USDOT Tier 1
University Transportation Center
on Improving Rail Transportation
Infrastructure Sustainability and Durability

Final Report UD-11

HIGHWAY-RAIL CROSSING ACCIDENT ANALYSIS USING BAYESIAN BELIEF NETWORKS

By

Tyler Bernstein
Undergraduate Research Assistant
Department of Civil & Environmental Engineering
University of Delaware

Joseph Palese, PhD, MBA, PE
Senior Scientist
Department of Civil & Environmental Engineering
University of Delaware
palesezt@udel.edu

Allan Zarembski PhD, PE, FASME, Hon. Mbr. AREMA
Professor and Director of Railroad Engineering and Safety Program
Department of Civil & Environmental Engineering
University of Delaware
dramz@udel.edu

September 10, 2021

Grant Number: 69A3551747132

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

ABSTRACT

The rate of highway-rail grade crossing collisions has steadily increased year over year since 2009, after a decades long period of decline, beginning in 1972. Several models exist that predict the likelihood and number of collisions at crossings. These models have decreased in accuracy as they have aged. This research employed Bayesian statistics and its graphical representation, Bayesian belief networks, to develop a new model that predicts the probability of a collision at a railway/highway grade crossing, as a function of known characteristics, readily available through open source data. The final model was found to be a relatively accurate predictor of collision likelihood but showed deficiencies that prevent its use in practical application. Despite these deficiencies, utilizing Bayesian statistics remains a promising method of predicting collision likelihood at a grade crossing, and further study into this application should be conducted.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Highway-rail grade crossings occur when a railroad intersects a roadway on the same plane, commonly referred to as "at-grade". A typical grade crossing for a two-lane road and single railway track with two quadrant gate protection is shown in

Figure 1. These crossings are a common form of infrastructure throughout the United States, located in urban, rural, and suburban communities in every state. More than 200,000 grade crossing exist in the US. Table 1 documents the number of crossings in all states and the District of Columbia according to the Federal Railroad Administration (FRA).

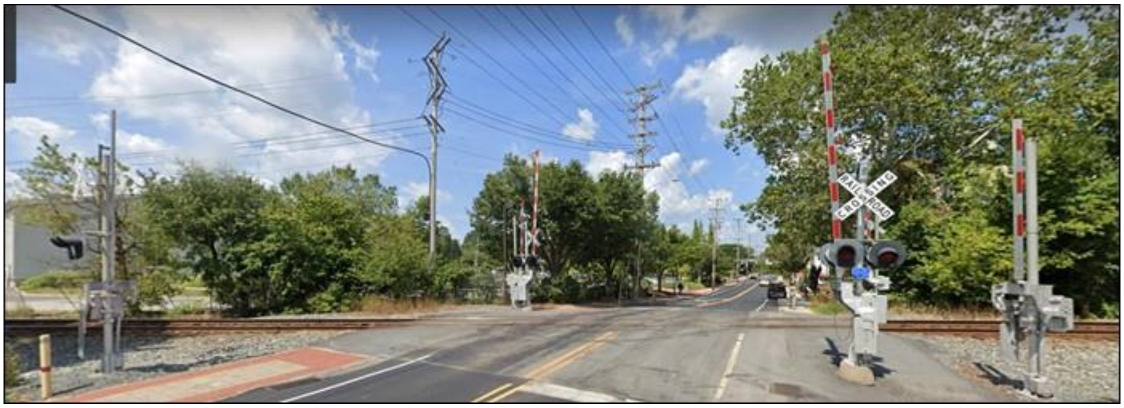


Figure 1: Typical Highway-Rail Grade Crossing with Gates

Table 1: Number of Crossings in Every State

State	# of Crossings	State	# of Crossings
District Of Columbia	6	Mississippi	3917
Hawaii	8	Kentucky	4104
Rhode Island	108	Virginia	4191
Alaska	237	Alabama	4309
Delaware	405	North Dakota	4359
New Hampshire	523	Tennessee	4463
Nevada	562	Washington	4665
Connecticut	606	Florida	4731
Vermont	850	Nebraska	4793
Arizona	1120	Oklahoma	4926
Wyoming	1154	Louisiana	5039
Massachusetts	1162	New York	5159
New Mexico	1223	Missouri	5338
Maryland	1253	Wisconsin	5964
Utah	1254	Pennsylvania	6097
Maine	1550	Minnesota	6344

New Jersey	1980	Iowa	6522
Idaho	2226	Michigan	6594
Colorado	2765	North Carolina	6779
South Dakota	2786	Kansas	7177
Montana	3026	Georgia	7222
West Virginia	3079	Indiana	7478
Oregon	3548	California	8429
Arkansas	3579	Ohio	8576
South Carolina	3789	Illinois	10891
		Texas	13863
Total			**200,729**

Though common, highway-rail grade crossings also present serious safety challenges. It is often difficult for motorists to see oncoming rail vehicles and impatient motorists will often circumvent crossing protection. Train operators that see an obstruction cannot come to a complete stop to avoid collision; due to the time it takes to stop a moving train. Additionally, Rail/Highway vehicle collisions can result in extensive property damage, environmental hazards, and loss of life.

Figure 2 shows the number of total accidents by U.S state since the year 2000. States with higher numbers of accidents have deeper blue coloring, while those with lower accident rates have lighter blue coloring. The state of Texas, with 5,431 incidents in the past 20 years has the greatest number of collision by far. This amounts to more than 270 incidents per year, and 0.02 incidents per crossing per year, or a 2% risk of an incident occurring.

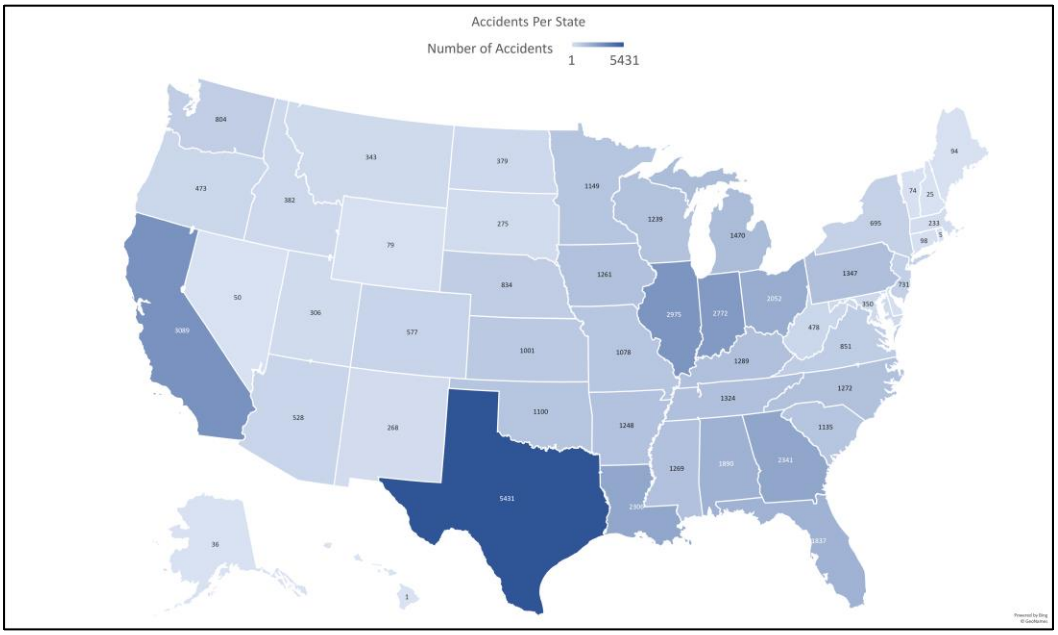


Figure 2: Number of Accidents by State Since 2000.

In addition to the danger they pose, highway-rail crossings also account for a sizable portion of all railroad incidents nationally. Nearly 19% of all train incidents/accidents involve highway rail crossings (Federal Railroad Administration Office of Safety Analysis). Due to the large amount of incidents involving this single piece of infrastructure, there has been considerable investment in methods of reducing incident occurrence. Operation Lifesaver, a public awareness campaign developed in 1972, sought to improve public awareness and safety at highway-rail grade crossings (Horton 2009). Since the creation of operation lifesaver, highway-rail crossing incidents saw deep declines for nearly 40 years, as shown in

Figure 3. Despite these reductions, the total incidents per year began to plateau in 2009 and even showed a slight increase. Though Operation Lifesaver was highly successful in reducing collision rates, it alone could not eliminate all instances of collision.

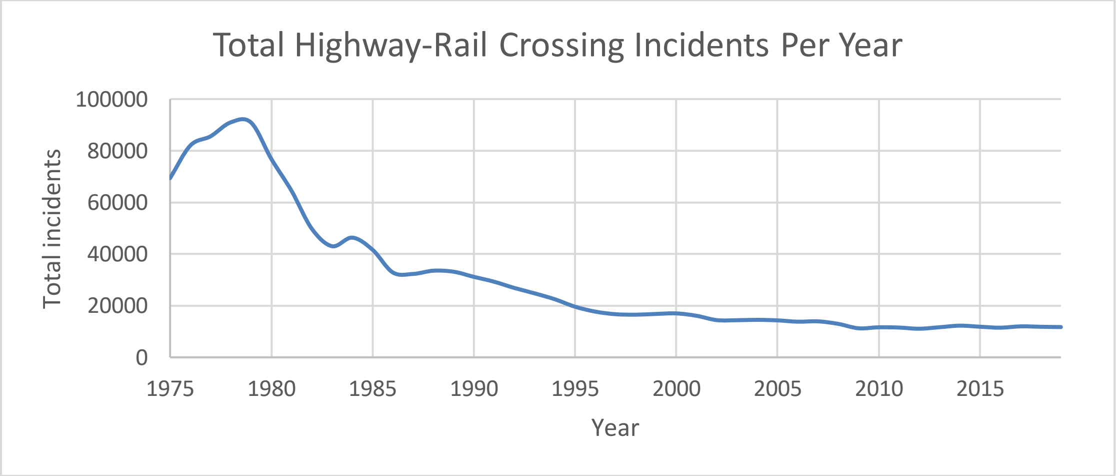


Figure 3: Total Rail-Highway Incidents Per Year

One means of reducing collisions at a crossing is to either close it or grade separate, have the highway vehicle pass over or under the track. Though effective from a crash prevention perspective, these options are not always feasible, since they require either large capital investments in the form of bridge/underpass construction, or restricted mobility from one side of the tracks to the other. As such, in environments where at-grade crossings are essential, it is critical that those responsible for them can assess the risk of collision and ensure protection is at an acceptable level. Capital improvements that increase protection involve the installation of crossbucks, stop signs, pavement markings, bells, flashing lights, train-activated gates, and other physical infrastructure.

The objective of this research was to analyze grade crossings, their properties, and the collisions that occur at them from a risk perspective. The product of this research is a risk-model that predicts the probability of a collision occurring at a grade crossing. The model’s outputs were calculated using principles of Bayesian statistics and conditional probabilities.

Developing this model consisted of analyzing open source data; crossing incident inventory, and crossing property inventory data retrieved from the FRA Office of Safety (Federal Railroad Administration Office Of Safety Analysis). The two inventory databases were consolidated and cleaned for analysis. Distributions of variables that were hypothesized to influence collision likelihood, such as daily train-movements and annual average daily traffic (AADT), were created to develop an understanding of the data and how best to further proceed with analysis. After this initial analysis, secondary analyses were conducted to show clear trends in collision occurrence with respect to crossing properties. These trends were then incorporated into Bayesian network models that sought to predict collision occurrence using conditional probabilities and Bayesian statistics. Though the models' efficacy was limited, clear trends were identified between collision occurrence and protection type.

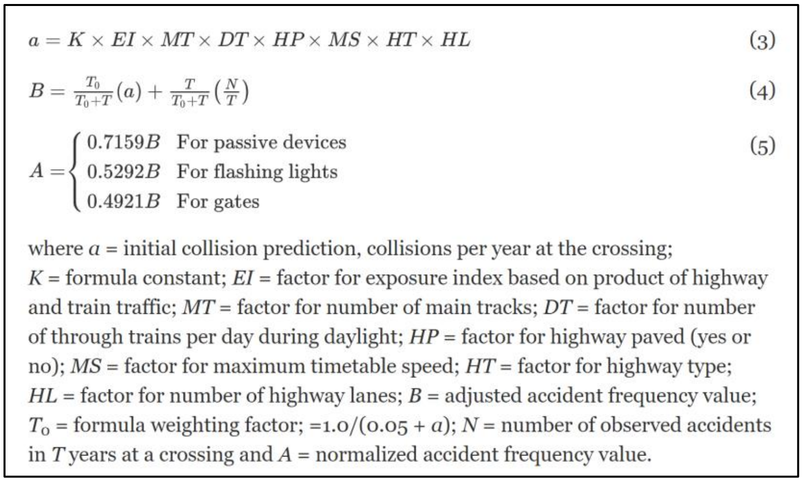
Upon completion of this research, there remains substantial room for further analysis of grade crossing collisions using Bayesian statistics. The issue of grade crossing accidents will likely grow in importance as states, regional governments and private industry expand light rail and interurban passenger services, whose rights of way tend to have many at-grade crossings. As a result, ensuring safety along these rights of way will increasingly become a matter of preventing loss of life, as more passenger services, such as Florida's Brightline higher speed passenger rail service, come online.

LITERATURE REVIEW

Highway-Rail Grade Crossings

Samantha Chadwick and her team at the University of Illinois performed research relating to the safety challenges of at-grade crossings along shared high-speed rail and freight operations. In her study, "Highway-rail grade crossing safety challenges for shared operations of high-speed passenger and heavy freight rail in the U.S.", she provides an in-depth analysis of the literature currently available to model collision likelihood and overall safety at rail-highway crossings. These models are often one of two types: relative or absolute. Relative formulae analyze current conditions at grade crossings to develop a comparative ranking of safety, from least dangerous crossing to most dangerous (Chadwick et al. 2014). Absolute formulae analyze current conditions to predict the number of collisions that will occur at a given crossing.

There have been several models produced. The most widely utilized model nationally is the U.S Department of Transportation Accident Prediction Model (APS), developed in the 1980s by Faghri and Demetsky, then later expanded on by Austin and Carson (Chadwick et al., 2014). This model predicts the expected number of yearly collisions at crossings, from factors such as train volume, highway type and crossing device; it is an absolute formula (Faghri & Demetsky, 1986). Both Chadwick et al. and Faghri & Demetsky's works look to understand the overall safety of crossings and how to predict/improve the rate of collision occurrence. The APS model is shown in
Figure 4.



$$a = K \times EI \times MT \times DT \times HP \times MS \times HT \times HL \quad (3)$$

$$B = \frac{T_0}{T_0+T}(a) + \frac{T}{T_0+T}\left(\frac{N}{T}\right) \quad (4)$$

$$A = \begin{cases} 0.7159B & \text{For passive devices} \\ 0.5292B & \text{For flashing lights} \\ 0.4921B & \text{For gates} \end{cases} \quad (5)$$

where a = initial collision prediction, collisions per year at the crossing; K = formula constant; EI = factor for exposure index based on product of highway and train traffic; MT = factor for number of main tracks; DT = factor for number of through trains per day during daylight; HP = factor for highway paved (yes or no); MS = factor for maximum timetable speed; HT = factor for highway type; HL = factor for number of highway lanes; B = adjusted accident frequency value; T_0 = formula weighting factor; =1.0/(0.05 + a); N = number of observed accidents in T years at a crossing and A = normalized accident frequency value.

Figure 4: APS Model (Chadwick, Zhou, & Saat, 2014)

Though the current APS model is the most widely used, its accuracy has declined over the 25 years since it was initially developed. Brod & Gillen (2020) have identified this decline in accuracy and, in the process of developing a new model, explain why the decline has occurred. According to their study, while the FRA updates normalized accident frequency values (A) based on new accident data, these updates do not consider environmental, technological and policy changes that influence accident prediction. Such changes include increased freight train lengths and increased intermodal traffic (Brod & Gillen, 2020). These changes present new challenges in accident prediction. The increase in train lengths and intermodal traffic result in longer and more frequent wait times for automobiles at crossings, which subsequently incentivizes more “risky behavior,” driving around closed gates to avoid the wait (Brod & Gillen, 2020). Additionally, since the APS model was developed, the presence of ridesharing vehicles, delivery vehicles, and SUVs on the road has increased substantially. Brod and Gillen (2020) speculate that these changes contributed to the decline in accuracy of the APS model.

In addition to the APS model, several state departments of transportation (DOT) have developed their own prediction models. The New Hampshire Hazard Index Model, shown in Equation 1, was among the first developed by a state entity and continues to be widely used due to its simplicity (Chadwick et al. 2014). Unlike the APS, the New Hampshire Model is a relative formula, which ranks the crossings most in need of upgrades. It has been shown to have similar accuracy to other, more complex relative models (Chadwick et al., 2014). Faghri and Demestky (1998) conducted a survey of 45 state DOTs and found that there were currently 13 separate models in use. Also from the survey, roughly a third of DOTs prefer to use their own models and a third use the APS (Abioye et al., 2020).

$$Hazard\ Index = VTP_f \quad (1)$$

where V = average 24-hour highway traffic volume, T = average 24-hour train volume and P_f = protection factor (0.1 for gates; 0.6 for flashing lights; 1.0 for signs only) (Chadwick et al., 2014)

In addition to state DOT models and the APS, the Peabody-Dimmick absolute formula, Equation 2, predicts the number of collisions at a crossing over 5 years (Chadwick et al., 2014).

$$A_5 = 1.28\frac{(V^{0.170})(T^{0.151})}{P^{0.171}} + K \tag{2}$$

where A_5 = 5-year accident count, V = AADT, T = daily train-movements, P = protection coefficient, and K = a smoothing parameter (Chadwick et al., 2014)

There are many prediction models that have been developed, each with their own strengths and weaknesses in terms of accuracy, simplicity, and application. While the APS model remains the most widely used, it continues to decline in accuracy. Thus, the research undertaken herein investigates new modelling techniques to develop a model that can potentially replace the APS. The model proposed in this research takes advantage of Bayesian statistics and its principles of conditional probability to determine the likelihood of collision occurrence, given known variables.

Bayesian Statistics

Bayesian statistics is a method of applying conditional probability to statistical problems. Koch provides an introductory explanation of Bayesian statistics, its mathematics, and its application. It is different from traditional statistics in that it is derived from Bayes Theorem, and can be easily deployed to test hypotheses and develop confidence regions for unknown parameters. Such hypotheses and confidence regions are developed via probability density functions for unknown parameters via application of Bayes Theorem (Koch, 2007). As a result of the ease with which hypothesis testing and confidence region computation can be performed via Bayesian Statistics, its use has spread rapidly since its initial development (Koch, 2007) .

Both Bayesian statistics and traditional statistics (frequentist) are foundationally derived from probability, whose component makeup is uncertainty and plausibility. Uncertainty is the degree to which the outcome of an event or statement is known, and plausibility is the likelihood that an event or statement will occur and is considered an expression of probability. Traditional statistics computes probability via analysis of random events, such as the results of random experimentation. In comparison to traditional statistics, Bayesian statistics is not relegated to computing the probability of random events, but also the probability of statements or propositions, which can be more general than in traditional statistics (Koch, 2007). Several rules of probability are combined to develop Bayes theorem and Bayesian networks.

The communitive, associative, distributive and DeMorgan’s laws of probability (Koch, 2007) are shown below, respectively. These laws, which are an algebraic application of probability, are used to derive Bayes Theorem.

$$A + B = B + A \quad \text{and} \quad AB = BA \tag{3}$$

$$(A + B) + C = A + (B + C) \quad \text{and} \quad (AB)C = A(BC) \tag{4}$$

$$(\overline{A+B}) = \bar{A}\bar{B} \quad \text{and} \quad \overline{AB} = \bar{A}+\bar{B} \tag{5}$$

Conditional probability allows for more generalized propositions and is the basis for Bayesian statistics. In conditional probability, a proposition is dependent on the results of an additional proposition. Conditional probability statements are denoted as *P(A|B)*, or the probability of event *A* given event *B*.

The product rule and sum rule of probability, adapted for conditional probabilities as used in Bayesian statistics, are shown in Equations 6 and 7 according to Koch (2007), respectively. The product rule derives a relation between statement *A* and statement *AB*, given that statement *C* is true. Note that *P(S|C)* is the probability of the sure statement. The sum rule provides a relationship between the probability of statement *A* and $\bar{A}$ (Koch, 2007).

$$\begin{gathered} P(AB|C) = P(A|C)P(B|AC) = P(B|C)P(A|BC) \\ P(S|C) = 1 \end{gathered} \tag{6}$$

$$P(A|C) + P(\bar{A}|C) = 1\ . \tag{7}$$

Using the product rule and sum rule, a generalized sum rule is developed in Equation 8 which is used to derive Bayes theorem. The generalized sum rule, Equation 9, denotes the probability of *A* + *B* given *C*, in relation to *A* given *C*, *B* given *C*, and the joint probability *AB*, given *C*.

$$\begin{aligned} P(A+B|C) &= P(\overline{\bar{A}\bar{B}}|C) = 1 - P(\bar{A}\bar{B}|C) = 1 - P(\bar{A}|C)P(\bar{B}|\bar{A}C) \\ &= 1 - P(\bar{A}|C)[1 - P(B|\bar{A}C)] = P(A|C) + P(\bar{A}B|C) \\ &= P(A|C) + P(B|C)P(\bar{A}|BC) = P(A|C) + P(B|C)[1 - P(A|BC)]\ . \end{aligned} \tag{8}$$

$$P(A+B|C) = P(A|C) + P(B|C) - P(AB|C)\ . \tag{9}$$

The chain rule of probability, Equation 10 expresses the formulation of conditional joint probability. The chain rule in Equation (10) is key for the development of Bayesian networks.

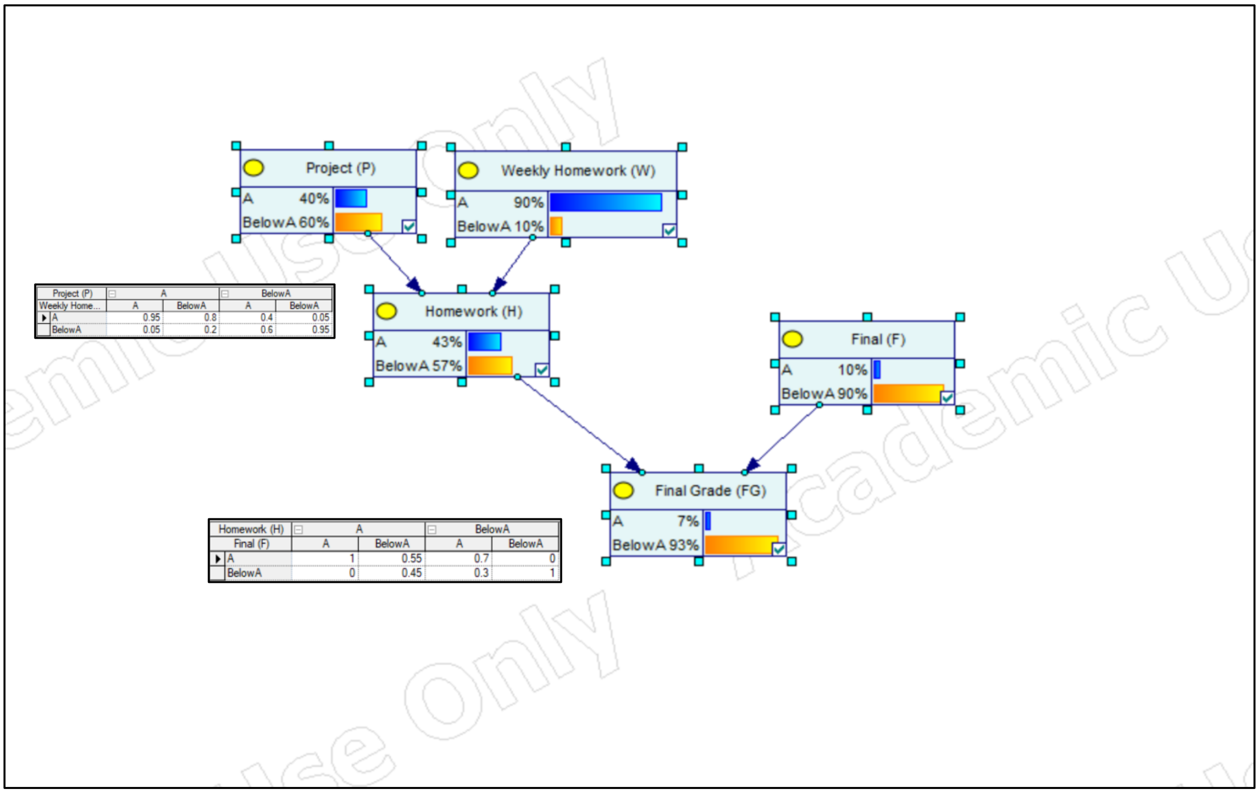
$$P(A_1A_2A_3|C) = P(A_3|A_1A_2C)P(A_1A_2|C) \tag{10}$$

$$\boldsymbol{P(A|B) = \frac{P(B|A)\cdot P(A)}{P(B)}} \tag{11}$$

Bayes Theorem, shows that the probability of hypothesis *A* before receiving new information, *P(A)* and the probability of hypothesis *A* after receiving new information, the posterior probability, *P(A|B)*, are relationally dependent. Bayes theorem also does not indicate that *P(A)* and *P(B)* cannot be independent of one another. For example, if *P(A|B)* is equal to *P(A)*, then the events are independent of one another and event *B* provides no useful information about the likelihood of event *A* occurring. As such, the closer *P(A)* and *P(A|B)* are in value, the less influence prior information has on the posterior probability.

This research takes advantage of Bayesian statistics and Bayes Theorem to develop a Bayesian network that predicts the likelihood of a rail/highway vehicle collision occurring at a grade crossing, given prior information about that crossing.

Bayesian networks were first developed by Pearl and Russel at the University of California, Los Angeles. They are a graphical method of representing joint probability distributions that allow for simplistic modelling of complex, interconnected systems (Pearl & Russel, 2000). Bayesian networks are "directed acyclic graphs" that contain nodes and linkages, where the "nodes represent variables of interest, and the links represent informational or causal dependencies" whose level of influence "is represented by conditional probabilities" that are shown for each parent-child cluster (Pearl & Russel, 2000) . The mathematical determination of those conditional probabilities are factorizations of joint probability distributions. A sample Bayesian network and its associated conditional probability formulae are shown in
Figure 5.



Project (P)	A		BelowA	
Weekly Home...	A	BelowA	A	BelowA
▶ A	0.95	0.8	0.4	0.05
BelowA	0.05	0.2	0.6	0.95

Homework (H)	A		BelowA	
Final (F)	A	BelowA	A	BelowA
▶ A	1	0.55	0.7	0
BelowA	0	0.45	0.3	1

Figure 5: Sample Bayesian Network

The sample Bayesian network displayed in Figure 5 is a graphical representation of the conditional probability that a student will earn an A in a course, given their prior performance in that course. The values used in this sample are for demonstration purposes only and do not reflect any collected data. This network shows that the student's final grade is dependent on their scores on the final and homework grade. Additionally, the homework grade is dependent on the project and weekly homework grades. The tables adjacent to the child nodes Final Grade and Homework display the conditional probabilities of occurrence given the information from their respective parent nodes. The sample network was developed using GeNIe driver software, a product of BayesFusion, LLC, and can calculate probability distributions via factorization (GeNIe modeler user manual (2020). A sample joint probability formula for the node, Final Grade (FG) is shown in Equation 12.

$$P(P, W, F, H, FG) = P(FG|H, F) \cdot P(H|P, HW) \cdot P(F) \cdot P(P) \cdot P(W) \qquad (12)$$

The GeNIe driver can update its joint posterior predictions for nodes as new information becomes available by taking advantage of the same factorization as Equation (12).

A sample GeNIe driver calculation is shown when the student already knows their performance on all assignments excluding the final in Figure 6. Using factorization and the product rule of probability the model predicts that given the student's performance on all assignments other than the final was an A, it is likely that they will receive an A for the final grade.

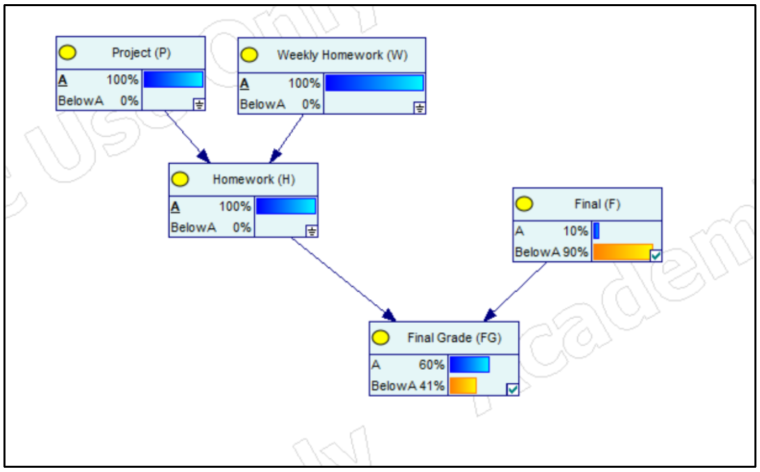


Figure 6: Updated Joint Probabilities for Sample Network

This research takes advantage of similar modeling and Bayesian techniques shown in the sample model. The primary difference between the sample model and the final models developed are that the final models' predictions will be developed using readily available crossing inventory and collision data.

METHODOLOGY

Data Collection

Consolidation of Grade Crossing and Collision Databases

The FRA Office of Safety provides open-source access to their grade crossing inventory and rail-highway grade crossing collision database (Federal Railroad Administration Office of Safety Analysis). The current crossing data for each state, last updated in July of 2020, was downloaded as a comma separated value (CSV) text file. The highway-rail accident data reports (structure 6180.57) were downloaded for each year from 2000 to 2020. The two databases were consolidated into a single Microsoft Access file (Microsoft Access 2019). The yearly collision reports were appended together to form a single collision incident table over 20 years. The crossing inventory and collision database were queried together to develop data regarding crossing collisions and the properties of the associated crossing.

Selected Variables and Data Cleaning

The parameters in the database for each crossing used for analysis are shown in Table 2.

Table 2: Parameter Descriptions

Parameter	Description
AADT	Average Annual Daily Auto-Traffic
DayThru	Daily Through Trains
NightThru	Nightly Through Trains
TypXing	Type of Crossing: 2 = Private 3 = Public
PosXing	Position of Crossing 1 = At Grade 2 = RR Under 3 = RR Over
CrossingClosed	Is Crossing Closed: “Yes” = Closed “No” = Open
WdCode	Protection Type 1 = No signs or signals 2 = Other signs or signals 3 = Crossbucks 4 = Stop signs 5 = Special Active Warning Devices 6 = Highway traffic signals, wigwags, bells, or other activated 7 = Flashing lights 8 = All other Gates 9 = Four Quad (full barrier) Gates
HwyClassrdtpID	Highway Type

	11 = Interstate 12 = Other Freeways and Expressways 13 = Other Principal Arterial 16 = Minor Arterial 17 = Major Collector 18 = Minor Collector 19 = Local
DevelTypID	Land Use 11 = Open Space 12 = Residential 13 = Commercial 14 = Industrial 15 = Institutional 16 = Farm 17 = Recreational 18 = RR Yard

Crossings in the database that report 0 AADT and 0 total daily thru trains were removed from analysis, as this meant data was unavailable and would skew the modeling results. Private crossings, above or below grade crossings, and closed crossings were also removed from the analysis. Additionally, crossing entries that contain N/A values for any of the listed parameters were removed from analysis (See Appendix A). This resulted in 78,403 crossings from the inventory, with a corresponding 30,744 total number of accidents over the 20-year history.

APPROACH AND RESULTS

AADT and Train-Movement Distributions

Annual average daily highway traffic (AADT) and the number of daily train movements at crossings were expected to significantly influence the frequency of collision, because as more highway vehicles and trains pass over a crossing, the more possibilities exist for them to interact or collide. As a result, distributions of these two parameters were created to determine what patterns they follow. Both train-movements and AADT were observed to follow exponential distributions, as shown in
Figure 7 and
Figure 8, and exponential equations were overlayed. With the understanding that these parameters follow exponential distributions, their influence on collision occurrence was investigated.

The distribution of Train Movements and AADT at crossings that have had at least one collision were compared to the distribution of those parameters for all crossings in the database, shown again in
Figure 7 and
Figure 8. These figures show that crossings with at least one collision have higher average through trains and a higher average AADT. These findings align with expectations, allowing train-movements and AADT to be utilized as normalizing factors in the secondary exploratory data analysis.

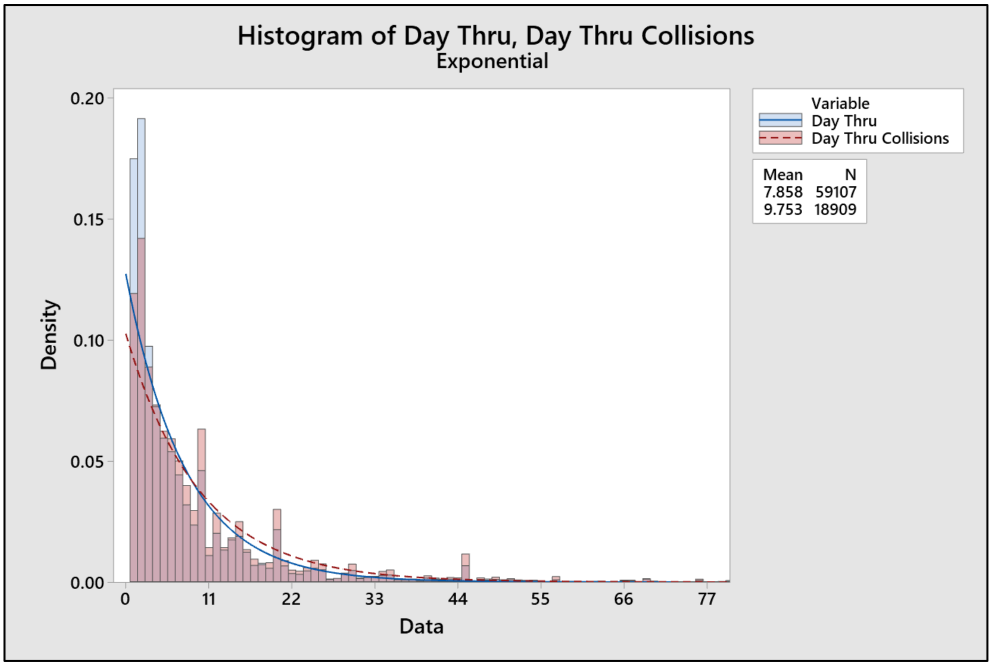


Figure 7: Distribution of Daily Through Trains at All Crossings in the Database and Distribution of Daily Through Trains Crossings in the Database that Have Had a Collision Since 2000

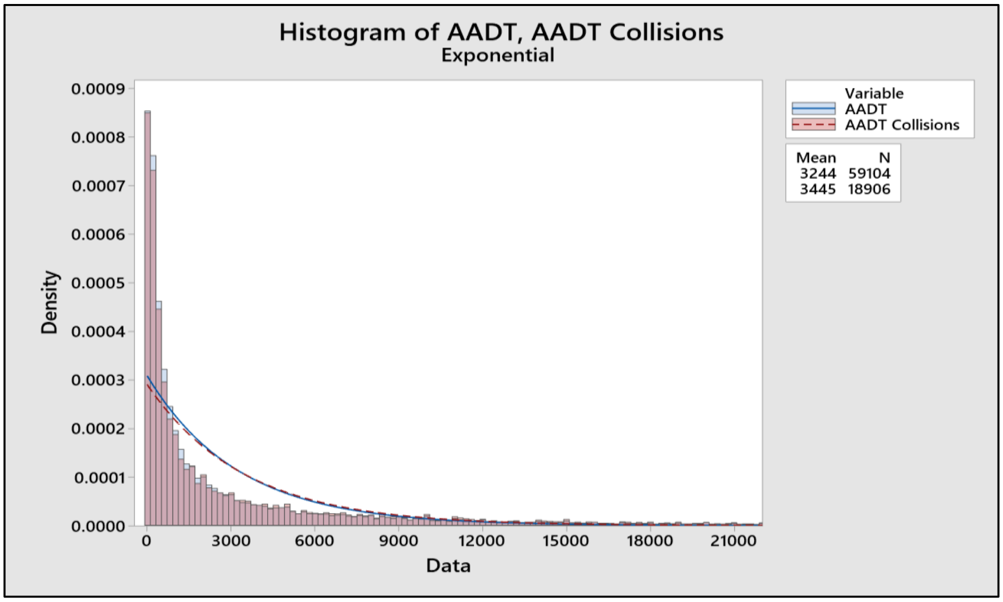


Figure 8: Distribution of AADT at All Crossings in the Database and Distribution of AADT at Crossings in the Database that Have Had a Collision Since 2000

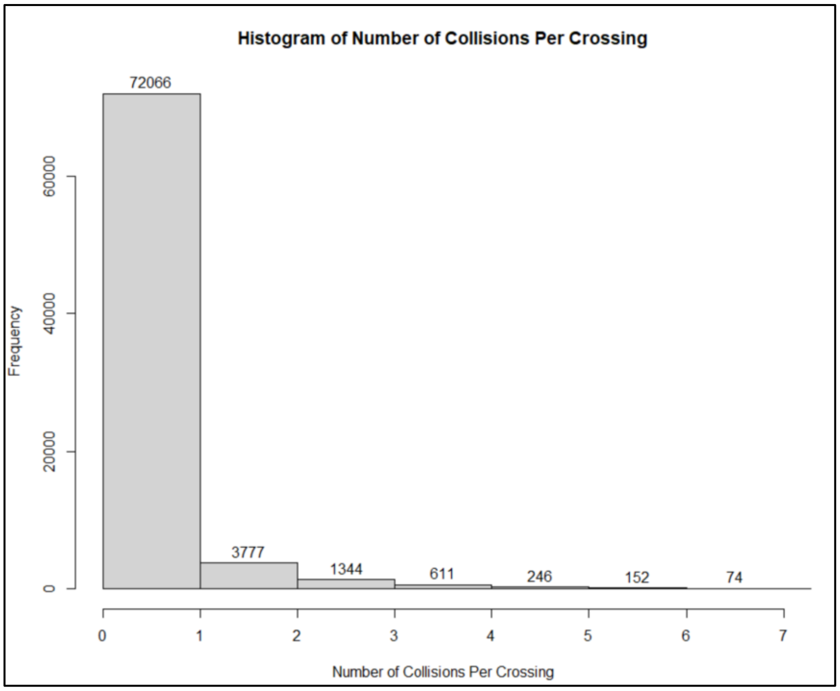
Initial Exploratory Data Analysis

To further understand the crossing and collision data, an exploratory data analysis (EDA) was performed. Analysis consisted of creating several charts and graphics that allow one to easily visualize relationships between aspects of the data. To undergo the analysis, the collision and crossing databases were queried to determine relational aspects of key crossing properties and their collisions. The queries were then exported to Microsoft Excel and R[1], where bar charts, histograms, scatterplots, and map charts of the data were created (RStudio 2021). Once performing the initial exploratory data analysis, Brod and Gillen's publication was consulted to discern the best methods for grouping collision risk by protection type.

Initial Exploratory Data Analysis Results

The results of the initial exploratory data analysis provided insight on the collision data and its trends. Trends identified contextualize the data and provided an understanding of the best approach for analysis of the data. For example,

Figure 9 indicates that there are many crossings with 0 reported collisions in the period of analysis, which is not unexpected as collisions are infrequent/catastrophic events. This presented challenges for analysis, as crossings with 0 collisions do not necessarily have 0 risk or 0 probability of collision. As such, a latency period between collision risk at a crossing and a collision occurring exists.



[1] Open-source statistical programming software

Figure 9: Histogram of Number of Collisions Per Crossing

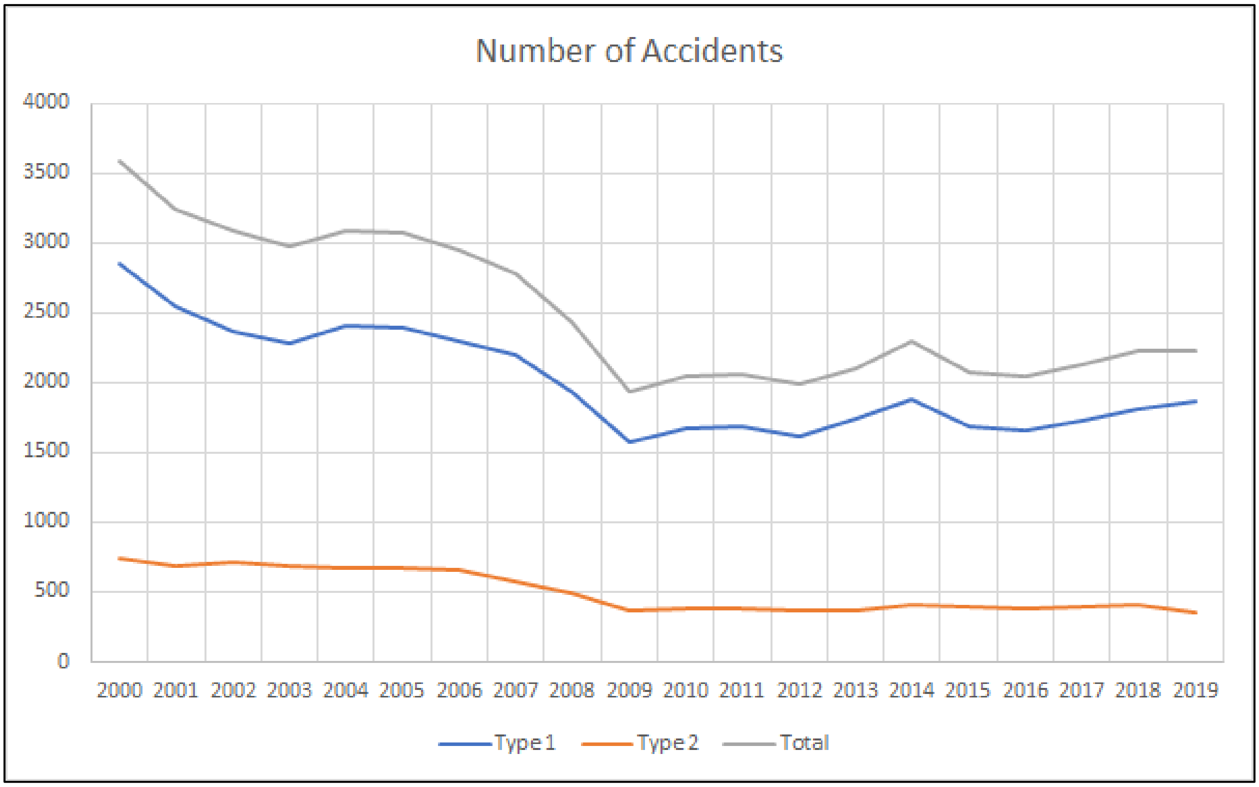


Figure 10: Total number of accidents and number of Type 1 and Type 2 accidents per year. Source FRA collision data.

Figure 10 shows the total number of accidents in the U.S per year, starting in 2000 and ending in 2019.

Figure 10 also shows the total number of accidents per year, total Type 1 accidents, and total Type 2 accidents. Type 1 accidents occur when a vehicle collides with a train. Type 2 accidents occur when a train collides with a vehicle. Type 2 accidents show a decline from 2000 to 2009 then plateau afterwards. Type 1 accidents show a decline from 2000 to 2009 then slowly increases. As a result, the increase in total accidents starting in 2009 is likely the result of an increase in Type 1 accidents only.

Among the most useful findings of this initial analysis is the clear correlation between the number of crossings in each U.S state and the total number of crossing collisions that occur in that state as shown in

Figure 11. In addition to

Figure 11, an accompanying map was created in

Figure 12, which maps those values.

Figure 12 better reflects the states in need of improved collision prevention strategies. Louisiana, for example, does not have many total collisions, but experiences a higher collision rate given its relatively small number of crossings. This is also shown in
Figure 11, as it deviates substantially from the trendline.

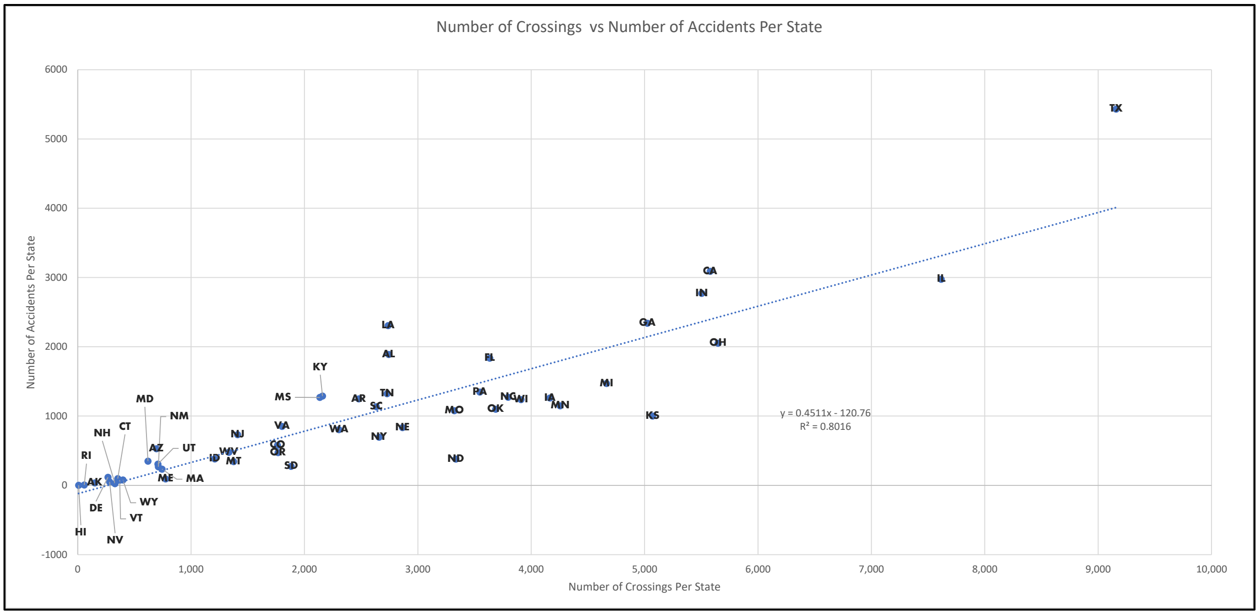


Figure 11: Number of incidents per state vs number of crossings per state

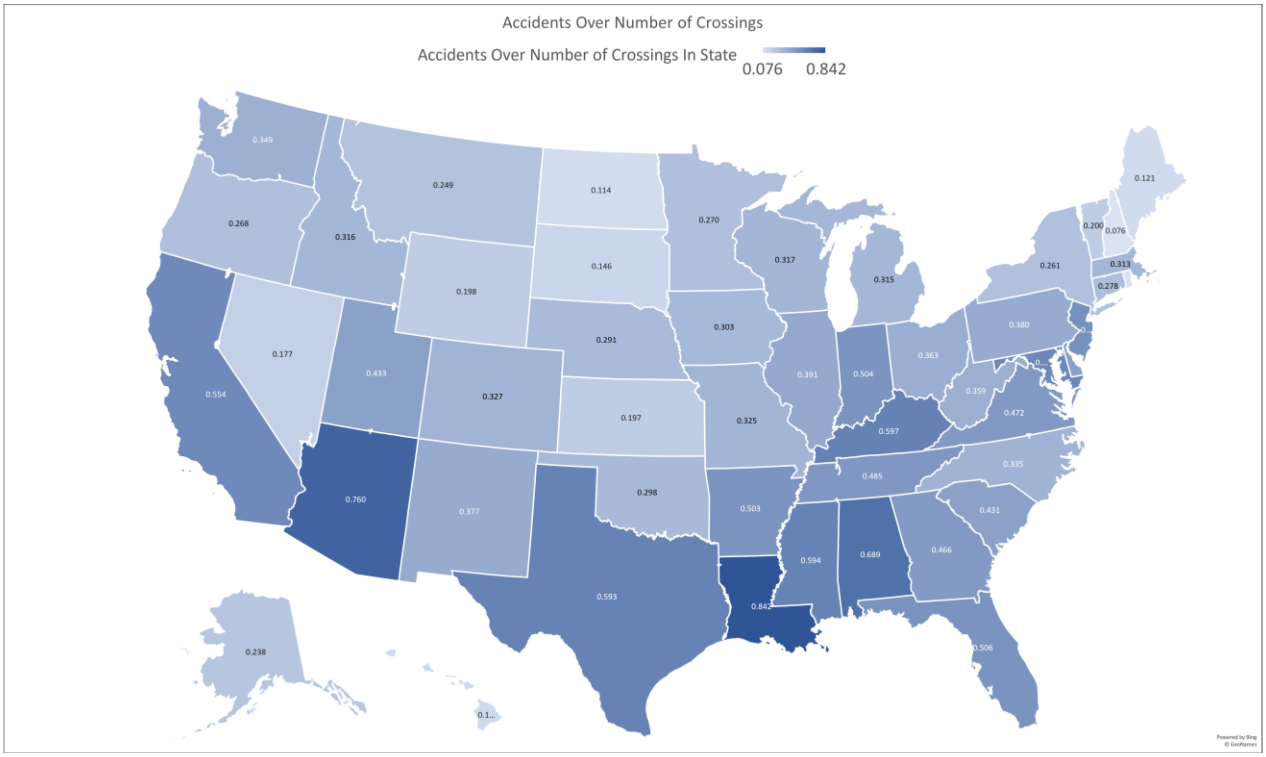


Figure 12: Heat Map of Accidents/Number of Crossings by State

Secondary Exploratory Data Analysis Using Normalized Collision Rates

The results of the initial exploratory data analysis indicated that collision rates should be normalized by the number of opportunities for collision, or exposures. This normalization was used in the secondary EDA to determine the influence of crossing properties on collision risk. In the secondary analysis, an exposure is a combined variable for AADT and daily through trains at a crossing over a 20-year period (2000-2020). Normalizing exposures allows for a more accurate understanding of how protection influences risk. For example, two crossings equipped with the same protection but differing AADT will have varying possibilities for collision and consequently, greater risk. The equation for exposures is modeled after that developed by Brod and Gillen and has been adjusted to account for a different period of analysis which is 20 years (Brod & Gillen, 2020). The exposure equation is shown below:

$$E = N/(A*T*300*20) \text{ X } 10^6 \qquad (13)$$

where E is the number of collisions per crossing per million exposures, N is the number of collisions over 20 years, A is AADT, T is the combined number of daily and nightly through trains, 300 is the number of annual number of traffic days, and 20 is the period of analysis in years.

Secondary Exploratory Data Analysis Using Normalized Collisions for Protection Types

Using the exposure equation defined in (13), the level of exposure for each crossing was calculated. These values were subsequently used to normalize the number of incidents at each crossing. The normalized average incidents per million exposures was plotted against the categorical parameters, Highway Classification (HwyClasstpID), Land Use Classification (DevelTypID), and Type of Protection (WdCode), as shown in
Figure 13 and
Figure 14.

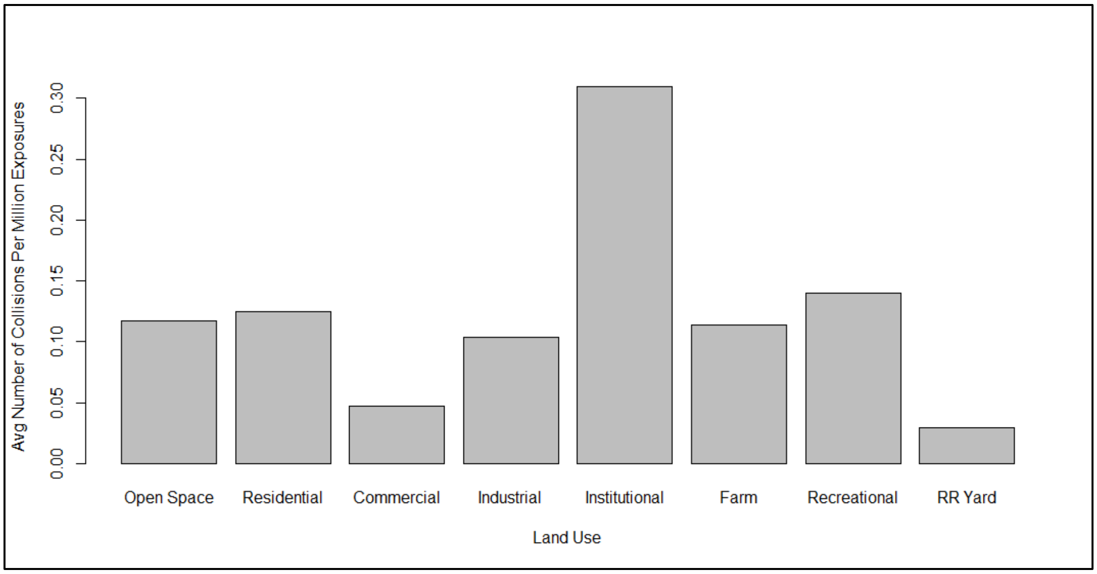


Figure 13: Average Number of Collisions Per Million Exposures for Land Use Categories

Figure 13 shows that institutional land uses have the highest likelihood of collision; nearly two times the likelihood of collision than the next highest land use, recreational.
Figure 14 plots the average normalized collision rates against highway classification.

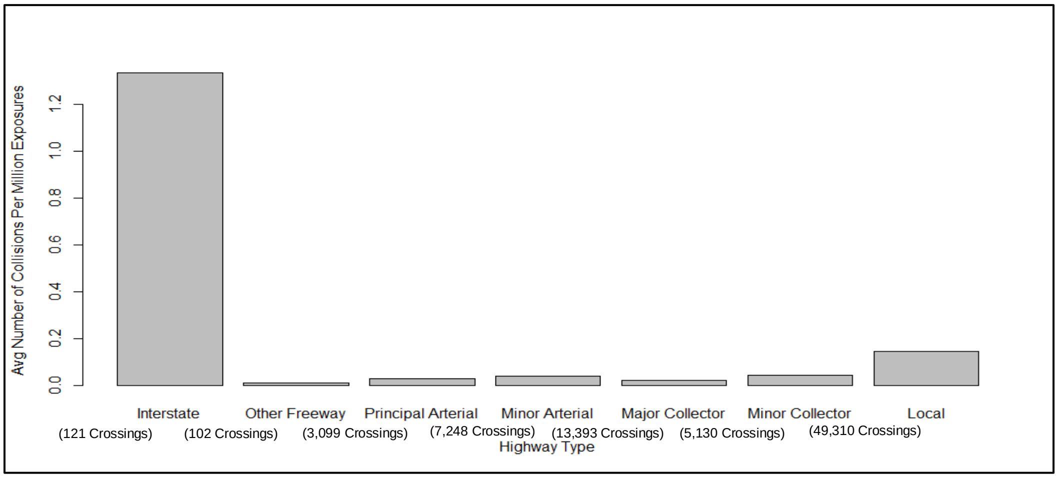


Figure 14: Average Number of Collisions Per Million Exposures for Highway Classification

Figure 14 shows that interstate highways have a much larger number of collisions per million exposures than all other highway classifications. Additionally, local highways have the second highest number of collisions per million exposures, though lower than interstates by a factor of 6.

The results of
Figure 13 and

Figure 14, though insightful, can be accounted for in the exposure normalization. While interstate highway crossings have a substantially higher normalized collision rate than all other road classifications, they also have the highest AADT. Additionally, highway-rail crossings at interstates are an extreme rarity, with only 121 total crossings in the entire United States experiencing a collision, so the application of these findings are limited. All other roadway types have similar collision rates.

For land use classifications, though institutional land uses have the highest rates of collisions, there are likely external factors, such as large young driver populations. Institutional land uses include universities, which have large concentrations of young drivers who are statistically more prone to car crashes (Rates of motor vehicle crashes, injuries and deaths in relation to driver age, united states, 2014-2015.2017). Such higher populations may account for the increase, however data to support this claim is not readily available for this type of analysis. Like interstate crossings, institutional crossings are also rare. Other land use types excluding commercial and rail yards have similar normalized accident rates. Though commercial and rail yards have somewhat different collision rates, the distinction is not substantial enough to analyze further in this research.

Figure 15 presents the most useful findings of the secondary exploratory data analysis. It shows that higher-quality, active, protection, such as 4-quadrant gates correspond with lower risks of collision than lower quality, passive protection, such as stop signs. The findings in
Figure 15 are the basis for further analysis using Bayesian belief networks. A Bayesian belief network model was developed with the expectation that its results would agree with those of
Figure 15.

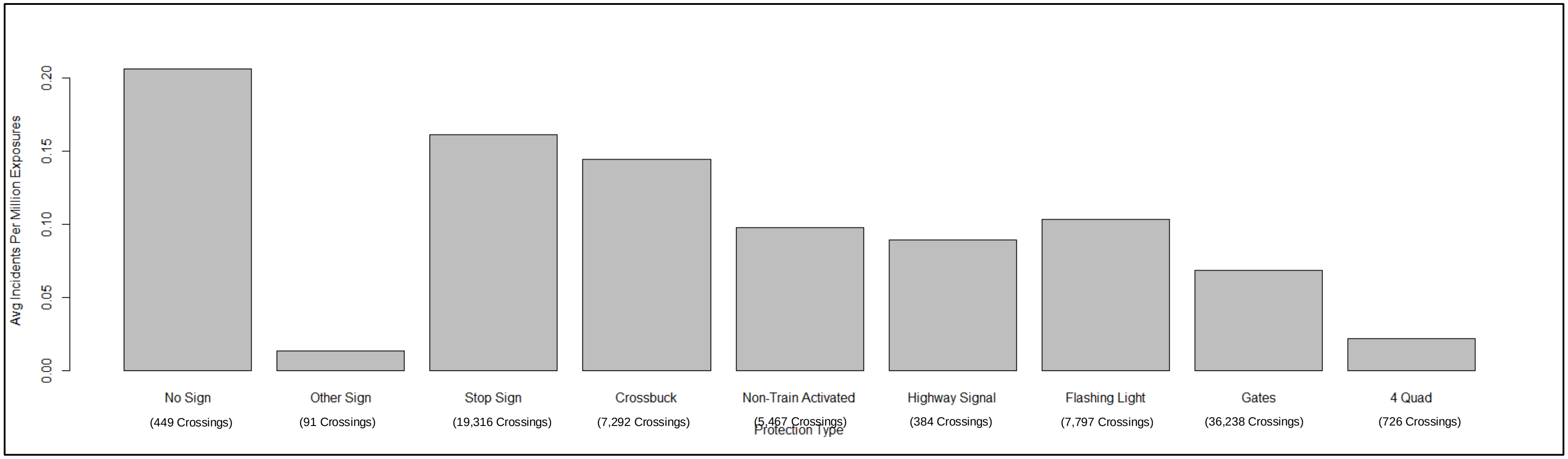


Figure 15: Average Number of Collisions Per Million Exposures for Protection Type

BAYESIAN NETWORK MODEL DESIGN

Using the findings of the exploratory data analysis, several modeling approaches and resulting models were developed to predict the risk of collision at a crossing based on its properties. Three models were developed in BayesFusion and are defined as Model 1, Model 2 and Model 3. Each model is a Bayesian network, and uses key parameters such as; number of main tracks, through trains, highway lanes, timetable speed, and protection type as either parent or child nodes. How these parameters are connected defines the differences in the model permutations. Additionally, whether a crossing has had a collision in the period of analysis is a child node, and the primary output, for Model 1 and Model 3. In Model 2, it is a parent node.

The probabilities shown for each variable are determined via input of the cleaned collision data described previously. The models' accuracy is calculated by inputting crossings from the FRA inventory and identifying the percentage of correct predictions of collision likelihood via a confusion matrix. Model 1 and Model 2 had intermediate accuracy, peaking at 46%, but showed severe flaws that made them unusable. Model 3 resolved the primary flaws of Model 1 and Model 2 but displayed a sharp drop in accuracy. As a result, the models, as currently designed, require further development.

Model 1

In Model 1, collision occurrence is the primary output, since its conditional probability is dependent on key crossing properties identified, shown graphically in Figure 16. As such, the accuracy and reliability of Model 1 increases as its ability to correctly predict the likelihood of collision increases. The nodes, connections and variables used in Model 1 are shown in **Error! Reference source not found.**. Initial accuracy for Model 1 was low, but after incremental adjustment of variables, parameters and connections, had a maximum accuracy of 46%, higher than Model 3. The primary flaw of Model 1 was its inability to predict increased collision likelihood with decreased crossing protection. Identifying this flaw early in the model development process helped guide the designs of Model 2 and Model 3.

Table 3: Model 1 Nodes and Connections

Node	Variable	Parent	Child
Main Tracks	One	Night Through Trains Day Through Trains	Collision
Main Tracks	More Than One		
Day Through Trains	Ten or Less	-	Collision Main Tracks
Day Through Trains	More Than Ten		
Night Through Trains	None	-	Collision Main Track
Night Through Trains	One or More		
Max Timetable Speed	Less than forty	-	Collision
Max Timetable Speed	More than forty		
Traffic Lanes	Two or Less	AADT	Collision

Traffic Lanes	More Than two		
Protection Type (Wd code)	4 Quadrant	-	Collision
Protection Type (Wd code)	Crossbucks		
Protection Type (Wd code)	Flashing Lights		
Protection Type (Wd code)	Gates		
Protection Type (Wd code)	Highway Traffic Signal		
Protection Type (Wd code)	No Sign or Signal		
Protection Type (Wd code)	Non-train Activated Special		
Protection Type (Wd code)	Other Sign or Signal		
Protection Type (Wd code)	Stop Sign		
Protection Type (Wd code)	Unknown		
Collision	Collision	Night Through Trains Main Track Day Through Trains AADT Max Timetable Speed Traffic Lanes Protection Type (WdCode)	-
Collision	No Collision		

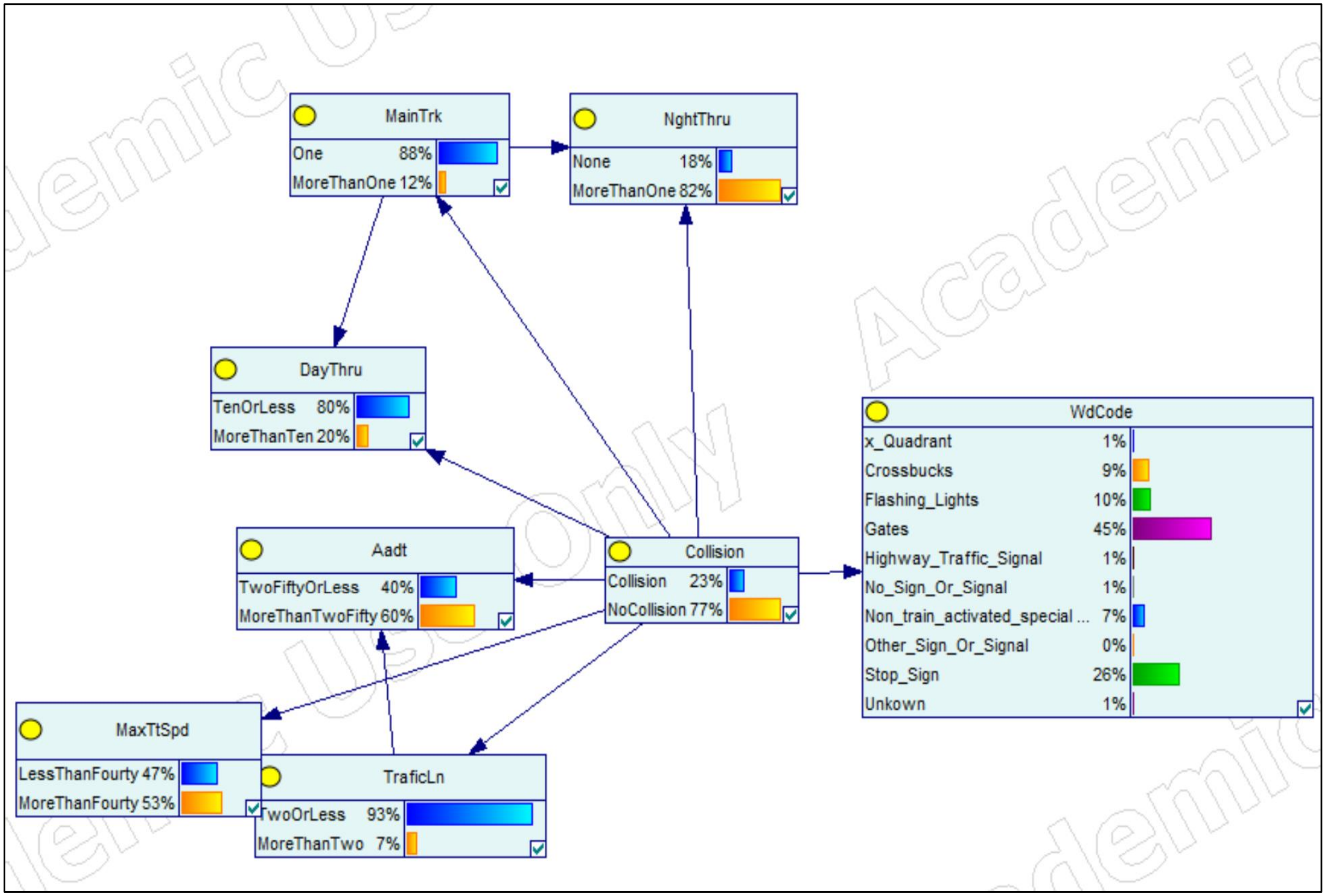


Figure 16: Model 1

This led to the development of new iterations and alterations of Model 1 to optimize its accuracy, using GeNIe software's validate function. The validate function compares the predicted collision output probability (either collision is more likely, or no collision is more likely) of a model to a given dataset. It then produces a confusion matrix, as shown in **Error! Reference source not found.**,that displays the counts of all the model's accurate and inaccurate predictions. Each new iteration of Model 1 sought to improve its collision prediction accuracy. The iteration number, change description and confusion matrix for each iteration are shown in
Table 5.

Table 4: Model 1 Confusion Matrix

		Predicted	
		Collision	No Collision
Act	Collision	5,633	13,601
	No Collision	6,850	56,827

Table 5: Model 1 Iterations

Iteration	**Change Description**	**Accuracy**		**Confusion Matrix**		
1	*Add all 9 WdCodes*	0.74			Predicted	
					No Collision	Collision
			Actual	No Collision	**59459**	4260
				Collision	15346	**3883**
2	*add aadt*	0.76			Predicted	
					No Collision	Collision
			Actual	No Collision	**57427**	6202
				Collision	13842	**5387**
3	*reduce aadt categories*	0.74			Predicted	
					No Collision	Collision
			Actual	No Collision	**54987**	8732
				Collision	9162	**3632**
4	*add all lane categories*	0.75			Predicted	
					No Collision	Collision
			Actual	No Collision	**56773**	6946
				Collision	13580	**5649**
5	*add passive lights or gates*	0.74			Predicted	
					No Collision	Collision
			Actual	No Collision	**55291**	8428
				Collision	12906	**6323**
6	*disconnect aadt and lanes, daythru and main trk*	0.74			Predicted	
					No Collision	Collision
			Ac tua	No Collision	**55163**	8556

				Collision	12812	**6417**
7	*add night thru*	0.7			Predicted	
					No Collision	Collision
			Actual	No Collision	**49264**	14513
				Collision	10327	**8907**
8	*add whistban - removed*	0.74			Predicted	
					No Collision	Collision
			Actual	No Collision	**54235**	9012
				Collision	12556	**6653**
9	*add night thru categories - removed*	0.7			Predicted	
					No Collision	Collision
			Actual	No Collision	**49226**	14021
				Collision	10374	**8835**
10	*add transit movement - removed*	0.68			Predicted	
					No Collision	Collision
			Actual	No Collision	**43177**	13600
				Collision	9852	**8481**
11	*add development type - removed*	0.71			Predicted	
					No Collision	Collision
			Actual	No Collision	**50371**	13406
				Collision	10609	**8625**

Figure 17 shows how accuracy improved with each iteration. The “overall rate” series of the model represents the combined rate of correct “collision” and “no collision” predictions. The “correct positive rate” series represents the rate of correct “collision” predictions and the “correct negative rate” series represents the rate of correct “no collision” predictions. After each iteration, the overall rate and correct negative rate remained substantially higher than the correct positive rate, with neither series dipping below 70% accuracy. This is likely because there are more instances of “no collision” than instances of “collision” in the data to train the model as shown in Figure 18, so it is an easier output for the model to accurately predict. This is a result on the imbalanced data. As such, the goal of each iteration was to improve the accuracy of “collision” prediction. The maximum accuracy that could be produced from Model 1 was 46%. Though this accuracy was lower than expected, it represents a notable improvement from an initial accuracy of only 20%.

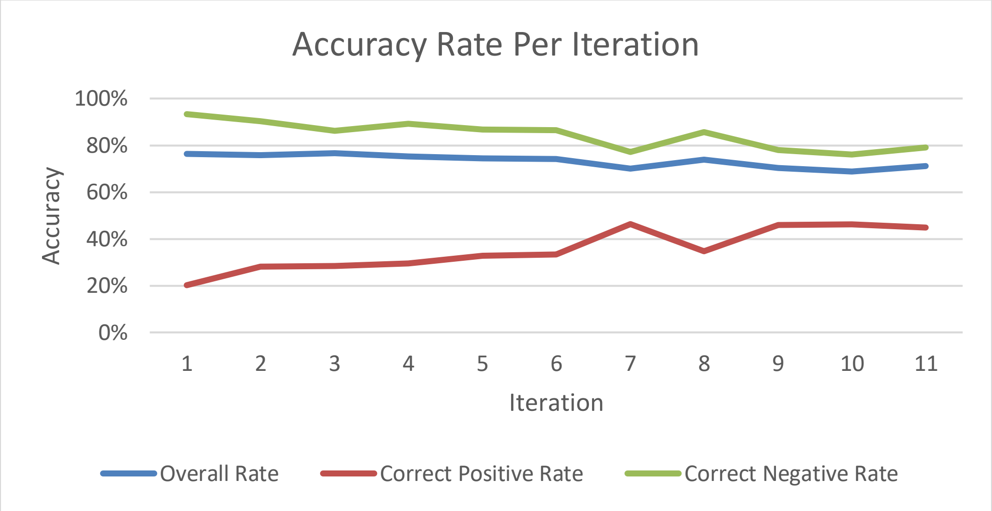


Figure 17: Prediction Accuracy of Model 1 Iterations

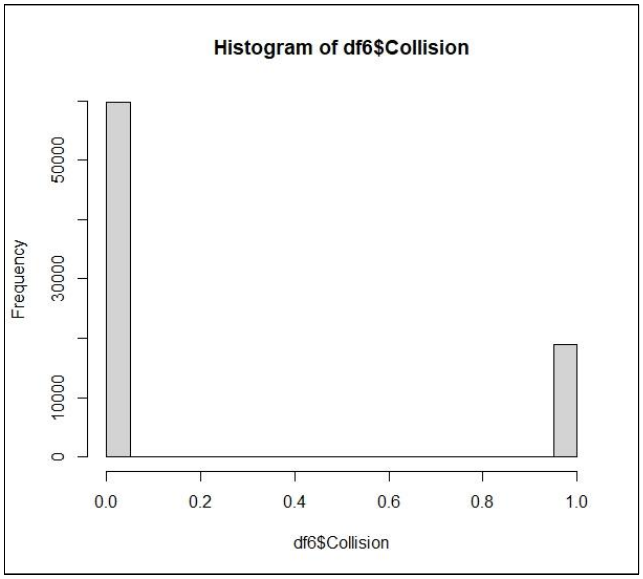


Figure 18: Total Number of Instances of Collision and No Collision. 0 indicates no collisions and 1 indicates collisions.

Problems with Model 1

Despite the gains in accuracy, further sensitivity and parametric analysis of Model 1 revealed that some of its outputs are inconsistent with both the literature and the exploratory data analysis.

Figure 19 overlays the prediction results of Model 1 onto the results of the secondary data analysis in

. It shows that the two results do not agree with one another and are inversely proportional. The results of

show that as the level of protection increases (from left to right), the number of collisions per exposure, an empirical measurement of collision likelihood, decreases. In contrast, the results of Model 1 predict that as the level of protection increases, the likelihood of collision increases. The results of Model 1 were expected to agree with

as well as the conclusions of Brod and Gillen’s study, which found that increase in protection type lowers the collision rate per exposure. The results of Model 1 are also logically inconsistent, as installing more robust and costly protection devices, such as four quadrant gates, should reduce collision risk, not increase it. Upon discovery of the flaws with Model 1, a second structure was developed in Model 2.

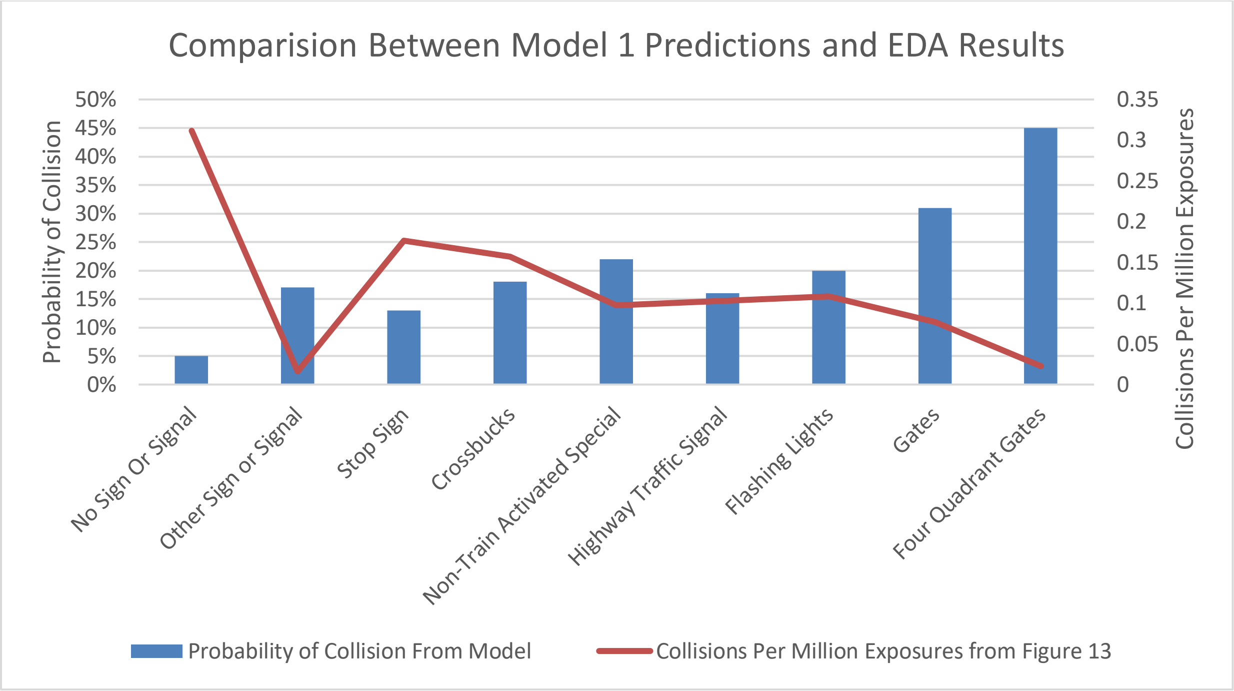


Figure 19: Model 1 vs

Results

Model 2

Model 2, as shown in Figure 20, was developed in response to Model 1. While Model 1 is developed to predict the likelihood of collision based on key crossing properties, Model 2 is developed to predict the protection type based on whether a collision has occurred on other properties. Note that the protection type is consolidated from its original 9 categories to 3, as shown in Table 6**Error! Reference source not found.**. They were consolidated because of the limited available data for specific protection types, such as no sign or signal. Additionally, day through trains, night through trains, and AADT were consolidated into the single exposure parameter that was developed during the exploratory data analysis using RStudio (See Appendix A). Increasing states (State0 to State11) reflect an increase in exposure.

Table 6: Protection Type Consolidation

Protection Type Model 1	**Protection Type Model 2**
No Sign or Signal Other Sign or Signal Stop Signs Crossbucks	Passive (P)
Non-Train Activated Highway Traffic Signal Flashing Lights	Lights (L)
Gates Four Quadrant Gates	Gates (G)

A parametric analysis of Model 2, shown in
Figure 21, shows that its results still do not agree with the implications of the literature and exploratory data analysis.
Figure 21 shows that the likelihood of a crossing having passive protection increases if a collision did not occur and decreases if a collision did occur. Conversely, the likelihood of a crossing having gates decreases if a collision did not occur and increases if a collision occurred.

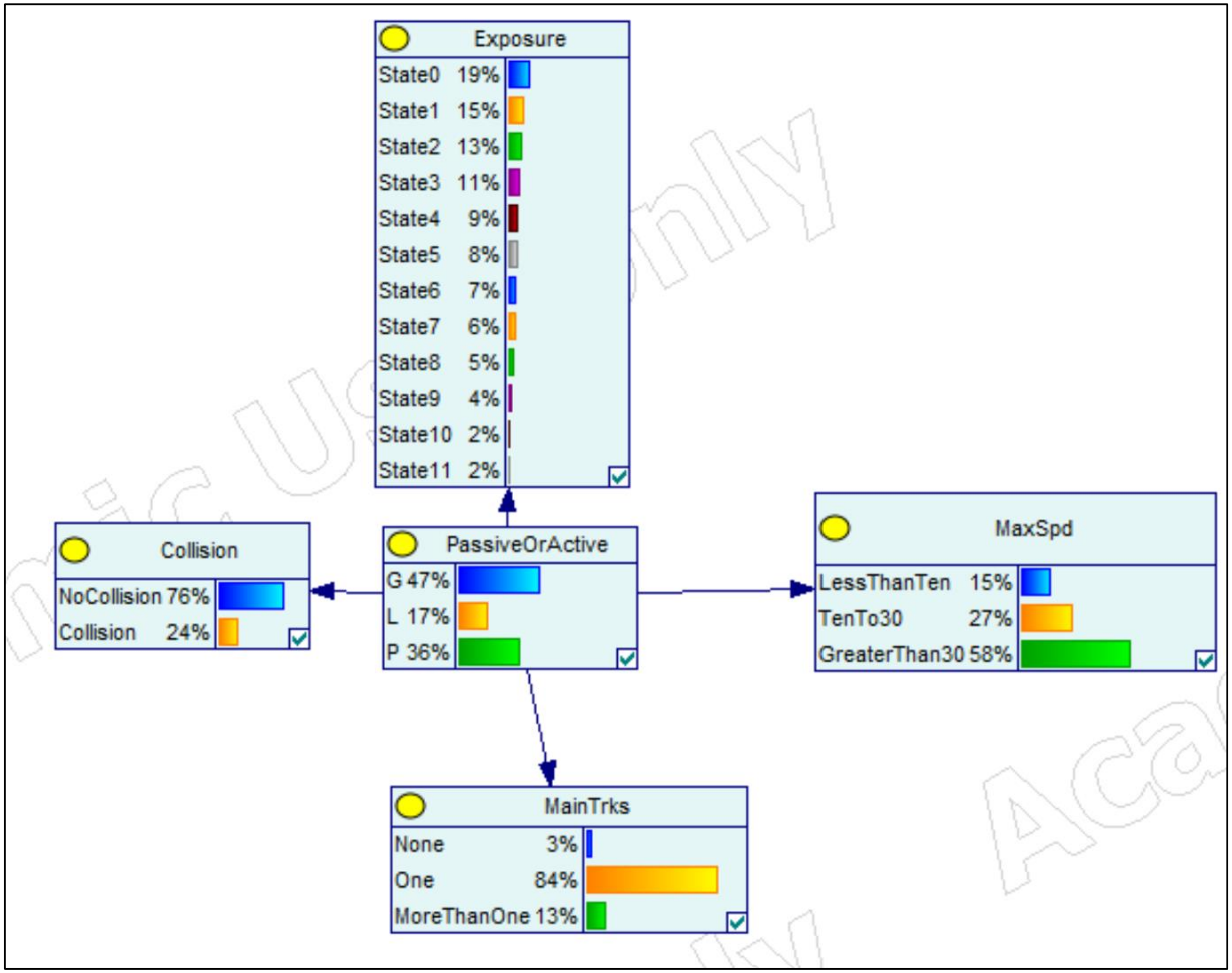


Figure 20: Model 2

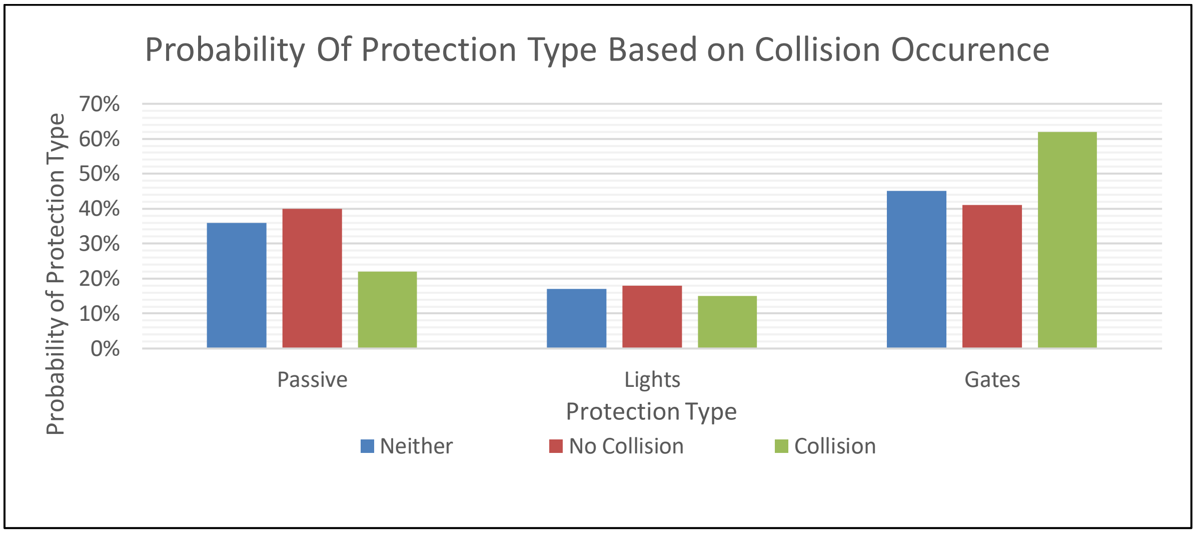


Figure 21: Model 2 Parametric Analysis

A confusion matrix was developed for Model 2 with protection type as the variable being analyzed, and is shown in **Error! Reference source not found.**. The overall accuracy of Model 2 was 66%, with gates and passive being the most accurately predicted protection types, at 77% and 79% respectively. The accuracy of lights was nearly a fifth that of both gates and passive at 13%. The accuracy of predicting lights is substantially lower than gates and passive protection. This is likely due to lights being an intermediary category that shares characteristics with both gates and passive protection.

Table 7: Model 2 Confusion Matrix

		Predicted		
		G	L	P
Act	G	**28,845**	1,647	6,587
	L	5,775	**1,800**	6,140
	P	4,697	1,185	**22,143**

Model 2 sought to improve upon and correct the flaws of Model 1 but encountered similar issues. It improved prediction accuracy for Passive and Gated crossings over Model 1. Like Model 1 however, its predictions are inversely related to what is anticipated for protection type. While increasing protection should decrease collision likelihood, both Model 1 and Model 2 predict the opposite.

Revision of Data

Given Model 1 and Model 2's flaws, the data was revisited to confirm that greater protection reduces collision occurrence.
Figure 22 revisits the findings in

Figure 15 to confirm that gated crossings have lower total collisions with increasing exposure than passive crossings. The exposure levels shown are the same as those in Model 2. Its results largely agree with what was anticipated, though as the level of exposure increases, it appears that the difference between gates and passive crossings becomes increasingly insignificant. Still the data supports that gated crossings coincide with lower collision rates than passive, which the first two model's did not predict.

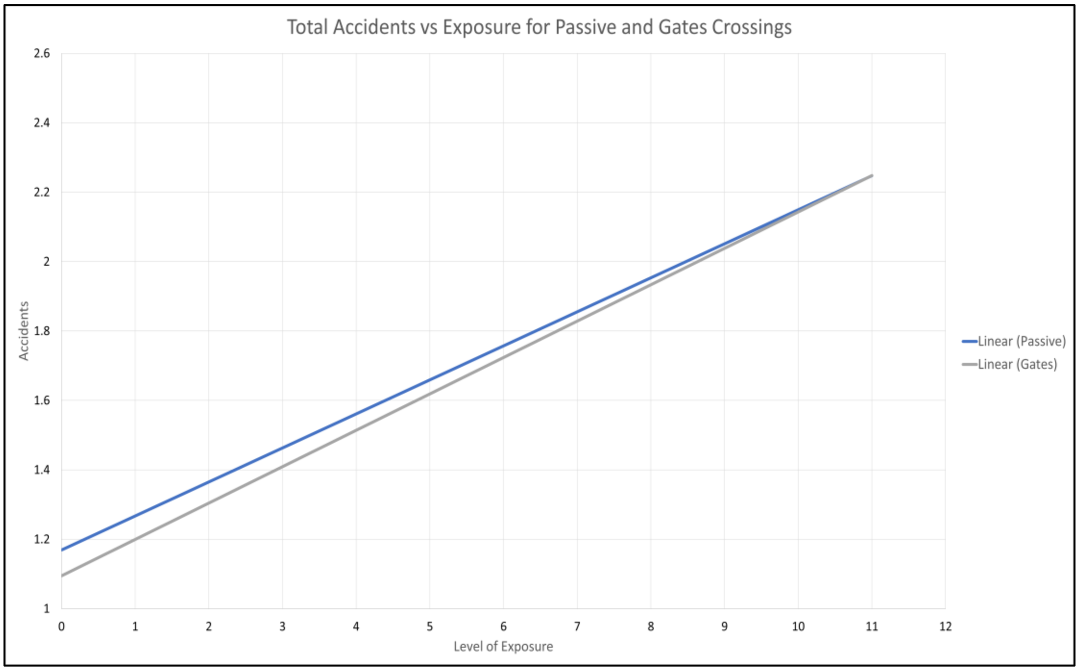


Figure 22: Total Accidents Vs Exposure for Passive and Gates Crossings

Though the exact reason for these inverse predictions for Model 1 and Model 2 are not entirely understood, upon revisiting the data, there are some patterns that provide insight as to what is driving the disconnect.

Figure 23 charts the average number of yearly incidents by protection type. It shows that as the level of protection increases, the average number of yearly incidents also increases, likely because crossings with higher annual collisions are those chosen for upgrades. The results plotted in Figure 23 align with those of both Model 1 and Model 2. Crossings with greater protection have more instances of collision and, without normalization by exposures, would be predicted to have greater collision likelihood.

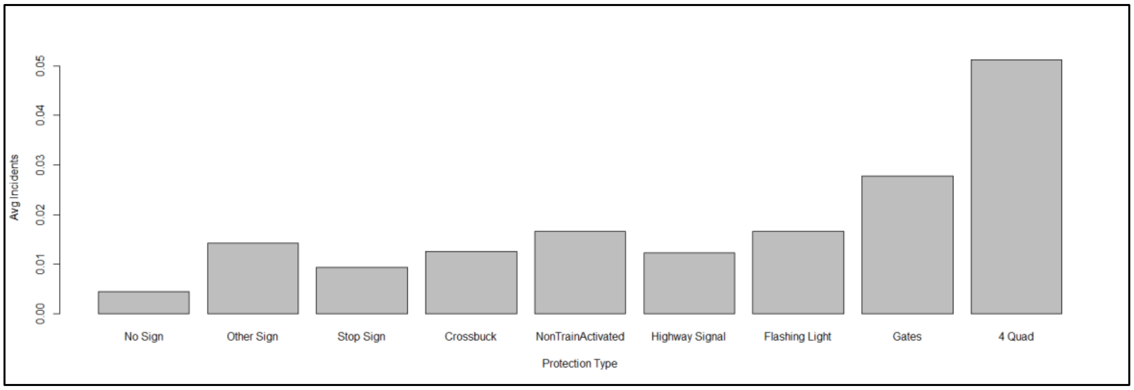


Figure 23: Average yearly incidents by protection type

In the designs of Model 1 and Model 2, it was anticipated that including exposure parameters (AADT, Thru Trains) as nodes would act as a normalization.
Figure 24 shows that crossings with increased protection have more exposures. It was thought that, if an input had greater protection and higher exposure, the models would predict a lower likelihood of collision. In contrast, the models consistently predicted lower collision likelihood with higher exposures and consistently predicted higher collision likelihood with greater protection.

It was hypothesized that the reason for this disconnect was an imbalanced data set. As
Figure 24 shows, crossings with increased protection tend to have higher exposure, which is expected, to produce higher collision rates, regardless of protection type. As such, it was believed that Model 1 and Model 2 predicted that crossings with greater protection are more likely to have a collision because they have more exposures than crossings with less protection. A third model was developed that sought to mitigate the imbalance through normalization of exposures.

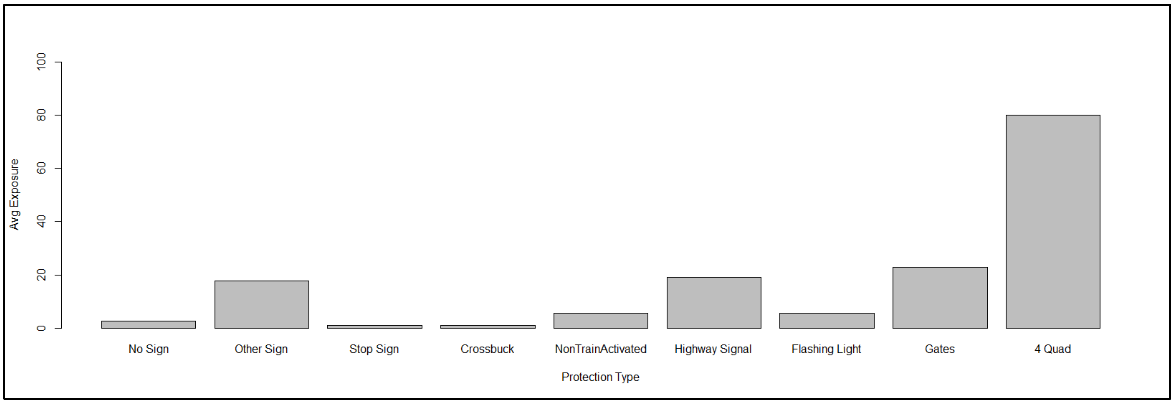


Figure 24: Average yearly exposure by protection type

Model 3

A final model was produced that sought to include normalization. Model 3, shown in Figure 25, incorporates aspects of both Model 1 and Model 2, including all nodes as Model 1 but

consolidating protection type categories as in Model 2. Unlike Model 1 and Model 2, Model 3 does not predict the likelihood of collision, but rather predicts the likelihood that a crossing will have a higher normalized collision rate than the weighted mean for all crossings. The weighed mean was chosen as the threshold between an acceptable collision rate and an unacceptable collision rate. Parametric analysis of Model 3, shown in
Figure 26, agrees with the results of the EDA and literature review. As the level of protection increases, Model 3 predicts that the crossing is more likely to have lower than average normalized collision rates.

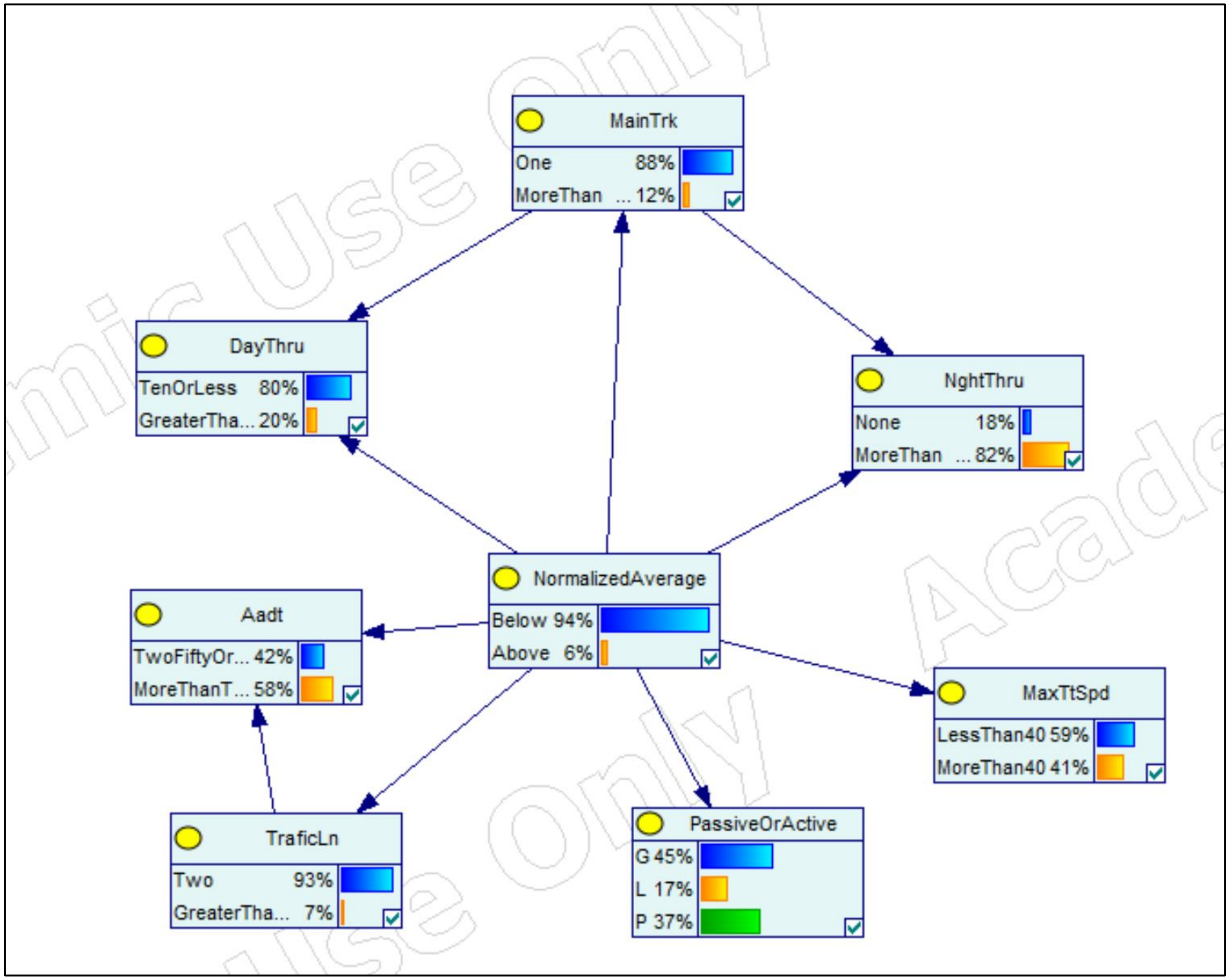


Figure 25: Model 3

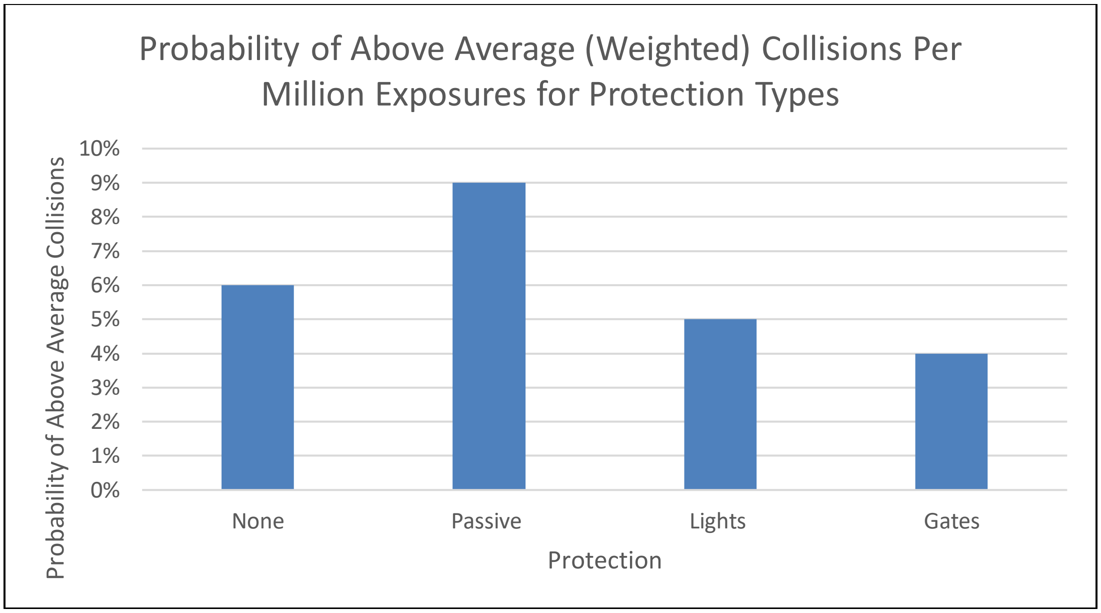


Figure 26: Model 3 Protection Type Parametric Analysis

Further parametric analysis of timetable speed for Model 3 is shown in
Figure 27. Model 3 predicts that crossings with higher timetable speeds are more likely to be above the acceptability threshold. This aligns with the results of the exploratory data analysis and literature review. Higher speeds at crossings result in less reaction time for the car driver, leading to collision. Model 3 confirms the increased collision risk for higher speed crossings.

The limitations of BayesFusion software made it difficult to determine the accuracy of Model 3. The confusion matrix of Model 3 is shown in **Error! Reference source not found.**. showing that BayesFusion identified no correct above-threshold predictions. This is likely due to most crossings having below average collision rates, in part due to most having 0 total collisions. As a result, Model 3 cannot predict a crossing is more likely to have above average collision rates than below average (never above 50% likelihood). Consequently, no crossing was predicted to have above average collision rates. As such, despite the incorporation of normalization in Model 3, its application is limited. Further research should be undertaken that identifies an acceptable level of normalized collisions and enables model 3 to predict that crossings are more likely to be beyond the acceptability threshold than below. One such threshold could be the random probability of the data.

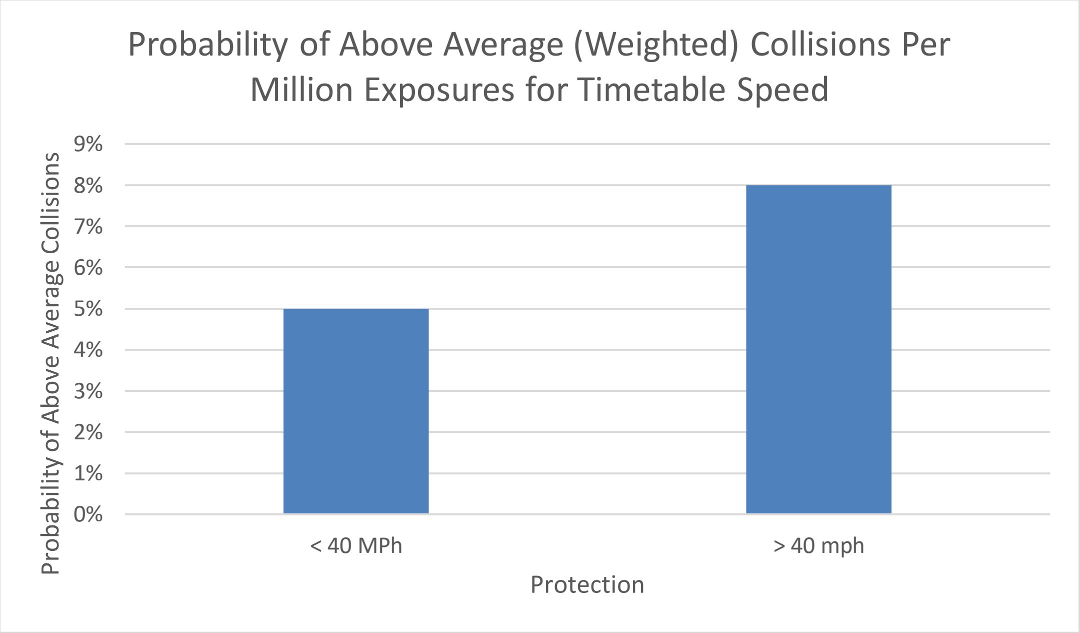


Figure 27: Model 3 Timetable Speed Parametric Analysis

Table 8: Model 3 Confusion Matrix

		Predicted	
		Below	Above
Act	Below	**77,832**	0
	Above	5,179	**0**

CONCLUSION AND RECOMMENDATIONS FOR FURTHER RESEARCH

This research sought to develop a new model for predicting collision likelihood at highway-rail grade crossings. The model most widely used today, the USDOT APS model, developed in the 1980s, has declined in accuracy as the data has changed, and does not consider modern statistical analytic methods.

New models were developed herein using the principles of Bayesian statistics, taking advantage of Bayes Theorem and its associated ability to update probabilities when new information. The new models developed, Model 1 and Model 2, and Model 3 are Bayesian networks, which are graphical representations of joint probability distributions.

Model 1 was designed to predict the probability of collision at a crossing, given prior knowledge about that crossing. Model 1 was determined to be fairly accurate at predicting collisions, with a maximum accuracy rate of 46%, but showed results that were counterintuitive to the results of the exploratory data analysis and literature review. Model 1 predicts that the more protection a crossing has, the more likely it is to have had a collision, which is not reflected in real world conditions or the EDA. Upon additional research, it was decided that a second model be developed that worked in the reverse, predicting the most likely protection type, given that a collision existed, along with other relevant factors. Model 1 provided insight as to how to develop a Bayesian network for Model 2 and the flaws encountered. It was hoped that the second model would accurately predict that more protection decreases collision likelihood.

Model 2 sought to improve upon the strengths of Model 1 while incorporating the reverse prediction plan. Model 2 consolidated the variable categories from Model 1 to develop a more streamlined Bayesian network, while still attempting to retain accuracy. Model 2 predicts the likelihood of a crossing having either lights, gates or passive protection based on collision occurrence and other factors. Model 2 was found to be as accurate, if not more than Model 1, depending on the protection type predicted. Model 2 predicted a gate crossing with 77% accuracy, a passive crossing with 79% accuracy and a lights crossing with 31% accuracy. Model 2 showed the same critical flaw as Model 1, predicting that crossings with collisions are more likely to have increased protection.

Model 3 combined aspects of Model 1, Model 2, and the findings of the exploratory data analysis. Rather than predicting the likelihood of collision, it predicts the likelihood of an above average normalized collision rate at a crossing, where above average is considered unacceptable. Model 3 successfully predicts that more protection lowers the likelihood of unacceptable collision rates but was interpreted by the software to have 0% accuracy. This is the result of BayesFusion software identifying a correct prediction only if the probability of above average collision rates is over 50%. However, the results can be used effectively for a relative based risk analysis.

Both Model 1 and Model 2 can somewhat accurately predict the likelihood of collision or likelihood of protection but show flaws that should be addressed with further research. Model 3 successfully addresses this flaw by predicting the likelihood of an unacceptable collision rate, above the weighted mean. It however cannot predict that an unacceptable collision rate is more likely than an acceptable collision rate for any crossing in the database. The primary goal of further research should be to identify a new acceptability threshold that can be predicted with a greater degree of certainty than Model 3 is currently capable of. Such a threshold may be the random probability of collision based on the data. Once this threshold has been identified, a second focus of further research should be to increase the overall accuracy of the model and to determine if there are other influential factors, such as human behavior, that are not currently considered.

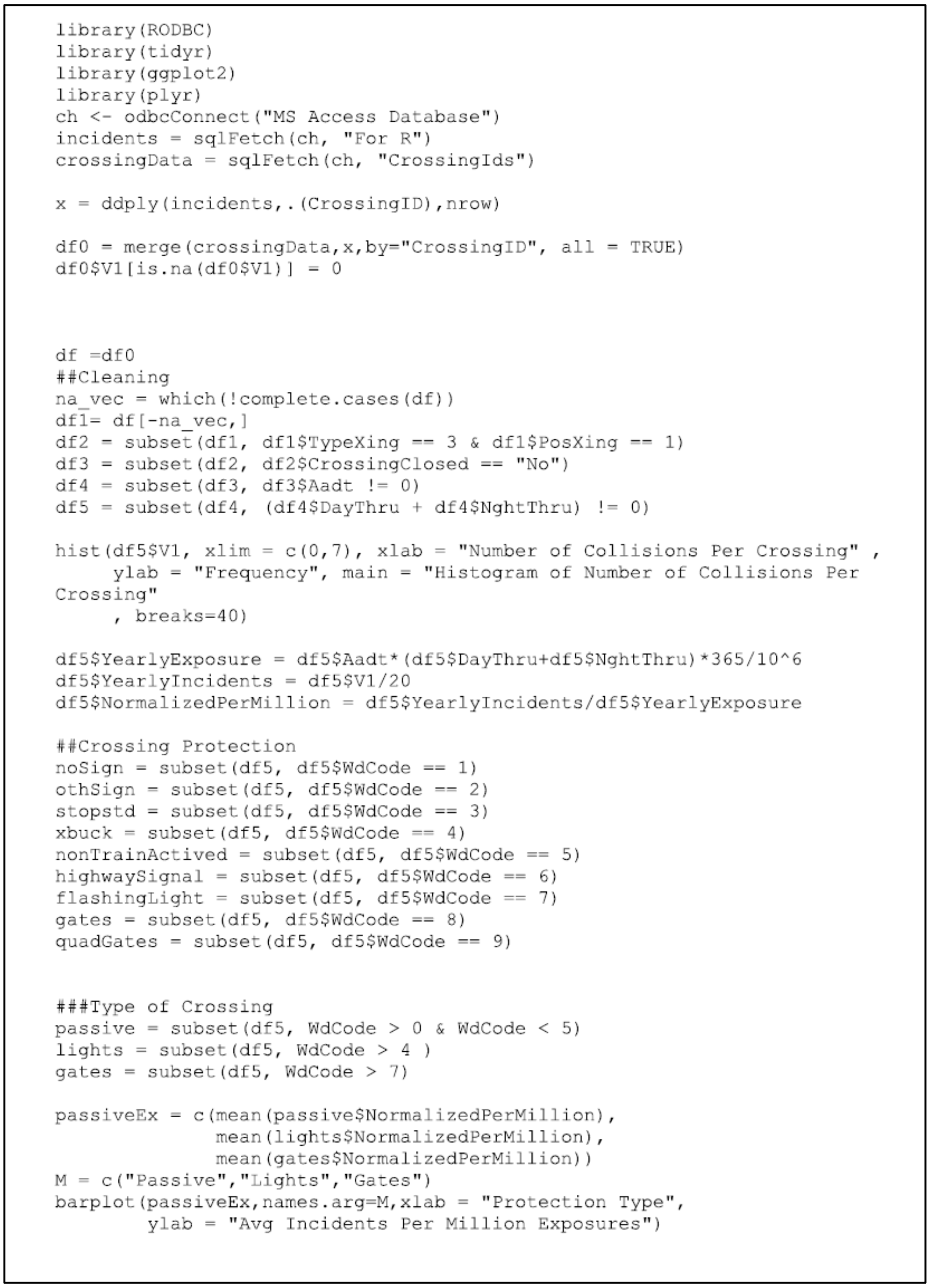
Though the resulting has limitations, this research has shown that Bayesian statistics, specifically Bayesian Networks, offer a promising method for determining the likelihood of collision risk at grade crossings. As the APS model ages and continues to decline in accuracy, the need for the development of a new model will grow. Additionally, should the rate of rail-highway crossing incidents continue to increase, more sophisticated modelling techniques and countermeasures will need to be developed to reverse the trend. As such, researchers should continue to pursue the development of new prediction models that can eventually replace the APS as the primary accident predication model.

REFERENCES

1. Abioye, O. F., Dulebenets, M. A., Pasha, J., Kavoosi, M., Moses, R., Sobanjo, J., & Ozguven, E. E. (2020). Accident and Hazard Prediction Models for Highway–Rail Grade Crossings: A State-of-the-Practice Review for the USA. *Railway Engineering Science, 28*(3), 251-274. doi:10.1007/s40534-020-00215-w
2. Austin, R. D., & L. Carson, J. (2002). *An Alternative Accident Prediction Model For Highway-Rail Interfaces* doi:https://doi.org/10.1016/S0001-4575(00)00100-7
3. Brod, D., & Gillen, D. (2020). *New Model for Highway-Rail Grade Crossing Accident Prediction And Severity*. U.S. Department of Transportation. doi:10.13140/rg.2.2.17487.71849
4. Chadwick, S. G., Zhou, N., & Saat, M. R. (2014). *Highway-Rail Grade Crossing Safety Challenges for Shared Operations of High-Speed Passenger and Heavy Freight Rail in the U.S.* Safety Science 68 (2014) 128–137.
5. Faghri, A., & Demetsky, M. J. (1986). Evaluation of Methods for Predicting Rail-Highway Crossing Hazards. Final Report. Virginia Highway and Transportation Council.
6. Federal Railroad Administration Office of Safety Analysis. Retrieved from https://safetydata.fra.dot.gov/OfficeofSafety/Default.aspx
7. GeNIe Modeler User Manual [computer software] (2020).
8. Hall, S. (2007). Reducing Risk at Automatically Operated Level Crossings on Public Roads. IET Seminar on Reducing Risk at the Road Rail Interface.
9. Horton, Suzanne M., United States. Federal Railroad Administration. Office of Research and Development, and John A. Volpe National Transportation Systems Center (U.S.). 2009. *Success Factors in the Reduction of Highway-Rail Grade Crossing Incidents from 1994 to 2003*. Washington, D.C.: U.S. Dept. of Transportation Federal Railroad Administration, Office of Research and Development. Retrieved May 5, 2021 (INSERT-MISSING-URL).
10. Koch, K. (2007). *Introduction to Bayesian Statistics* (2. Aufl. ed.). Berlin, Heidelberg: Springer-Verlag. doi:10.1007/978-3-540-72726-2
11. Microsoft access [computer software] (2019).
12. Mok, S. C., & Savage, I. (2005). Why Has Safety Improved at Rail-Highway Grade Crossings? *Risk Analysis: An International Journal, 25*(4), 867-881.
13. RStudio [computer software] (2021).
14. Pearl, J., & Russel, S. (2000). Bayesian Networks. *Handbook of Brain Theory and Neural Networks, R-277*
15. Foundation for Traffic Safety, Rates of Motor Vehicle Crashes, Injuries and Deaths in Relation to Driver Age, United States, 2014-2015. (2017). Retrieved from https://aaafoundation.org/rates-motor-vehicle-crashes-injuries-deaths-relation-driver-age-united-states-2014-2015/
16. Soleimani, S., Mousa, S. R., Codjoe, J., & Leitner, M. (2019). A Comprehensive Railroad-Highway Grade Crossing Consolidation Model: A Machine Learning Approach. *Accident Analysis & Prevention, 128*, 65-77.
17. Washington, S., & Oh, J. (2006). *Bayesian Methodology Incorporating Expert Judgment for Ranking Countermeasure Effectiveness under Uncertainty: Example Applied to at Grade Railroad Crossings in korea* doi:https://doi.org/10.1016/j.aap.2005.08.005

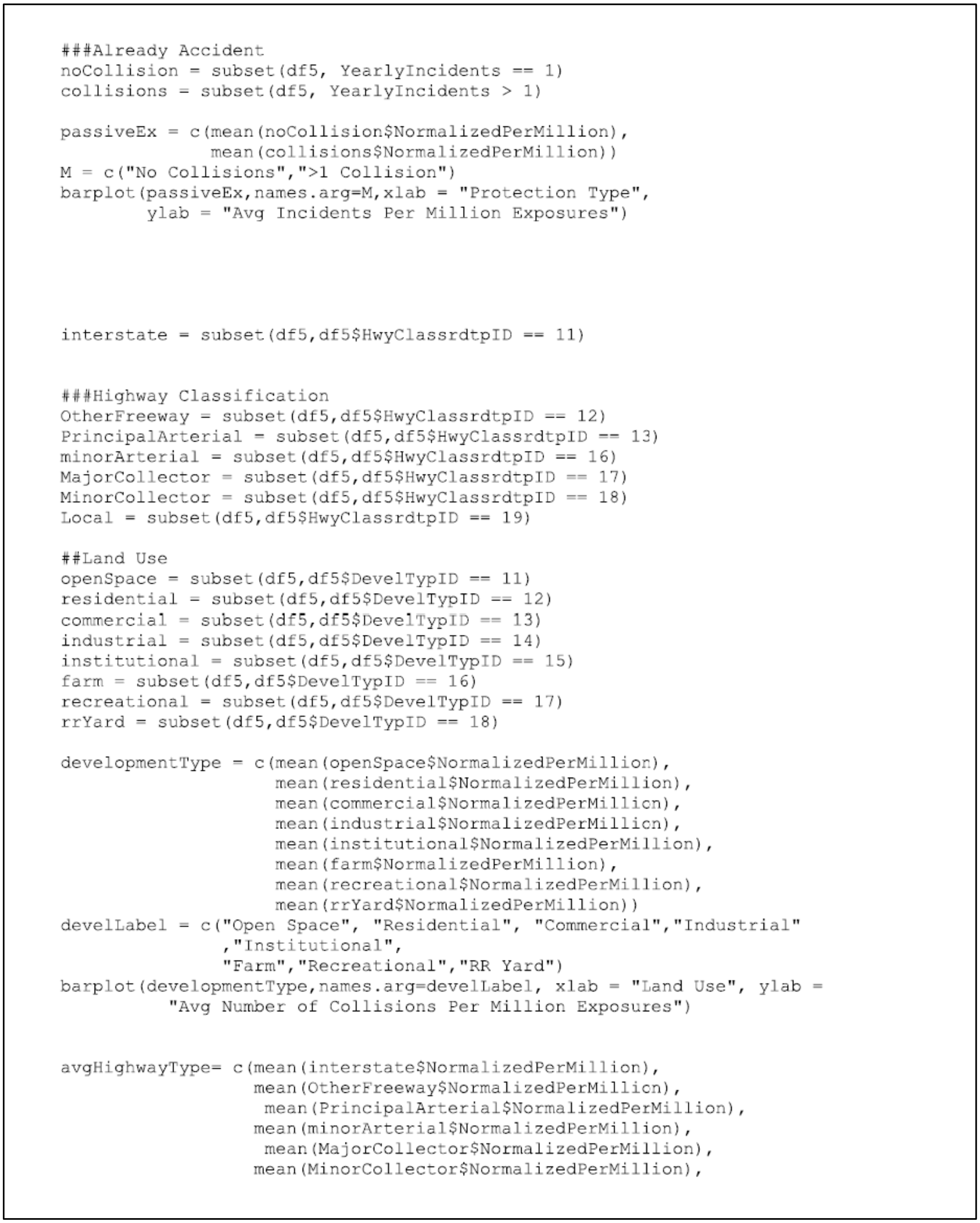
Appendix A

R Scripts



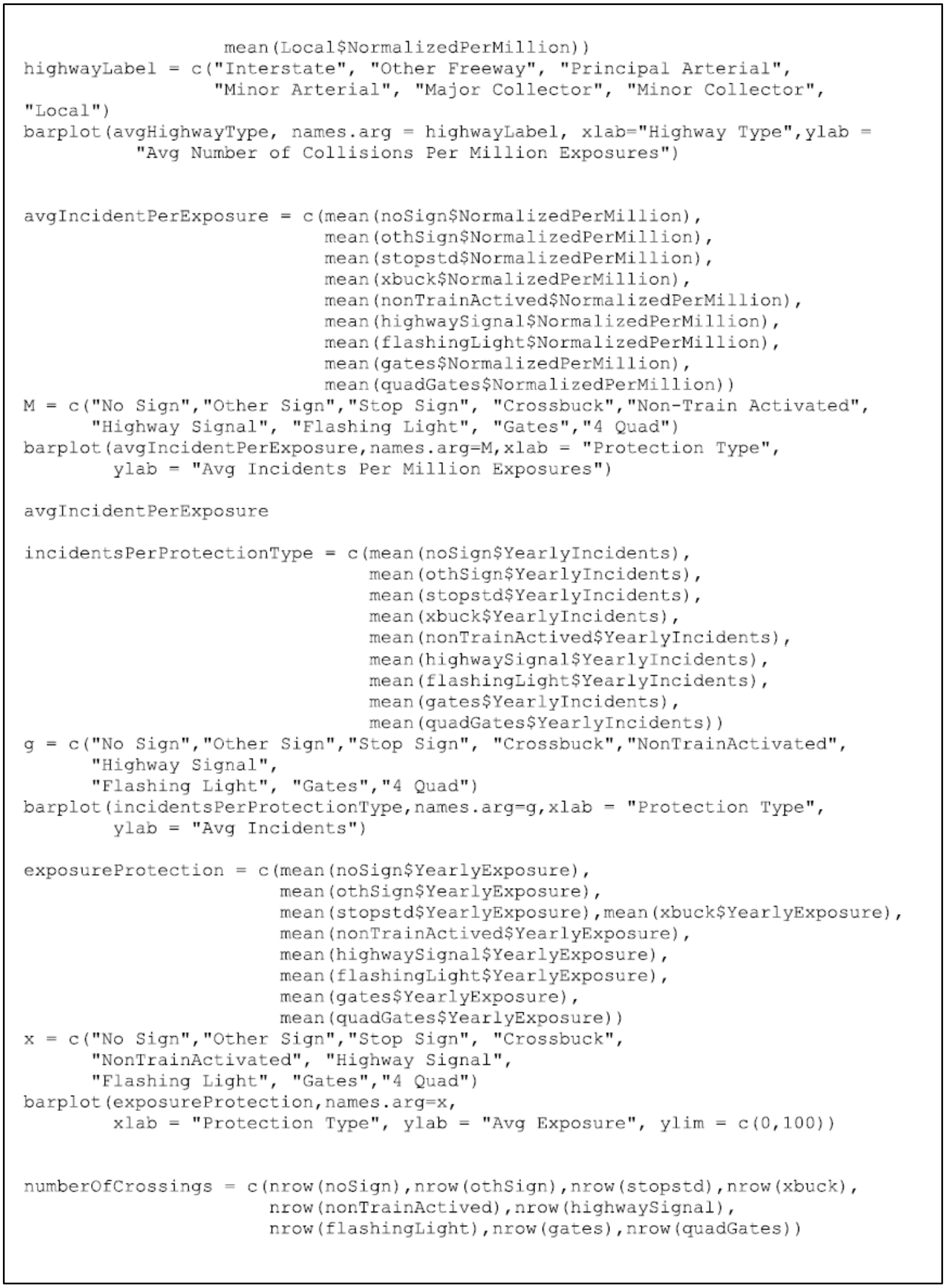
```
library(RODBC)
library(tidyr)
library(ggplot2)
library(plyr)
ch <- odbcConnect("MS Access Database")
incidents = sqlFetch(ch, "For R")
crossingData = sqlFetch(ch, "CrossingIds")

x = ddply(incidents,.(CrossingID),nrow)

df0 = merge(crossingData,x,by="CrossingID", all = TRUE)
df0$V1[is.na(df0$V1)] = 0



df =df0
##Cleaning
na_vec = which(!complete.cases(df))
df1= df[-na_vec,]
df2 = subset(df1, df1$TypeXing == 3 & df1$PosXing == 1)
df3 = subset(df2, df2$CrossingClosed == "No")
df4 = subset(df3, df3$Aadt != 0)
df5 = subset(df4, (df4$DayThru + df4$NghtThru) != 0)

hist(df5$V1, xlim = c(0,7), xlab = "Number of Collisions Per Crossing" ,
     ylab = "Frequency", main = "Histogram of Number of Collisions Per
Crossing"
     , breaks=40)

df5$YearlyExposure = df5$Aadt*(df5$DayThru+df5$NghtThru)*365/10^6
df5$YearlyIncidents = df5$V1/20
df5$NormalizedPerMillion = df5$YearlyIncidents/df5$YearlyExposure

##Crossing Protection
noSign = subset(df5, df5$WdCode == 1)
othSign = subset(df5, df5$WdCode == 2)
stopstd = subset(df5, df5$WdCode == 3)
xbuck = subset(df5, df5$WdCode == 4)
nonTrainActived = subset(df5, df5$WdCode == 5)
highwaySignal = subset(df5, df5$WdCode == 6)
flashingLight = subset(df5, df5$WdCode == 7)
gates = subset(df5, df5$WdCode == 8)
quadGates = subset(df5, df5$WdCode == 9)


###Type of Crossing
passive = subset(df5, WdCode > 0 & WdCode < 5)
lights = subset(df5, WdCode > 4 )
gates = subset(df5, WdCode > 7)

passiveEx = c(mean(passive$NormalizedPerMillion),
              mean(lights$NormalizedPerMillion),
              mean(gates$NormalizedPerMillion))
M = c("Passive","Lights","Gates")
barplot(passiveEx,names.arg=M,xlab = "Protection Type",
        ylab = "Avg Incidents Per Million Exposures")
```

Figure 28: Exploratory Data Analyses Page 1



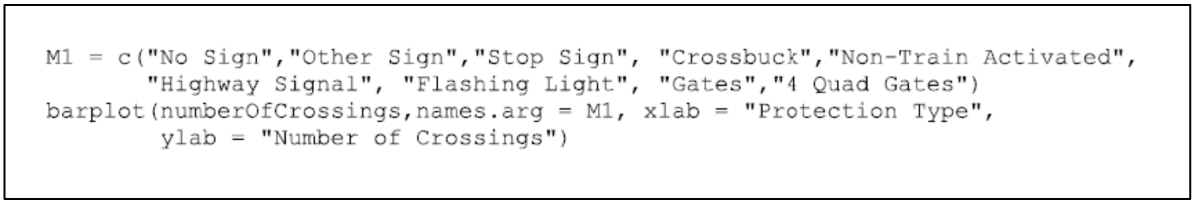
```
###Already Accident
noCollision = subset(df5, YearlyIncidents == 1)
collisions = subset(df5, YearlyIncidents > 1)

passiveEx = c(mean(noCollision$NormalizedPerMillion),
              mean(collisions$NormalizedPerMillion))
M = c("No Collisions",">1 Collision")
barplot(passiveEx,names.arg=M,xlab = "Protection Type",
        ylab = "Avg Incidents Per Million Exposures")




interstate = subset(df5,df5$HwyClassrdtpID == 11)


###Highway Classification
OtherFreeway = subset(df5,df5$HwyClassrdtpID == 12)
PrincipalArterial = subset(df5,df5$HwyClassrdtpID == 13)
minorArterial = subset(df5,df5$HwyClassrdtpID == 16)
MajorCollector = subset(df5,df5$HwyClassrdtpID == 17)
MinorCollector = subset(df5,df5$HwyClassrdtpID == 18)
Local = subset(df5,df5$HwyClassrdtpID == 19)

##Land Use
openSpace = subset(df5,df5$DevelTypID == 11)
residential = subset(df5,df5$DevelTypID == 12)
commercial = subset(df5,df5$DevelTypID == 13)
industrial = subset(df5,df5$DevelTypID == 14)
institutional = subset(df5,df5$DevelTypID == 15)
farm = subset(df5,df5$DevelTypID == 16)
recreational = subset(df5,df5$DevelTypID == 17)
rrYard = subset(df5,df5$DevelTypID == 18)

developmentType = c(mean(openSpace$NormalizedPerMillion),
                    mean(residential$NormalizedPerMillion),
                    mean(commercial$NormalizedPerMillion),
                    mean(industrial$NormalizedPerMillion),
                    mean(institutional$NormalizedPerMillion),
                    mean(farm$NormalizedPerMillion),
                    mean(recreational$NormalizedPerMillion),
                    mean(rrYard$NormalizedPerMillion))
develLabel = c("Open Space", "Residential", "Commercial","Industrial"
               ,"Institutional",
               "Farm","Recreational","RR Yard")
barplot(developmentType,names.arg=develLabel, xlab = "Land Use", ylab =
          "Avg Number of Collisions Per Million Exposures")


avgHighwayType= c(mean(interstate$NormalizedPerMillion),
                  mean(OtherFreeway$NormalizedPerMillion),
                   mean(PrincipalArterial$NormalizedPerMillion),
                  mean(minorArterial$NormalizedPerMillion),
                   mean(MajorCollector$NormalizedPerMillion),
                  mean(MinorCollector$NormalizedPerMillion),
```

Figure 29: Exploratory Data Analyses Page 2



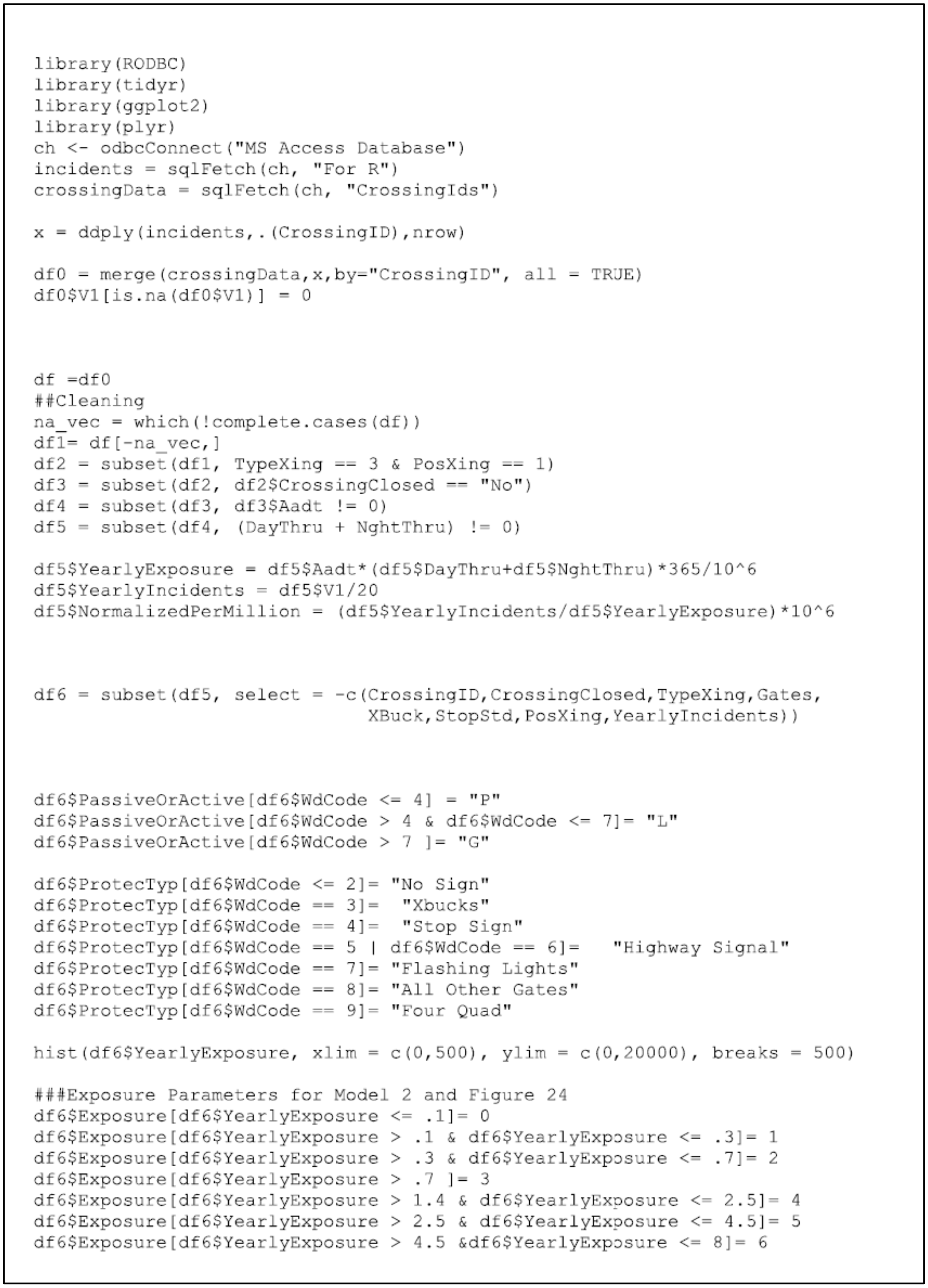
```
                  mean(Local$NormalizedPerMillion))
highwayLabel = c("Interstate", "Other Freeway", "Principal Arterial",
                 "Minor Arterial", "Major Collector", "Minor Collector",
"Local")
barplot(avgHighwayType, names.arg = highwayLabel, xlab="Highway Type",ylab =
          "Avg Number of Collisions Per Million Exposures")


avgIncidentPerExposure = c(mean(noSign$NormalizedPerMillion),
                           mean(othSign$NormalizedPerMillion),
                           mean(stopstd$NormalizedPerMillion),
                           mean(xbuck$NormalizedPerMillion),
                           mean(nonTrainActived$NormalizedPerMillion),
                           mean(highwaySignal$NormalizedPerMillion),
                           mean(flashingLight$NormalizedPerMillion),
                           mean(gates$NormalizedPerMillion),
                           mean(quadGates$NormalizedPerMillion))
M = c("No Sign","Other Sign","Stop Sign", "Crossbuck","Non-Train Activated",
      "Highway Signal", "Flashing Light", "Gates","4 Quad")
barplot(avgIncidentPerExposure,names.arg=M,xlab = "Protection Type",
        ylab = "Avg Incidents Per Million Exposures")

avgIncidentPerExposure

incidentsPerProtectionType = c(mean(noSign$YearlyIncidents),
                               mean(othSign$YearlyIncidents),
                               mean(stopstd$YearlyIncidents),
                               mean(xbuck$YearlyIncidents),
                               mean(nonTrainActived$YearlyIncidents),
                               mean(highwaySignal$YearlyIncidents),
                               mean(flashingLight$YearlyIncidents),
                               mean(gates$YearlyIncidents),
                               mean(quadGates$YearlyIncidents))
g = c("No Sign","Other Sign","Stop Sign", "Crossbuck","NonTrainActivated",
      "Highway Signal",
      "Flashing Light", "Gates","4 Quad")
barplot(incidentsPerProtectionType,names.arg=g,xlab = "Protection Type",
        ylab = "Avg Incidents")

exposureProtection = c(mean(noSign$YearlyExposure),
                       mean(othSign$YearlyExposure),
                       mean(stopstd$YearlyExposure),mean(xbuck$YearlyExposure),
                       mean(nonTrainActived$YearlyExposure),
                       mean(highwaySignal$YearlyExposure),
                       mean(flashingLight$YearlyExposure),
                       mean(gates$YearlyExposure),
                       mean(quadGates$YearlyExposure))
x = c("No Sign","Other Sign","Stop Sign", "Crossbuck",
      "NonTrainActivated", "Highway Signal",
      "Flashing Light", "Gates","4 Quad")
barplot(exposureProtection,names.arg=x,
        xlab = "Protection Type", ylab = "Avg Exposure", ylim = c(0,100))


numberOfCrossings = c(nrow(noSign),nrow(othSign),nrow(stopstd),nrow(xbuck),
                      nrow(nonTrainActived),nrow(highwaySignal),
                      nrow(flashingLight),nrow(gates),nrow(quadGates))
```

Figure 30: Exploratory Data Analyses Page 3



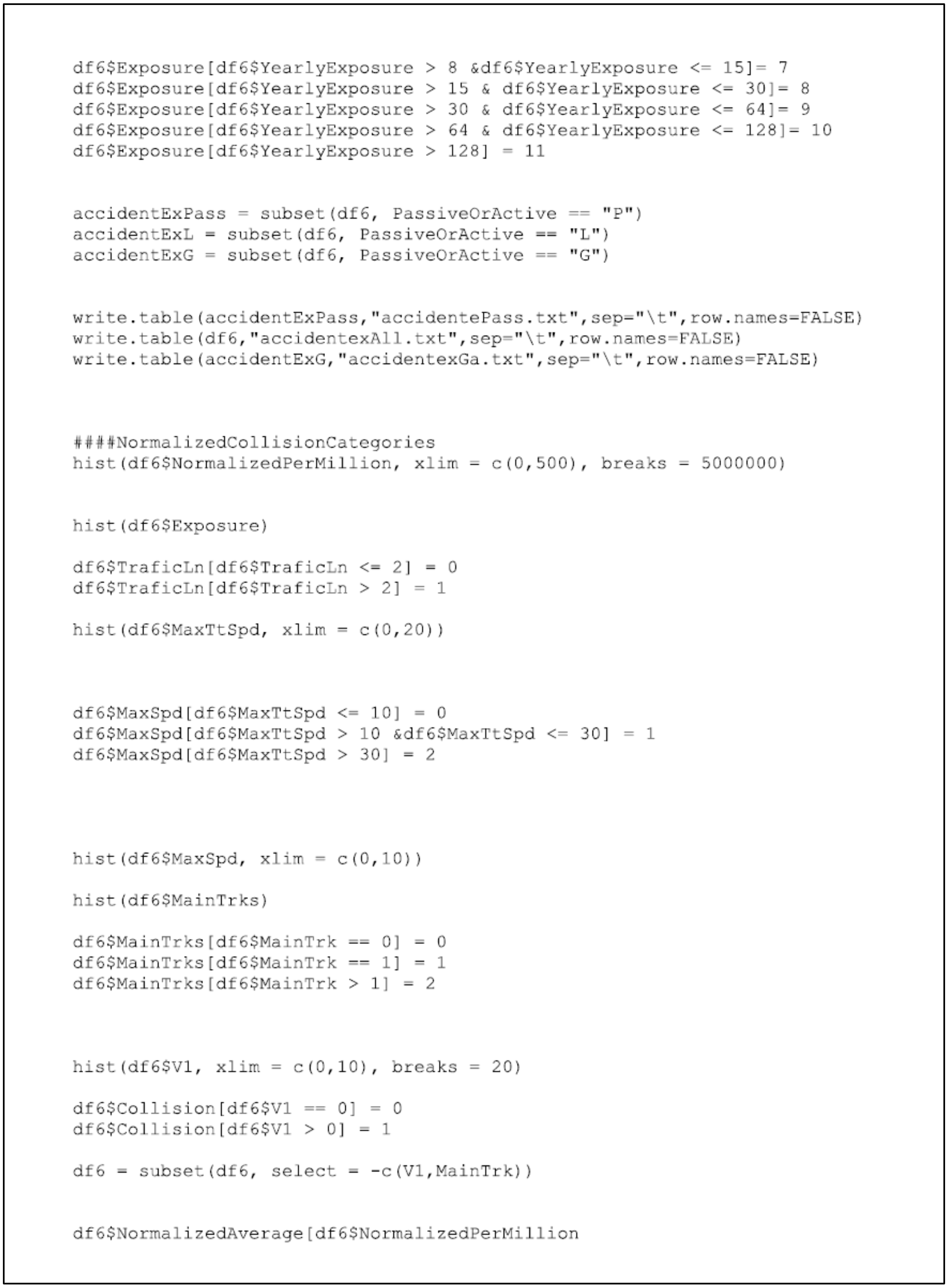
```
M1 = c("No Sign","Other Sign","Stop Sign", "Crossbuck","Non-Train Activated",
       "Highway Signal", "Flashing Light", "Gates","4 Quad Gates")
barplot(numberOfCrossings,names.arg = M1, xlab = "Protection Type",
        ylab = "Number of Crossings")
```

Figure 31: Exploratory Data Analyses Page 4



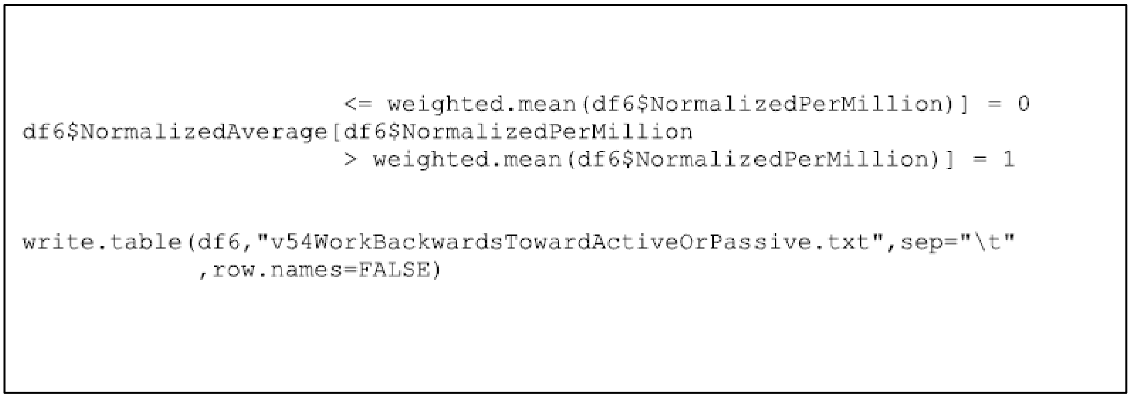
```
library(RODBC)
library(tidyr)
library(ggplot2)
library(plyr)
ch <- odbcConnect("MS Access Database")
incidents = sqlFetch(ch, "For R")
crossingData = sqlFetch(ch, "CrossingIds")

x = ddply(incidents,.(CrossingID),nrow)

df0 = merge(crossingData,x,by="CrossingID", all = TRUE)
df0$V1[is.na(df0$V1)] = 0



df =df0
##Cleaning
na_vec = which(!complete.cases(df))
df1= df[-na_vec,]
df2 = subset(df1, TypeXing == 3 & PosXing == 1)
df3 = subset(df2, df2$CrossingClosed == "No")
df4 = subset(df3, df3$Aadt != 0)
df5 = subset(df4, (DayThru + NghtThru) != 0)

df5$YearlyExposure = df5$Aadt*(df5$DayThru+df5$NghtThru)*365/10^6
df5$YearlyIncidents = df5$V1/20
df5$NormalizedPerMillion = (df5$YearlyIncidents/df5$YearlyExposure)*10^6



df6 = subset(df5, select = -c(CrossingID,CrossingClosed,TypeXing,Gates,
                              XBuck,StopStd,PosXing,YearlyIncidents))



df6$PassiveOrActive[df6$WdCode <= 4] = "P"
df6$PassiveOrActive[df6$WdCode > 4 & df6$WdCode <= 7]= "L"
df6$PassiveOrActive[df6$WdCode > 7 ]= "G"

df6$ProtecTyp[df6$WdCode <= 2]= "No Sign"
df6$ProtecTyp[df6$WdCode == 3]=  "Xbucks"
df6$ProtecTyp[df6$WdCode == 4]=  "Stop Sign"
df6$ProtecTyp[df6$WdCode == 5 | df6$WdCode == 6]=   "Highway Signal"
df6$ProtecTyp[df6$WdCode == 7]= "Flashing Lights"
df6$ProtecTyp[df6$WdCode == 8]= "All Other Gates"
df6$ProtecTyp[df6$WdCode == 9]= "Four Quad"

hist(df6$YearlyExposure, xlim = c(0,500), ylim = c(0,20000), breaks = 500)

###Exposure Parameters for Model 2 and Figure 24
df6$Exposure[df6$YearlyExposure <= .1]= 0
df6$Exposure[df6$YearlyExposure > .1 & df6$YearlyExposure <= .3]= 1
df6$Exposure[df6$YearlyExposure > .3 & df6$YearlyExposure <= .7]= 2
df6$Exposure[df6$YearlyExposure > .7 ]= 3
df6$Exposure[df6$YearlyExposure > 1.4 & df6$YearlyExposure <= 2.5]= 4
df6$Exposure[df6$YearlyExposure > 2.5 & df6$YearlyExposure <= 4.5]= 5
df6$Exposure[df6$YearlyExposure > 4.5 &df6$YearlyExposure <= 8]= 6
```

Figure 32: Model Development Page 1



```
df6$Exposure[df6$YearlyExposure > 8 &df6$YearlyExposure <= 15]= 7
df6$Exposure[df6$YearlyExposure > 15 & df6$YearlyExposure <= 30]= 8
df6$Exposure[df6$YearlyExposure > 30 & df6$YearlyExposure <= 64]= 9
df6$Exposure[df6$YearlyExposure > 64 & df6$YearlyExposure <= 128]= 10
df6$Exposure[df6$YearlyExposure > 128] = 11


accidentExPass = subset(df6, PassiveOrActive == "P")
accidentExL = subset(df6, PassiveOrActive == "L")
accidentExG = subset(df6, PassiveOrActive == "G")


write.table(accidentExPass,"accidentePass.txt",sep="\t",row.names=FALSE)
write.table(df6,"accidentexAll.txt",sep="\t",row.names=FALSE)
write.table(accidentExG,"accidentexGa.txt",sep="\t",row.names=FALSE)



####NormalizedCollisionCategories
hist(df6$NormalizedPerMillion, xlim = c(0,500), breaks = 5000000)


hist(df6$Exposure)

df6$TraficLn[df6$TraficLn <= 2] = 0
df6$TraficLn[df6$TraficLn > 2] = 1

hist(df6$MaxTtSpd, xlim = c(0,20))



df6$MaxSpd[df6$MaxTtSpd <= 10] = 0
df6$MaxSpd[df6$MaxTtSpd > 10 &df6$MaxTtSpd <= 30] = 1
df6$MaxSpd[df6$MaxTtSpd > 30] = 2




hist(df6$MaxSpd, xlim = c(0,10))

hist(df6$MainTrks)

df6$MainTrks[df6$MainTrk == 0] = 0
df6$MainTrks[df6$MainTrk == 1] = 1
df6$MainTrks[df6$MainTrk > 1] = 2



hist(df6$V1, xlim = c(0,10), breaks = 20)

df6$Collision[df6$V1 == 0] = 0
df6$Collision[df6$V1 > 0] = 1

df6 = subset(df6, select = -c(V1,MainTrk))


df6$NormalizedAverage[df6$NormalizedPerMillion
```

Figure 33: Model Development Page 2



```
                      <= weighted.mean(df6$NormalizedPerMillion)] = 0
df6$NormalizedAverage[df6$NormalizedPerMillion
                      > weighted.mean(df6$NormalizedPerMillion)] = 1


write.table(df6,"v54WorkBackwardsTowardActiveOrPassive.txt",sep="\t"
            ,row.names=FALSE)
```

Figure 34: Model Development Page 3

ACKNOWLEDGEMENT

This study was conducted with the support from the USDOT Tier 1 University Transportation Center on Railroad Sustainability and Durability.

ABOUT THE AUTHORS

Tyler Bernstein

Tyler Bernstein was a undergraduate Research Assistant when he worked on this research project. He obtained his Bachelor's degree from the University of Deleware.

Joseph W. Palese, MCE, PE

Mr. Palese is a Senior Scientist and Program Manager of Railroad Engineering and Safety Program at the University of Delaware. He has over 28 years of experience in track component design and analysis, failure analysis and component life forecasting algorithm specifications, and development of inspection systems. Throughout his career, Mr. Palese has focused on acquiring and utilizing large amounts of track component condition data for planning railway maintenance activities.

Mr. Palese has a Bachelor's Degree of Civil Engineering, and a Master's Degree of Civil Engineering, both from the University of Delaware, along with a MBA from Rowan University. He is currently pursuing his PhD in Civil Engineering at the University of Delaware. He is a registered Professional Engineer in the state of New Jersey.

Allan M. Zarembski, Ph.D., P.E., Hon. Mbr. AREMA, FASME

Dr. Zarembski is an internationally recognized authority in the fields of track and vehicle/track system analysis, railway component failure analysis, track strength, and maintenance planning. Dr. Zarembski is currently Professor of Practice and Director of the Railroad Engineering and Safety Program at the University of Delaware's Department of Civil and Environmental Engineering, where he has been since 2012. Prior to that he was President of ZETA-TECH, Associates, Inc. a railway technical consulting and applied technology company, he established in 1984. He also served as Director of R&D for Pandrol Inc., Director of R&D for Speno Rail Services Co. and Manager, Track Research for the Association of American Railroads. He has been active in the railroad industry for over 40 years.

Dr. Zarembski has PhD (1975) and M.A (1974) in Civil Engineering from Princeton University, an M.S. in Engineering Mechanics (1973) and a B.S. in Aeronautics and Astronautics from New York University (1971). He is a registered Professional Engineer in five states. Dr. Zarembski is an Honorary Member of American Railway Engineering and Maintenance of way Association (AREMA), a Fellow of American Society of Mechanical Engineers (ASME), and a Life Member of American Society of Civil Engineers (ASCE). He served as Deputy Director of the Track Train Dynamics Program and was the recipient of the American Society of Mechanical Engineer's Rail Transportation Award in 1992 and the US Federal Railroad Administration's Special Act Award in 2001. He was awarded The Fumio Tatsuoka Best Paper Award in 2017 by the Journal of Transportation Infrastructure Geotechnology

He is the organizer and initiator of the **Big Data in Railroad Maintenance Planning Conference** held annually at the University of Delaware. He has authored or co-authored over 200 technical papers, over 120 technical articles, two book chapters and two books.

Tyler Bernstein
Undergraduate Research Assistant
Department of Civil & Environmental Engineering
University of Delaware

Joseph Palese, PhD, MBA, PE
Senior Scientist
Department of Civil & Environmental Engineering
University of Delaware
palesezt@udel.edu

Allan Zarembski PhD, PE, FASME, Hon. Mbr. AREMA
Professor and Director of Railroad Engineering and Safety Program
Department of Civil & Environmental Engineering
University of Delaware
dramz@udel.edu