Causes of Injuries and Fatalities in Traffic Accidents in Chicago

Group 76: Kevin Fernandez, Christina Goodwin, Landon Lowe, Steph Wagner, Jake Weeren

Background

Crash fatalities in Illinois have been on the rise over the past five years, peaking in 2021. The city of Chicago provides traffic crash data from 2015 through the present. This data includes posted speed limit, weather conditions, lighting conditions, roadway conditions, injuries, fatalities, time of day, and location of the accident.

Traffic accidents put a strain on the city's infrastructure, healthcare system, and municipal workers, while profoundly affecting the lives of those involved in the accident itself. By investigating the most significant factors that lead to accidents in Chicago, a business case can be developed to use the city budget most effectively in the areas most impacted by these accidents. Each factor can be analyzed for its significance, and the cost of correcting these issues for specific high congestion and high accident rate areas within the city.

Problem Statement

What are the main factors that lead to injuries or fatalities in number and/or probability for traffic accidents in Chicago?

Business Justification

Traffic accidents put a strain on the city's infrastructure, healthcare system, municipal workers, while profoundly affecting the lives of those involved in the accident itself. By investigating the most significant factors that lead to accidents in Chicago, a business case can be developed to use the city budget most effectively in the areas most impacted by these accidents. Each factor can be analyzed for its significance, and the cost of correcting these issues for specific high congestion and high accident rate areas within the city.

Dataset

Traffic Crashes – Shows data on crashing in on the streets of Chicago. Data comes from electronic crash reporting system (E-Crash) at Chicago Police Department. About half of all crash reports, mostly minor crashes, are self-reported at the police district by the driver(s) involved and the other half are recorded at the scene by the police officer responding to the crash. A traffic crash within the city limits for which CPD is not the responding police agency, typically crashes on interstate highways, freeway ramps, and on local roads along the city boundary, are excluded from this dataset. As per Illinois statute, only crashes with a property damage value of $1,500 or more or involving bodily injury to any person(s) and that happen on a public roadway and that involve at least one moving vehicle, except bike dooring, are considered reportable crashes. However, CPD records every reported traffic crash event, regardless of the statute of limitations. Time Period: 2015 to present (All police districts September 2017 - present) Frequency: Daily Rows: 703K, Columns: 49, Each row is a Traffic Crash

Literature Survey

The paper from Comi et al discusses factors that lead to severity increase of accidents. The observed city is a similarly heavily populated city in Rome so findings there should be greatly correlated with findings in a city like Chicago, although public transport may be slightly more used in Rome. Comi et al find the importance of signals, size and type of cars, and the conditions of road itself along with road signals, to be the largest factors associated with amount and severity of accidents. They come to this conclusion via K- means and network algorithms and the approach itself matches some results we'd hypothesize ourselves. Comi et al note the lack of availability of road condition and other data to the extent necessary to be truly complete analysis wise and we feel that our dataset can answer more questions than answered in there paper while using methodology that produces similar findings on the commonalities.

The analysis by the National Institute of Health uses traffic data in Korea to assess the major reasons for accidents. It finds in its analysis the common refrain that men tend to be far more reckless drivers than women, and thus cause more accidents and higher severity accidents. And also finds that a large proportion (disproportionate #) of accidents of course comes from areas with larger speed limits like highways that tend to be highly congested by nature. It also suggests the efficacy of solutions to the issues causing these accidents like speed cameras, bans of mobile usage while driving, and more are very high, thus suggesting that the findings we have made in our analysis can be similarly hugely impactful. This analysis reads as more of a preliminary fact to the data we tasked ourselves with investigating here along with a suggestion of solutions.

Methodology

Our methodology involved the use of logistic and standard multiple regression models to analyze the most significant factors that contribute to traffic accidents. In order to perform this analysis, we converted the various conditions of the road into categorical variables. Our model evaluation approach consisted of conducting partial F-tests in order to decide which variables contribute greatly to the model, chi squared tests for independence, as well as comparing the models' AIC (Akaike Information Criterion), AUC, ROC Curves, along with simple graphs/charts for understanding. The resulting prediction and regression models allowed us to determine the likelihood of a fatal accident occurring and better understand the potential causes. In order to isolate factors and due to the presence of many sets of typically associated factors within factors, we grouped the factors that were related in a sense to do analysis (hence the finding labels) since those results would give more actionable solutions and observations. With there being lots of levels of each factor, a full model would include thousands of coefficients and thus wasn't desired. By using this approach, we aimed to not only identify the specific factors that played a significant role in causing accidents, but also to measure the importance of different factors and their potential interactions in preventing or reducing the occurrence of accidents. By doing so, we've helped assist the development of strategies to minimize the risks associated with this issue.

Findings

Condition Related Findings

Looking at the Conditions data effect on the response variable Crash_Type, which was converted to binary 0 being no injury and 1 injury. The different variables on conditions that are explored are Device Condition, Traffic Control Device, Weather Condition, and Lighting Condition. Before getting into their effect on the response variables a missing data analysis was done. There were no NA present in the independent variables. However, each of them has a factor 'unknown' and 'other' except lighting conditions only had an unknown factor. The reason to point this out is because those factors will not add value to our analysis. All the rows where these values are present only account for 3.61% of the data so not enough to worry about but I just removed those rows because they don't add any value to the analysis.

Next, running a stepwise logistic regression both ways forward and backward the following variables proved to be significant in determining if there was an injury in the crash. "LIGHTING_CONDITIONDARKNESS, LIGHTED ROAD" and "LIGHTING_CONDITIONDAWN" both had negative coefficients at -.52 and -.23 respectively. Meaning that these factors lead to crashes with injuries. "LIGHTING_CONDITIONDAYLIGHT", , "TRAFFIC_CONTROL_DEVICENO CONTROLS", "TRAFFIC_CONTROL_DEVICEOTHER RAILROAD CROSSING", "TRAFFIC_CONTROL_DEVICEPOLICE/FLAGMAN", "TRAFFIC_CONTROL_DEVICERAILROAD CROSSING GATE", "TRAFFIC_CONTROL_DEVICETRAFFIC SIGNAL"," DEVICE_CONDITIONFUNCTIONING PROPERLY", "DEVICE_CONDITIONMISSING" are all significant factors as well. These factors lead to crashes with no injuries.

From this we are now aware of certain conditions that influence crashes with injuries. None of the weather conditions seem to be significant which is surprising to me because I know that the weather can be very cold in the winter in Chicago. Also we see that certain traffic control devices/signs can help lead to less crashes with injuries so we could recommend to add more of these signs in areas with more crashes.

There was a problem with the amount of unknowns and others around 10% which is not ideal. Ideally we would like it to be below 5%. The next step for this analysis could be to look at just the crash regardless of an injury or not and see if the different conditions have an effect. I'd like to think that weather would then be significant but as we discovered above maybe the people of chicago are used to driving in for example snow and so that could be the reason for weather not being significant.

City Design Related Findings

Several city design related factors could be contributing to injury or fatal accidents within Chicago. Of the fields that are reported, if the accident occurred as the result of an intersection, because of an issue with right of way, or what type of trafficway all relate to the design of the city and are potential causes of accidents. Each of these factors were turned into indicator variables, 1 for yes, 0 for no. The types of trafficways are ALLEY, DIVIDED - W/MEDIAN (NOT RAISED), DIVIDED - W/MEDIAN BARRIER, FIVE POINT OR MORE, FOUR WAY, NOT DIVIDED, NOT REPORTED, ONE-WAY, OTHER, PARKING LOT, ROUNDABOUT, T-INTERSECTION, TRAFFIC ROUTE, UNKNOWN, and UNKNOWN INTERSECTION TYPE.

Total injuries and fatal accidents were reported in the INJURIES_TOTAL, and INJURIES FATAL. This field was numeric, so was converted into a binary field to represent if the accident resulted in any injury or fatality. These were used as the dependent variables in logistic regression.

Predicting Total Injuries

The likelihood that injuries would result from a crash was modeled using logistic regression. The first model used INTERSECTION_RELATED_I, NOT_RIGHT_OF_WAY_I, HIT_AND_RUN_I, and all trafficway type indicator variables. The intercept was the only statistically significant factor, which represents the base case, of “OTHER” trafficway, not intersection related, not right of way related, and not a hit and run. Hit and run, four-way trafficway type, and not divided were above the alpha threshold of 0.05.

Removing hit and run as a factor seemed logical, as it is an outcome after the fact of an accident and not a cause that could be changed. Running the logistic regression again without hit and run, the intercept is even more statistically significant: alpha changed from 0.0127 to 0.000951. Both four-way and not divided were still not statistically significant, but the closest to significant of any of the factors.

Predicting Fatalities

An issue with predicting fatalities using this data, is that once data was cleansed to remove rows with missing data for these predictors, only one fatality remained. A logistic regression model was run, but did not result in a quality model and should not be used. As there were 211,736 observations for trafficway type, that factor was used in a logistic regression to determine the likelihood of a fatal accident for the five most common trafficway types and the other group. In the table below, the “Other” traffic way stop had both the highest fatality rate and had the highest level of confidence of the model. The model also shows that parking lots had the lowest level of fatal incidents with a high degree of confidence.

Trafficway Type	Percent	Pr(>|z|)	
Other	0.2063%	< 2e-16	***
Divided Not Raised	0.1834%	0.52141	
Four Way	0.1702%	0.352831	
Not Divided	0.1315%	0.003339	**
One Way	0.1022%	0.002388	**
Parking Lot	0.0076%	0.000997	***

Likelihood of Fatality and Level of Confidence by Trafficway Type

	Reference Non-Fatality	Reference Fatality
Predicted Non-Fatality	179509	236
Predicted Fatality	31925	66

Confusion Matrix of Fatalities by Trafficway Type

As evident from the confusion matrix, traffic type alone is not enough to accurately predict fatalities in traffic accidents. This is supported by the model’s AIC of over 4500. Trafficway type is not a strong enough indicator to determine fatalities.

Work Related Findings

Looking into work related variables can provide insight into how the presence of roadwork and the type of roadwork affects the likelihood of severe accidents. We suspect since roadwork interrupts the usual flow of traffic in the form of detours, speed limit changes, road closures, etc., there might be an increased chance of an accident in areas undergoing roadwork. From our dataset, we have three work-related variables: WORK_ZONE_I, whether the crash occurred in an active work zone, WORK_ZONE_TYPE, the type of work zone (i.e. Construction, Maintenance, Utility, and Unknown), and WORKERS_PRESENT_I, whether or not there were construction workers present in an active work zone at crash location. Due to data limitations, we only considered observations with no missing values in any of these fields, thus limiting our analysis to 1065 observations. This reduction also resulted in the WORK_ZONE_I variable to only contain values of “Yes” and was therefore removed since there was no variability we can learn from. The response, CRASH_TYPE, was converted to 0 and 1 for “no injury” and “injury”, respectively. Once the predictors were converted to factors, we ran three logistic regressions: the first with CRASH_TYPE versus WORK_ZONE_TYPE, the second with CRASH_TYPE versus WORKERS_PRESENT_I, and the third with CRASH_TYPE versus WORK_ZONE_TYPE and WORKERS_PRESENT_I.

The first regression looked at WORK_ZONE_TYPE in isolation. The results showed that the work zones of Maintenance, Utility, and Unknown were not significant at the 95% confidence level, but the base case, Construction, was very significant. This makes sense since Construction work tends to be more involved and ends up displacing and diverting more drivers in general than routine Utility work or surface level Maintenance work. The second regression looked at WORKERS_PRESENT_I in isolation. The results showed that having workers present was not significant in predicting the likelihood of a severe accident at the 95% confidence level but was significant at the 90% confidence level. The base case of not having workers present was significant at the 95% confidence level and with a negative coefficient implying not having workers present led to a lower likelihood of severe accidents. Lastly, the third regression considering both these predictors resulted in all the factors being less significant with the base case of a Construction zone with no workers present being the only significant variable. Moving forward, it could be useful to look at interactions between these work-related variables as well as with the other variables considered in our analysis.

An additional model was considered using the interaction between WORK_ZONE_TYPE and WORKERS_PRESENT_I. In this case, the intercept remained significant at a 95% confidence level as did the interaction between work WORK_ZONE_TYPEUNKNOWN and WORKERS_PRESENT_IY. This suggests that the severity of accidents is higher when there are workers present, particularly when the work zone type is unknown. Since unknown is merely a filler category, it is difficult to interpret the meaning of this finding other than perhaps if the work zone type is unknown due to several factors that themselves contribute to a work zone that might be more challenging to navigate. Since we have several models each with slight variations in variables, we can check AICs to determine which model fits the data best. The model with the lowest AIC was the model with just WORKERS_PRESENT_I as a predictor. We compare how each additional variable affects the fit of the model relative to the model with just WORKERS_PRESENT_I. The likelihood ratio test between these two models yielded a p-value of 0.62 which means we fail to reject the null hypothesis and conclude that there’s no significant difference and the addition of this variable did not improve the fit. Notably, if we do a likelihood ratio test with a model only having WORK_ZONE_TYPE and the full model, we get a p-value of 0.08 meaning the addition of WORKERS_PRESENT_I greatly improves the fit of the model. Overall, it seems like WORKERS_PRESENT_I is the variable with the most explanatory power

and should be the only we consider when building a full more comprehensive model incorporating other variables.

Location Related Findings

With Regard to Location of Accident related observations/findings, it seemed natural to observe the relations in regards to attributes about the streets themselves (Road Defects and Roadway Surface Conditions), along with the police beat(zone of city accident occurred in). These were measured against whether or not a fatal accident occurred or not on that road.

In terms of descriptions of data, there were 275 different Beats (Zones) included in this accident data. (See below with the beats layout for the city.) There were 7 different Roadway Surface Conditions which could be also refactored into two (Dry or not): the levels were the following:

```
[1] DRY             UNKNOWN         WET             ICE             SNOW OR SLUSH   OTHER
[7] SAND, MUD, DIRT
```

There were 7 different Road Defect Types which could be also refactored into two (None or not): the levels were the following:

```
[1] UNKNOWN           NO DEFECTS        OTHER             SHOULDER DEFECT   WORN SURFACE
[6] DEBRIS ON ROADWAY RUT, HOLES
```

Lastly, one must note that since the fatality happening is labeled 1, the observation of no fatality is 0. Also, the sample size for nonfatal accidents was about three times that of fatal ones for known data values for each of the variables so that may slightly bias the coefficients in a sense although the large sample size should compensate.

The first of the things noticed when putting the large full logistic regression model with all factors and fatality as dependent variable together was the weight of importance of location and Impact of certain road conditions on likelihood of fatality. When looking at the logistic fit we can see that certain factor levels had huge impacts on likelihood of fatality. Interestingly, snow or slush affected standard asphalt roads seemed to be the most labeled(non-Other) fatality prone perhaps due to non-obvious slipperiness and asphalt itself fatality (.18498… coefficient). Lastly, we can see neighborhood cluster commonality in location of fatal accidents. Beats that follow each other, say 522 and 523 (i.e neighboring geographically), have fairly similar coefficients with regards to impact on fatality of accidents. This could suggest the presence of better/worse roads and/or common congested main roads etc. and/or condition commonalities between close areas road surface wise, but is clearly an important thing to consider in terms of local funds and resource allocation. Also, road defects had massive impacts on fatality as we see by the factor size of the coefficients. The model fit and coeffs. generated are shown below.

The next findings involve taking parts of the models.

The first thing of interest found when analyzing this project was that there was no independence between roads having defects and location of police beats. A chi squared test of independence resulted in a P value of 0 meaning we can reject the null and assume there is a dependence between variables. Neighborhoods presumably with larger tax bases did in fact have better

roads. Of course, the natural implication of this is that presuming there is an effect on probability of accidents on roads given quality of road, the improvement of roads in lower income areas would in fact have massive impact on the probability of fatality on the roads due to car accidents. Also, we now know in which areas most investment is needed.

Naturally, we can also observe the relations between whether fatality occurred or not and the road having defects or not. So, we went on to refactor the Road Defects value as the presence (1) or lack of presence of defects (0). A table of these results is shown below. Clearly the presence of defects affects the likelihood of accidents, the percent difference was around .05%, but of course that is the equivalent 5 deaths per ten thousand people . To put it in perspective, this is a bigger percentage than the rightly criticized US Maternal Mortality Rate (around .0329%). The chi squared test of significance was ran on the table and had a pvalue of .055 which is significant at the 10% significance level and near the .05 cutoff so clearly dependence is likely to exist as expected.

	FATAL	
Defected	0	1
0	167110	252
1	46626	52

A linear model could be built with both the Roadway Surface Condition and Road Defects as factors, along with the interaction of the two and fatality happening.The most important findings as expected happen with the interaction of the two location/ road quality factors. The largest increases in odds of fatal accident occurring when the surface itself was wet and Road Defect was with regards to the road shoulder(.03492 coefficient). The fact that interactions involving the road being wet or icy along with the presence of a defect are the most dangerous, makes sense when one considers the existence of hydroplaning, very frequently warned about to drivers, and the idea that its hard enough to stop on a good surface if that's the case, not to mention a worn out surface or defective shoulder lane.

The factor Roadway Surface Conditions can also be made binary (Bad = 1, Ideal = 0) and this binification is contained in the BadSurface variable. Using this we can create a simplified GLM in order to have a straightforward observation about the importance of the variables (Defected, BadSurface). As expected the weather and the road defects aren't correlated so the interaction variable actually worsens the model when we look at the major variables as simple binary. Thus, the final model with these two variables as IVs leaves interaction out. The interesting find here is the lack of significance of bad surface(due to current weather) in and of itself, while the presence of road defects is significant. This is in alignment with the Chi Squared test from earlier which tested defects or not vs fatality or not. Roadway surface condition seems to only matter significantly in specific cases addressed earlier in interaction with road defects. What may be the rationale is that bad surfaces tend to induce drivers to slow down more than good ones but crashes where road defects exist on bad surface conditions tend to be far more deadly.

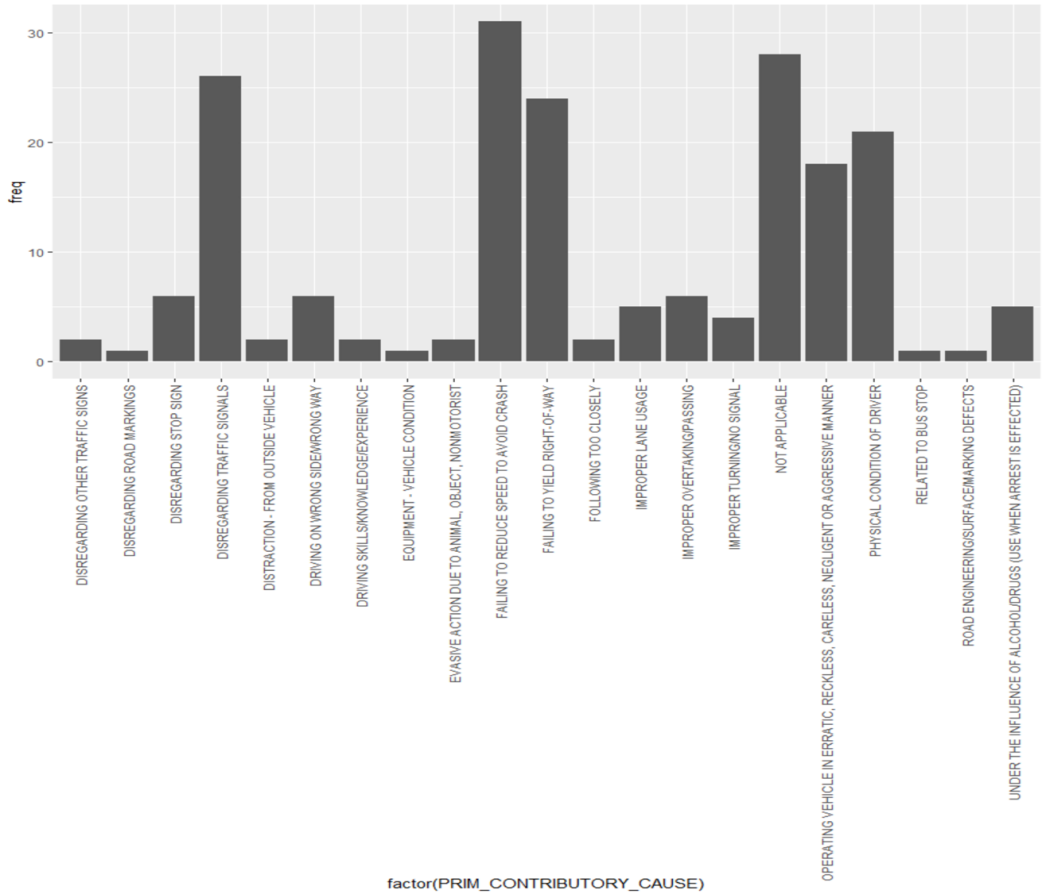
```
Call:
glm(formula = FATAL ~ BadSurface + Defected, family = (binomial(link = "logit")),
    data = locationdf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.0573  -0.0544  -0.0544  -0.0544   3.6927

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.5141     0.0682 -95.511   <2e-16 ***
BadSurface    0.1036     0.1513   0.685   0.4936
Defected     -0.3030     0.1524  -1.988   0.0468 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4594.2  on 214039  degrees of freedom
Residual deviance: 4589.5  on 214037  degrees of freedom
AIC: 4595.5

Number of Fisher Scoring iterations: 9
```

Cause Related Find

One interesting thing that we can also look at general stated causes of fatal accidents. In terms of primary causes, a chart showing the most common primary causes is shown below:

```
> plyr::count(primCauseDF, "PRIM_CONTRIBUTORY_CAUSE")
                                                    PRIM_CONTRIBUTORY_CAUSE freq
1                                          DISREGARDING OTHER TRAFFIC SIGNS    2
2                                               DISREGARDING ROAD MARKINGS    1
3                                                   DISREGARDING STOP SIGN    6
4                                             DISREGARDING TRAFFIC SIGNALS   26
5                                        DISTRACTION - FROM OUTSIDE VEHICLE    2
6                                           DRIVING ON WRONG SIDE/WRONG WAY    6
7                                         DRIVING SKILLS/KNOWLEDGE/EXPERIENCE    2
8                                              EQUIPMENT - VEHICLE CONDITION    1
9                                EVASIVE ACTION DUE TO ANIMAL, OBJECT, NONMOTORIST    2
10                                         FAILING TO REDUCE SPEED TO AVOID CRASH   31
11                                          FAILING TO YIELD RIGHT-OF-WAY   24
12                                                  FOLLOWING TOO CLOSELY    2
13                                                    IMPROPER LANE USAGE    5
14                                           IMPROPER OVERTAKING/PASSING    6
15                                            IMPROPER TURNING/NO SIGNAL    4
16                                                          NOT APPLICABLE   28
17 OPERATING VEHICLE IN ERRATIC, RECKLESS, CARELESS, NEGLIGENT OR AGGRESSIVE MANNER   18
18                                                PHYSICAL CONDITION OF DRIVER   21
19                                                          RELATED TO BUS STOP    1
20                                       ROAD ENGINEERING/SURFACE/MARKING DEFECTS    1
21                                                          UNABLE TO DETERMINE  110
22              UNDER THE INFLUENCE OF ALCOHOL/DRUGS (USE WHEN ARREST IS EFFECTED)    5
```

Notably, speeding and not following road rules were the most common observed issues when fatalities happened, along with “Not Applicable” suggesting non-human road/car quality causes. This may suggest need for greater punishment for not following road rules etc along with road/car improvement since those worsen an already occurring accident as shown earlier in the paper. The graph below shows that the top 5 factors or so were responsible for most accidents observed when the cause was determined and noted. A graph showing this massive disparity can be seen below:



.

Works Cited

(n.d.). *Traffic Crashes - Crashes*. Chicago Data Portal. https://data.cityofchicago.org/Transportation/Traffic-Crashes-Crashes/85ca-t3if

Comi, A., Polimeni, A., & Balsamo, C. (n.d.). *Road Accident Analysis with Data Mining Approach: Evidence from Rome*. ScienceDirect. https://www.sciencedirect.com/science/article/pii/S2352146522002265

Rawat, S. (n.d.). *USA Accidents Data Analysis*. Towards Data Science. https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6785079/