fix some spelling issues. Missing data has either been removed or, more often, in case of the variable at issue already having an "UNKNOWN" modality, the nan values have been turned to "UNKNOWN", to preserve information coming from the non-nan of variables in the same row. Most of the types have been changed to a more appropriate type, often categorical.

Ultimately, from the dataset of people involved in crashes, a sub-dataset has been extracted using the "INSTANCEID" variable, generating a second dataset useful to extract insight about car crashes, and not people involved, without the information being distorted by the fact that in one single car crash multiple people can be involved. The two datasets contain respectively 258448 people involved in a car crash and 133431 crashes, with the following features:

Crashes dataset:

- **CRASHDATE**: Day on which the crash occurred.
- **CRASHLOCATION**: Type of location where the crash occurred.
- **CRASHSEVERITY**: Indicator of how severe the crash was (worst condition among the people involved).
- **DAYOFWEEK**: Day of the week on which the crash occurred.
- **LIGHTCONDITIONSPRIMARY**: Light conditions when the crash occurred.
- **MANNEROFCRASH**: How the crash occurred.
- **ROADCONDITIONSPRIMARY**: Reported road conditions of when the crash occurred.
- **ROADCONTOUR**: The type of road contour the crash occurred on.
- **ROADSURFACE**: The type of road surface the crash occurred on.
- **SNA_NEIGHBORHOOD**: The statistical approximation of where the crash occurred.
- **WEATHER**: The weather at the time of the crash.

People involved dataset: all the variables above with the addition of the following:

- **AGE**: Age of the person.
- **GENDER**: gender of the person.
- **INJURIES**: Injury suffered by the person because of the crash.
- **INSTANCEID**: Unique ID or the crash.
- **TYPEOFPERSON**: type of person involved in the crash.
- **UNITTYPE**: Type of vehicle on which the person was.

DESCRIPTIVE STATISTICS AND VISUALIZATION:

The chosen approach for this section has been to remove the unknown values from a single variable when performing analysis on it. This allowed to retain much more information at the cost of more computational power.

CRASHES:

Frequencies and bar charts for categorical variables:

Category	Absolute Frequency	Relative Frequency (%)
NOT AN INTERSECTION	65278	65.26429449815538
FOUR-WAY INTERSECTION	18520	18.516111616560522
T-INTERSECTION	9878	9.875926055528339
DRIVEWAY/ALLEY ACCESS	2331	2.330510592775517
ON RAMP	1353	1.3527159296547724
OFF RAMP	1239	1.238739864628428
Y-INTERSECTION	941	0.9408024314893872
FIVE-POINT, OR MORE	249	0.2489477209785945
CROSSOVER	84	0.08398236370362223
SHARED-USE PATHS OR TRAILS	75	0.07498425330680558
TRAFFIC CIRCLE/ROUNDABOUT	47	0.04699013207226482
RAILWAY GRADE CROSSING	26	0.025994541146359268

<u>Table 1:</u> Frequency table of CRASHLOCATION

Category	Absolute Frequency	Relative Frequency (%)
PROPERTY DAMAGE ONLY	105654	79.18249881961464
INJURY	20074	15.044479918459727
MINOR INJURY SUSPECTED	3854	2.888384258530626
INJURY POSSIBLE	3204	2.401241090900915
SERIOUS INJURY SUSPECTED	417	0.3125210783101378
FATAL INJURY	161	0.12066161536674386
FATAL	67	0.05021321881721639

<u>Table 2:</u> Frequency table of CRASHSEVERITY

Category	Absolute Frequency	Relative Frequency (%)
FRI	22401	16.788452458574092
THU	20161	15.109682157819398
TUE	20106	15.068462351327652
WED	20096	15.060967841056428
MON	19068	14.2905321851 7436
SAT	17340	12.99548081030645
SUN	14259	10.68642219574162

Table 3: Frequency table of DAYOFWEEK

Category	Absolute Frequency	Relative Frequency (%)
DAYLIGHT	89756	68.03046954939933
DARK	33263	25.211657255466708
DUSK	3349	2.5383711676204195
DAWN	2767	2.09724485542123
DARK - ROADWAY NOT LIGHTED	1961	1.4863379694546557
DARK - UNKNOWN ROADWAY LIGHTING	839	0.6359192026376624

Table 4: Frequency table of LIGHTCONDITIONSPRIMARY

Category	Absolute Frequency	Relative Frequency (%)
CLEAR	86246	65.30918232897666
CLOUDY	23140	17.5226037044329
RAIN	19365	14.664011267776281
SNOW	2778	2.103621136167441 7
SLEET, HAIL	265	0.20066940283814688
FOG, SMOG, SMOKE	190	0.14387617561980343
SEVERE CROSSWINDS	27	0.020445561798603643
BLOWING SAND, SOIL, DIRT, SNOW	25	0.018931075739447818
FREEZING RAIN OR FREEZING DRIZZLE	22	0.016659346650714082

Table 5: Frequency table of WEATHER

Category	Absolute Frequency	Relative Frequency (%)
DRY	99834	75.40218425703539
WET	29210	22.06160027794142
SNOW	1939	1.464479388528874
ICE	1203	0.9085965468799565
SLUSH	91	0.06873007960604825
SAND, MUD, DIRT, OIL, GRAVEL	81	0.0611773236053836
WATER (STANDING, MOVING)	44	0.03323212640292443

Table 6: Frequency table of ROADSCONDITIONSPRIMARY

Category	Absolute Frequency	Relative Frequency (%)
STRAIGHT LEVEL	91293	68.48841309256773
STRAIGHT GRADE	26590	19.947935812508906
CURVE GRADE	8863	6.6490618693594
CURVE LEVEL	6551	4.914589225563966

Table 7: Frequency table of ROADCONTOUR

Category	Absolute Frequency	Relative Frequency (%)
BLACKTOP, BITUMINOUS, ASPHALT	111537	83.66048859519506
CONCRETE	21298	15.974977685435904
BRICK/BLOCK	295	0.22127046751824545
SLAG, GRAVEL, STONE	134	0.10050929710998269
DIRT	57	0.042753954740813525

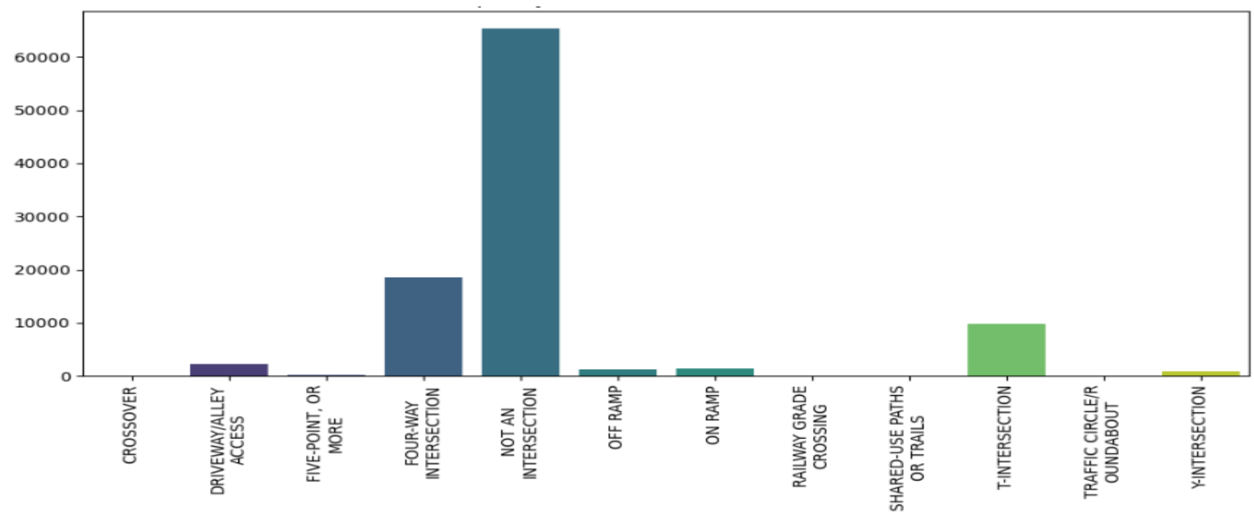
Table 8: Frequency table of ROADSURFACE

Category	Absolute Frequency	Relative Frequency (%)
ANGLE	34536	26.632324930404007
REAR-END	33608	25.91670072564911
NOT COLLISION BETWEEN TWO MOTOR VEHICLES IN TRANSPORT	29844	23.01410427446656
SIDESWIPE, SAME DIRECTION	21135	16.29818703393817
BACKING	4880	3.763196249142099
SIDESWIPE, OPPOSITE DIRECTION	2868	2.2116489431433486
HEAD-ON	2311	1.782120190935941
REAR-TO-REAR	495	0.3817176523207662

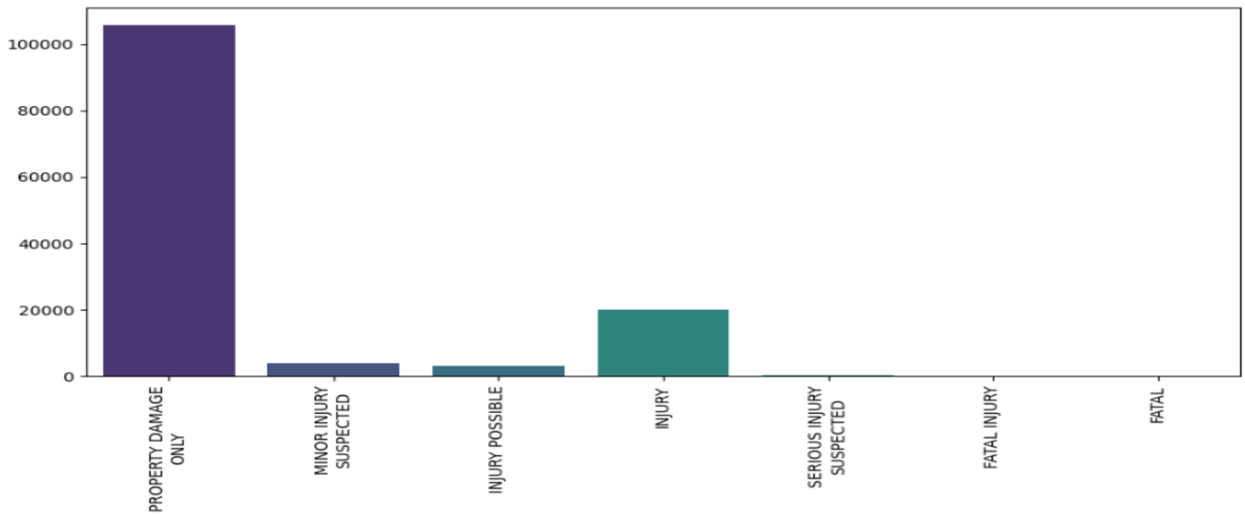
Table 9: Frequency table of MANNEROFCRASH

Category	Absolute Frequency	Relative Frequency (%)
WESTWOOD	10300	7.89090630506397
DOWNTOWN	9086	6.960851911437984
AVONDALE	6720	5.1482417834980465
CUF	6459	4.9482877499425415
WEST PRICE HILL	5556	4.256492760284992
EAST PRICE HILL	5448	4.173753160193059
BOND HILL	4681	3.5861487780586834
CAMP WASHINGTON	4673	3.5800199187926145
CLIFTON	4505	3.4513138742051632
WALNUT HILLS	4411	3.3792997778288516
OAKLEY	3940	3.018463188539033
WEST END	3921	3.003907147782119
CORRYVILLE	3619	2.7725427104880107
OVER-THE-RHINE	3513	2.691335325212595
COLLEGE HILL	3494	2.676779284455681
NORTHSIDE	3277	2.5105339768635564
ROSELAWN	3146	2.41017390638167 5
QUEENSGATE	3096	2.371868535968743
MT. AIRY	3089	2.3665057841109323
SOUTH FAIRMOUNT	2723	2.086110472688271
HYDE PARK	2667	2.043208457825787
MADISONVILLE	2590	1.984218187389872
EVANSTON	2457	1.8823259020914733
SPRING GROVE VILLAGE	2107	1.61418830920095
CARTHAGE	2069	1.5850762276871218
MT. AUBURN	1950	1.4939094461043438
HARTWELL	1815	1.3904849459894277
NORTH AVONDALE - PADDOCK HILLS	1800	1.378993334865548
PLEASANT RIDGE	1597	1.2234735309890448
WINTON HILLS	1261	0.9660614418141424
EAST END	1212	0.9285221788094692
MT. WASHINGTON	1118	0.8565080824331571
MT. ADAMS	1029	0.7883245230981384
EAST WALNUT HILLS	972	0.7446564008273959
PENDLETON	949	0.7270359304374473
LOWER PRICE HILL	919	0.7040527081896882
RIVERSIDE	878	0.6726423044510841
EAST WESTWOOD	834	0.6389335784877039
NORTH FAIRMOUNT	793	0.6075231747490998
VILLAGES AT ROLL HILL	773	0.592201026583927
SOUTH CUMMINSVILLE	743	0.569217804336168
MILLVALE	695	0.5324446487397533
LINWOOD	680	0.5209530376158737
CALIFORNIA	585	0.44817283383130313
MT. LOOKOUT	577	0.44204397456523403
SEDAMSVILLE	449	0.3439822263081284
KENNEDY HEIGHTS	438	0.3355550448172834
COLUMBIA TUSCULUM	436	0.3340228300007661
SAYLER PARK	334	0.25587987435838505
ENGLISH WOODS	146	0.11185168160576112

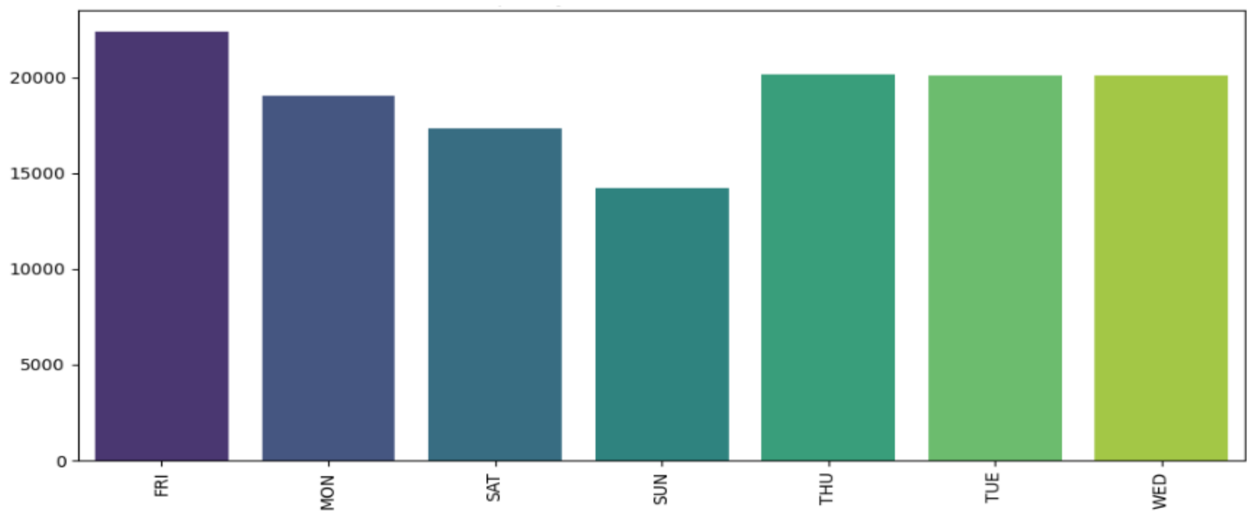
Table 10: Frequency table of SNA_NEIGHBORHOOD



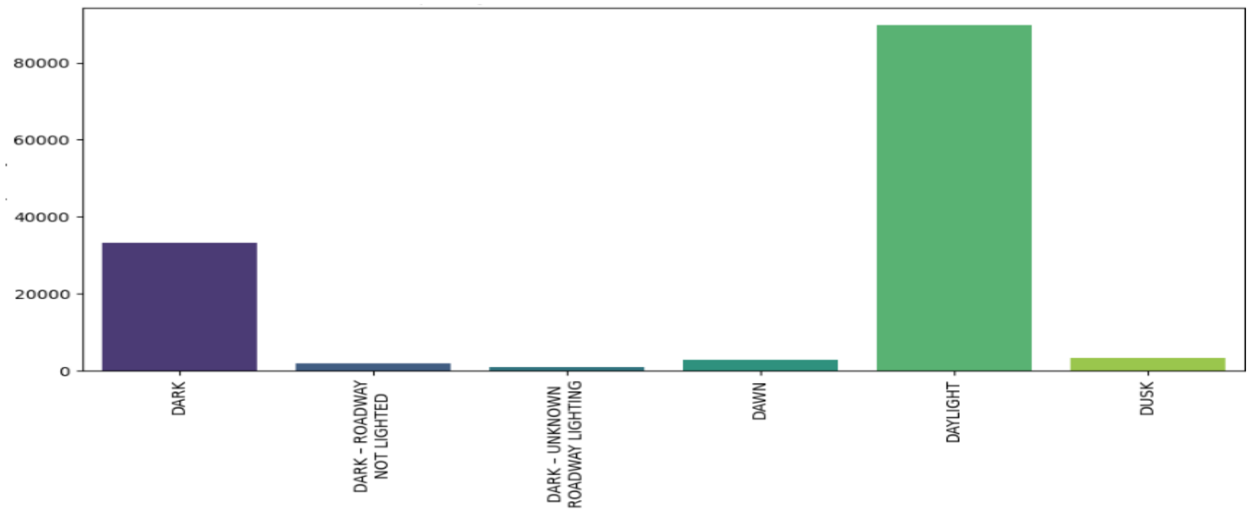
Plot 1: Bar chart of CRASHLOCATION



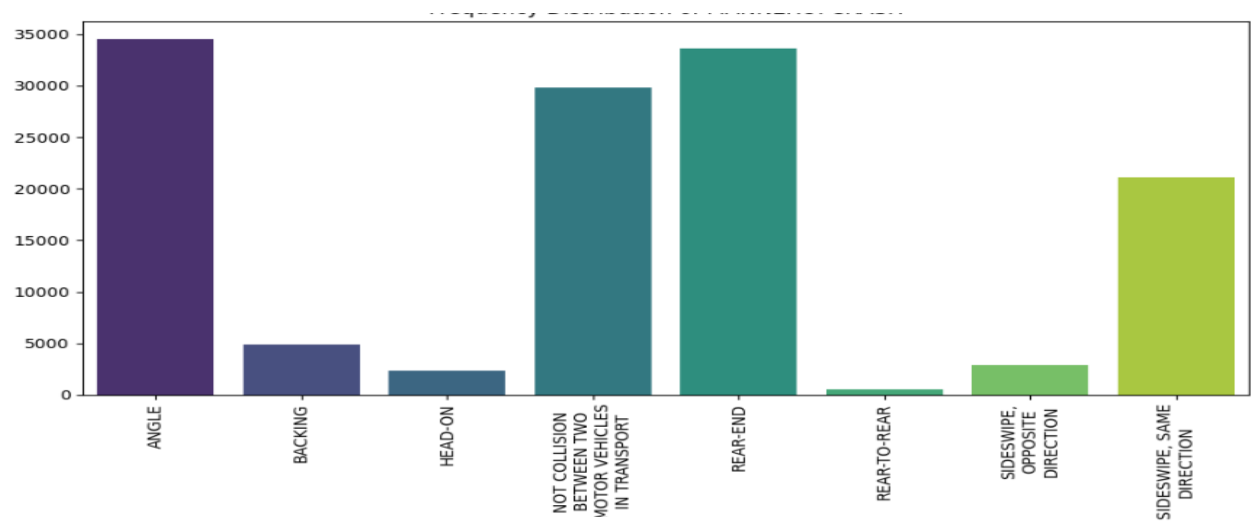
Plot 2: Bar chart of CRASHSEVERITY



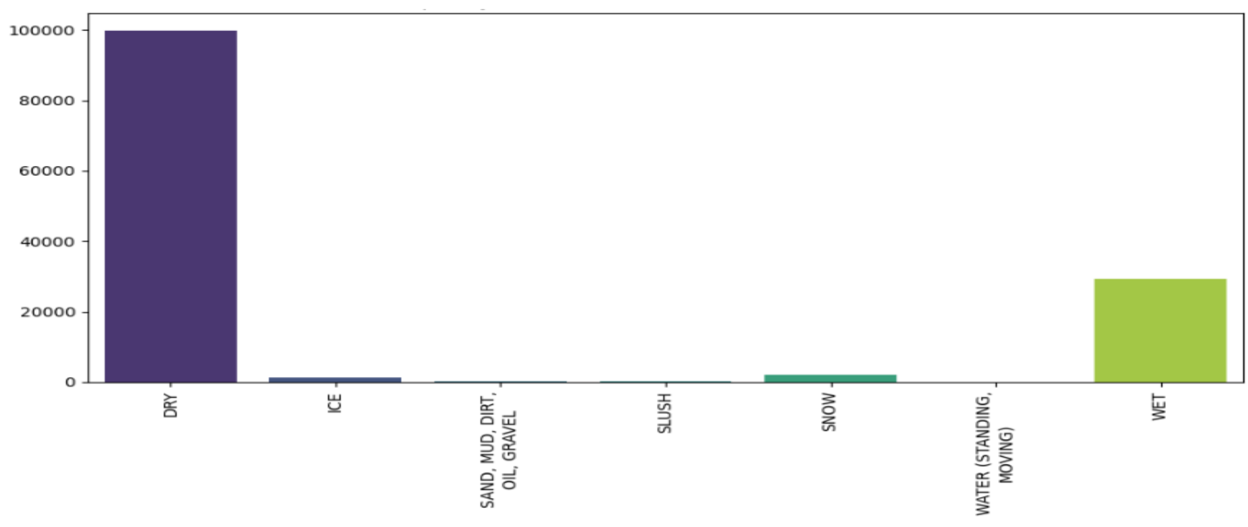
Plot 3: Bar chart of DAYOFWEEK



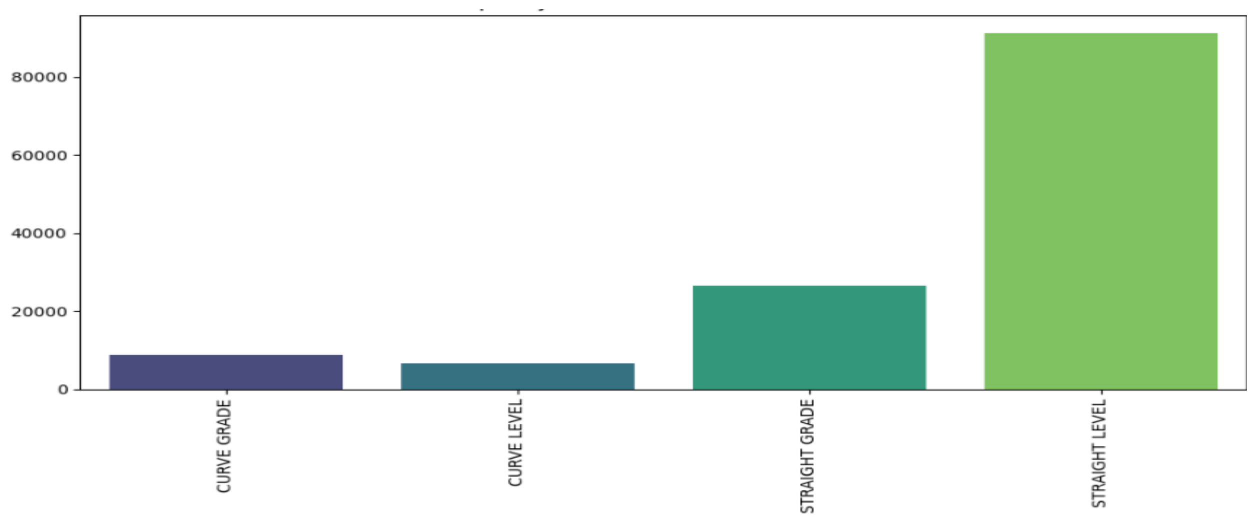
Plot 4: Bar chart of LIGHTCONDITIONSPRIMARY



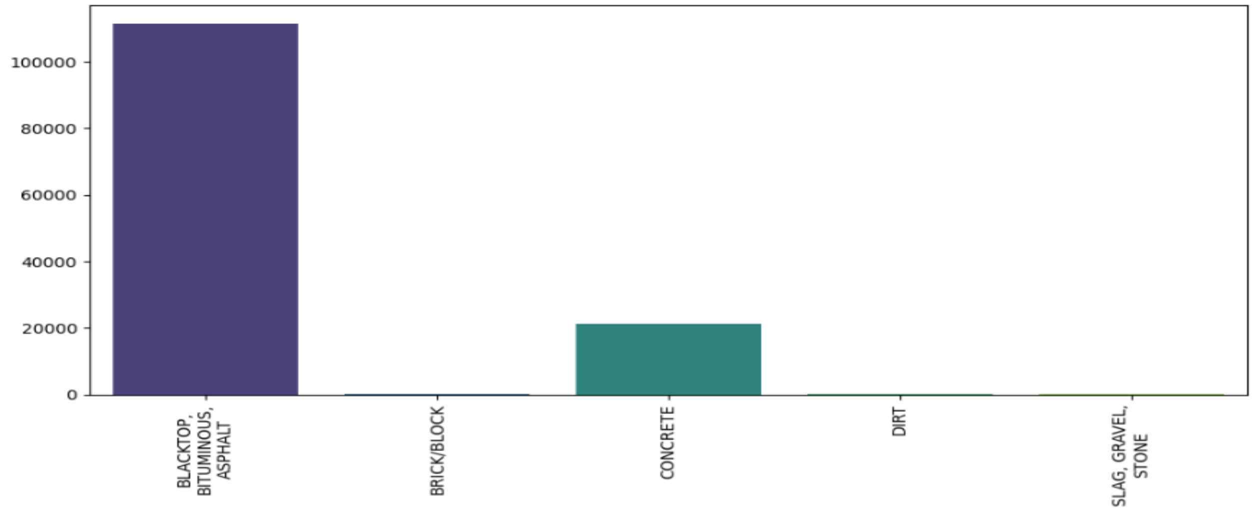
Plot 5: Bar chart of MANNEROFCRASH



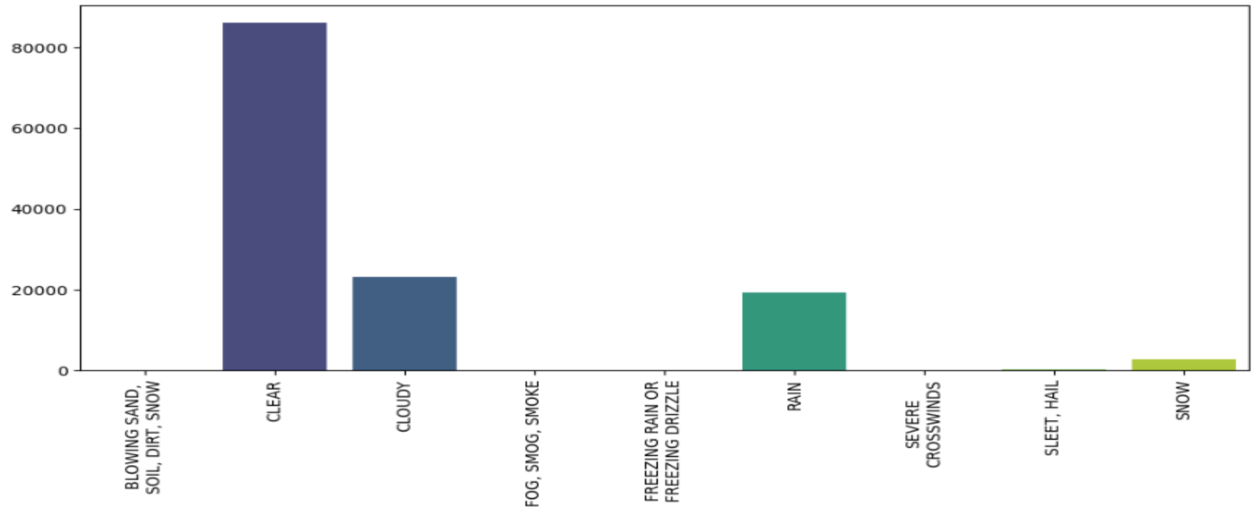
Plot 6: Bar chart of ROADCONDITIONSPRIMARY



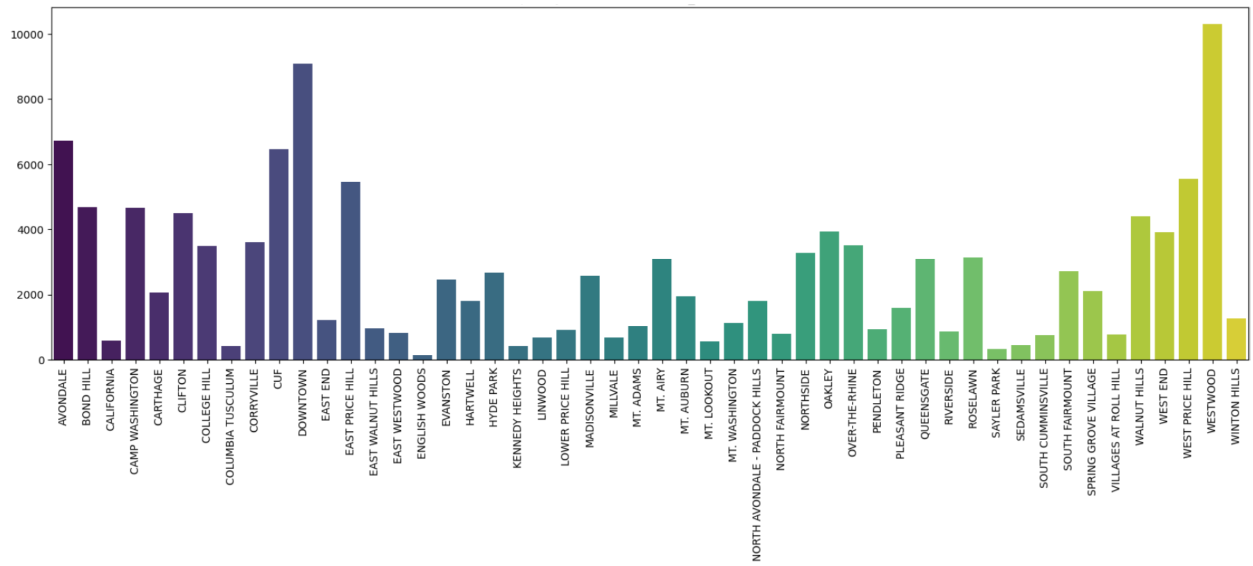
Plot 7: Bar chart of ROADCONTOUR



Plot 8: Bar chart of ROADSURFACE



Plot 9: bar chart of WEATHER



<u>Plot 10:</u> bar chart of SNA_NEIGHBORHOOD

Most crashes in Cincinnati did not occur at an interception (65.2%; *Table 1*, *Plot 1*). About 85% of the crashes resulted in no harm to any of the people involved, with only damages to properties reported, and less than 0.2% of crashes resulting in a fatality (*Table 2*, *Plot 2*). The distribution of crashes over the days of the week is uniform, with the exception of weekend days, with fewer crashes (on average) on Saturday and even fewer on Sunday (*Table 3*, *Plot 3*), this might be because people work less during the weekend, and are therefore less prone to take the car, and fewer cars going around means fewer car crashes (to be noted that this last assumption does not come from the available data, and comes exclusively from my own thoughts, under the hypothesis that more crashes happen where more cars are on the streets). Most crashes occur in daylight (68%; *Table 4*, *Plot 4*), with clear weather (65%; *Table 5*, *Plot 9*), on dry roads (75.4%; *Table 6*, *Plot 6*), on a straight level of road contour (68.4%; *Table 7*, *Plot 7*), and on asphalt/blacktop/bituminous (83.6%, *Table 8*, *Plot 8*); and again this might be due to these being the most common conditions (again, my assumption, not coming from the available data). The two neighborhoods where there are the most car crashes are Westwood and Downtown (7.8% and 6.9%; *Table 10*, *Plot 10*), and the most common type of crashes are angle crashes and rear-ends (26.6% and 25.9%; *Table 9*, *Plot 5*).

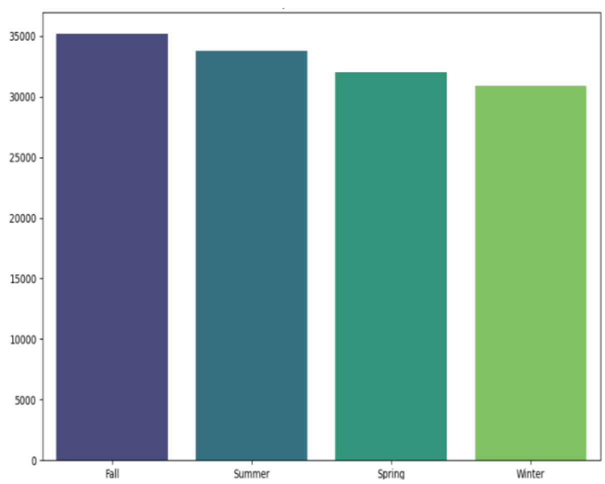
Frequencies and bar charts for temporal variables:

Season	Absolute Frequency	Relative Frequency (%)
Fall	35173	26.660552266749537
Summer	33796	25.616809041226716
Spring	32051	24.294127902129176
Winter	30909	23.428510789894563

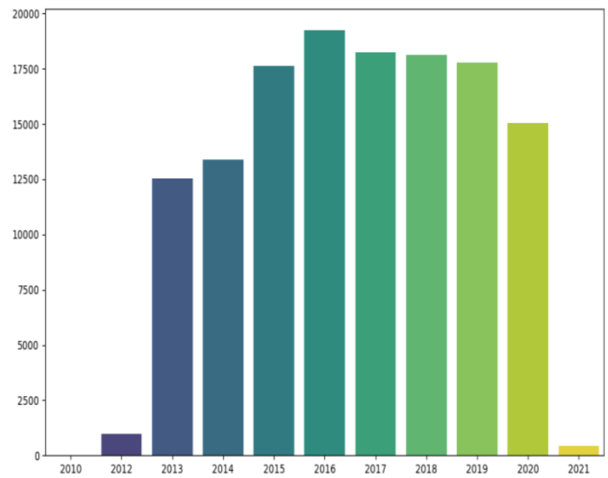
Table 11: Frequency table of crashes per season

Year	Absolute Frequency	Relative Frequency (%)
2010.0	2.0	0.0014989020542452653
2012.0	986.0	0.7389587127429158
2013.0	12549.0	9.404860939361917
2014.0	13371.0	10.020909683656722
2015.0	17628.0	13.211322706117768
2016.0	19231.0	14.412692702595347
2017.0	18256.0	13.681977951150781
2018.0	18131.0	13.588296572760452
2019.0	17792.0	13.334232674565879
2020.0	15051.0	11.279987409222745
2021.0	434.0	0.3252617457712226

Table 12: Frequency table of crashes per year



Plot 11: Bar chart of crashes per season



Plot 12: Bar chart of crashes per year

Without taking into consideration the first 2 years available (2010 and 2012) and the last one (2021) because of the incompleteness of the data, Cincinnati has experienced a stable growth in reported car crashes from 2013 to 2016. From 2017 until 2019 the number of crashes decreased slightly, and decreased more in 2020 (*Table 12*, *Plot 12*), possibly due to COVID-19 restrictions (personal assumption).
Seasonally, the differences in number of crashes are small, Fall is the season that counts the most of them in the considered years (26.6%) followed by Summer (25.6%), Spring (24.3%), and Winter (23.4%) (*Table 11*, *Plot 11*). These data are originated from the variable CRASHDATE.

PEOPLE INVOLVED:

Frequencies and bar charts for categorical variables:

Category	Absolute Frequency	Relative Frequency (%)
18-25	52972	23.27835857953322
31-40	45444	19.970205529115525
41-50	33528	14.733761354198252
26-30	31361	13.781480846725463
51-60	29940	13.157027408276534
61-70	16230	7.132216260398402
UNDER 18	10494	4.6115512899951225
OVER 70	7590	3.335398731757 4783

Table 13: Frequency table of AGE

Category	Absolute Frequency	Relative Frequency (%)
DRIVER	232230	89.85559957902557
OCCUPANT	23368	9.04166408716647
PEDESTRIAN	2850	1.1027363338079612

Table 14: Frequency table of TYPEOFPERSON

Category	Absolute Frequency	Relative Frequency (%)
NO INJURY / NONE REPORTED	165914	64.19627932891723
NO APPARENTY INJURY	53956	20.876926886646444
POSSIBLE INJURY	20247	7.834071070389402
NON-INCAPACITATING	10371	4.01279947997276
SUSPECTED MINOR INJURY	5140	1.9887946511483936
INCAPACITATING	2067	0.7997740357828267
SUSPECTED SERIOUS INJURY	513	0.19849254008543304
FATAL	240	0.09286200705751255

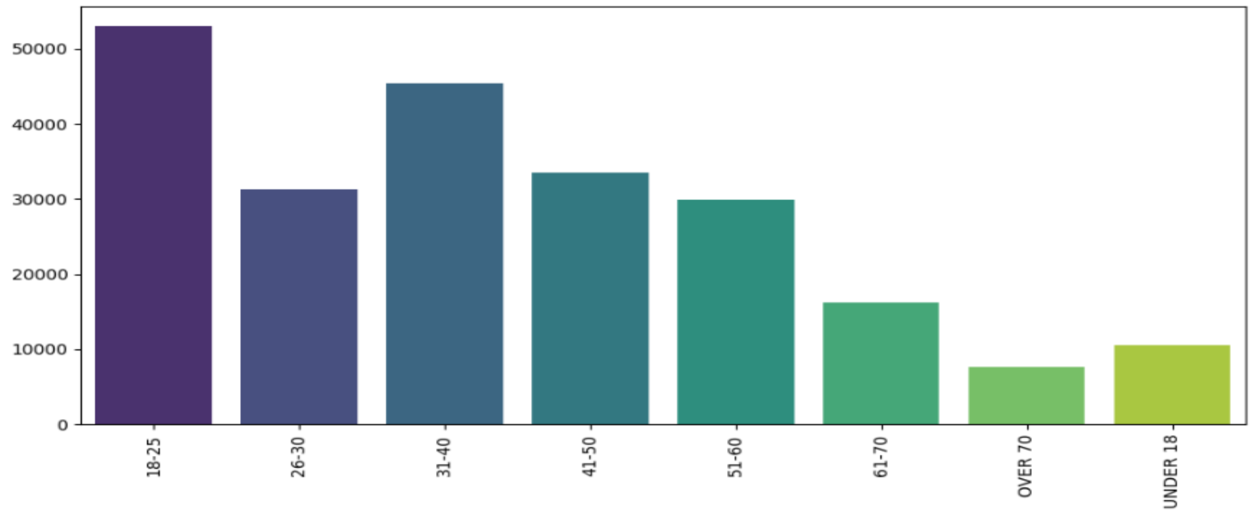
Table 15: Frequency table of INJURIES

Category	Absolute Frequency	Relative Frequency (%)
MALE	124324	54.15327252611313
FEMALE	105254	45.84672747388687

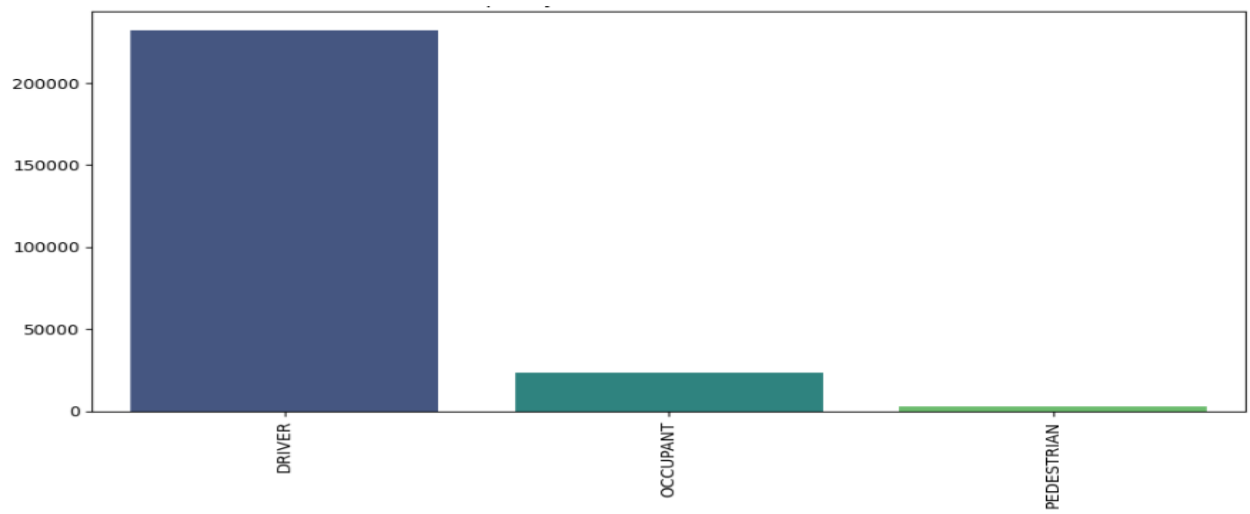
Table 16: Frequency table of GENDER

Category	Absolute Frequency	Relative Frequency (%)
MID SIZE	69735	27.88641489514852
SPORT UTILITY VEHICLE	46151	18.455380136602844
PASSENGER CAR	40815	16.321560535534335
COMPACT	23666	9.46382583937169
FULL SIZE	16216	6.484636178959323
PICKUP	12033	4.8118911656029555
MINIVAN	8210	3.2831069948973877
VAN	3925	1.5695730761232947
PICK UP	3644	1.4572036406097542
TRACTOR/SEMI-TRAILER	2829	1.1312922884975287
PEDESTRIAN/SKATER	2809	1.1232944639058176
BUS (16+ SEATS, INCLUDING THE DRIVER)	2639	1.0553129548762736
UNKNOWN OR HIT/SKIP	2551	1.020122526672745
SINGLE UNIT TRUCK OR VAN 2 AXLES, 6 TIRES	2361	0.94414319305149
PASSENGER VAN (MINIVAN)	2189	0.8753619015627749
SUB-COMPACT	1585	0.633827598893101
SEMI-TRACTOR	1098	0.43908057008493695
SINGLE UNIT TRUCK	832	0.3327095030151799
CARGO VAN	821	0.3283106994897388
SINGLE UNIT TRUCK; 3+ AXLES	770	0.3079162467808756
BUS (16+ PASSENGERS)	761	0.3043172257146056
BUS /VAN (9-15 SEATS INCLUDING THE DRIVER)	719	0.2875217940720124
MOTORCYCLE	691	0.2763248396436169
SINGLE UNIT TRUCK / TRAILER	607	0.2427339763584305
OTHER MED/HEAVY VEHICLE	442	0.17675192347681432
BICYCLE/PEDACYCLIST	378	0.15115888478333894
OTHER VEHICLE	318	0.12716541100820578
MOTORCYCLE 2 WHEELED	256	0.10237215477390149
VAN (9-15 SEATS)	242	0.09677367755970376
OTHER PASSENGER VEHICLE	211	0.08437704944255163
TRUCK/TRACTOR (BOBTAIL)	125	0.049986403698194085
HEAVY EQUIPMENT	99	0.03958923172896972
OTHER NON-MOTORIST	98	0.03918934049938417
BICYCLE	87	0.03479053697394309
TRACTOR/DOUBLES	42	0.016795431642593214
MOTORIZED BICYCLE	28	0.011196954428395476
MOTORHOME	18	0.0071980421325399495
MOPED OR MOTORIZED BICYCLE	14	0.005598477214197738
ALL TERRAIN VEHICLE (ATV/UTV)	11	0.00439880352544108
SNOWMOBILE/ATV	11	0.00439880352544108
TRACTOR/TRIPLES	8	0.0031991298366844216
GOLF CART	6	0.0023993473775133166
MOTORCYCLE 3 WHEELED	4	0.0015995649183422108
AUTOCYCLE	3	0.0011996736887566583
WHEELCHAIR (ANY TYPE)	3	0.0011996736887566583
ANIMAL WITH BUGGY, WAGON, SURREY	2	0.0007997824591711054
TRAIN	2	0.0007997824591711054
LIMO (LIVERY VEHICLE)	2	0.0007997824591711054
ANIMAL WITH RIDER OR ANIMAL DRAWN VEHICLE	1	0.0003998912295855527

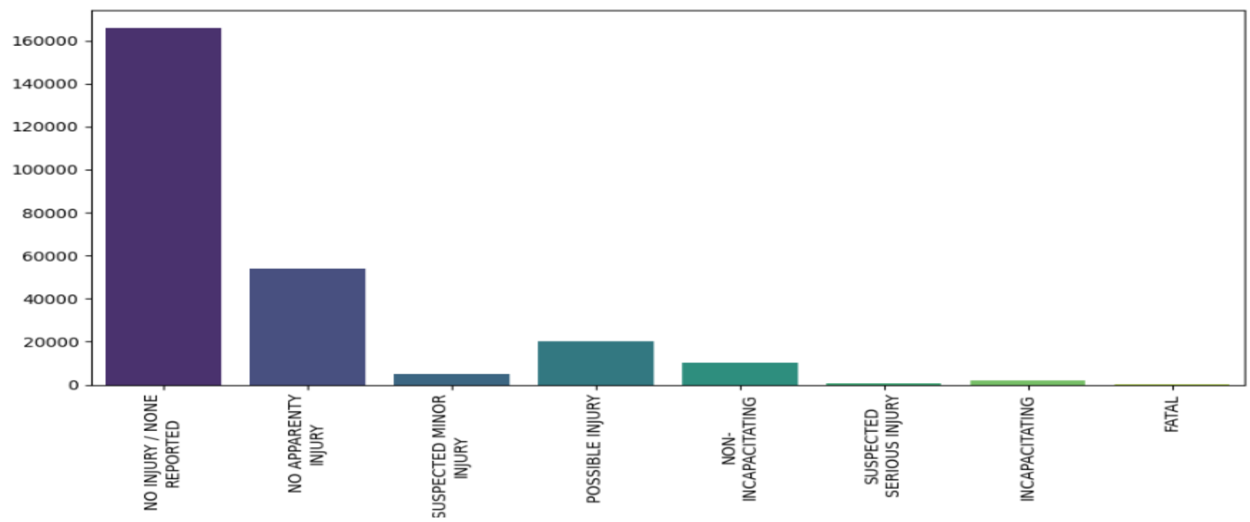
Table 17: Frequency table of UNITTYPE



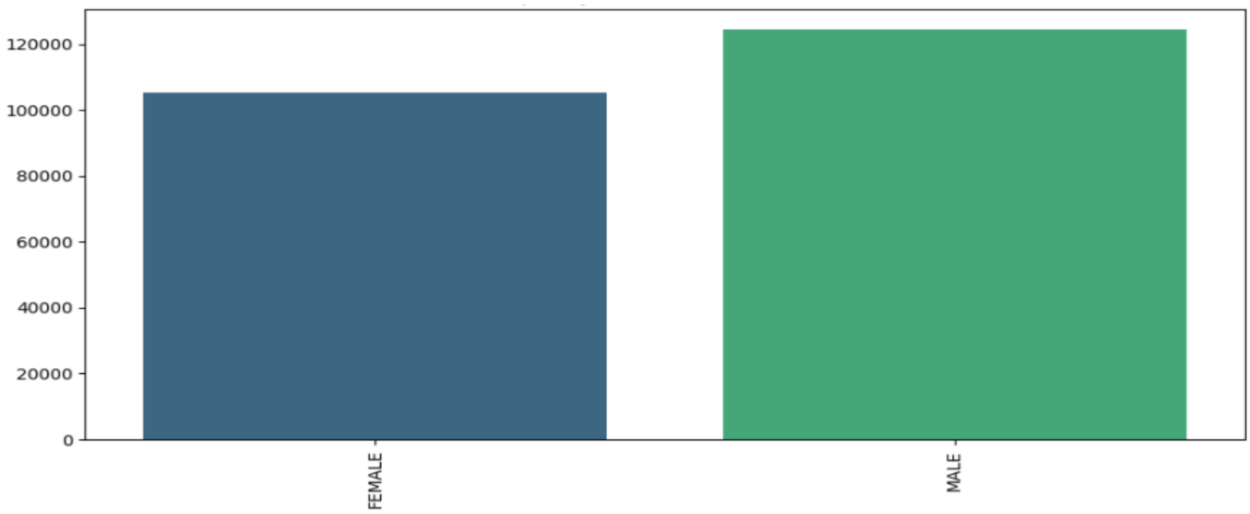
Plot 13: bar chart of AGE



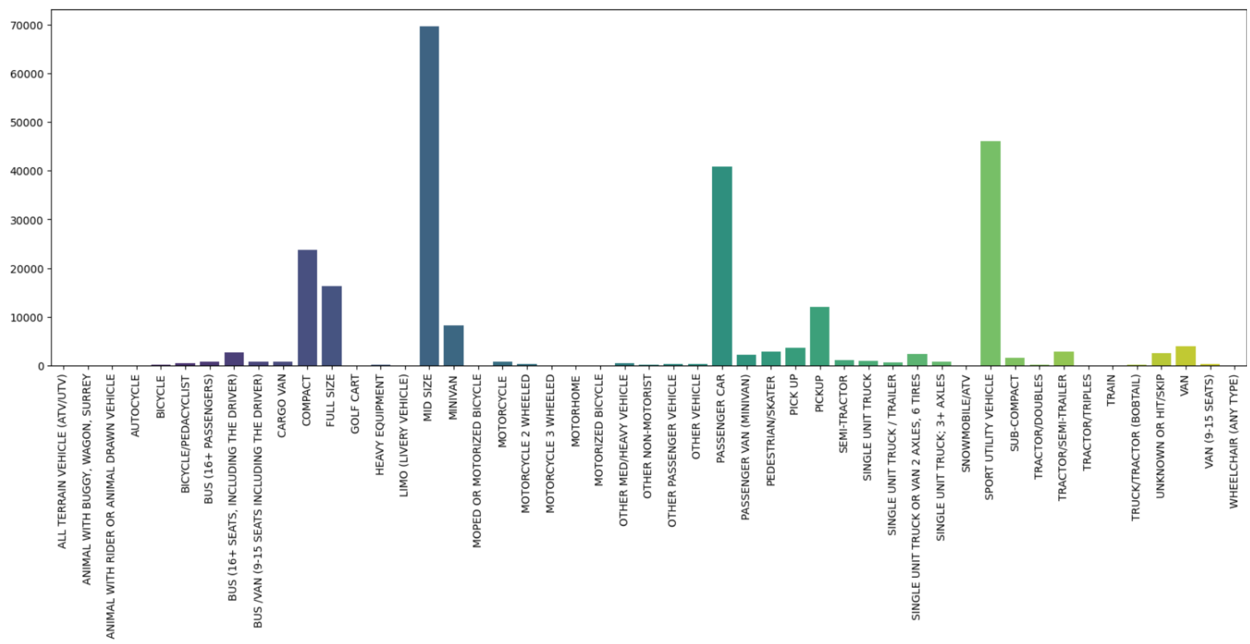
Plot 14: bar chart of TYPEOFPERSON



Plot 15: bar chart of INJURIES



Plot 16: bar chart of GENDER



Plot 17: bar chart of UNITTYPE

Most people belong either to the age range 18-25 (23.2%) or 31-40 (19.9%), while the least involved age ranges are under 18 (4.6%) and over 70 (3.3) (*Table 13*, *Plot 13*), with a slightly higher involvement of male figures (54%) compared to female ones (46%) (*Table 16*, *Plot 16*). In about 85% of the cases, people involved either suffered no injuries or minor ones, with fatal events occurring only about 0.09% of the times (*Table 15*, *Plot 15*). In 90% of cases the involved people are drivers (to be considered that drivers are the only figure that will always be involved in a crash), in 9% of the cases they are passengers and in about 1% of the cases they are pedestrians (*Table 14*, *Plot 14*). The most involved types of vehicles are mid-size units (27%), followed by sport utility vehicles (18%), and by passenger cars (16%) (*Table 17*, *Plot 17*).

PEOPLE ANALYSIS:

This section focuses on providing insights about the people involved in the crashes.

CONTINGENCY TABLES:

The presented contingency tables are computed as the difference, in percentual, between the contingency tables of the expected frequencies and the actual contingency tables, therefore suggesting how much each pair of occurrences diverges from the case of complete independence between the two variables at issue. In the computation of these tables, rows that present UNKNOWN values in either of the two variables have been removed.

Table 18: Contingency table of AGE and INJURIES

The data shows that fatal occurrences are more frequent than expected in the case of older people (71% more for 61-70 and 203% more for over 70) and for under 18 people (46.39% more). Under 18 people suffer harder injuries more often than expected and small/no injuries less than expected (*Table 18*).

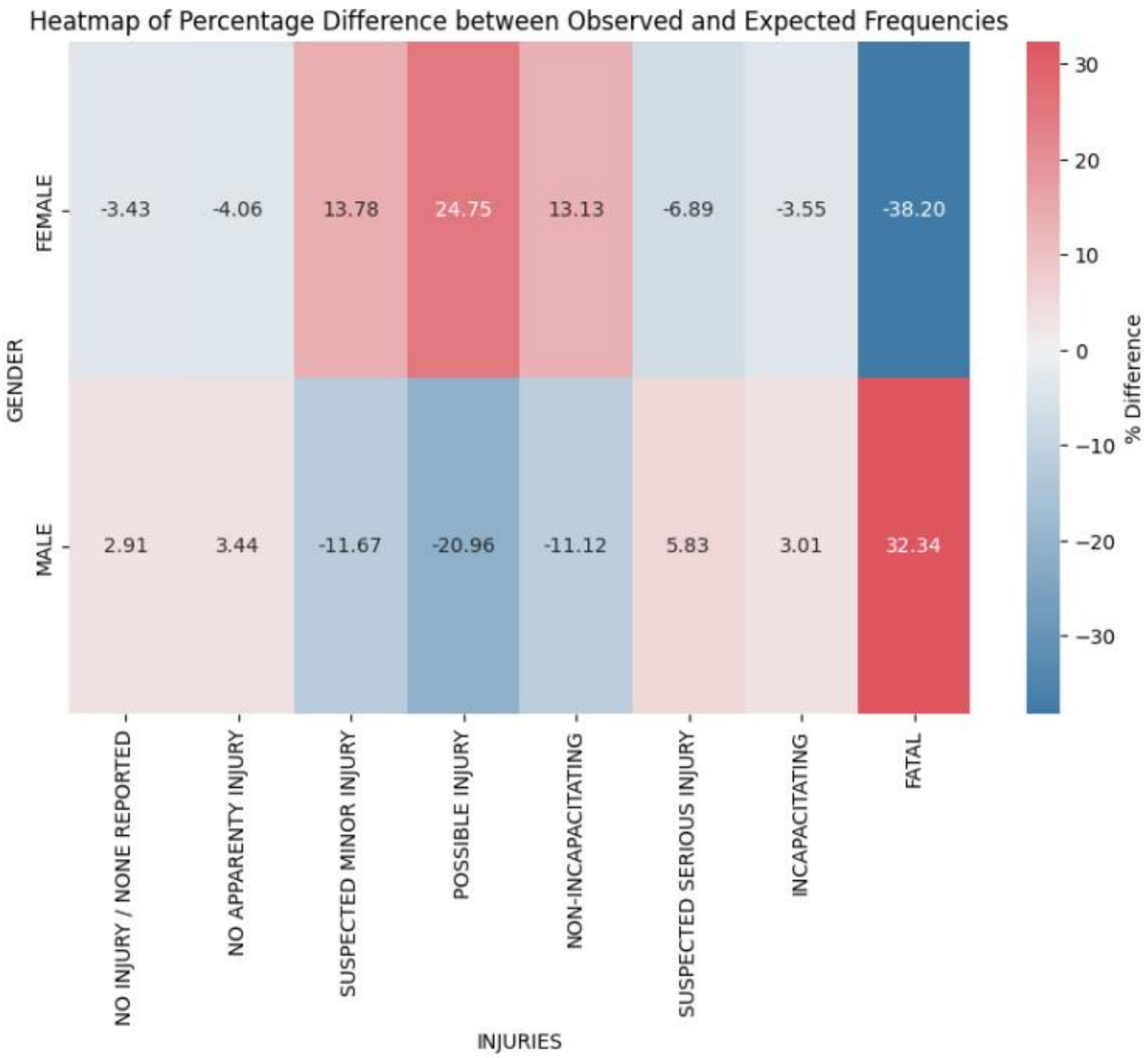
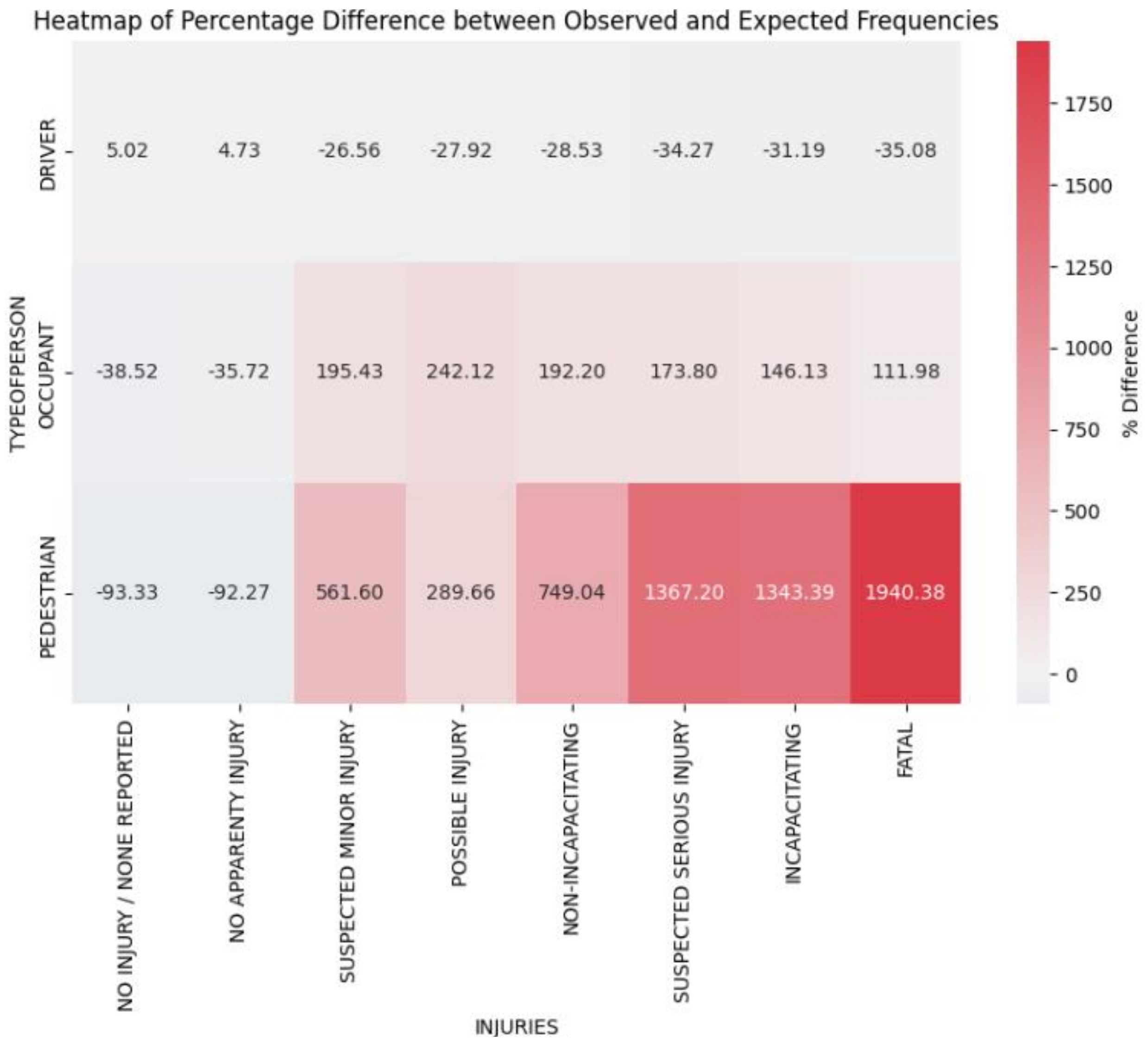


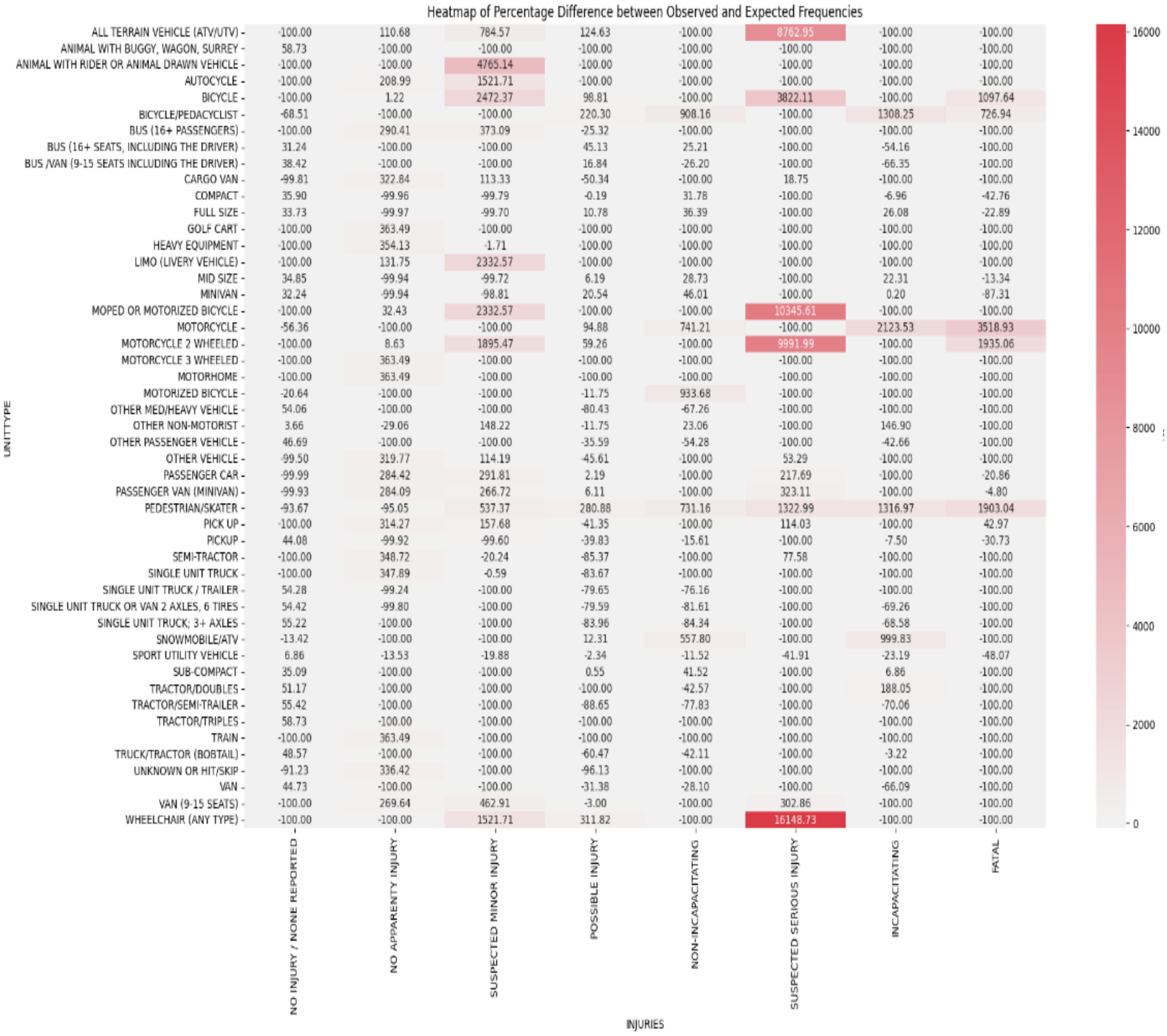
Table 19: Contingency table of GENDER and INJURIES

The data shows evidence of females suffering possible injuries more than expected (25% more) and fatal injuries less than expected (38% less), with the opposite being true for males (21% fewer possible injuries and 32% more fatal injuries) (*Table 19*). To be noted that this could be due to third factors.



<u>Table 20:</u> Contingency table of TYPEOFPERSON and INJURIES

The data shows that pedestrians often suffer much harder consequences than statistically expected from car crashes, with up to 2000% more fatal injuries, and 93% fewer "no injuries". Occupants also show a (lighter) tendency to suffer harder consequences from crashes (*<u>Table 20</u>*).



Heatmap of Percentage Difference between Observed and Expected Frequencies

UNITTYPE \ INJURIES	NO INJURY / NONE REPORTED	NO APPARENTY INJURY	SUSPECTED MINOR INJURY	POSSIBLE INJURY	NON-INCAPACITATING	SUSPECTED SERIOUS INJURY	INCAPACITATING	FATAL
ALL TERRAIN VEHICLE (ATV/UTV)	-100.00	110.68	784.57	124.63	-100.00	8762.95	-100.00	-100.00
ANIMAL WITH BUGGY, WAGON, SURREY	58.73	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
ANIMAL WITH RIDER OR ANIMAL DRAWN VEHICLE	-100.00	-100.00	4765.14	-100.00	-100.00	-100.00	-100.00	-100.00
AUTOCYCLE	-100.00	208.99	1521.71	-100.00	-100.00	-100.00	-100.00	-100.00
BICYCLE	-100.00	1.22	2472.37	98.81	-100.00	3822.11	-100.00	1097.64
BICYCLE/PEDACYCLIST	-68.51	-100.00	-100.00	220.30	908.16	-100.00	1308.25	726.94
BUS (16+ PASSENGERS)	-100.00	290.41	373.09	-25.32	-100.00	-100.00	-100.00	-100.00
BUS (16+ SEATS, INCLUDING THE DRIVER)	31.24	-100.00	-100.00	45.13	25.21	-100.00	-54.16	-100.00
BUS /VAN (9-15 SEATS INCLUDING THE DRIVER)	38.42	-100.00	-100.00	16.84	-26.20	-100.00	-66.35	-100.00
CARGO VAN	-99.81	322.84	113.33	-50.34	-100.00	18.75	-100.00	-100.00
COMPACT	35.90	-99.96	-99.79	-0.19	31.78	-100.00	-6.96	-42.76
FULL SIZE	33.73	-99.97	-99.70	10.78	36.39	-100.00	26.08	-22.89
GOLF CART	-100.00	363.49	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
HEAVY EQUIPMENT	-100.00	354.13	-1.71	-100.00	-100.00	-100.00	-100.00	-100.00
LIMO (LIVERY VEHICLE)	-100.00	131.75	2332.57	-100.00	-100.00	-100.00	-100.00	-100.00
MID SIZE	34.85	-99.94	-99.72	6.19	28.73	-100.00	22.31	-13.34
MINIVAN	32.24	-99.94	-98.81	20.54	46.01	-100.00	0.20	-87.31
MOPED OR MOTORIZED BICYCLE	-100.00	32.43	2332.57	-100.00	-100.00	10345.61	-100.00	-100.00
MOTORCYCLE	-56.36	-100.00	-100.00	94.88	741.21	-100.00	2123.53	3518.93
MOTORCYCLE 2 WHEELED	-100.00	8.63	1895.47	59.26	-100.00	9991.99	-100.00	1935.06
MOTORCYCLE 3 WHEELED	-100.00	363.49	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
MOTORHOME	-100.00	363.49	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
MOTORIZED BICYCLE	-20.64	-100.00	-100.00	-11.75	933.68	-100.00	-100.00	-100.00
OTHER MED/HEAVY VEHICLE	54.06	-100.00	-100.00	-80.43	-67.26	-100.00	-100.00	-100.00
OTHER NON-MOTORIST	3.66	-29.06	148.22	-11.75	23.06	-100.00	146.90	-100.00
OTHER PASSENGER VEHICLE	46.69	-100.00	-100.00	-35.59	-54.28	-100.00	-42.66	-100.00
OTHER VEHICLE	-99.50	319.77	114.19	-45.61	-100.00	53.29	-100.00	-100.00
PASSENGER CAR	-99.99	284.42	291.81	2.19	-100.00	217.69	-100.00	-20.86
PASSENGER VAN (MINIVAN)	-99.93	284.09	266.72	6.11	-100.00	323.11	-100.00	-4.80
PEDESTRIAN/SKATER	-93.67	-95.05	537.37	280.88	731.16	1322.99	1316.97	1903.04
PICK UP	-100.00	314.27	157.68	-41.35	-100.00	114.03	-100.00	42.97
PICKUP	44.08	-99.92	-99.60	-39.83	-15.61	-100.00	-7.50	-30.73
SEMI-TRACTOR	-100.00	348.72	-20.24	-85.37	-100.00	77.58	-100.00	-100.00
SINGLE UNIT TRUCK	-100.00	347.89	-0.59	-83.67	-100.00	-100.00	-100.00	-100.00
SINGLE UNIT TRUCK / TRAILER	54.28	-99.24	-100.00	-79.65	-76.16	-100.00	-100.00	-100.00
SINGLE UNIT TRUCK OR VAN 2 AXLES, 6 TIRES	54.42	-99.80	-100.00	-79.59	-81.61	-100.00	-69.26	-100.00
SINGLE UNIT TRUCK; 3+ AXLES	55.22	-100.00	-100.00	-83.96	-84.34	-100.00	-68.58	-100.00
SNOWMOBILE/ATV	-13.42	-100.00	-100.00	12.31	557.80	-100.00	999.83	-100.00
SPORT UTILITY VEHICLE	6.86	-13.53	-19.88	-2.34	-11.52	-41.91	-23.19	-48.07
SUB-COMPACT	35.09	-100.00	-100.00	0.55	41.52	-100.00	6.86	-100.00
TRACTOR/DOUBLES	51.17	-100.00	-100.00	-100.00	-42.57	-100.00	188.05	-100.00
TRACTOR/SEMI-TRAILER	55.42	-100.00	-100.00	-88.65	-77.83	-100.00	-70.06	-100.00
TRACTOR/TRIPLES	58.73	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
TRAIN	-100.00	363.49	-100.00	-100.00	-100.00	-100.00	-100.00	-100.00
TRUCK/TRACTOR (BOBTAIL)	48.57	-100.00	-100.00	-60.47	-42.11	-100.00	-3.22	-100.00
UNKNOWN OR HIT/SKIP	-91.23	336.42	-100.00	-96.13	-100.00	-100.00	-100.00	-100.00
VAN	44.73	-100.00	-100.00	-31.38	-28.10	-100.00	-66.09	-100.00
VAN (9-15 SEATS)	-100.00	269.64	462.91	-3.00	-100.00	302.86	-100.00	-100.00
WHEELCHAIR (ANY TYPE)	-100.00	-100.00	1521.71	311.82	-100.00	16148.73	-100.00	-100.00

Table 21: Contingency table of UNITTYPE and INJURIES

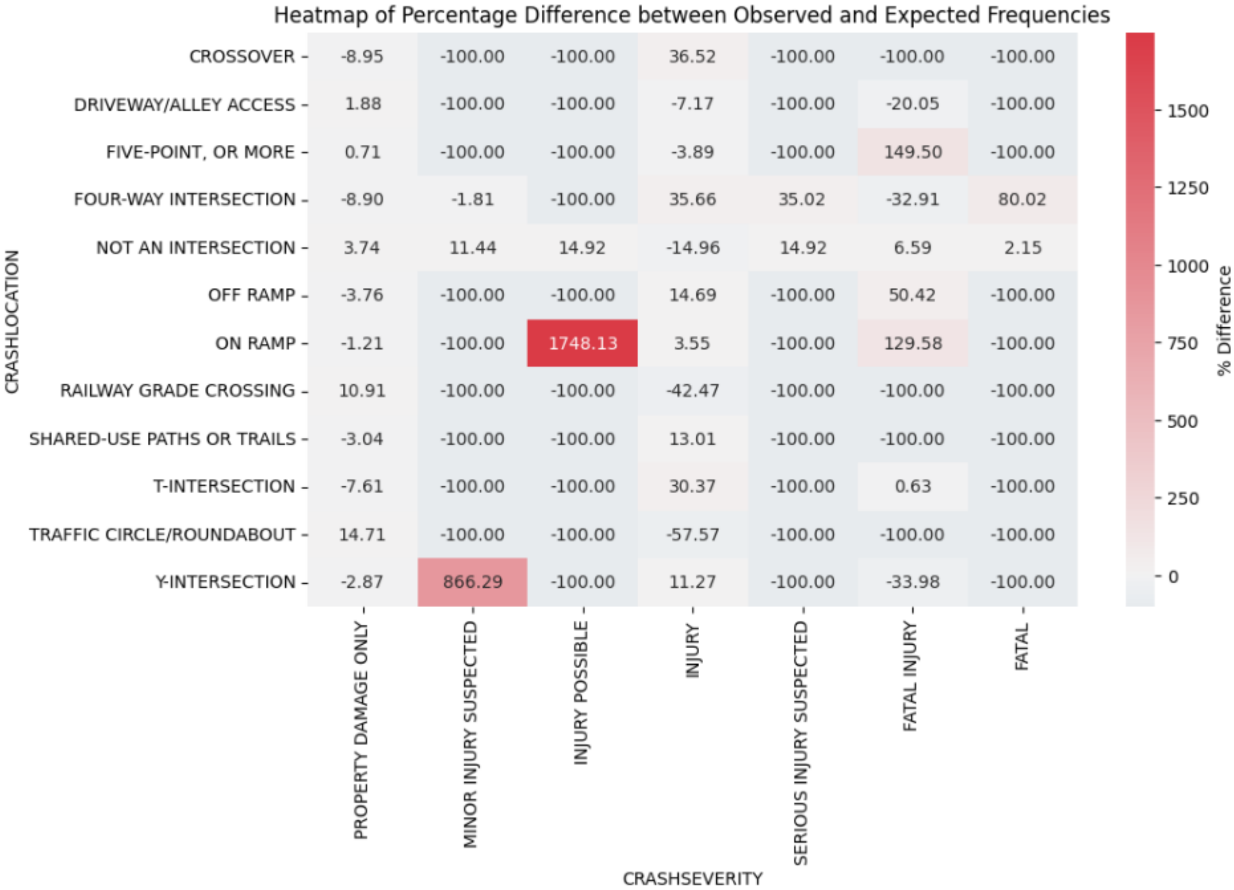
Wheelchairs and two-wheeled vehicles suffer more "suspected serious injuries" than expected (*Table 21*).

CRASHES ANALYSIS:

This section focuses on providing insights about the crashes in Cincinnati.

CONTINGENCY TABLES:

Again, the presented contingency tables are computed as the difference, in percentual, between the contingency tables of the expected frequencies and the actual contingency tables, therefore suggesting how much each pair of occurrences diverges from the case of complete independence between the two variables at issue. In the computation of these tables, rows that present UNKNOWN values in either of the two variables have been removed.



CRASHLOCATION	PROPERTY DAMAGE ONLY	MINOR INJURY SUSPECTED	INJURY POSSIBLE	INJURY	SERIOUS INJURY SUSPECTED	FATAL INJURY	FATAL
CROSSOVER	-8.95	-100.00	-100.00	36.52	-100.00	-100.00	-100.00
DRIVEWAY/ALLEY ACCESS	1.88	-100.00	-100.00	-7.17	-100.00	-20.05	-100.00
FIVE-POINT, OR MORE	0.71	-100.00	-100.00	-3.89	-100.00	149.50	-100.00
FOUR-WAY INTERSECTION	-8.90	-1.81	-100.00	35.66	35.02	-32.91	80.02
NOT AN INTERSECTION	3.74	11.44	14.92	-14.96	14.92	6.59	2.15
OFF RAMP	-3.76	-100.00	-100.00	14.69	-100.00	50.42	-100.00
ON RAMP	-1.21	-100.00	1748.13	3.55	-100.00	129.58	-100.00
RAILWAY GRADE CROSSING	10.91	-100.00	-100.00	-42.47	-100.00	-100.00	-100.00
SHARED-USE PATHS OR TRAILS	-3.04	-100.00	-100.00	13.01	-100.00	-100.00	-100.00
T-INTERSECTION	-7.61	-100.00	-100.00	30.37	-100.00	0.63	-100.00
TRAFFIC CIRCLE/ROUNDABOUT	14.71	-100.00	-100.00	-57.57	-100.00	-100.00	-100.00
Y-INTERSECTION	-2.87	866.29	-100.00	11.27	-100.00	-33.98	-100.00

Table 22: Contingency table of CRASHLOCATION and CRASHSEVERITY

On-ramp crashes show a much higher tendency of having crash severity as "injury possible" (1748% more), being the only location together with "not an intersection" to have this modality. Y intersection results much more often than expected in "minor injury suspected" (866% more than expected) (*Table 22*).

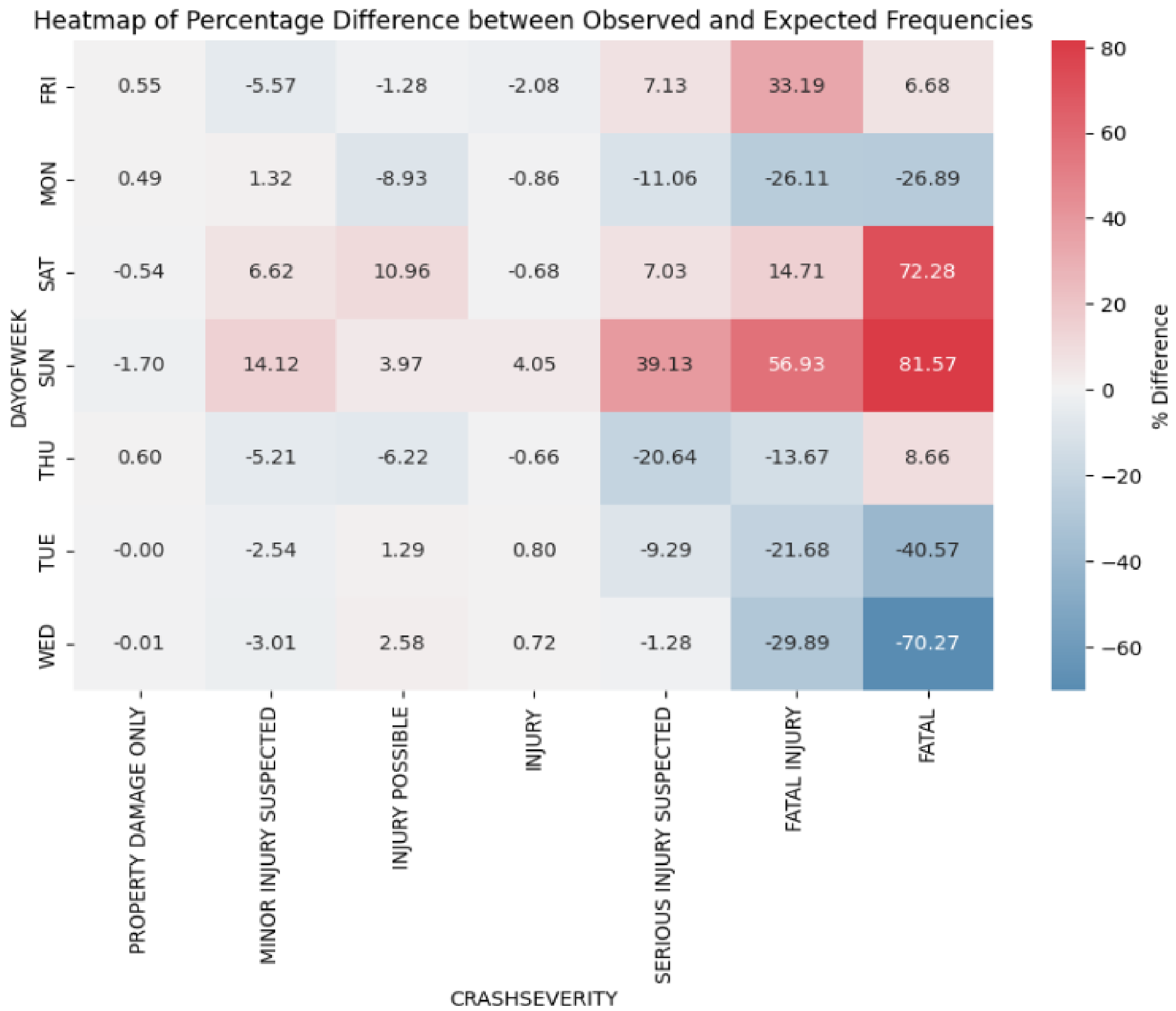
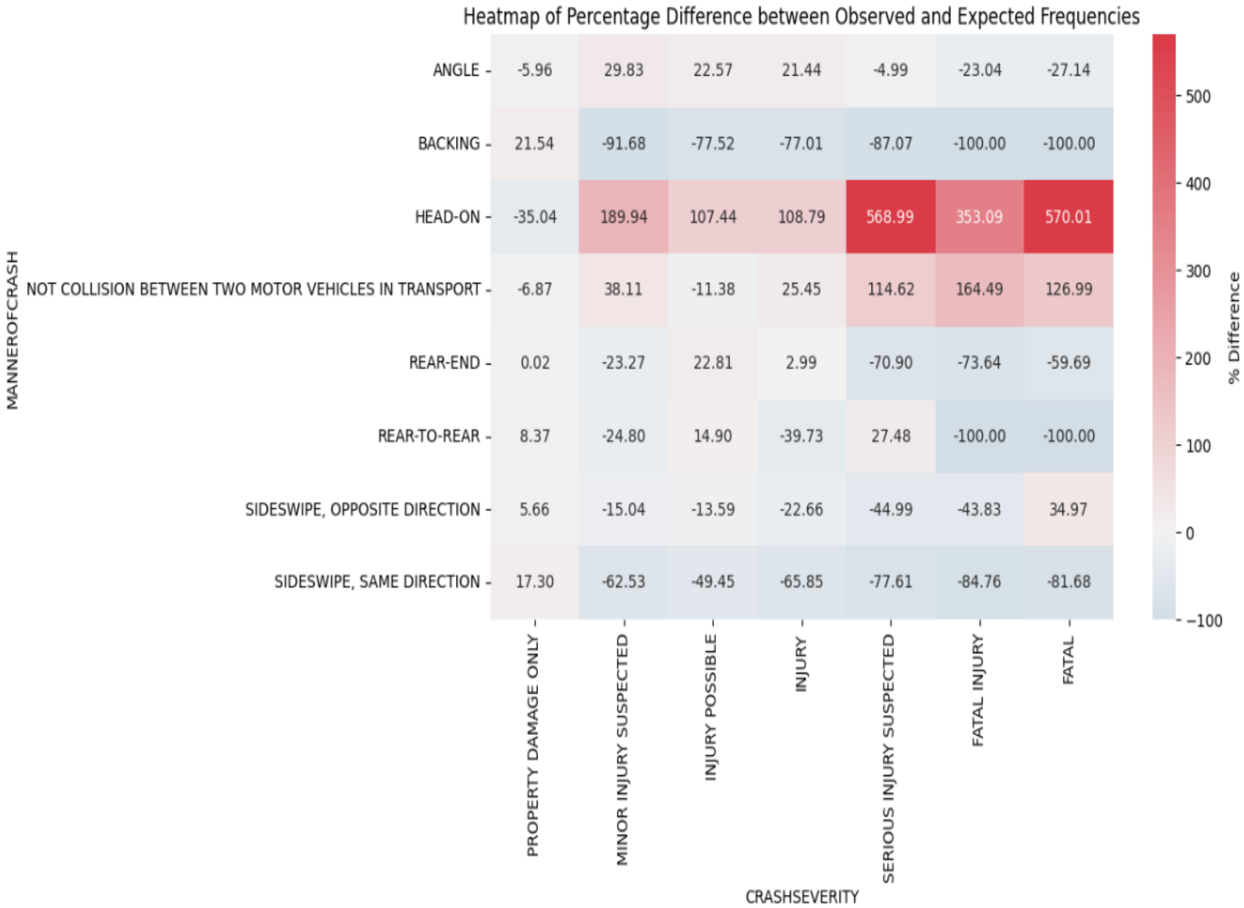


Table 23: Contingency table of DAYOFWEEK and CRASHSEVERITY

Fatal events are much more common on a Saturday (72% more) and Sunday (81% more) and less common on a Wednesday (70% less) and on a Tuesday (40% less). Fatal injuries follow a similar trend (*Table 23*).

Table 24: Contingency table of LIGHTCONDITIONSPRIMARY and CRASHSEVERITY

Dark conditions of the road generally result in more severe crashes than expected (*Table 24*).



<u>Table 25:</u> Contingency table of MANNEROFCRASH and CRASHSEVERITY

Head-on crashes show much higher counts of injuries, and especially severe injuries (570% more fatal events than expected, 569% more serious injuries suspected than expected, 353% more fatal injuries than expected) and fewer "property damage only" crashes than expected (35% less). The modality "Not collision between two motor vehicles in transport" follows a similar, less pronounced, trend (164% more fatal injuries than expected) (*<u>Table 25</u>*).

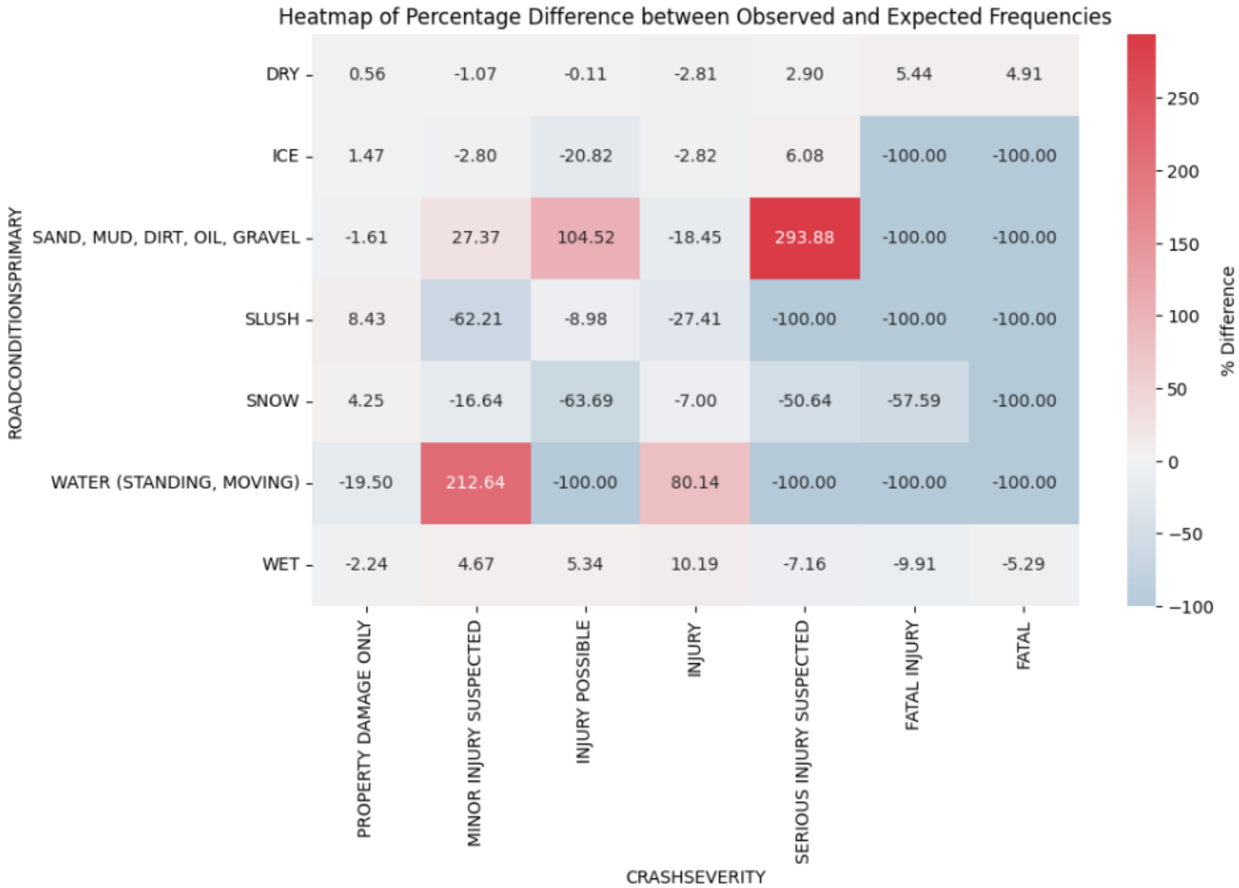


Table 26: Contingency table of ROADCONDITIONSPRIMARY and CRASHSEVERITY

Roads covered in water result in a higher number of minor injuries suspected (212% more than expected) but do not show any serious injury suspected nor fatal occurrences. Sand, mud, dirt, oil, and gravel roads are slightly more dangerous, resulting in 293% more serious injuries suspected than expected, but still no fatal occurrences (*Table 26*).

Table 27: Contingency table of ROADCONTOUR and CRASHSEVERITY

Curved road contours are much more dangerous than straight ones, with both curve grade and curve level road contours showing more serious injuries suspected (105% and 114% more, respectively), fatal injuries (226% and 165% more, respectively), and fatal (214% and 294% more, respectively) than expected (*Table 27*).

<u>Table 28:</u> Contingency table of ROADSURFACE and CRASHSEVERITY

Brick, block, dirt, slag, gravel, and stone are surfaces that do not show the highest levels of injuries (*<u>Table 28</u>*).

<u>Table 29:</u> Contingency table of WEATHER and CRASHSEVERITY

Severe crosswinds result in fewer property damage only situations (11% less than expected) and in way more minor injuries suspected (535% more than expected), with no serious injury suspected, fatal injuries, and fatal events. Fog, smog, and smoke are the most dangerous conditions, with 331% more fatal injuries than expected. Snow conditions are also dangerous, with 42% more fatal events than expected (*<u>Table 29</u>*).

Heatmap of Percentage Difference between Observed and Expected Frequencies

SNA_NEIGHBORHOOD \ CRASHSEVERITY	PROPERTY DAMAGE ONLY	MINOR INJURY SUSPECTED	INJURY POSSIBLE	INJURY	SERIOUS INJURY SUSPECTED	FATAL INJURY	FATAL
AVONDALE	-1.75	-10.45	8.88	9.54	8.44	12.06	-4.47
BOND HILL	-1.46	3.44	-3.54	7.67	-18.78	43.00	-8.57
CALIFORNIA	-3.47	-23.14	0.06	16.76	8.31	472.12	631.57
CAMP WASHINGTON	-3.38	25.08	21.68	9.27	22.04	-10.47	-54.21
CARTHAGE	-1.02	23.70	19.23	-3.20	37.81	21.32	3.42
CLIFTON	-1.01	5.18	14.15	2.31	5.49	-25.71	-52.50
COLLEGE HILL	-4.26	10.87	29.23	15.21	17.88	19.74	-38.76
COLUMBIA TUSCULUM	1.99	-52.40	-52.05	7.99	-27.33	91.91	-100.00
CORRYVILLE	3.29	-12.08	-15.66	-10.39	-64.98	-76.88	-40.87
CUF	6.42	-25.57	-26.85	-24.33	-16.61	-9.32	65.65
DOWNTOWN	4.25	-15.49	-28.68	-13.36	-54.67	-26.33	-29.35
EAST END	1.32	-0.12	-51.70	1.77	-47.72	38.07	76.55
EAST PRICE HILL	-1.89	19.99	7.44	5.29	-6.95	-7.85	-60.72
EAST WALNUT HILLS	2.29	-7.48	-52.68	-3.12	30.38	72.17	-100.00
EAST WESTWOOD	-4.87	-0.47	50.39	16.89	13.96	100.65	-100.00
ENGLISH WOODS	-14.34	89.52	71.82	31.72	334.00	473.10	1365.64
EVANSTON	1.45	-2.87	-16.62	-4.45	54.74	-100.00	-100.00
HARTWELL	-1.02	6.71	35.91	-0.98	-12.72	-100.00	135.79
HYDE PARK	6.82	-33.86	-35.73	-23.41	-16.85	25.49	-19.77
KENNEDY HEIGHTS	1.24	-21.04	-33.18	4.47	-100.00	91.03	-100.00
LINWOOD	0.14	22.07	-32.37	0.45	-53.41	23.05	214.68
LOWER PRICE HILL	-2.67	-13.44	13.74	11.85	106.85	82.10	-100.00
MADISONVILLE	5.80	-43.91	-16.06	-17.55	-63.30	-35.39	-100.00
MILLVALE	-8.75	79.15	68.44	13.55	310.27	20.39	-100.00
MT. ADAMS	3.99	-52.94	-26.86	-3.97	-100.00	-18.69	-100.00
MT. AIRY	-7.78	41.08	51.59	22.16	105.13	62.52	-30.73
MT. AUBURN	3.33	-9.54	-20.67	-10.90	-51.26	-14.18	-100.00
MT. LOOKOUT	11.44	-46.05	-49.28	-43.68	-45.09	190.03	-100.00
MT. WASHINGTON	5.20	-28.85	-36.43	-15.77	13.35	-25.16	-100.00
NORTH AVONDALE - PADDOCK HILLS	-1.88	21.05	16.14	1.32	111.21	-7.03	-100.00
NORTH FAIRMOUNT	-6.33	22.12	31.81	22.10	-60.05	427.57	-100.00
NORTHSIDE	-3.00	11.88	-15.79	15.35	25.68	-23.40	95.90
OAKLEY	5.59	-25.38	-29.96	-19.38	-3.51	-15.05	-45.69
OVER-THE-RHINE	4.83	-18.28	-29.78	-16.37	-9.81	-76.18	21.82
PENDLETON	4.23	2.05	-20.70	-16.84	-66.62	-100.00	-100.00
PLEASANT RIDGE	6.48	-22.03	-24.08	-27.75	38.87	57.18	167.98
QUEENSGATE	-0.76	9.48	17.49	-0.61	2.33	-18.92	38.23
RIVERSIDE	-11.37	65.45	23.81	38.98	80.42	281.20	143.72
ROSELAWN	-2.06	5.54	-0.33	8.35	40.99	32.98	104.05
SAYLER PARK	-5.82	13.91	25.18	23.10	-5.14	150.52	-100.00
SEDAMSVILLE	-6.02	107.98	2.43	4.87	323.37	-100.00	-100.00
SOUTH CUMMINSVILLE	-7.16	67.58	35.05	16.03	70.56	237.85	-100.00
SOUTH FAIRMOUNT	-3.50	9.23	18.23	14.71	-18.56	-100.00	57.17
SPRING GROVE VILLAGE	-2.27	37.89	19.06	0.40	5.26	-20.58	407.79
VILLAGES AT ROLL HILL	-4.39	38.70	51.45	6.38	63.94	8.24	-100.00
WALNUT HILLS	-0.59	-10.61	-15.64	7.95	-13.81	-5.15	-2.98
WEST END	-0.70	12.91	-11.50	2.83	-11.12	6.70	118.29
WEST PRICE HILL	0.59	-24.68	10.62	0.86	-3.06	-84.94	-61.49
WESTWOOD	-2.53	6.78	30.71	7.59	-20.03	-10.64	3.88
WINTON HILLS	-7.03	61.82	22.68	17.80	75.87	298.13	-100.00

Table 30: Contingency table of SNA_NEIGHBORHOOD and CRASHSEVERITY

English Woods seems to be the neighborhood with the highest values of severe injuries compared to what was expected, followed by the California one (*Table 30*).

PREDICTIONS:

In this section it is attempted to develop models to predict with some degree of effectiveness the gravity of a crash given the conditions in which it occurs, to get valuable insights to prevent the most critical situations.
Before developing the model, the crashed dataset was filtered of all crashes having at least one UNKNOWN value, other than having the CRASHDATE variable removed.
Because of the significant reduction in information, some crash severity classes have been merged to have more populated and different classes, the following:

- **Property damage only:** includes Property damage only
- **Injury:** includes Injury, Minor injury suspected, Injury possible, Serious injury suspected
- **Fatal injury:** includes Fatal injury, Fatal

Because of the high imbalance in class populations, the chosen valuation metric is the f1 score, computed as the harmonic mean of precision and recall, where precision is the ratio of true positives to the sum of true and false positives, and recall is the ratio of true positives to the sum of true positives and false negatives, attenuating misleadingly high scores from one dominant class.

K nearest neighbors:

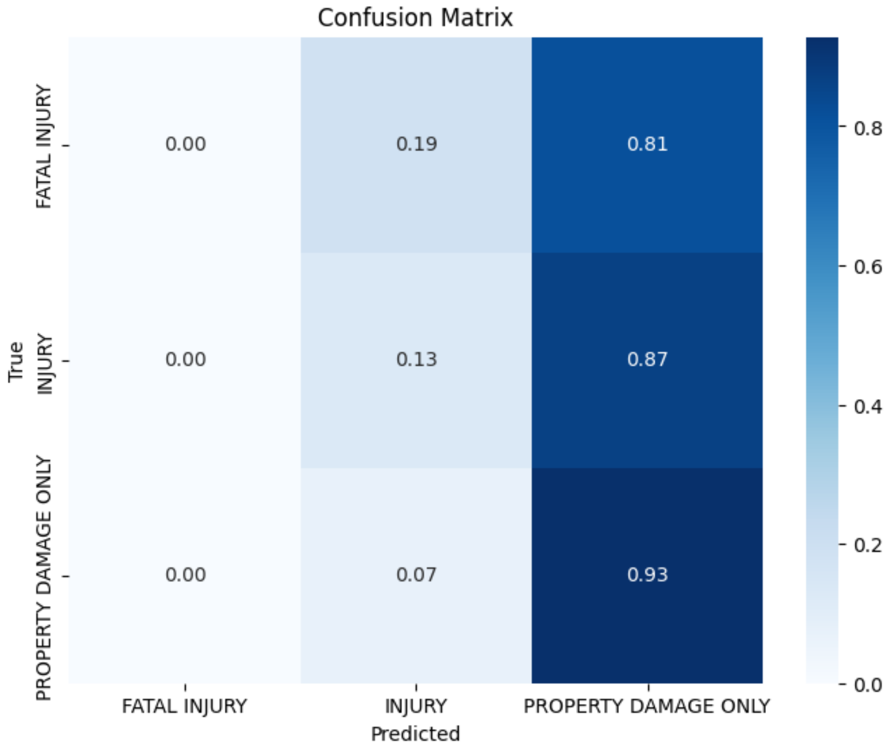


Table 31: Confusion matrix of the knn model's performance; f1 score: 0.7198

The reason behind the good score of this model is that it tends to always predict property damage only, which is statistically good but not relevant for the purpose of this work, especially given that the higher interest of the analysis is posed on more severe crashes (*Table 31*).

Gradient boosting algorithm:

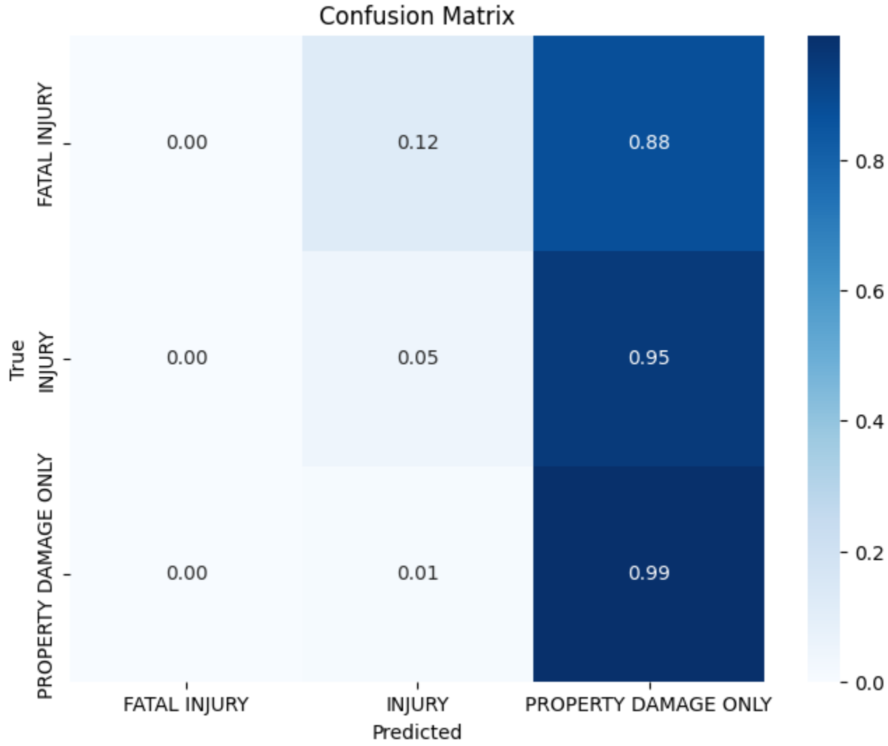


Table 32: Confusion matrix of the gba model's performance; f1 score: 0.7180

The Gradient boosting algorithm shows the same problem of "always predict property damage only" as knn, here even more accentuated (*Table 32*).

Random forest:

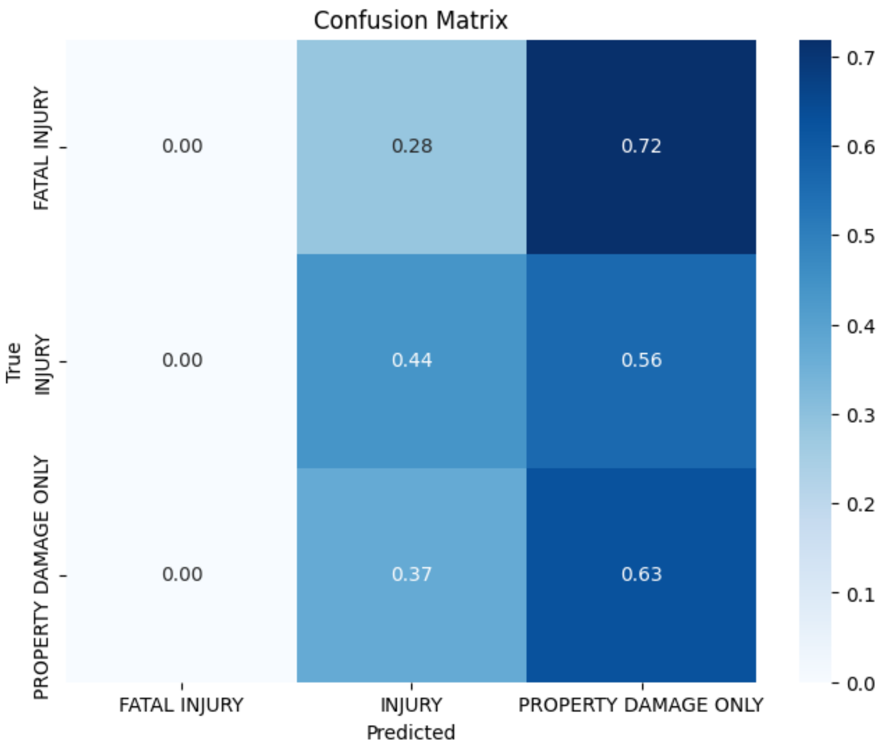


Table 33: Confusion matrix of the random forest model's performance; f1 score: 0.6218

The random forest model has the option to assign a weight to the classes to impact the score of their performance, and this allows to partially overcome the previously presented issue. In this case, this allowed to predict more accurately the injury class, which is the second most common after property damage only, but still the model prefers to never predict fatal injury, as the reward it gets is just too little, even in case of higher class weight. I still value this model's performance compared to the knn one and the GBA one, regardless of a lower f1 score (*Table 33*).

Class weights: fatal injury:15, injury: 1, property damage only: 0.001

Logistic regression:

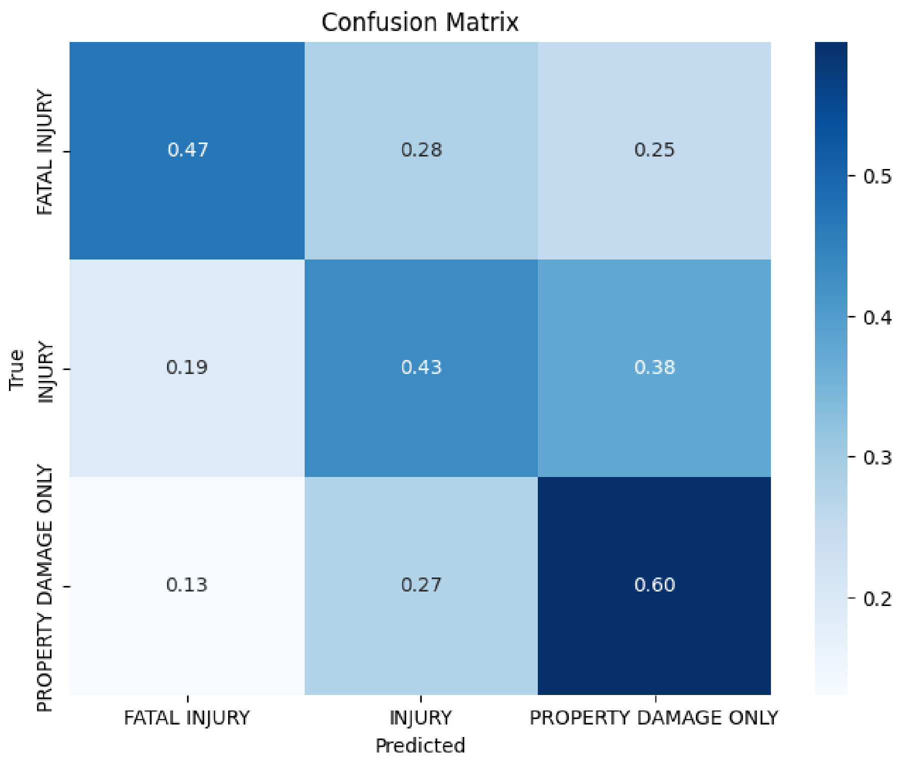


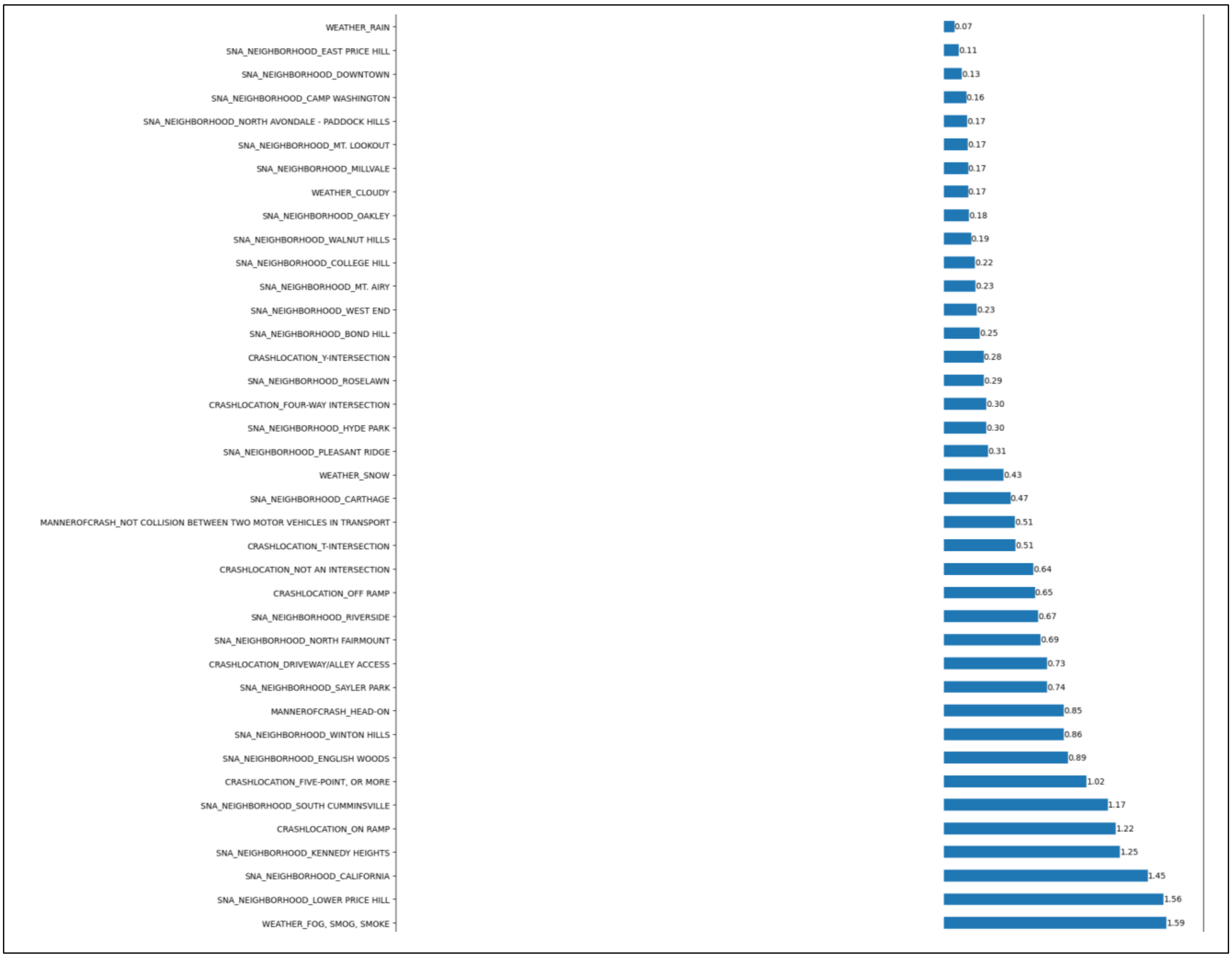
Table 34: Confusion matrix of the logistic regression model's performance; f1 score: 0.6274

Logistic regression allows, just like random forests, to have the classes weighted. Even with an f1 score of 0.62, this is the best developed model of the analysis, managing to capture in the best way among the ones attempted, the outcome of fatal injuries, other than injuries and property damage only (*Table 34*).

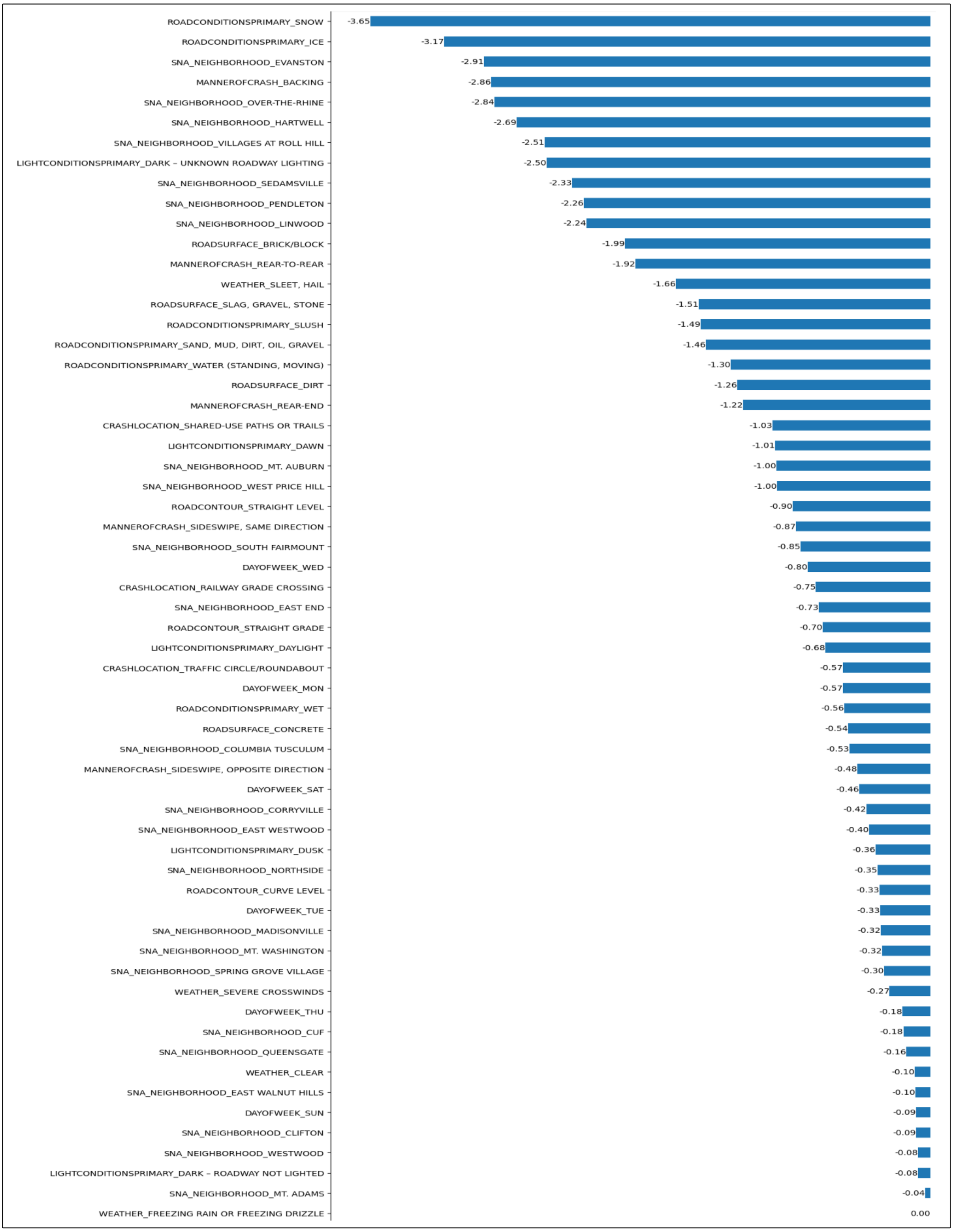
Class weights: fatal injury: 100, injury: 1, property damage only: 0.3

Features importance:

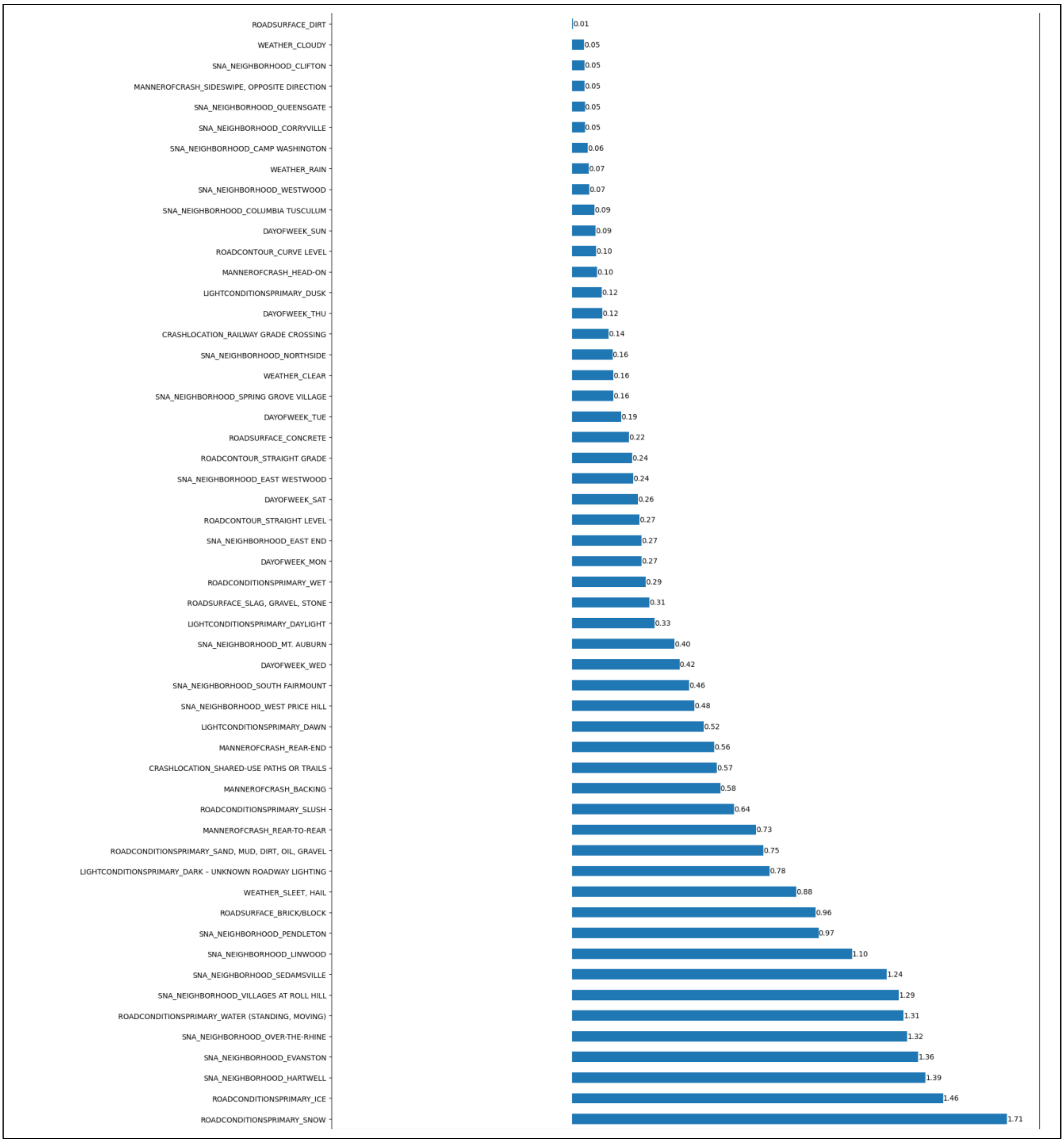
This analysis of the features refers to the chosen model, the logistic regression. It allows to have a look inside the model, to understand what is it that it uses to predict each class, to propend towards and away from it (the higher the absolute value of the coefficient higher the weight of the variable).



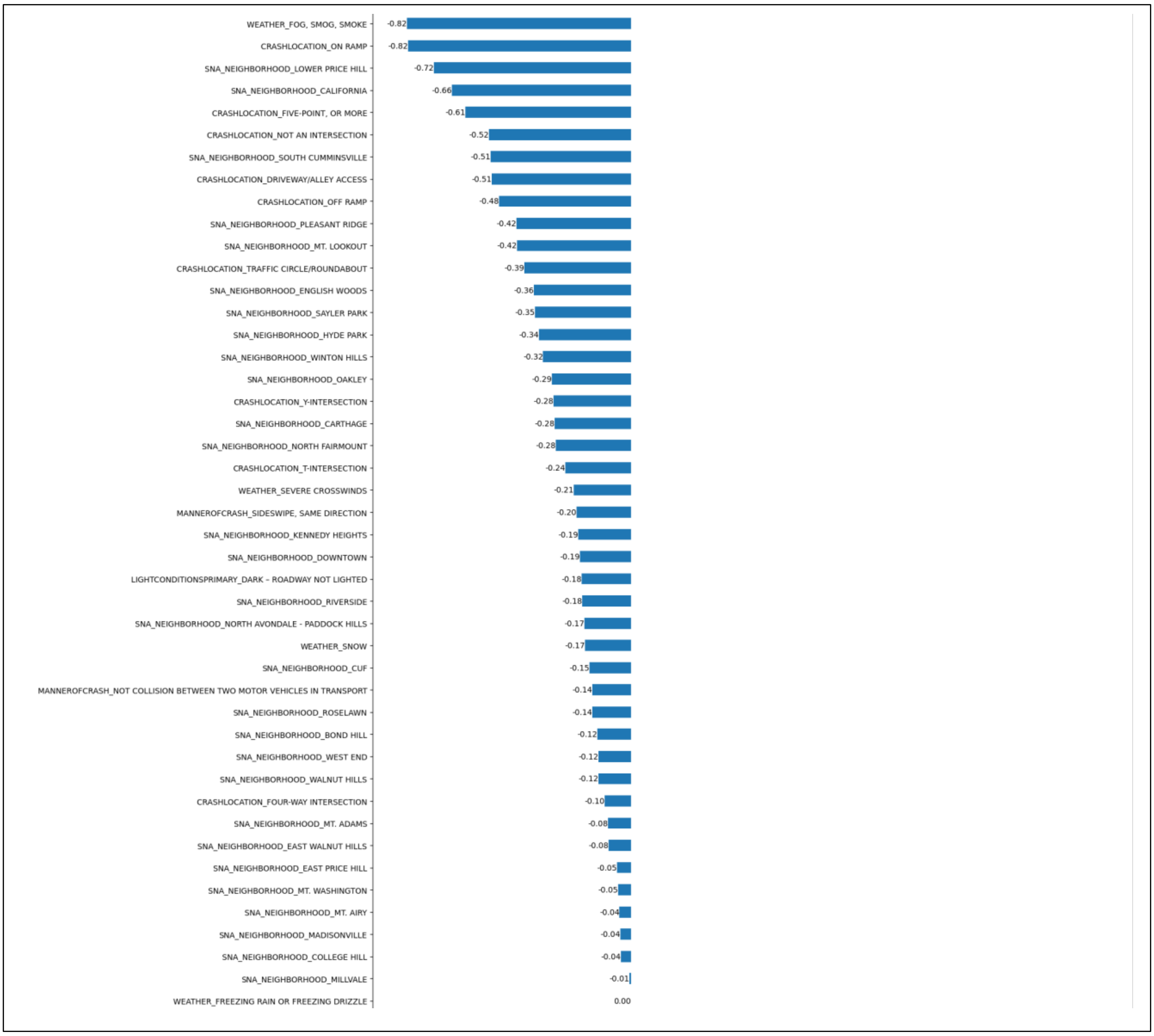
__Plot 18:__ bar chart of features in favor of the prediction of "fatal injury"



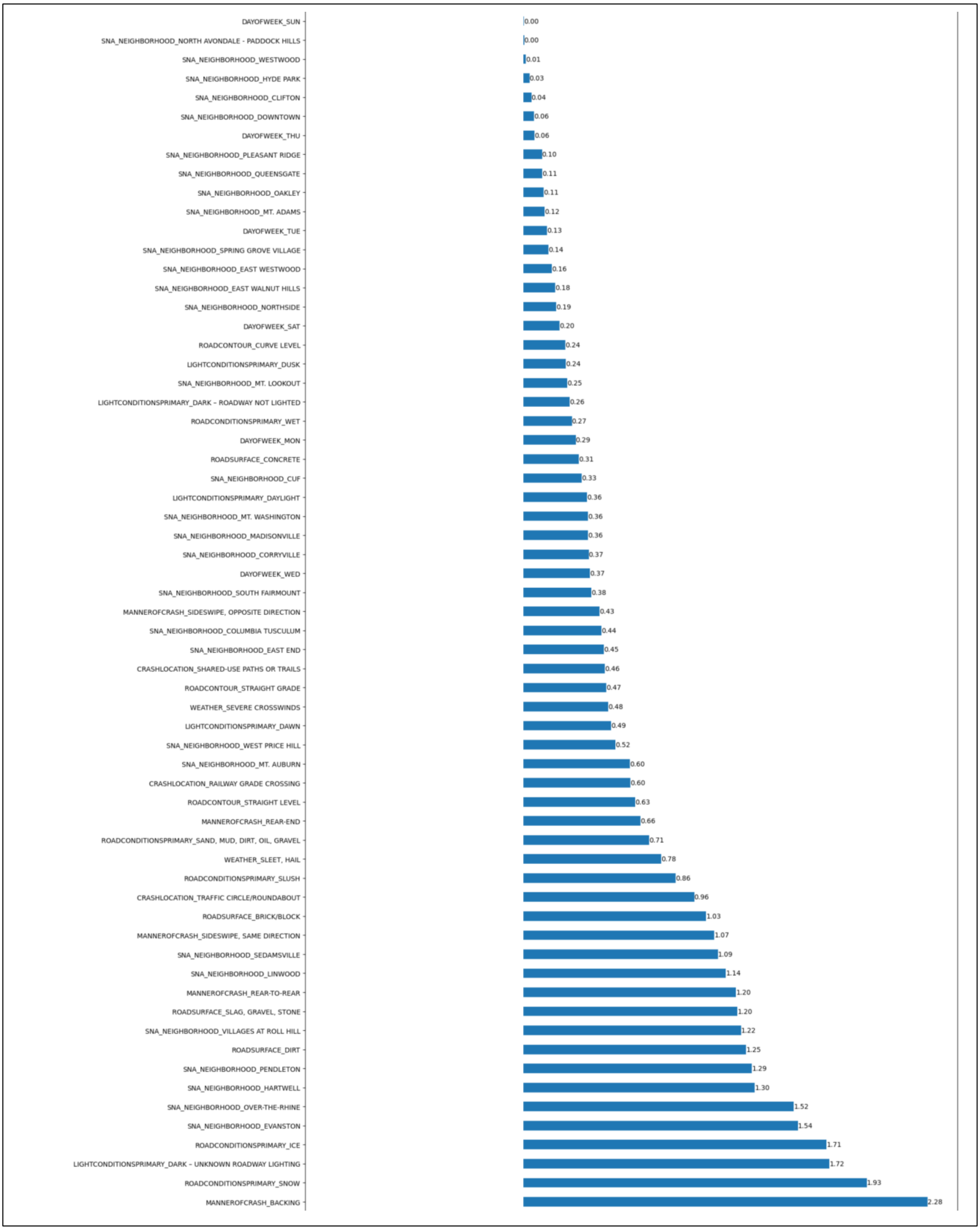
<u>Plot 19:</u> bar chart of features against the prediction of “fatal injury”



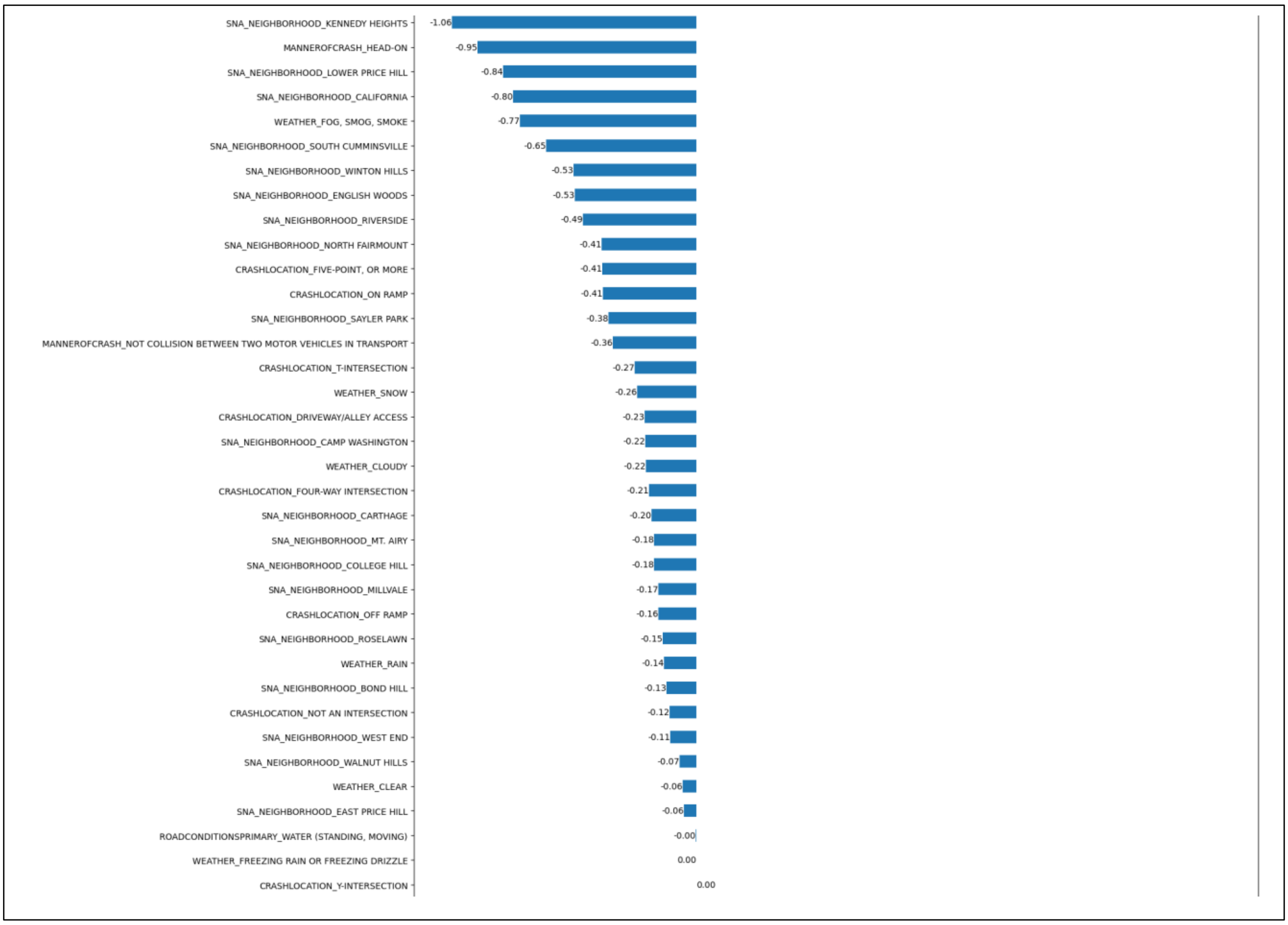
Plot 20: bar chart of features in favor of the prediction of “injury”



<u>Plot 21:</u> bar chart of features against the prediction of “injury”



<u>Plot 22:</u> bar chart of features in favor of the prediction of “property damage only”



<u>Plot 23:</u> bar chart of features against the prediction of “property damage only”

Conclusions:

There are not many strong indicators of the fact that a crash will result in either property damage only or in an injury. Strong indicators of the fact that a car might not result in a fatality are many neighborhoods, including Evanston, Over the Rhine, Hartwell, Pendleton, Villages at Roll Hill, Linwood, and Sedamsville, (that are also indicators of crashes resulting in property damage only or injuries) but also Mt auburn, and West price hill.
Other modalities that go against the prediction of a crash being fatal are unknown roadway lightning; backing or rear to rear manner of crash; snow, ice, and water conditions of the road; brick/block, slag/gravel/stone, or dirty surface of the road; all these are also features that make the model heavily propend towards prediction of non fatal crashes; and then there are also: dawn light of the street; rear end manner of crash, sandy/muddy/oily/gravel road surface; and shared used path or trails as crash location.
Elements to which it must be paid close attention, as they are strong indicators of fatal crashes are: on ramp and five point or more crash locations; Lower Price Hill, California, Kennedy Heights, and South Cumminsville neighborhoods: and the most impactful feature, which is fog/smoke/smog in the weather. weather conditions, namely snow and ice, are also the most important feature to predict crashes resulting in injuries, according to the model (*Plot 18*, *Plot 19*, *Plot 20*, *Plot 21*, *Plot 22*, *Plot 23*).

In conclusion, it would be advised to take action in the following neighborhoods to lower the number of fatal crashes: Lower Price Hill, California, Kennedy Heights, and South Cumminsville; for example, lowering the speed limit and imposing stricter traffic rules near ramps and in the proximity of five points (or more). It is advised to the drivers to avoid as much as possible to drive in case of altered conditions of the road surface (water, snow, or ice) and especially in case of adverse weather, including fog, smoke or smog in the air.
People under 18 and over 61 should pay extra attention while inside vehicles, making sure to adopt the mandatory and also the suggested safety measures always, as they have been found to be the most fragile people.
Increasing controls on weekends is also suggested, as these days tend to be the ones in which the worst crashes happen.