The Danger of Class: an Exploration of the Effect of Road Classification on Alcohol Related Crash Severity

Nicholas Champagne

Alden Felix

Will Kilcoyne

Jim Pan

December 9, 2022

Contents

Abstract

Project centers around determining the connection between types of roads and the severity of alcohol-related crashes within the United States. Data specifically targets the area of Collin County in Texas between the year 2018. Using regression and data visualization to come to the conclusion that there is a marked difference in severity of crashes based upon classifications of roadways. The research question is whether or not there is a discrepancy between the different classification of roadways. We find that lower injury levels are more common on county roads or smaller road types while significant injuries are roughly equivalent between smaller and larger road types. Tollways have the highest safety rating of non-trafficways, while places such as alleyways have predictably low likelihoods of injuries.

Introduction

Drunk driving has been an issue ever since automobiles were originally invented. Injuries and deaths occur year round due to drivers driving under the influence, and while illegal, there is very little sign of anything slowing this hazard down for other drivers. It is about time, then, to recognize the best course of action for looking for a solution for this problem is to better understand where solutions are most needed. Instead of looking at the wide range of drunk driving and the issues, attempting to plug up every single hole to keep the proverbial ship afloat, it would make much more sense to instead search and determine where the resources would best be allocated.

In order to do this, it must be determined where the most casualties or danger lies. The purpose of this data visualization is to do precisely that. Using regressions on data of crashes across within Texas, it can be determined the severity of crashes based on the road the crash occurred. It is important to note that there is not the question of drunkenness, as the data used explains that all events recorded include an inebriated driver of some kind, be it the instigator of the accident or simply a member. The primary dependent variable in this case is the classification of roads, ranging from smaller backroads that receive less traffic, to larger highways and freeways which see hundreds of vehicles a day. The secondary dependent variable is the severity of the crash itself, ranging from life-threatening injuries and death to minor fender benders where the only true harm came to the vehicle. It is important to have these two dependent variables to properly visualize the differences in severity across types of roadways.

All data used in this project comes directly from the state of Texas, specifically accidents which occurred in Collin County during the year 2018. The limited scope is meant to help keep the project within a reasonable amount of work while also providing ample data. Had the entire United States been taken into account, there would have been far too many variables as well. Some states have dangerous roads regardless of drunkenness or not thanks to the colder weather and more frigid climates, so in an effort to prevent data being skewed by icy roads, Texas was chosen where roads largely remain the same over the course of any given year. Furthermore, the year 2018 was chosen since that is the last complete set of data prior to Covid hitting. Covid caused a massive decrease in all forms of transportation usage, be it vehicular or otherwise, and would have skewed the data far too much to draw reasonable conclusions from. Thankfully the data is still relatively fresh and will provide ample analysis despite not being within the previous two years.

Literature Review

Multiple articles and papers have been written in regards to drunk driving and the effects it has on vehicular safety, both to the drunk driver itself and the drivers around them (Sethi et al., 2021). Papers have delved into the ways to increase safety in general, across all road types and encounters on roads (Pedan et al., 2004; WHO, 2007). As previously stated, drunk driving has been an issue for quite a long time now, and nothing has seemed to slow it down short of the Prohibition era of America. Papers and essays have extensively covered the generals of drunk driving, how the severity of drunkness might lead to an increase of crash severity. However, these works mostly take public policy approaches such as reduction in BAC levels (Albalate, 2006), harsher penalties (Zobeck & Williams, 1994), or media campaigns (Elder et al., 2004). Little papers or essays however have delved into the road types and the correlations between the types of roads and the crash severities. For example, modeling the effect of traffic barriers and whether collision with these barriers versus other objects would decrease the severity of a crash (Rezapour et al., 2019). The purpose of this analysis is to shrink the scope and apply a more focused look at drunk driving. By zeroing in on specifically the types of roads these incidents occur on and the severity attached to those roads, it will be easier to understand where safety measures are specifically needed rather than providing blanket solutions that hardly move the needle as it is. With the data and regressions provided here, safety measures can be applied to the more dangerous roads and make a true impact.

When seatbelts were originally introduced to motor vehicles and became the standard, fatal crashes went down yet overall crashes went up. There is reasonable belief within the field that safety measures, at least when originally implemented, can lead to an uptick in accidents. While there are other studies that delve deeper into this phenomenon, this study mostly looks to see if this uptick of accidents holds true. Put differently, the belief would be that more safety measures means that highways and freeways would be inherently safer than backroads. Recent research in safety technology and implementation would question whether or not this holds true in regards to automobiles. This study seeks to prove definitively if a higher number of safety features truly makes a difference or not.

It is important to note that there are similar papers out there. Some have discussed how severe crashes on certain roads are, some have discussed how drunk driving in certain areas such as rural or urban areas can result in more severe accidents. None have looked specifically at the relationship between these types of roads and the severity of accidents, or to put it differently, none have decided to compare and contrast the severity of accidents based upon the classification of road itself. That is where this research will fit into the current research on the topic, filling an admittedly niche sector of knowledge that has yet to be delved into.

Methods & Data

The type of regression used for this analysis is cumulative logit using a stopping ratio. The data, as previously stated, is taken from Collin County in the state of Texas during the year 2018. The data comes directly from the Texas Department of Transportation, specifically from their Crash Records Information System. This data includes classifications of roads, which receive a numerical designation dependent upon their classification. For example, a highway might receive designation 1 in the analysis, while a backroad or dirt road might receive a designation of 5 to help distinguish and differentiate the two. This also helps in regards to the regression itself, making the designations distinct numbers to help in comparing and contrasting them.

For the purpose of the model and data crash severity ranges from five different classifications. These are as follows: Not Injured, designated as N; Possible Injury, designated as C; Suspected Minor Injury, designated as B; Suspected Serious Injury, designated as A; Fatal Injury, designated as K. It is important to note within the data that these are not mutually exclusive designations. It is assumed within the dataset and the visualization that if there is the designation K, then the previous designations are all present as well. Put differently, if a crash has caused sufficient injury to be classified as 'Fatal', then the crash has also caused 'Serious', 'Minor', and 'Possible' injuries. This makes it easier to visualize, as will be shown later, because now the graphs and charts are not visualizing the presence of a certain designation, but rather the absence. This allows averages to be taken, thresholds to be shown, and more.

Other important data notations include types of roadways, speed limits of those roadways, the weather during the time of the accident, time of day, and what type of collision the accident was. For the type of roadway they were split into eight different variables, which are as follows: County Road, City Street, Farm to Market, Interstate, Non-Trafficway, Other, Tollway, US and State Highways. Non-Trafficways encompass all areas where cars are usually not found, these being alleyways or roads designated to not be for vehicles. While the other data variables are important to note, the main points of focus are the crash severity and the road classes themselves. All the variables play a role in the regression model itself, which will be discussed in length.

To briefly explain what cumulative logit regression models are, they simply use log proportional odds to predict an ordinal response. Put differently, the cumulative logit regression model effectively takes in all of the different classifications and severity of crashes and begins ranking them against one another. With there being five different categories of crash severity, there will be four different regressions. Each model takes the form of this equation:

$$ln\left(\frac{Pr(Y_i\ lej)}{Pr(Y_i > j)}\right) = \beta_1 + \beta_2\text{Road Class} + \beta_3\text{Speed Limit} + \beta_4\text{Weather}$$
$$+ \beta_5\text{Time of Day} + \beta_6\text{Type of Collision} + \beta_7\text{Intersection}$$

In regards to the equation, j represents the crash severity. The formula showcases very clearly where each of the individual variables come into play as well and attaches them to a beta. These betas are all added together and used for a prediction. This regression uses a stopping ratio, which means that the regression will predict the log odds of being at or lower than a current category versus being in a higher one. What this means is that when the beta is positive, the crash is less likely to be severe, and when the beta is negative, the crash is more likely to be severe. Put differently, the beta serves as a check for the probability of how severe a crash is. A positive beta means a lesser severity, and a negative beta means the opposite.

The actual visualization methods used for this project come in two different groupings, descriptive visualizations and inferential. For the descriptive grouping, there is a pie and bar graph detailing the frequency in which all the variables were found within the data itself, allowing for clean and easy visualization of the scope of the data. In inferential, there is a coefficient and a probability plot which allow for more in-depth discussion and focus on the regressions themselves.

Visual Analysis

Figure 1

Crash Severity	Frequency
99 - UNKNOWN	1152
A - SUSPECTED SERIOUS INJURY	1713
B - SUSPECTED MINOR INJURY	3848
C - POSSIBLE INJURY	4037
K - FATAL INJURY	1057
N - NOT INJURED	15352

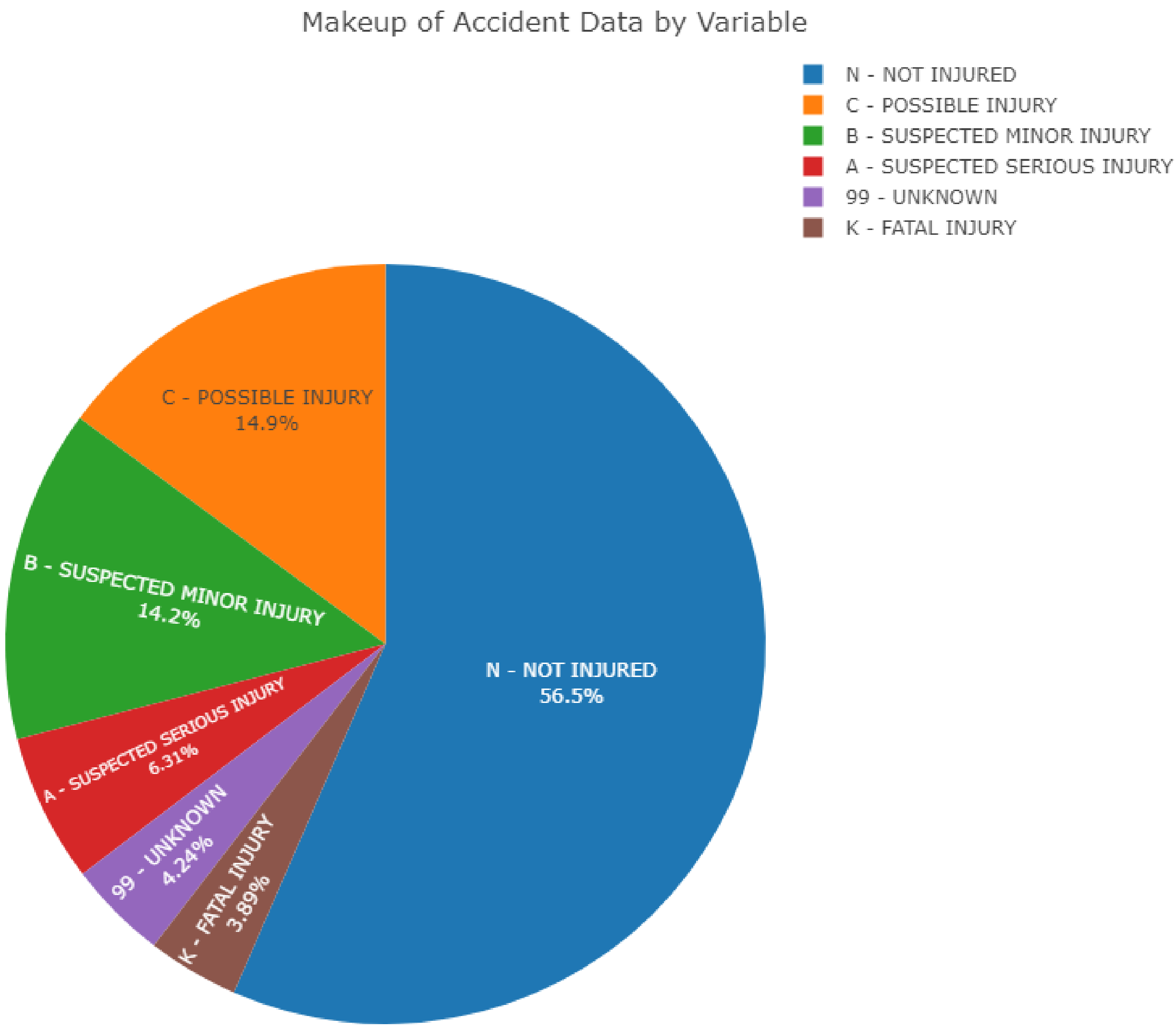
Road Class	Frequency
COUNTY ROAD	2329
CITY STREET	9726
FARM TO MARKET	3129
INTERSTATE	3555
NON TRAFFICWAY	1582
OTHER ROADS	36
TOLLWAY	347
US & STATE HIGHWAYS	6455

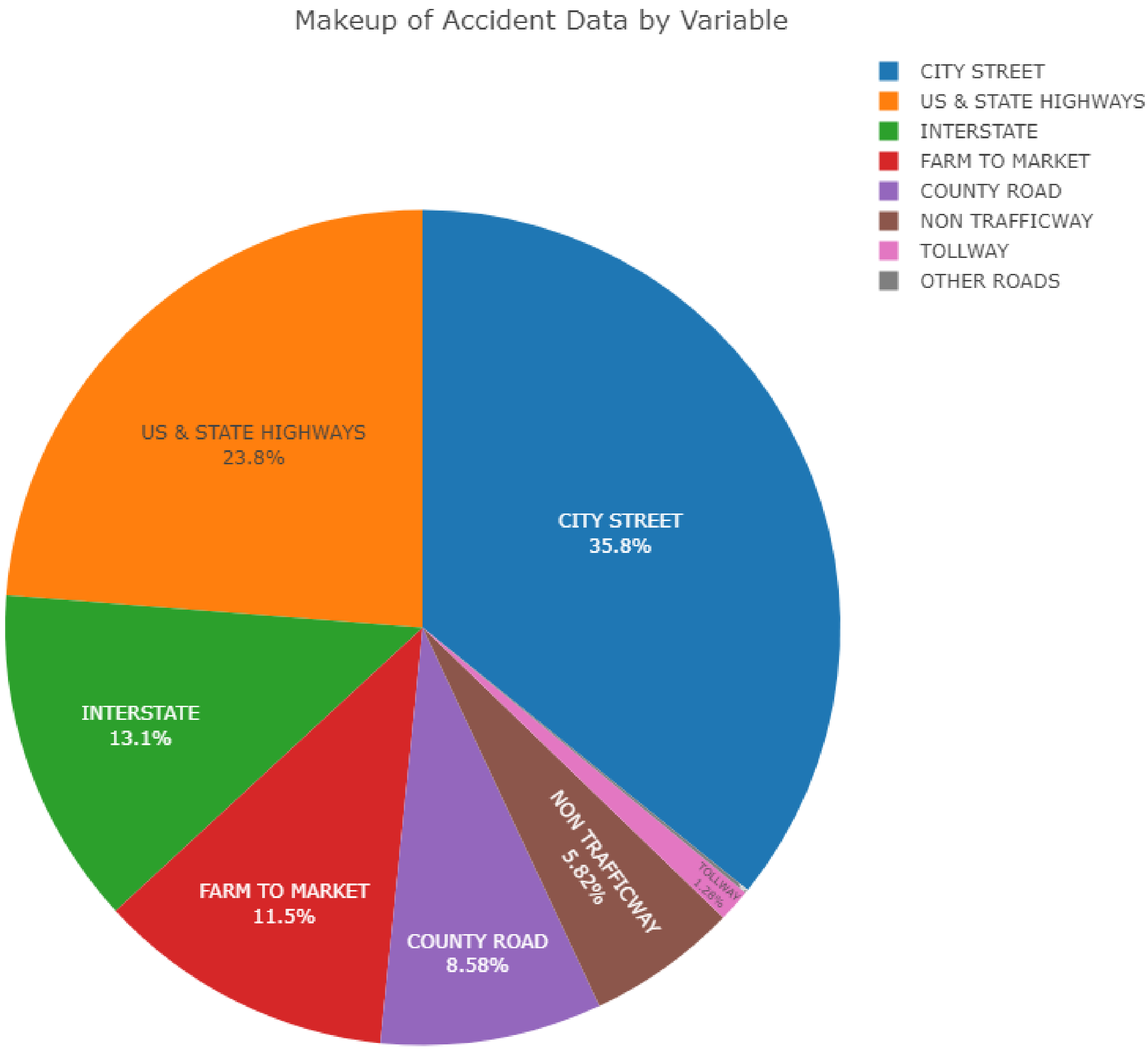
Weather	Frequency
Normal Weather	24682
Dangerous Weather	2403

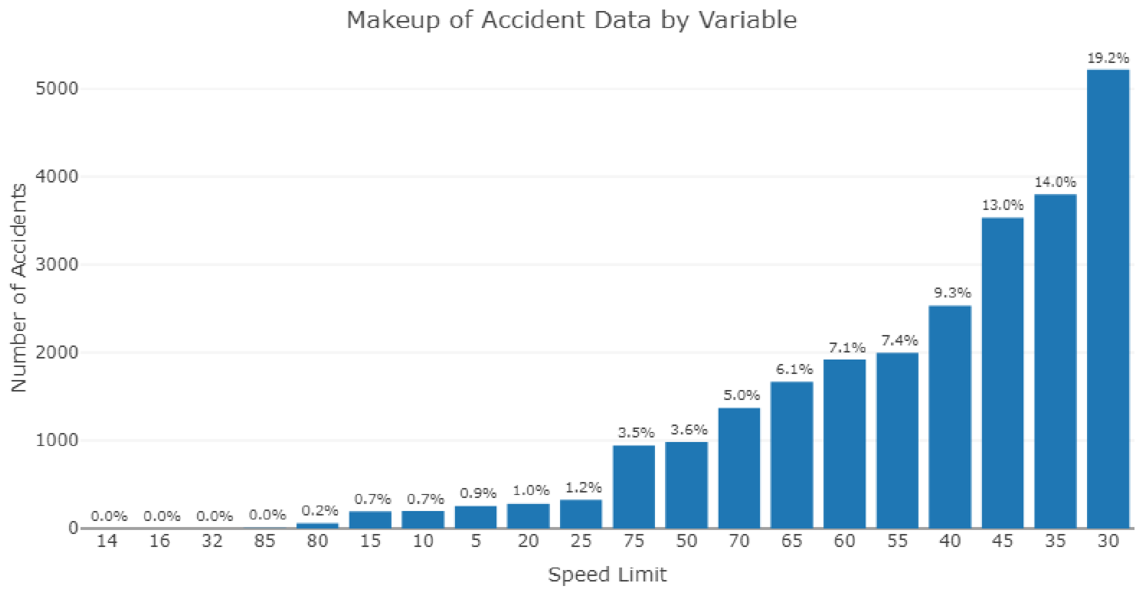
Listed here are some of the tables to showcase the data frequency found after running the app itself. The three most frequently found accident types were those that resulted in no injuries, possible injuries, and minor injuries. In regards to road types, the most common roadway was city streets with highways coming in second. With weather, it was important to show that the vast majority of the accidents were happening under the normal weather conditions, adding a constant to the data.

The importance of including these frequency tables is to show the breadth of the data. It also allows those using the application to quickly and easily see how extensive the data is, and how much they will remove should they want to take away variables during the actual inferential graph. Offering a simple, clean look at the actual numbers will help those looking to quickly understand the data and the meanings of the variables.

Figure 2


Makeup of Accident Data by Variable
N - NOT INJURED
C - POSSIBLE INJURY
B - SUSPECTED MINOR INJURY
A - SUSPECTED SERIOUS INJURY
99 - UNKNOWN
K - FATAL INJURY
C - POSSIBLE INJURY 14.9%
B - SUSPECTED MINOR INJURY 14.2%
A - SUSPECTED SERIOUS INJURY 6.31%
99 - UNKNOWN 4.24%
K - FATAL INJURY 3.89%
N - NOT INJURED 56.5%


Makeup of Accident Data by Variable
CITY STREET
US & STATE HIGHWAYS
INTERSTATE
FARM TO MARKET
COUNTY ROAD
NON TRAFFICWAY
TOLLWAY
OTHER ROADS
US & STATE HIGHWAYS
23.8%
CITY STREET
35.8%
INTERSTATE
13.1%
FARM TO MARKET
11.5%
COUNTY ROAD
8.58%
NON TRAFFICWAY
5.82%
TOLLWAY
1.28%


Makeup of Accident Data by Variable
Number of Accidents
Speed Limit
0
1000
2000
3000
4000
5000
14
16
32
85
80
15
10
5
20
25
75
50
70
65
60
55
40
45
35
30
0.0%
0.0%
0.0%
0.0%
0.2%
0.7%
0.7%
0.9%
1.0%
1.2%
3.5%
3.6%
5.0%
6.1%
7.1%
7.4%
9.3%
13.0%
14.0%
19.2%

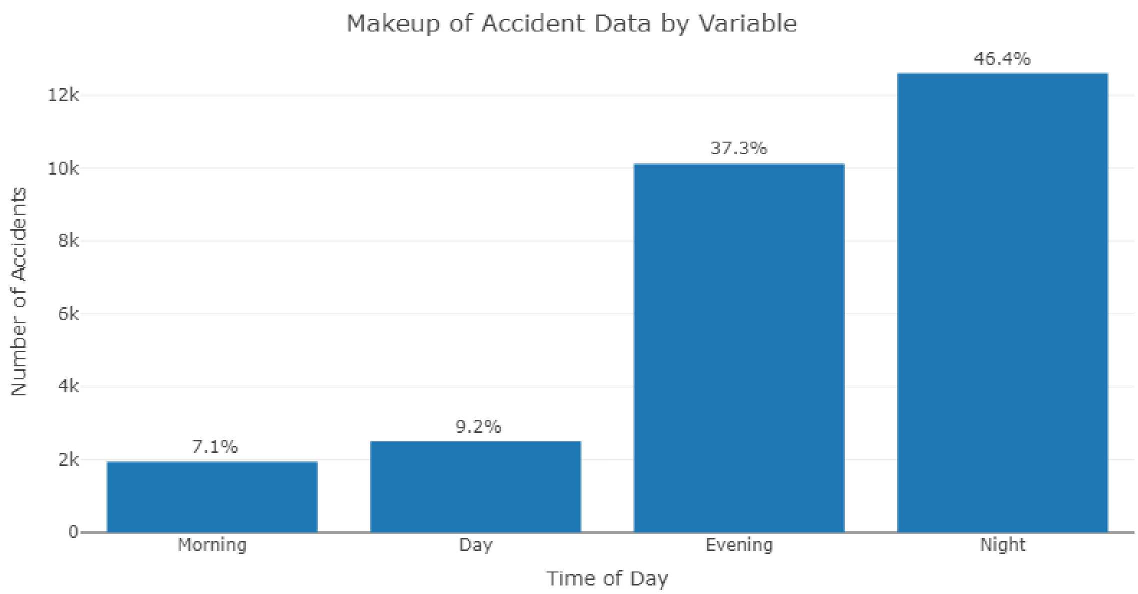
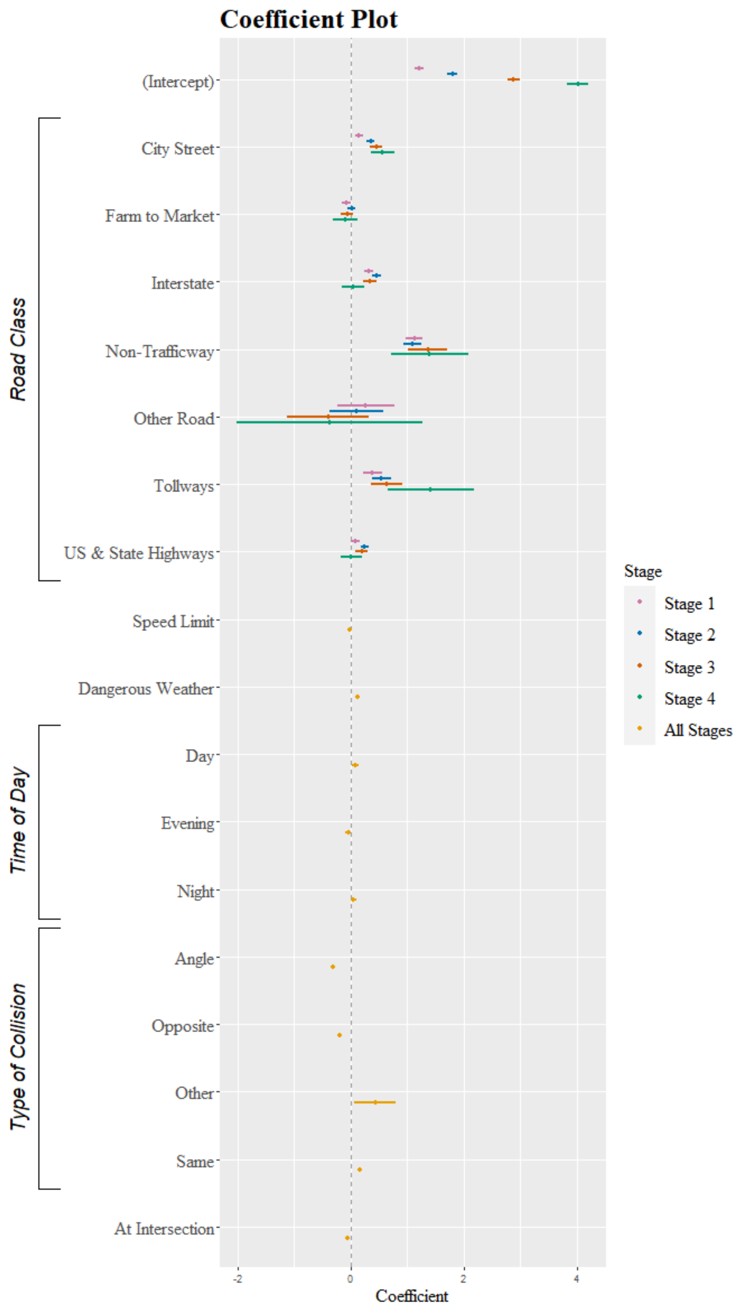

Makeup of Accident Data by Variable
Number of Accidents
Time of Day
0
2k
4k
6k
8k
10k
12k
Morning
Day
Evening
Night
7.1%
9.2%
37.3%
46.4%

Figure Two includes pie charts and bar charts which will allow a more colorful visualization and also a different representation of the data. In Figure One, the data is shown in flat terms with straight numbers. In Figure Two, the data is shown in terms of percentages, to help more clearly visualize the large percentages that these specific variables take up. Both are downloadable and able to be used in whichever way one would want, and both allow for different interpretations and visualizations of the data. Ultimately, Figures One and Two come down to personal preferences. They allow the data to be viewed quickly and succinctly in whichever way most benefits the user.

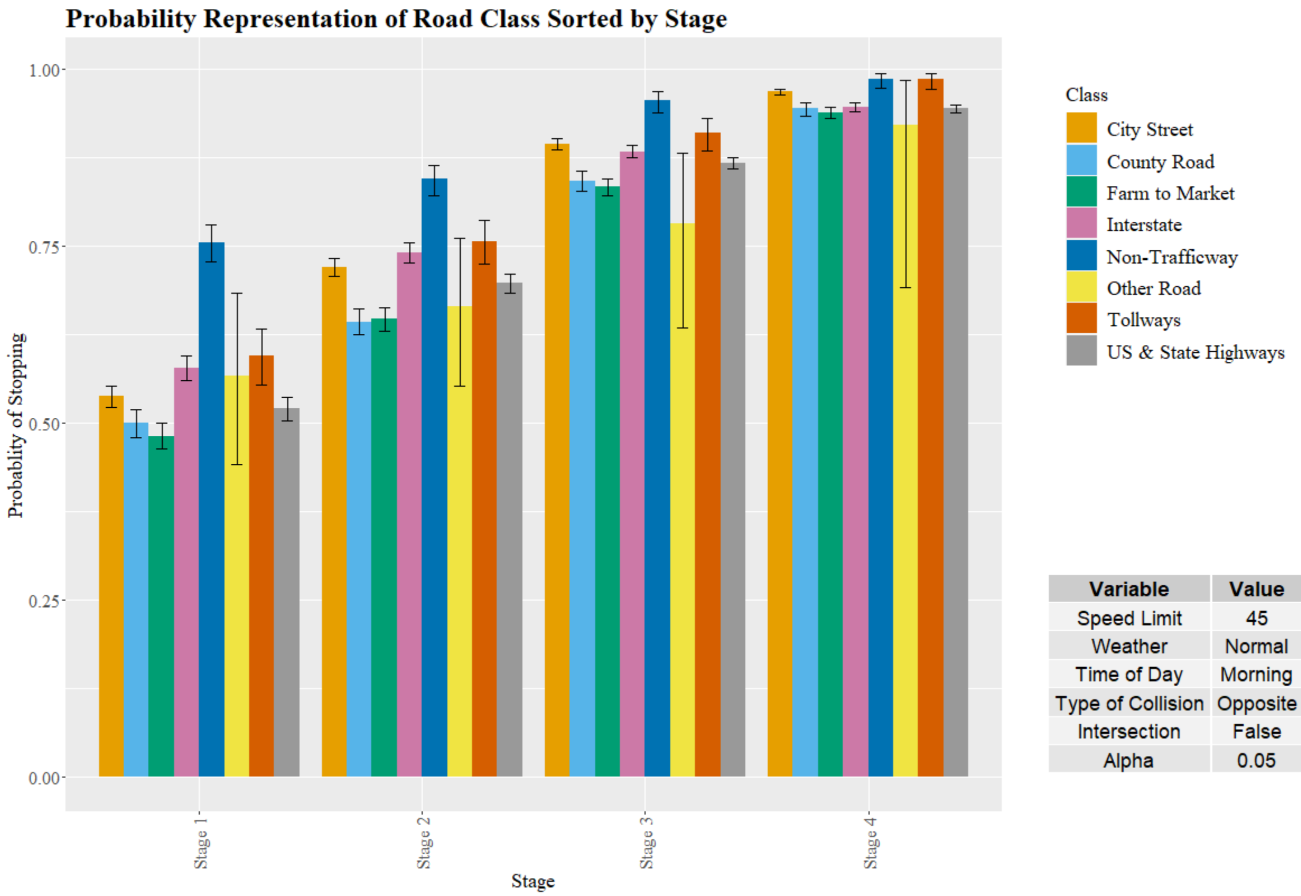
Figure 3



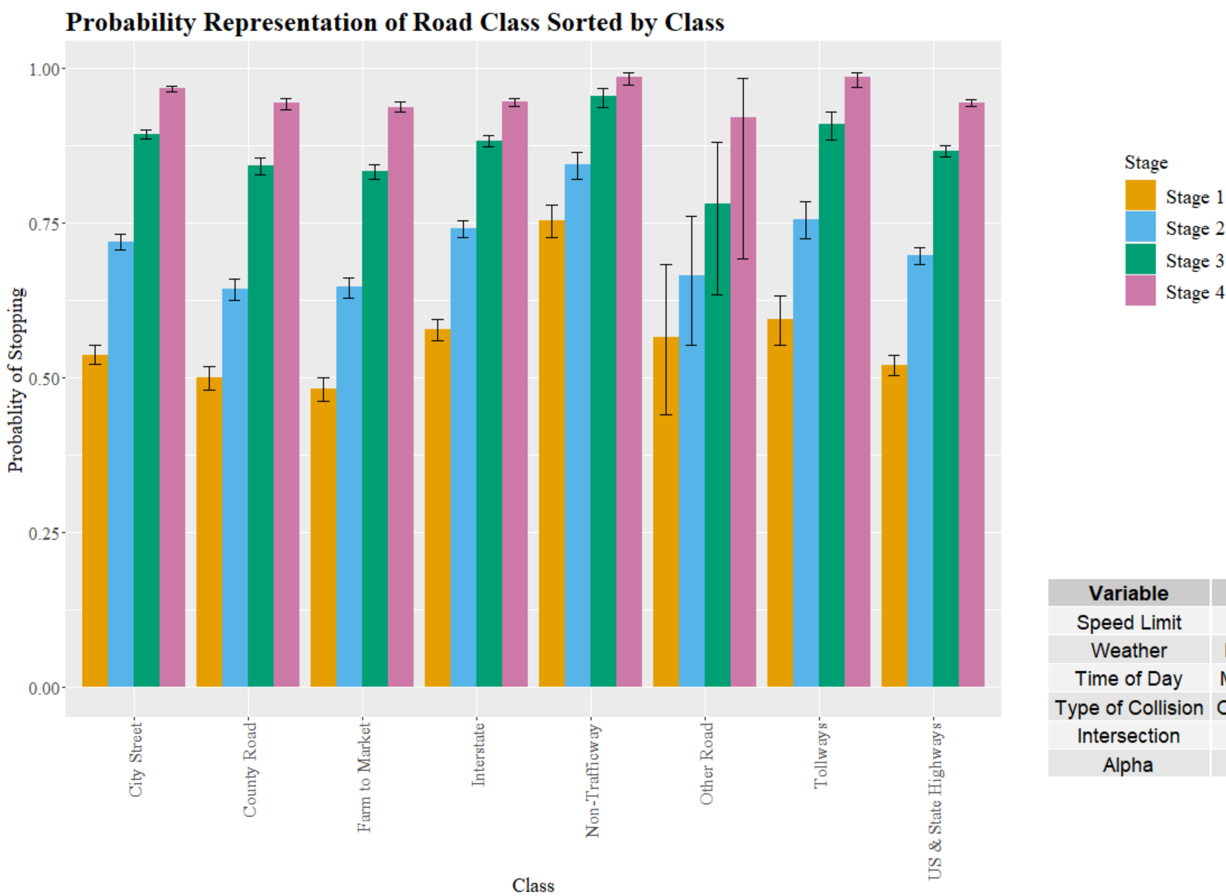
Coefficient Plot
(Intercept)
City Street
Farm to Market
Interstate
Non-Trafficway
Other Road
Tollways
US & State Highways
Speed Limit
Dangerous Weather
Day
Evening
Night
Angle
Opposite
Other
Same
At Intersection
Road Class
Time of Day
Type of Collision
-2
0
2
4
Coefficient
Stage
Stage 1
Stage 2
Stage 3
Stage 4
All Stages

Figure Four is the first of the inferential plots, this being the Coefficient Plot. What this plot does is list each of the coefficients and the effect they have on the log itself. The further away the point is from the dotted line set at zero, the larger of an impact that specific coefficient has on the log. For example, Tollways at Stage Four has a larger impact than Night at All Stages. It is important to note that a larger impact does not inherently mean a positive or negative outcome. This Coefficient Plot is not listing the type of impact, simply the magnitude of the impact itself. The dotted line exists to show when impact is effectively zero, basically serving as the barometer for all the coefficients to be measured against. The closer a coefficient comes to the line on the plot, the more irrelevant they are to the primary conclusion.

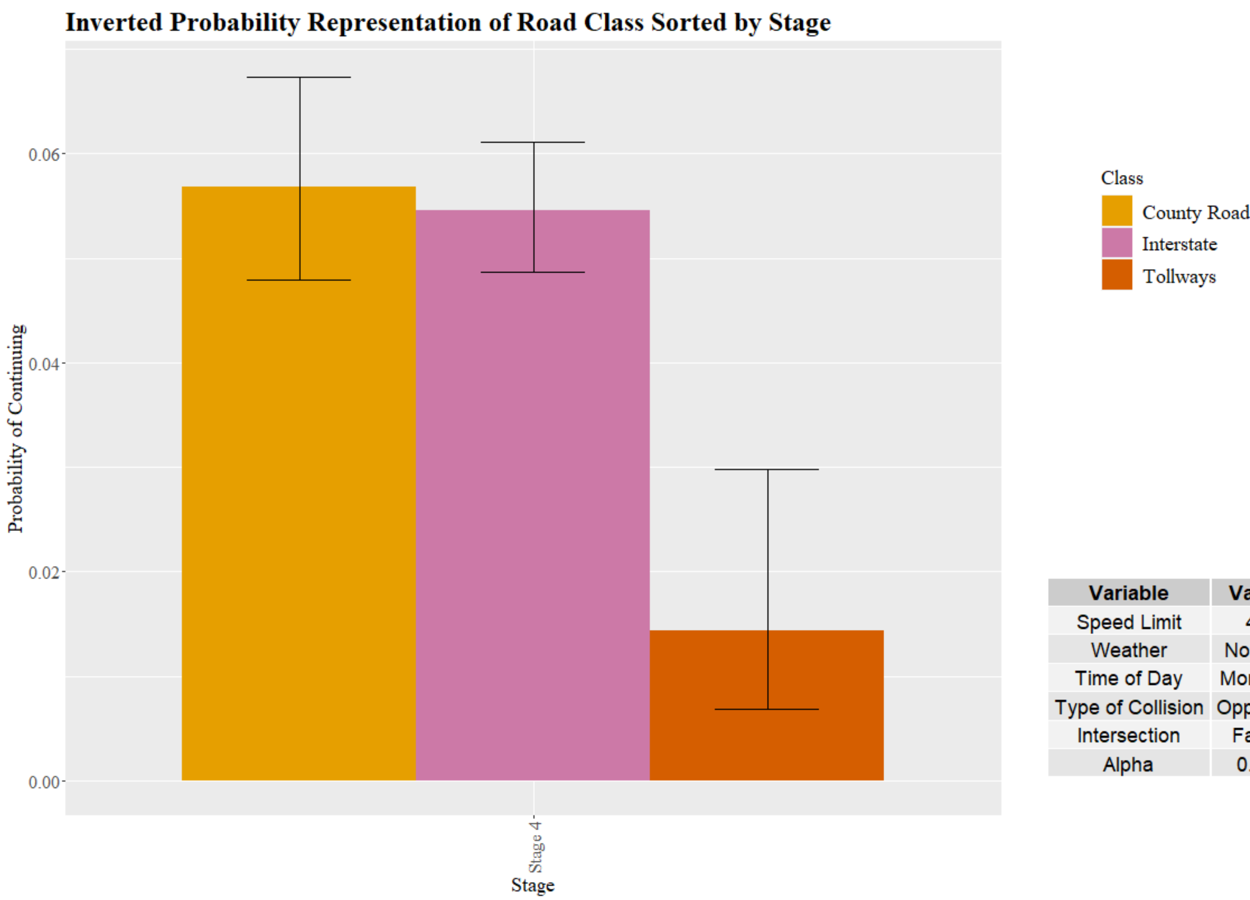
Figure 4



Variable	Value
Speed Limit	45
Weather	Normal
Time of Day	Morning
Type of Collision	Opposite
Intersection	False
Alpha	0.05



Variable	Value
Speed Limit	45
Weather	Normal
Time of Day	Morning
Type of Collision	Opposite
Intersection	False
Alpha	0.05



Variable	Value
Speed Limit	45
Weather	Normal
Time of Day	Morning
Type of Collision	Opposite
Intersection	False
Alpha	0.05

Finally, there is Figure Five and the Inferential plot attached to it, the Probability plot. These charts visualize the probability of a class of roadway to reach a next level. What is meant by that is that the crash severity is effectively put into tiers, with the first tier being the lack of injury, the second tier being minor, all the way up to the top tier which is fatal injury. Each stage shows the probability of a crash to 'stop' at that tier. What this means is that the crash will fail to move onto the next tier, or in this case the crash will not move from being a minor injury crash to a medium injury crash. For example, the probability for an accident which occurred on a non-trafficway to stop moving from stage one to stage two is 75%. In other words, there is only a 25% chance that an accident on a non-trafficway will be worse than no injury or not be a stage 1 crash. Obviously as the stages go on, these probabilities increase. In other words, there are only five possibilities for the results of a crash, and as such the probability of each crash stopping becomes higher and higher as they go up a stage.

Conclusion & Implications

The regressions and visualizations point to multiple conclusions. First, that the higher volume roadways, being highways and tollways, were less likely to suffer serious injury as opposed to the lower volume roadways. Put differently, backroads and the like were more likely to hit the threshold for being a minor, mild, or serious injury than the other roadways. While this theoretically could be considered a volume issue, as highways see a much larger amount of accidents and crashes and as such would have more variance, it is important to note that the regression solves this issue. With the regression, it does not matter the volume so much as the probability. The regression served as an equalizer of sorts, properly showcasing the likelihoods without taking into account volume. Even if the number of crashes were equalized between the two, backroads would still show a higher chance of serious injury than highways or tollways did.

Another conclusion that can be come to thanks to the visualizations and regression is that clearly safety features do make a difference. The primary difference between these roadways are the safety measures put in place upon them. Highways and tollways have significantly more safety measures put in place due to their higher volume of traffic. Even city streets, which see the highest volume, have considerably more safety features and built-in mechanisms to lessen the intensity of crashes. The data does not point directly to safety features being the cause of the discrepancy, though an educated guess makes it incredibly likely that these safety features are the root cause. There is very little else that could cause the difference, as the model already takes into account volume of traffic, conditions of the road, weather, time of day. And even with all of these variables listed and baked into the regression, the ultimate reality is that backroads and lower classifications of roadways suffer more serious injuries on average with their crashes.

The most surprising result ultimately was the fact that fatal accidents were roughly uniform across the board. Regardless of the classification of roadway or any other variable, when it came to fatal accidents it did not really matter what type of road the accident occurred, what time of day, the weather, or anything else. This points to there being another cause for fatal accidents besides all of these outside factors, something that could be touched upon with further research into the field. Perhaps some questions for future researchers could include a look into how drunk drivers with previous sober accidents on their records lead to more reckless future accidents as time has gone on, or perhaps a question of the types of vehicles involved in these accidents play a role more than the roadways themselves. It would not take a tremendous leap to reason that large semis or massive trucks are more likely to cause fatal accidents than those found on city streets or backroads, and the few times those semis are found on backroads in accidents the result may always be fatal. Ultimately the uniformity of fatal crashes regardless of other variables raises more questions than answers.

What this means for future research is fairly easy to understand as well. With the knowledge that crashes on lower roadways are more severe on average, researchers could study exactly what safety features are making such a difference in the numbers. It is entirely possible that, despite the number of variables at play, there is a potential explanation besides road safety features as well that could explain the discrepancy. That would be a topic best suited for another project or set of research however. The scope of this research and regression is simply to look at the connections, not to look at the outright causes. The conclusion that safety features are the reasoning for the discrepancy comes from an overarching view of the situation and less of any actual regression work.

So ultimately, the answer to the research question of whether or not drunk driving is more dangerous on larger roadways or smaller roadways points to smaller roadways being the answer. Despite being roughly equivalent once fatal injuries are reached as the threshold, the fact that minor and medium injuries are much more likely on smaller roadways means that the only true reasoning is safety features.

Nicholas Champagne - Team Coordinator

Nicholas worked primariy on running the initial model, and devloping the skeleton of the shiny applicaiton to send to the other memebers. Nicholas contribution to the shiny application was focused on UI parameter specification, and enabling easy references to other parts of the data, as well as importing proper model specification into the shiny applicaiton. As the team coordinator, Nicholas also lead and directed team members to work together to finish the application

Alden Felix

Alden was in charge of graphical presentation. Alden's contribution to the shiny application was generating all of the graphics shown in the application. Alden also customized all of the graphs and tested them to ensure that the applicaiton worked properly for presentation.

Will Kilcoyne

Will was in charge of write ups. Will focused on finding which graphics conveyed the most amount of information. Will also wrote up the final paper and was the one who created the final conclusions to focus on for the final presentation.

Jim Pan

Jim was in charge of finding, cleaning, and sorting the data to pass onto other memebrs. Jim contribution to the shiny applicaiton was focused on passing data with the correct information to Alden to ensure that graphics could easily be implimented wihout chaning the format of the data very much. Jim also created the presentation.

References

[1] Albalate D (2006). Lowering blood alcohol content levels to save lives: the European experience. Barcelona, Research Institute of Applied Economics (No. CREAP20006-07).

[2] Elder RW et al. (2004). Effectiveness of mass media campaigns for reducing drinking and driving and alcohol-involved crashes: a systematic review. American Journal of Preventive Medicine, 27:57–65.

[3] Peden M et al., eds. (2004). World report on road traffic injury prevention. Geneva, World Health Organization.

[4] Rezapour, Wulff, S. S., & Ksaibati, K. (2019). Examination of the severity of two-lane highway traffic barrier crashes using the mixed logit model. Journal of Safety Research, 70, 223–232. https://doi.org/10.1016/j.jsr.2019.07.010

[5] Sethi, D. & Mitis, Francesco. (2012). Alcohol and Road Safety Policy in Europe: Successes and Failures. Injury Prevention. 18. A31-A32. 10.1136/injuryprev-2012-040580c.28.

[6] WHO (2007). Drinking and driving. A road safety manual for decision-makers and practitioners. Geneva, World Health Organization, Global Road Safety Partnership.

[7] Zobeck TS, Williams GD (1994). Evaluation synthesis of the impacts of DWI laws and enforcement methods: final report. Rockville, MD, National Institute on Alcohol Abuse and Alcoholism, Office of Policy Analysis.