Predicting Fatal Crashes in Washington, DC

Courtney Claessens

General Assembly, Data Science, March 2017

Outline

- Vision Zero is a worldwide initiative with a goal to get all traffic related deaths down to zero
- Washington, DC wants to eliminate all traffic deaths by 2024
- Understanding what kinds of crashes result in fatalities is critical information to achieve this goal

VISION ZERØ

SAFE STREETS FOR WASHINGTON, DC

Summary

Goal: determine which factors lead to fatal crashes

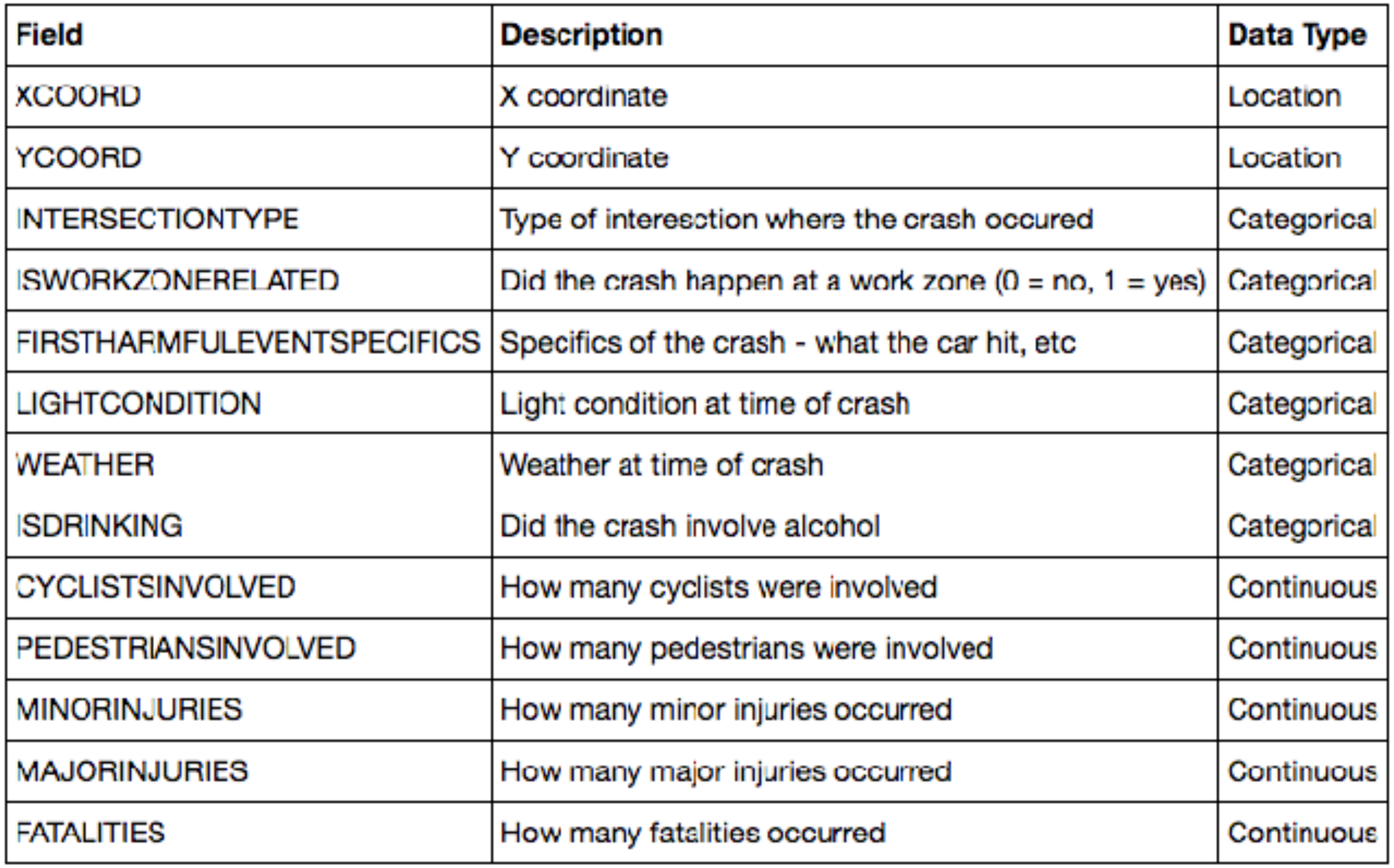
Steps:

1. Data acquisition
2. Data parsing and exploratory analysis
3. Data mining - creating dummy variables
4. Refine data - eliminating unnecessary fields
5. Build logistic regression model
6. Results

Summary

- Dataset: 152,744 crashes spanning from 2000 to 2016 with 63 fields, from DC's open data site

- This study used 13 fields



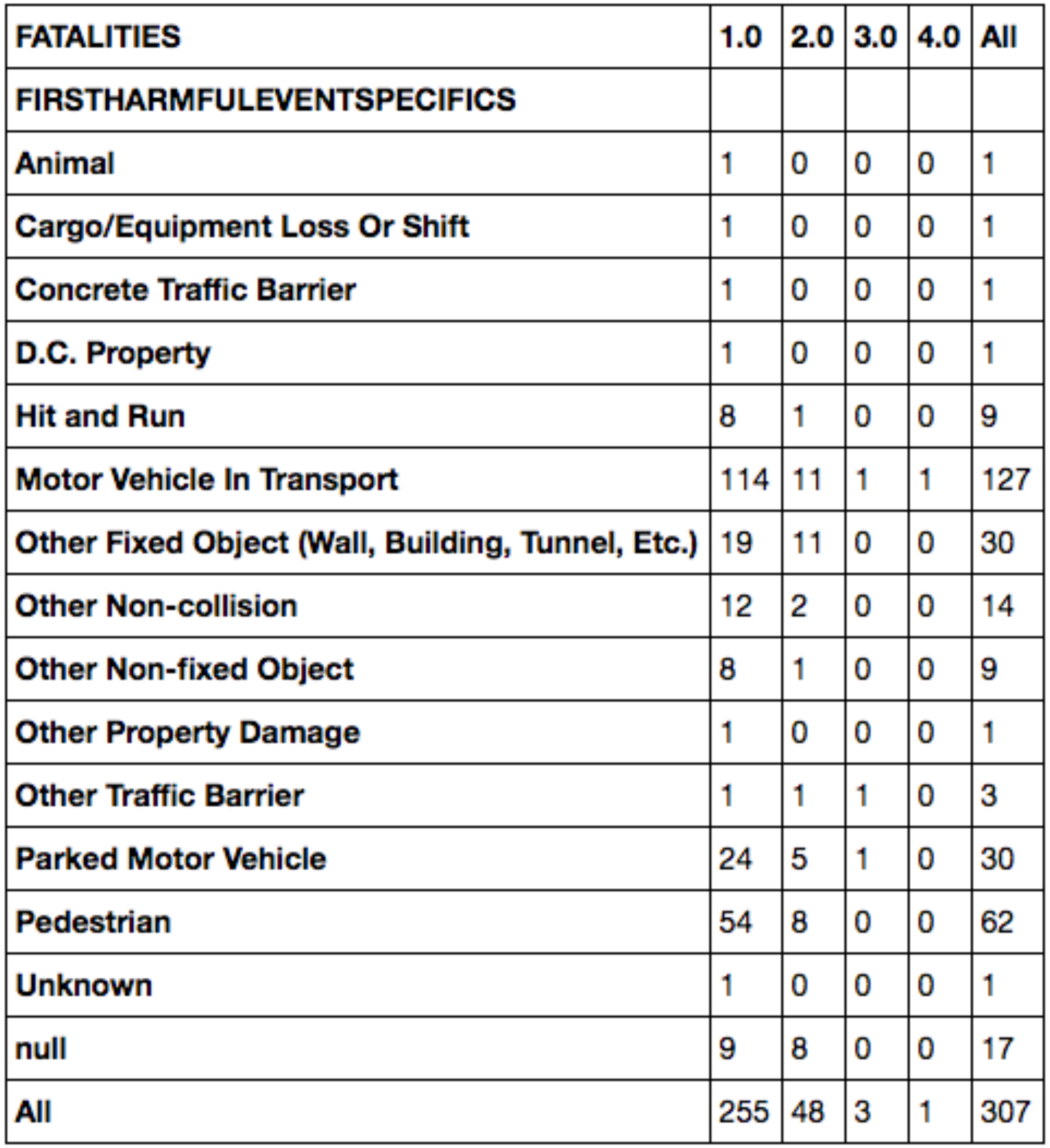
Field	Description	Data Type
XCOORD	X coordinate	Location
YCOORD	Y coordinate	Location
INTERSECTIONTYPE	Type of interesction where the crash occured	Categorical
ISWORKZONERELATED	Did the crash happen at a work zone (0 = no, 1 = yes)	Categorical
FIRSTHARMFULEVENTSPECIFICS	Specifics of the crash - what the car hit, etc	Categorical
LIGHTCONDITION	Light condition at time of crash	Categorical
WEATHER ISDRINKING	Weather at time of crash Did the crash involve alcohol	Categorical Categorical
CYCLISTSINVOLVED	How many cyclists were involved	Continuous
PEDESTRIANSINVOLVED	How many pedestrians were involved	Continuous
MINORINJURIES	How many minor injuries occurred	Continuous
MAJORINJURIES	How many major injuries occurred	Continuous
FATALITIES	How many fatalities occurred	Continuous

Summary

Out of 152,744 crashes, **only 307 were fatal**.

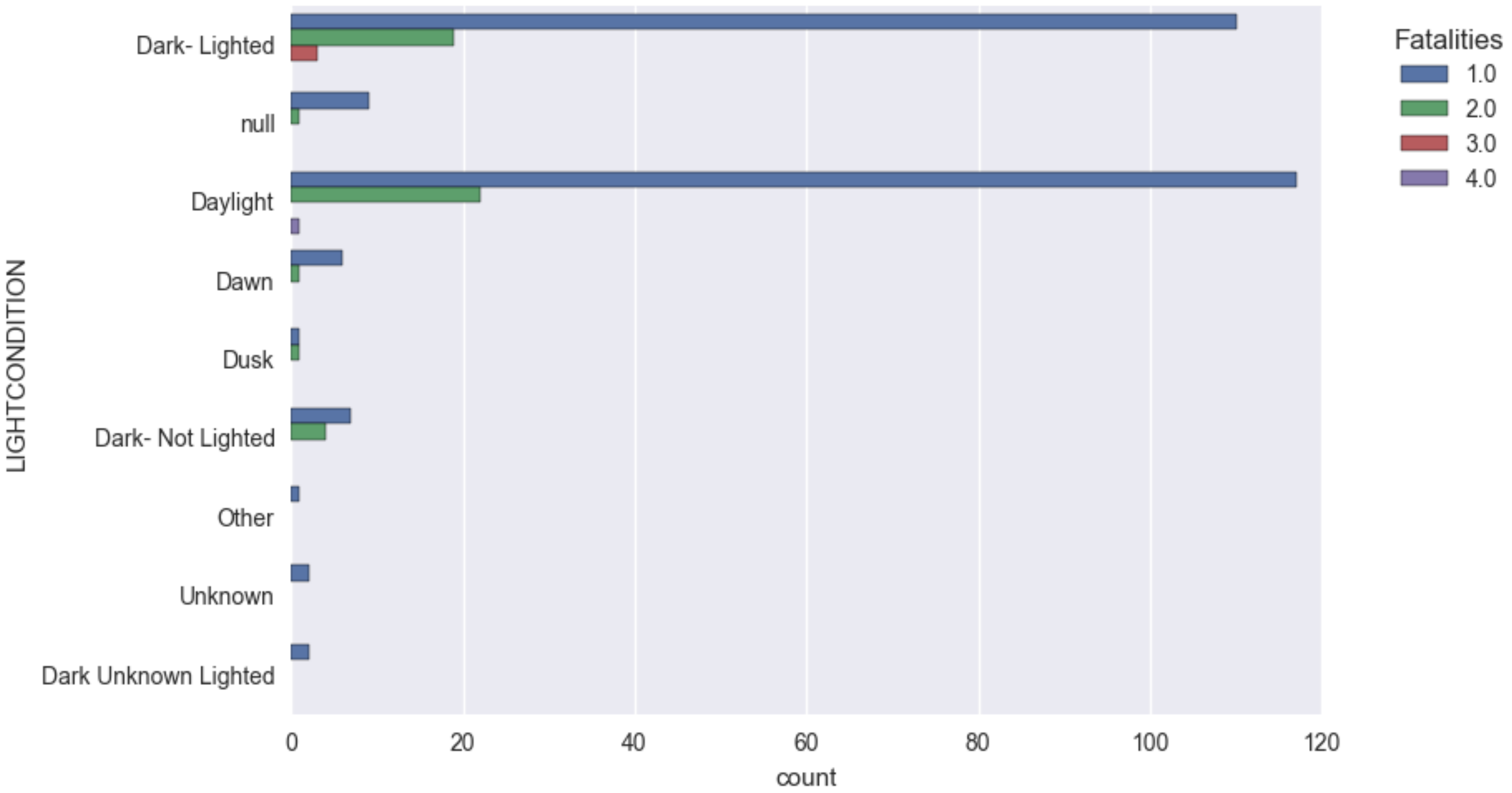
Most of the variables are categorical. We can create some tables and charts to get a better understanding of them.

These tables and charts represent the variables that are turned into dummies.

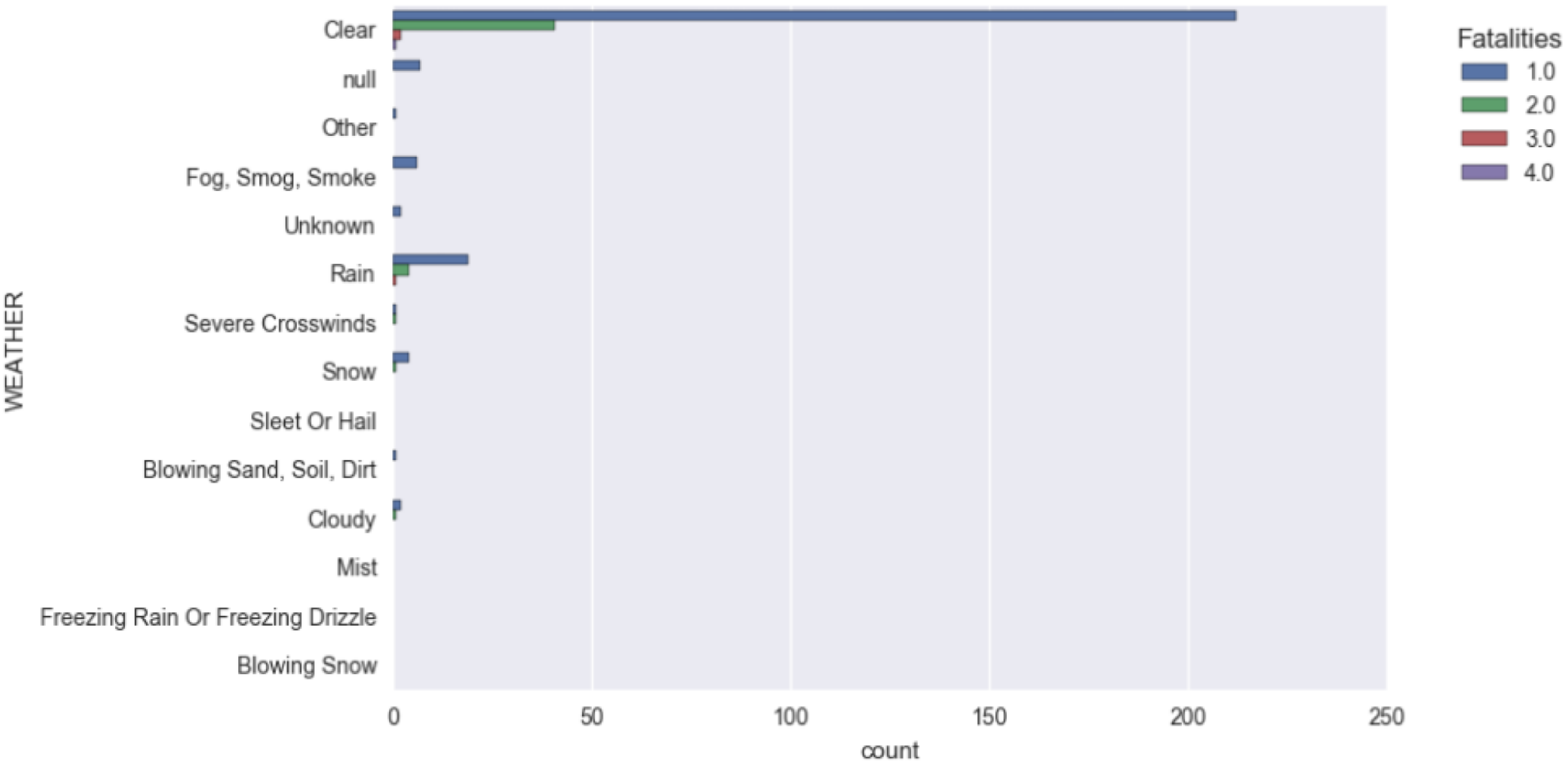


FATALITIES	1.0	2.0	3.0	4.0	All
FIRSTHARMFULEVENTSPECIFICS					
Animal	1	0	0	0	1
Cargo/Equipment Loss Or Shift	1	0	0	0	1
Concrete Traffic Barrier	1	0	0	0	1
D.C. Property	1	0	0	0	1
Hit and Run	8	1	0	0	9
Motor Vehicle In Transport	114	11	1	1	127
Other Fixed Object (Wall, Building, Tunnel, Etc.)	19	11	0	0	30
Other Non-collision	12	2	0	0	14
Other Non-fixed Object	8	1	0	0	9
Other Property Damage	1	0	0	0	1
Other Traffic Barrier	1	1	1	0	3
Parked Motor Vehicle	24	5	1	0	30
Pedestrian	54	8	0	0	62
Unknown	1	0	0	0	1
null	9	8	0	0	17
All	255	48	3	1	307

Summary

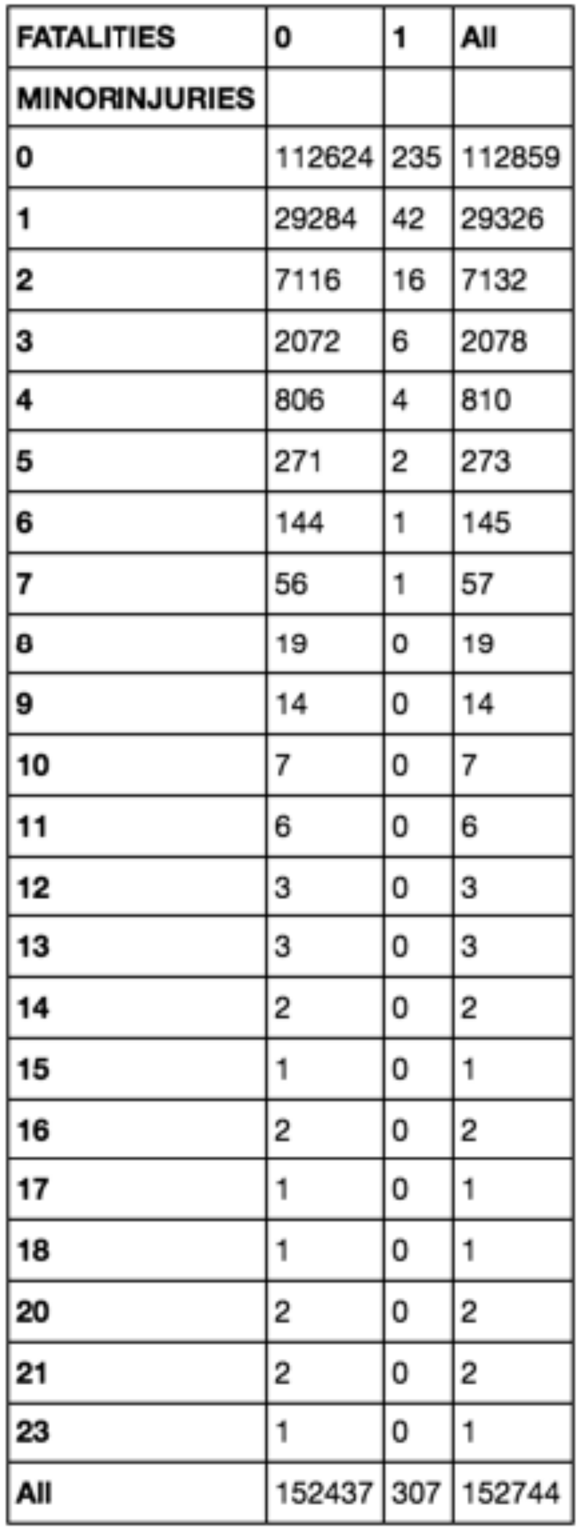

Dark- Lighted
null
Daylight
Dawn
Dusk
Dark- Not Lighted
Other
Unknown
Dark Unknown Lighted
LIGHTCONDITION
0
20
40
60
80
100
120
count
Fatalities
1.0
2.0
3.0
4.0

Summary

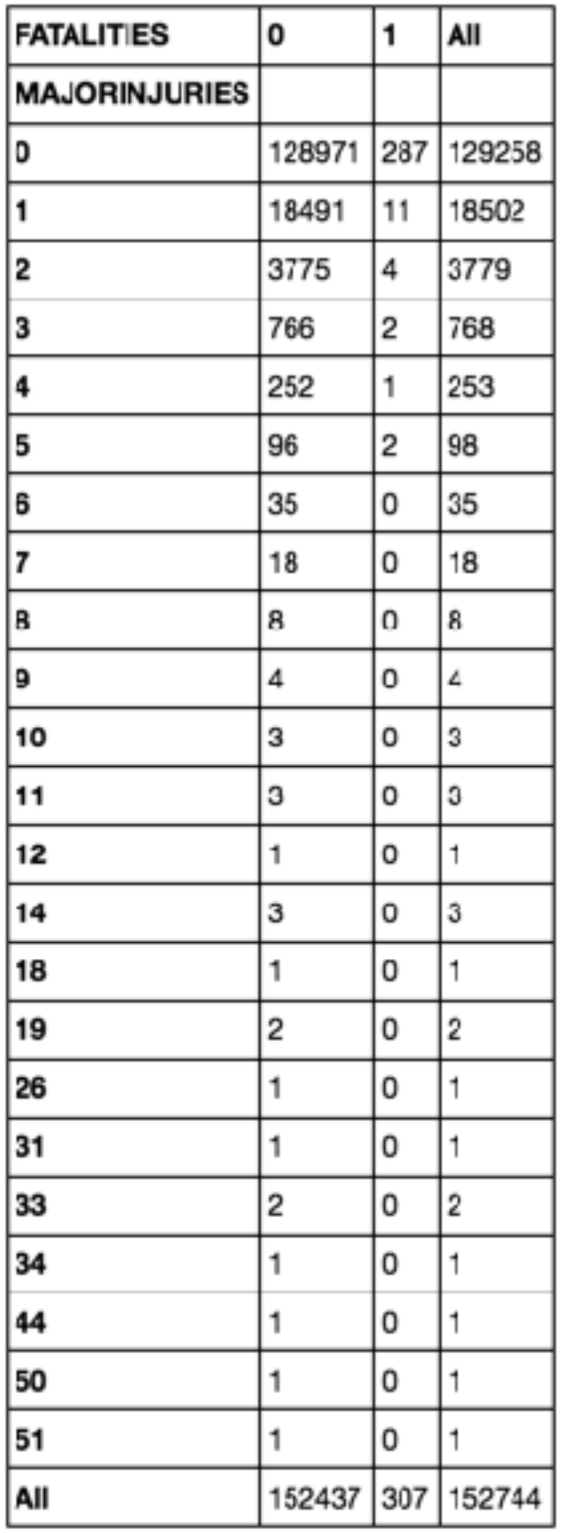

Clear
null
Other
Fog, Smog, Smoke
Unknown
Rain
Severe Crosswinds
Snow
Sleet Or Hail
Blowing Sand, Soil, Dirt
Cloudy
Mist
Freezing Rain Or Freezing Drizzle
Blowing Snow
WEATHER
0
50
100
150
200
250
count
Fatalities
1.0
2.0
3.0
4.0

Summary

From a glance, a crash that has more minor or major injuries does not lead to the crash being fatal

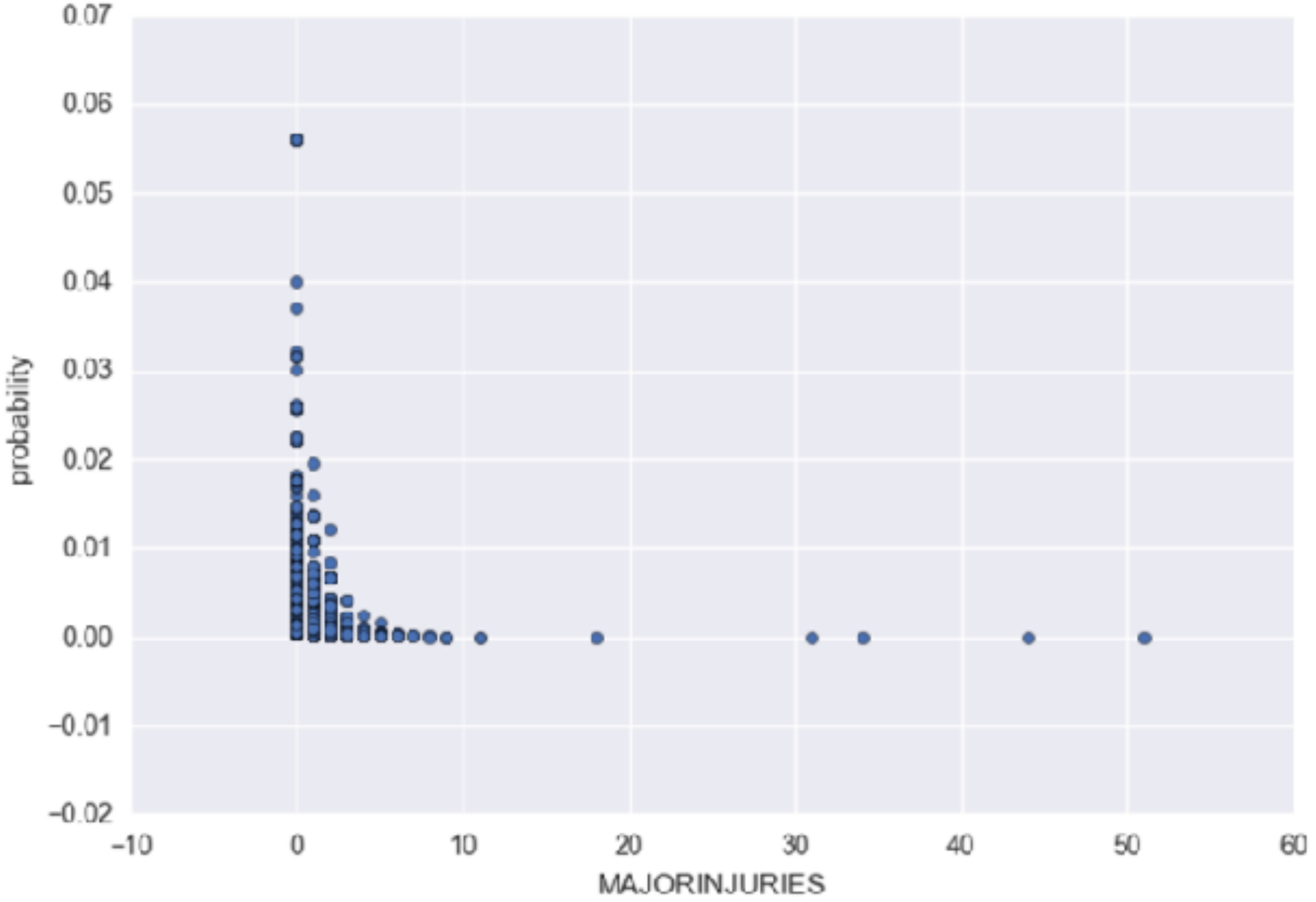


FATALITIES	0	1	All
MINORINJURIES			
0	112624	235	112859
1	29284	42	29326
2	7116	16	7132
3	2072	6	2078
4	806	4	810
5	271	2	273
6	144	1	145
7	56	1	57
8	19	0	19
9	14	0	14
10	7	0	7
11	6	0	6
12	3	0	3
13	3	0	3
14	2	0	2
15	1	0	1
16	2	0	2
17	1	0	1
18	1	0	1
20	2	0	2
21	2	0	2
23	1	0	1
All	152437	307	152744



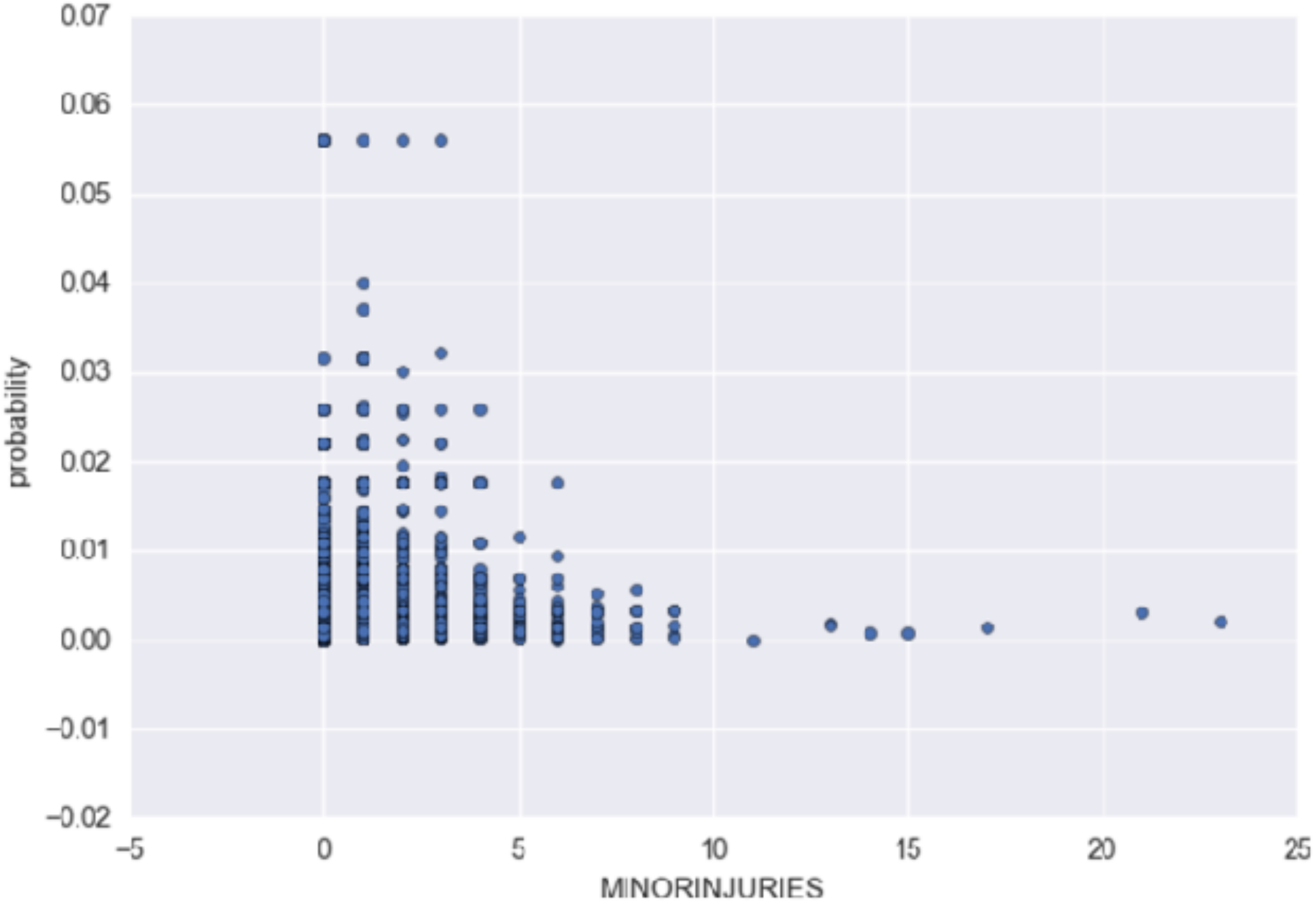
FATALITIES	0	1	All
MAJORINJURIES			
0	128971	287	129258
1	18491	11	18502
2	3775	4	3779
3	766	2	768
4	252	1	253
5	96	2	98
6	35	0	35
7	18	0	18
8	8	0	8
9	4	0	4
10	3	0	3
11	3	0	3
12	1	0	1
14	3	0	3
18	1	0	1
19	2	0	2
26	1	0	1
31	1	0	1
33	2	0	2
34	1	0	1
44	1	0	1
50	1	0	1
51	1	0	1
All	152437	307	152744

Modeling Insight



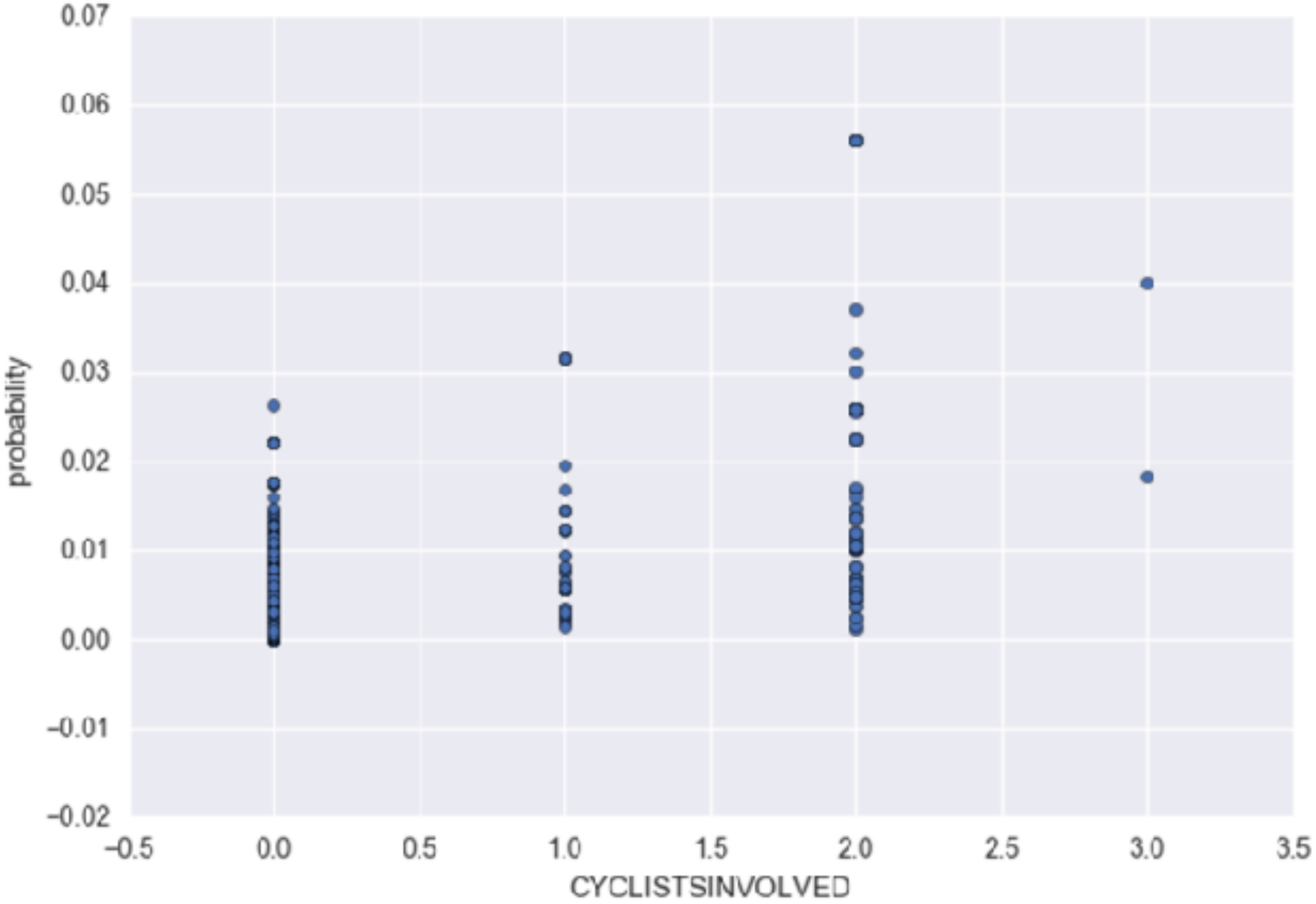
As the number of major injuries in a crash increases, the probability of that crash being fatal decreases.

Modeling Insight



For crashes with minor injuries under 3, the probability of that crash being fatal is about the same. For each increase in minor injuries for a crash thereafter, the probability of that crash being fatal decreases to the point where variations are negligible.

Modeling Insight



The more cyclists involved in a crash, the higher the probability that crash will be fatal, though the increases in probability are extremely slight.

Modeling Insight

Visualizing the calculated probabilities of all other variables saw that there was a decrease in the probability of a crash being fatal as the independent variable increased (for numeric) or was present (for categorical/dummy variables).

Modeling Approach

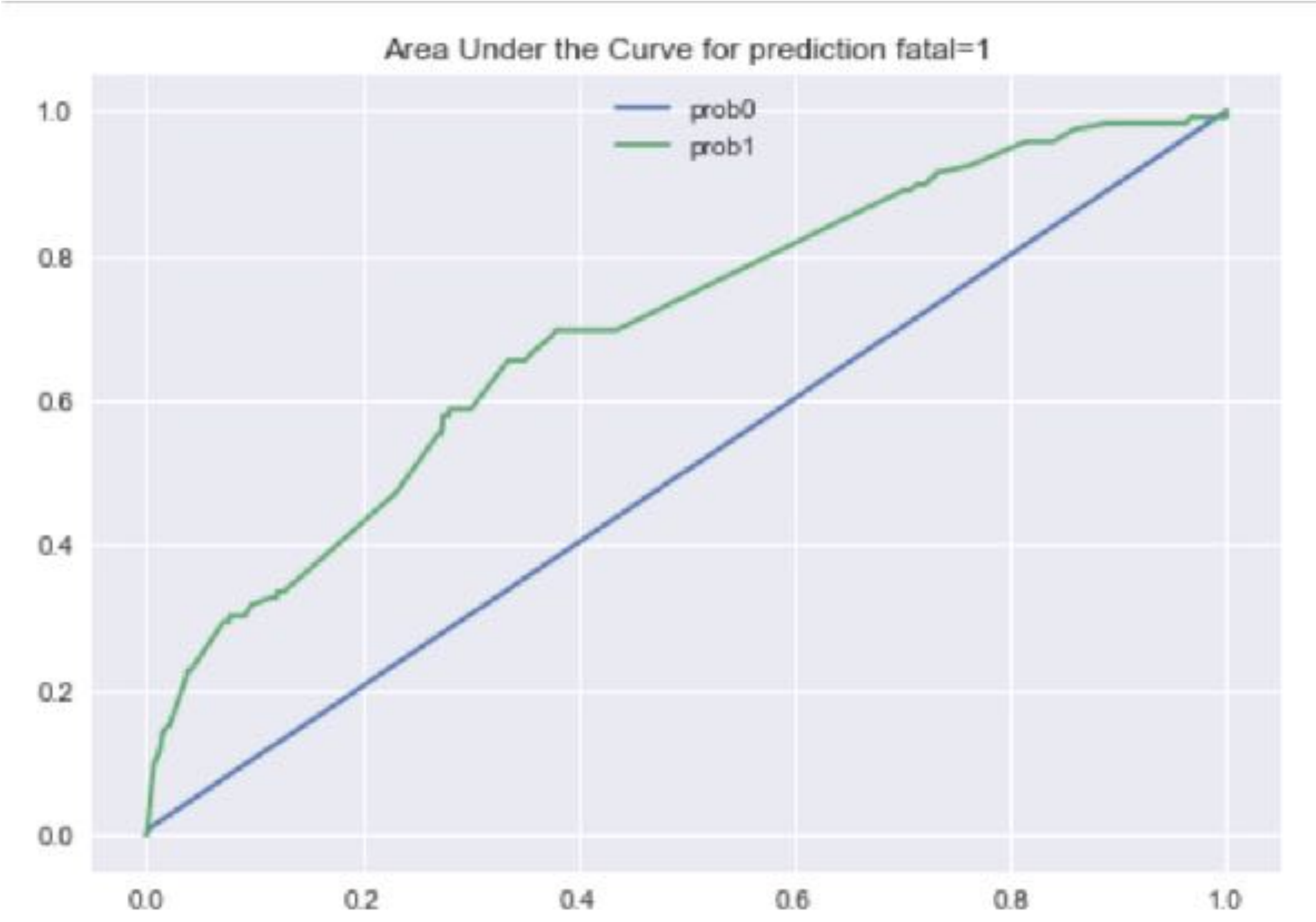
- This study involved created a logistic regression model

- The logistic regression model was optimizing for the area under the curve score (AUC score) - we’re trying to get it as high as possible

- The AUC summarizes the receiver operating characteristic (ROC) curve in one figure - used to evaluate the quality of the output of the model

Results

Area under the ROC curve score: **0.70 (the low end of fair)**



Conclusion

The analysis remains mostly inconclusive; results are poor and probabilities are low

This study cannot reliably suggest one way or another what variables would affect fatal crashes

Learned: many more data sources need to be considered for this kind of analysis

Next Steps

Other data that could augment this study include, but are not limited to:

- traffic data
- location (x and y coordinates)
- age of driver
- type of vehicle

With more data, we could see an improved analysis and come to a supported conclusion as to what factors contribute the most to fatal car crashes.

Next Steps

With significant results and a solid conclusion, this work could be used to inform appropriate advertising campaigns for Vision Zero.

When including location, the results could also be used to inform the District on places where traffic calming initiatives should be instated.