Inferring accident severity in Michigan

Ameen Alkhabbaz, Jordan Burton, David Emery, George McKenzie, Mahzad Zavareh

Abstract

Traffic accidents are an economic and societal burden that could be prevented through risk management strategies. To implement risk management strategies, factors that increase the risk of a traffic accident need to be well understood. For this purpose, we combined Michigan accident data from 2016-2019 with 2017 predicted census data. Plotting the total number of accidents in each county to a map of Michigan highlighted the most accident-prone counties. The top three counties with the most accidents in Michigan are Genesee, Wayne, and Kent counties. Surprisingly, the highest frequency of accidents in each of these counties occur on a highway, where road infrastructure like traffic lights are not present. The highest frequency of accidents from 2016 - 2019 in Michigan occurred at I-75 Exit 118 in Genesee County. This location was explored in our analysis.

Background

Traffic accidents account for over 38,000 fatalities and over 4.4 million injuries in the United States (U.S.) every year. Traffic accidents have an economical and societal impact costing over $850 billion tax dollars in the U.S. per year. Mitigating traffic accidents by reducing risk in the U.S. is critical because the number of fatalities is over 50% higher than other high-income countries including Canada, Australia, and Japan. Of all the U.S. traffic accidents in 2018 there were 905 fatal crashes in Michigan, resulting in 974 deaths. 53% of fatal crashes in Michigan involved more than one vehicle and/or alcohol use.

Factors that contribute to traffic accidents include poor management and road infrastructure, unsafe road user behaviors, unenforced or non-existent traffic laws, and non-road worthy vehicles. Understanding the contribution of these factors to traffic accidents and accident severity is important for predicting and preventing accidents through planning, management and evidence-based interventions. To further investigate the above factors, U.S. accident[a] data and predicted 2017 census[b] data was obtained from Kaggle. The accident data includes metrics such as *severity*, *location*, *weather condition*, *time*, and *roadway features* near an accident. The census data includes county-specific metrics such as *population*, *sex*, *ethnicity*, *income*, *employment*, and *transportation*. Michigan-specific traffic accident data from 2016 – 2019 was combined with 2017 predicted census data to determine if accident severity could be inferred from accident or census metrics, discern the cause of accidents in specific locations, and to find patterns in the accident data.

Our group faced several challenges when analyzing the Michigan dataset due to lack of information, categories with a low frequency of observations, and categories with a high number of levels. The Michigan dataset does not include information on either accident fatalities or road construction, businesses, and traffic laws near an accident. The Michigan dataset contains incomplete information both for accidents in the years 2016 and 2020 and for the traffic message channel code. The *severity* categories in the Michigan dataset contain two categories with a high number of observations and two categories with a low number of observations. The *city* and *county* variables in the Michigan dataset had too many levels so could not be used as an input in our decision tree model.

Methodology

The Michigan accident and census datasets were prepared for statistical analysis with the "dplyr", "tidyr" and "tidyverse" packages. To prepare both datasets we recoded duplicates in the *county* column and converted column variable types as needed. Duplicates, based on naming differences, in the *county* column were recoded to provide accurate statistics for data visualizations and inference as detailed below. Column variables were converted to the appropriate type (i.e. factors, dates, and strings) based on the contents of each column. The Michigan accident and census data were joined by the *county* variable to form the Michigan dataset. Factor levels were added to the Michigan dataset variables *accident severity*, *hour*, and *sunrise/sunset*. The *accident severity* variable factors in the Michigan dataset were reduced from 4 levels to 2 levels because two levels were infrequently used. Finally, "Per capita" were calculated for the *accident*, *employed*, *men*, *women*, and *voting age* variables to account for the total population of each county.

The R-packages “ggplot2”, “ggmaps”, “mapproj”, and “maps” were used to generate all figures for analysis of the Michigan dataset. The relationship between variables in the Michigan dataset were mapped to correlation plots and to a map of Michigan Counties. The R package “Corrplot” was used to build the correlation matrix. The matrix plots the correlation between each variable with possible values ranging from –1, an inverse relationship, to 1, variables move in the same direction, and 0, no relationship between variables. The “Corrplot” package nicely visualizes these relationships to identify variable relationships that warrant further analysis. These correlation plots were used to determine if there was a significant relationship between accidents per capita and the census *population, gender, race, voting age citizen, profession, and commute type* variables. Plotting variables from the Michigan dataset to a map of Michigan allowed us to delineate patterns in the accident data and to identify the frequency of accidents on specific streets. Visualizations not shown in this document are available in our GitHub repository at (https://github.com/JBB-bio/Group-4-Final-Report).

Linear regression analysis of the Michigan dataset allowed us to explore *accident* as a response to other variables in the dataset and build a prediction model. Linear regression assumes that a relationship between the response variable can be predicted with a linear equation in the slope intercept form with a normally distributed random error term with a mean of zero. In the case of qualitative predictors one of the variable levels is 0, where the equation is defined by the intercept, other predictor terms in the equation, and the error term. All other levels of the factor variable take on a positive or negative coefficient value that distinguish them from the variable level set to 0. The linear regression model was built using the qualitive variables *county* and *time of day* for a training set of 2017 and 2018 data. The linear regression model was then used to predict the monthly accident volume for counties in 2019.

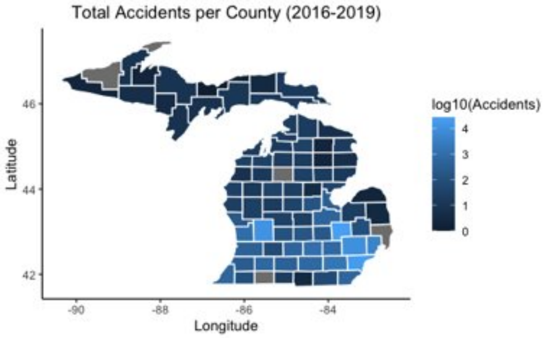
The Michigan dataset contains the variable *accident severity* which was reduced from 4 levels to the two levels “Less Severe” and “More Severe” as described above contains. We used the *accident severity* variable as a classification response to make inferences about which variables may be more predictive of *accident severity* for the 2017 – 2019 data. Inferences about *accident severity* were made using decision tree modeling. The “RPART” package was used to build the decision tree in the Rattle GUI, then R code was exported to review “RPART’s” built-in cross-validation in more detail.

In the classification setting, a decision tree predicts the response variable by assigning the most commonly occurring class in the training data set. The decision tree prediction procedure is repeated starting at the top with no data spits. At each node the split is made with the predictors that make the best split, without looking several steps ahead for a more optimal split. The node split decision is made with the objective of increasing node purity. The reason for selecting a decision tree model was to understand what kinds of inferences can be made about accident severity from our predictors and the decision tree model provides a simple way to explain the relationship between the response and predictor variables.

Normally, logistic regression is performed when the dependent variable is binary or continuous. However, an ordinal logistic regression model was developed with *accident severity* as the response variable in an ordinal logistic regression model. Ordinal logistic regression was chosen because *accident severity* consists of 4 ordered levels. The ordinal regression allows visualization and analysis of the relationship between all types of predictors with the response variable. We used ordinal logistic regression to model *accident severity* in Wayne County.

Finally, we performed an in-depth case study of I-75 Exit 118 in Flint, Michigan. This location is important as it has the most accident occurrences in Genesee County. The case study is a deep dive into factors that contribute to accidents at this location. The case study uses a combination of the above models and map visualizations to understand why accidents so frequently occur at Exit 118.

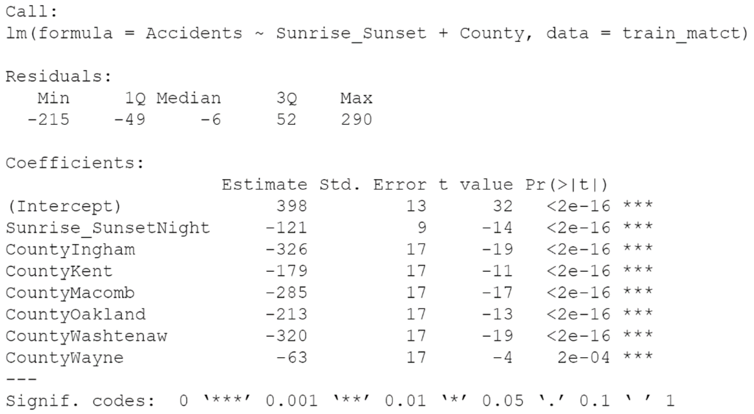
Results:



Maps and correlation plots were used to understand the relationship between different variables in the Michigan dataset. Maps, like the one shown to the left, were used to determine if there is a pattern between *accidents, accidents per capita,* and *population* in each Michigan county The figure on the left shows that Genesee, Wayne, and Kent counties have the highest number of accidents in Michigan and shows that counties above Latitude 44 have a low number of accidents. Correlation plots were used to visualize the relationship that the variables *gender*, *race*, *voting age citizen*, *profession*, and *commute types* have with *accidents per capita*. The correlation plot in the appendix displays how statistically related each variable is to other variables. *Accidents a*re positively correlated with *population,* but there were no other relevant strong correlations with *accidents*. Variables that are strongly correlated with *population* (e.g. *gender, voting age citizen*, *race*) were also positively correlated with *accidents*, though the relationship was weaker. Surprisingly, there was not a strong association between *accidents* and *commuting* variables (e.g *drive*, *transit*).

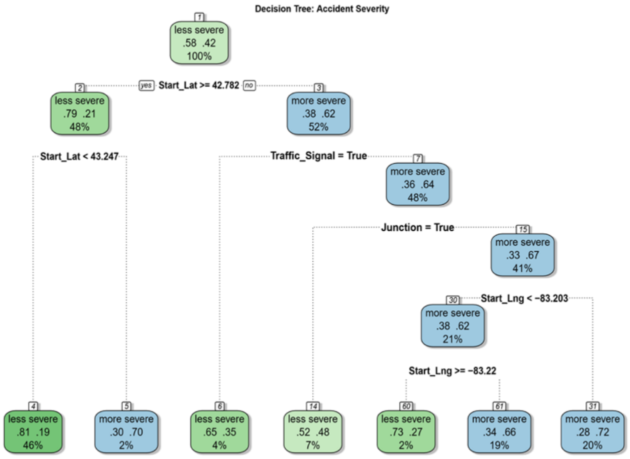
Linear regression analysis was used to predict the frequency of accidents using the predictors *time of day, month,* and county. Two linear regression models were considered for prediction, as shown in the table to the right. There was a lower test MSE (6,130) for the simpler model when no distinction between months was made; this is compared with the more complex model including each month as a predictor where the test MSE was higher (6,245). Both models are compared in the above table . Having had the test set and calculating the Test MSE was helpful in picking the better model, looking at both models Adjusted R-Squared the more complex model looks slightly more attractive as it received a higher Adjusted R-Squared (0.7262) versus the simpler model (0.7218).

Linear Regression Model	Test MSE	Adjusted R-Squared
lm(Accidents ~Sunrise_Sunset + County)	6,130	0.7218
lm(Accidents~Sunrise_Sunset + County + Month)	6,245	0.7262



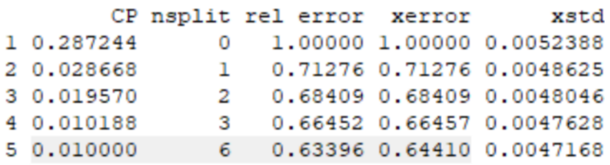
```
Call:
lm(formula = Accidents ~ Sunrise_Sunset + County, data = train_matct)

Residuals:
   Min     1Q Median     3Q    Max
  -215    -49     -6     52    290

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)              398         13      32   <2e-16 ***
Sunrise_SunsetNight     -121          9     -14   <2e-16 ***
CountyIngham            -326         17     -19   <2e-16 ***
CountyKent              -179         17     -11   <2e-16 ***
CountyMacomb            -285         17     -17   <2e-16 ***
CountyOakland           -213         17     -13   <2e-16 ***
CountyWashtenaw         -320         17     -19   <2e-16 ***
CountyWayne              -63         17      -4    2e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The table on the left, contains the simple linear regression model’s coefficients. Genesee county, where the city of Flint is located, was the county predicted to receive the highest number of accidents. In the model, Genesee county has a coefficient of 0 and daytime also has a coefficient of 0. The predicted monthly volume of daytime accidents in Genesee county is equal to the intercept coefficient at 398, with a prediction interval of 235 to 560 monthly accidents. The coefficient for night-time accidents is -121; there are a predicted 277 monthly night-time accidents in Genesee county with a prediction interval of 114 to 439 monthly accidents. The county coefficients for Wayne county –63 and Kent –179 subtract the least amount from the intercept indicating they are predicted to have the second highest accidents.

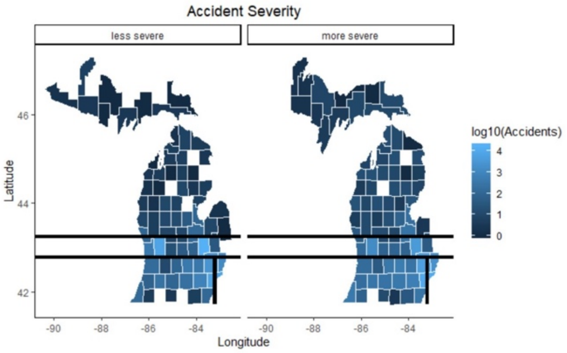


A decision tree model of the Michigan accident data, shown left, was used to segment the data for inference in a way that allows us to understand *accident severity* in Michigan based on a series of splitting rules. Variables considered in the model included *latitude, longitude, roadway features* (e.g. *traffic signal, junction*), *time* (e.g. *month, time of day*). The splitting criteria for the tree model from the above variables were *junction, latitude, longitude* and *traffic signal*. Model cross-validation and complexity parameter limited the need for tree pruning by avoiding unnecessary splits.

The printcp() table is displayed on the right and in Rattle. The lowest cross-validation error "xerror" .06452 is at the 6 splits "nsplit" and is at the default minimum RPART complexity parameter "CP" of 0.01. The decision tree is in fact pruned at this, the lowest displayed cross-validation error point. The lowest cross-validation error rate was chosen resulting in a tree depth to 6 splits (7 terminal nodes); this is also the number of splits at the minimum complexity threshold meaning a deeper tree could be considered if our goal was prediction accuracy. At the root node of the decision tree model, 58% of accidents are "less severe" and 42% are "more severe". The decision tree model was built with a training set consisting of 70% of the data, 15% was reserved for a validation set, and the model was tested using a test set that consists of the other 15% of the data resulting in an overall test error rate of 26%~~.~~



```
        CP nsplit rel error  xerror      xstd
1 0.287244      0   1.00000 1.00000 0.0052388
2 0.028668      1   0.71276 0.71276 0.0048625
3 0.019570      2   0.68409 0.68409 0.0048046
4 0.010188      3   0.66452 0.66457 0.0047628
5 0.010000      6   0.63396 0.64410 0.0047168
```



In the decision tree model, latitude or longitude were the splitting criteria (See appendix) for 4 of the 6 node splits. The splits are visualized in the map on the left. The first split off the decision tree's terminal node is made on latitude 42.782 splitting the state in half with more severe accidents happening in the southern part of the state. However, accidents near traffic signals and junctions are less severe in southern Michigan. The highest concentration of more severe accidents is in south-east Michigan (Detroit). As for the northern part of the state, less severe accidents more frequently happen above latitude 43.247 and there is a central band of Michigan with more severe accidents (Grand Rapids and Flint).

The ordinal logistic regression model produced the table shown in the appendix, that displays the predictors *Start_Lat*, *Start_Lng*, *Distance.mi.*, *SideR*, *Pressure.in.*, *Visibility.mi.*, *Wind_Speed.mph.*, *Crossing*, *Station*, and *Traffic_Signal* and their effect on *accident severity* level. All their p-values are statistically significant as they are less than the default value of alpha 0.05. The plots, also shown in the appendix, display the relationship between some of these predictors with *accident severity*. Since the dataset has larger amount of *accident severity* records that are level 2 and level 3, level 1 and level 4 did not have insightful plots. However, it is clear how the probability of *accident severity* decreases for level 3 and increases for level 2 when an accident takes place in a traffic signal. The same effect can also be seen for *accident severity* levels 2 and 3 as the start latitude increases which means that the location moves toward the north. On other hand, the probability of *accident severity* level 3 increases and level 2 decreases as the start longitude decreases which means that the location moves toward the east.

We set out to identify why accidents were occurring in a specific location using the variables in the Michigan dataset. The goal of this analysis is to provide feasible recommendations to reduce the number of accidents and accident severity at that specific location. The city of Flint in Genesee county had the highest number of accidents on any given year, so was used in our case study. The top three accident locations in Flint were identified using street filters and road infrastructure description, all three locations were highway roads. Of the top three accident locations, two accident sites were on I-75 Northbound at Exit 118 M-21 and I-75 Southbound at Exit 118 M-21. As this is the same location, we decided to explore other variables to discern the cause of these repetitive accidents.

Traffic variables in the Michigan dataset should provide evidence of what causes crashes on the roadway. However, at the I-75 Northbound at Exit 118 M-21 and I-75 Southbound at Exit 118 M-21 crash sites, the traffic variables (*Amenity, Bump, Crossing, Give Way, Traffic Signal*, etc.) were not applicable because these are highway locations. To explore the traffic variables further we identified the top accident sites in Wayne and Kent counties. However, the top accident sites in Wayne and Kent county were also on highways; so, the traffic variables were irrelevant to those locations. This allowed us to focus on the I-75 at Exit 118 locations in Flint.

Weather conditions during an accident and time of an accident may also contribute to accident frequency and severity in a location. The effect of weather conditions on accident frequency and severity was explored using a summary table. The summary table identified outliers in the weather variables at the crash site. However, most crashes took place under fair weather conditions. The time and day of the week played an integral role in accident frequency. Crashes were significantly higher Monday to Friday than on Saturday and Sunday. Furthermore, the count of accidents occurring between (7:00 – 9:00 am) and (4:00 – 6:00 pm) were drastically higher than any other hours of the day. This coincides directly with when traffic would be the most hectic for people driving to and from work.

We mapped the I-75 Exit 118 crash data to further understand the location and why accidents frequently occur there. Exit 118 is the main exit for Kettering University, is a very short drive to one of Flints largest employers and is less than a mile away from the I-69 and I-75 interchange. We argue that the quick decisions made by drivers to exit one highway and merge onto another, the speed limit of 70 miles an hour, and the two lanes of traffic creating a congested merging situation lead to an increased frequency of accidents at this location. These factors are amplified on the weekdays when everyone is in a rush to get to work.

Conclusion

Our analysis of Michigan accident data was funneled from a state-wide view to one specific accident location in Flint, Michigan. The Flint accident location case study provided further evidence to our county level linear regression analysis on time of day as a predictor for accidents. The Flint accident location at I-75 Exit 118 was the crash site with the highest frequency of crashes in Michigan. Other locations were also studied and at each location different variables could be identified as a factor underlying accidents. On an individual site basis, we believe the ability to filter data down, and graph accident frequency can give us a big picture view of what is causing accidents. Additionally, we believe that using outside resources and researching specific accident sites in congruence with using the data will provide the most in-depth look at accident risk in a location. All-in-all, we believe that this analysis could be used to build effective risk management strategies by understanding where crashes are occurring and the risks associated with each site.

DATA SOURCES:

a) Moosavi, S. (2020, January 17). US Accidents. Retrieved from https://www.kaggle.com/sobhanmoosavi/us-accidents

B) 2017 Population Estimates - United States Census Bureau. (n.d.). Retrieved from https://www.census.gov/data.html

REFERENCES:

An Introduction to corrplot Package. (n.d.). Retrieved from https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html

Fatality Facts 2018 State by State. (2019, December). Retrieved from https://www.iihs.org/topics/fatality-statistics/detail/state-by-state#rural-versus-urba

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* . Retrieved from http://faculty.marshall.usc.edu/gareth-james/ISL/

Michigan Department of Transportation. (n.d.). Retrieved from https://www.michigan.gov/mdot/

Michigan State Police. (n.d.). Retrieved from https://www.michigan.gov/msp/

NCSA Tools, Publications, and Data. (n.d.). Retrieved from https://cdan.nhtsa.gov/

R Package Documentation . (n.d.). Retrieved from https://cran.r-project.org/

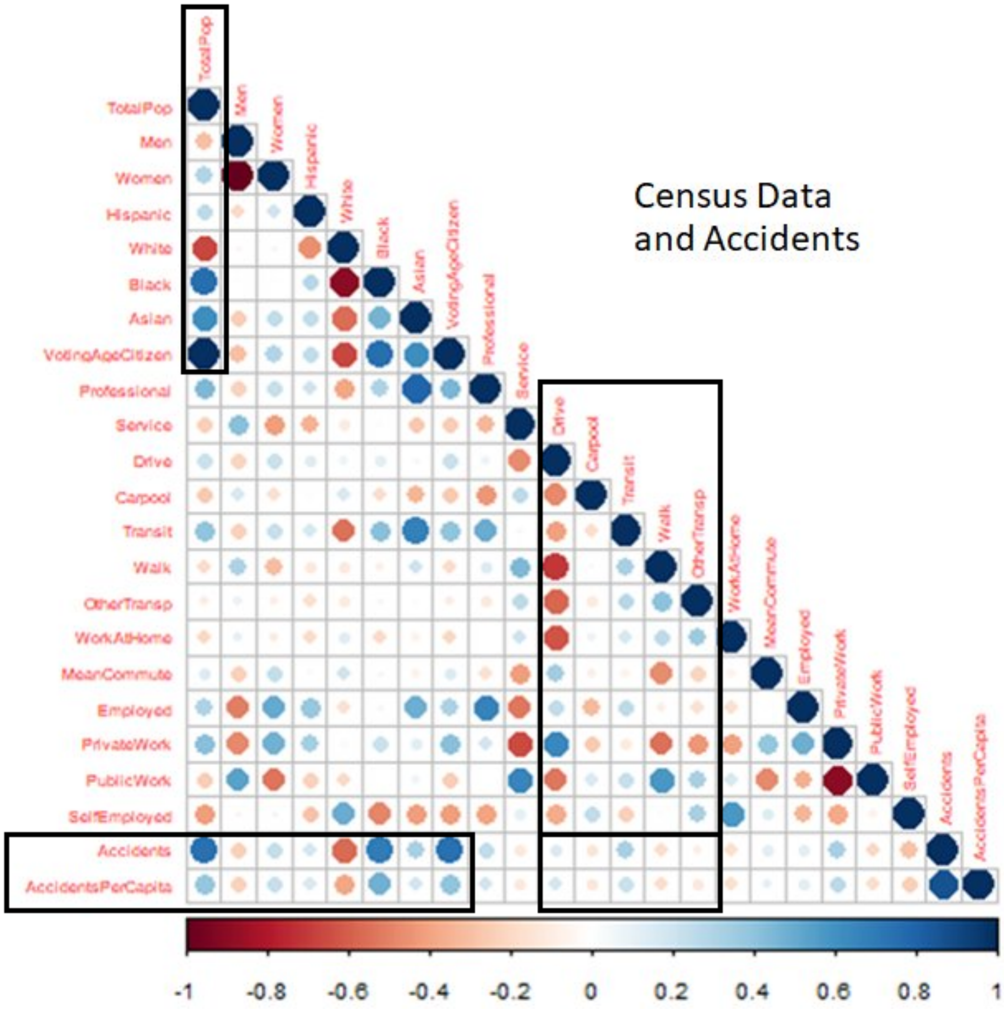
Road Safety Facts. (n.d.). Retrieved from https://www.asirt.org/safe-travel/road-safety-facts/

Therneau, T., & Atkinson, E. (2019, April 11). An Introduction to Recursive Partitioning Using the RPART Routines. Retrieved from https://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf

Tree Based Models. (n.d.). Retrieved from https://www.statmethods.net/advstats/cart.html

APPENDICIES

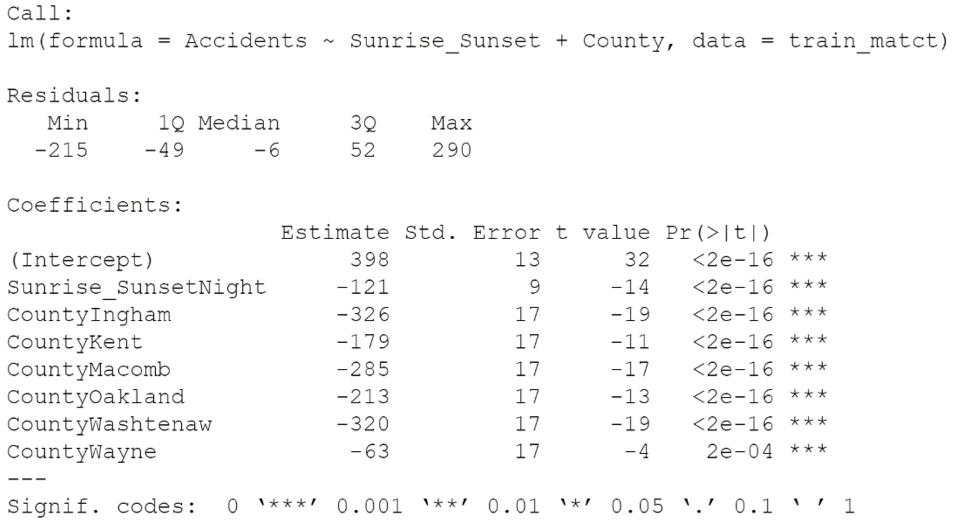
Correlation plot Appendix



Correlation plot were used to visualize the relationship that the variables *gender*, *race*, *voting age citizen*, *profession*, and *commute types* have with *accidents per capita*. In figure # of the appendix, the correlation plot displays how statistically related each variable is to other variables. *Accidents a*re positively correlated with *population,* but there were no other strong correlations with *accidents*. Variables that are strongly correlated with *population* (e.g. *gender, voting age citizen*, *race*) were also positively correlated with *accidents*, though the relationship was weaker. Surprisingly, there was not a strong association between *accidents* and *commuting* variables (e.g *drive*, *transit*).

Linear Regression Appendix

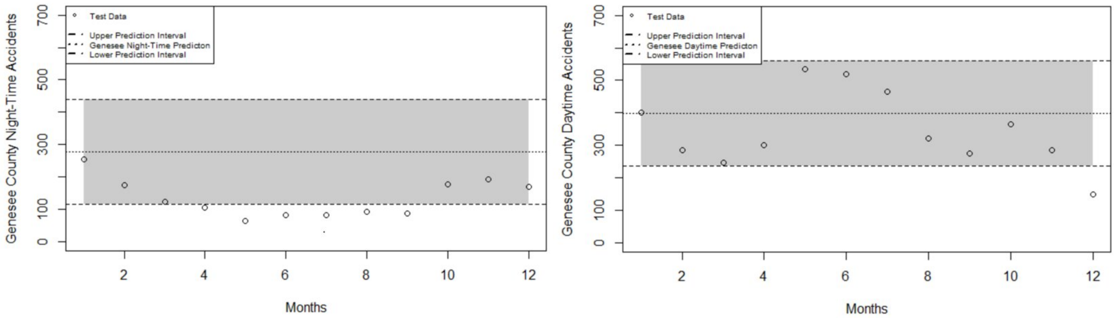
Linear regression with accidents as the response and time of day and county as the predictors



```
Call:
lm(formula = Accidents ~ Sunrise_Sunset + County, data = train_matct)

Residuals:
   Min     1Q Median     3Q    Max 
  -215    -49     -6     52    290 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)              398         13      32   <2e-16 ***
Sunrise_SunsetNight     -121          9     -14   <2e-16 ***
CountyIngham            -326         17     -19   <2e-16 ***
CountyKent              -179         17     -11   <2e-16 ***
CountyMacomb            -285         17     -17   <2e-16 ***
CountyOakland           -213         17     -13   <2e-16 ***
CountyWashtenaw         -320         17     -19   <2e-16 ***
CountyWayne              -63         17      -4    2e-04 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Genesee

County Accident Prediction by Day and Night



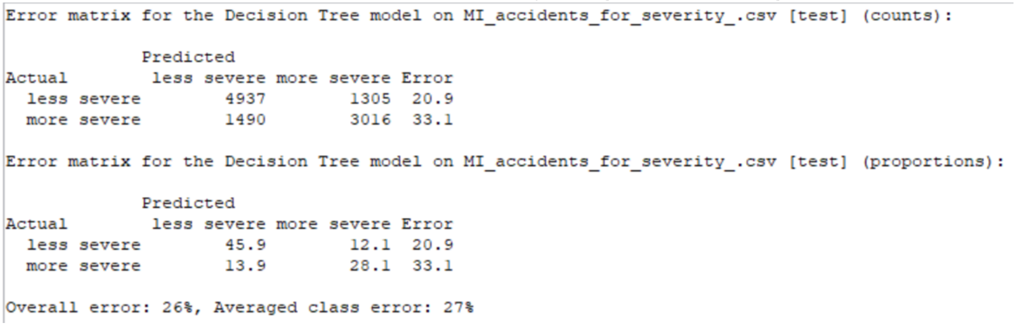
Inferring accident severity in Michigan

Ameen Alkhabbaz, Jordan Burton, David Emery, George McKenzie, Mahzad Zavareh

Decision Tree Appendix

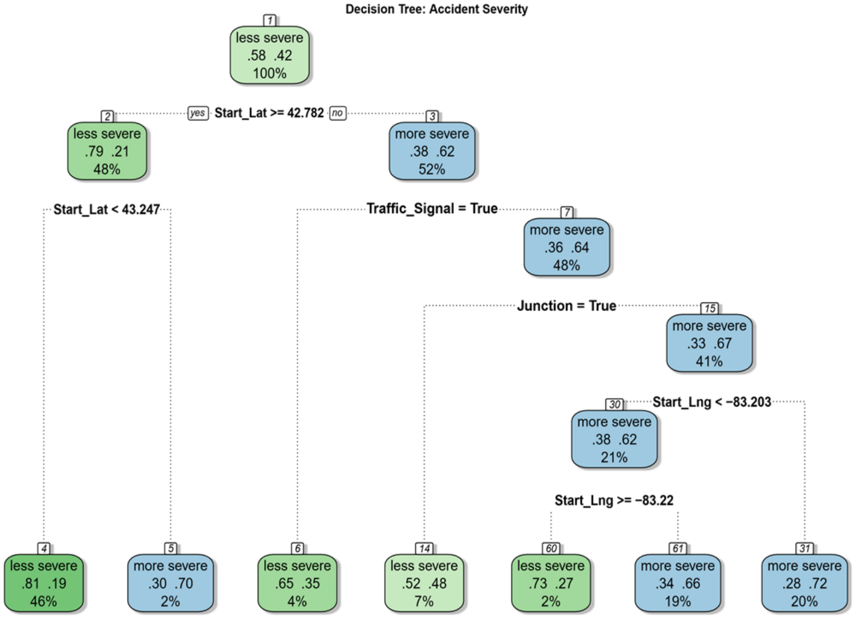
Roadway features considered as predictors but not used in the tree include: Bump, Crossing, Give Way, No Exit, Railway, Roundabout, Station, Stop, and Traffic Calming
Time features considered but not used by the tree model: Hour, month of year, and day vs night

Error matrix for decision tree model comparing 70% training data to 15% test set



```
Error matrix for the Decision Tree model on MI_accidents_for_severity_.csv [test] (counts):

             Predicted
Actual        less severe more severe Error
  less severe        4937        1305  20.9
  more severe        1490        3016  33.1

Error matrix for the Decision Tree model on MI_accidents_for_severity_.csv [test] (proportions):

             Predicted
Actual        less severe more severe Error
  less severe        45.9        12.1  20.9
  more severe        13.9        28.1  33.1

Overall error: 26%, Averaged class error: 27%
```

State of Michigan Decision Tree

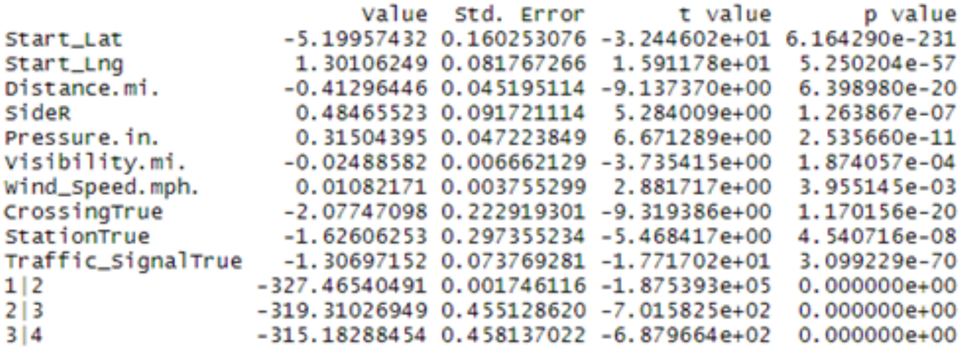


Listed Decision Tree Fields Below:

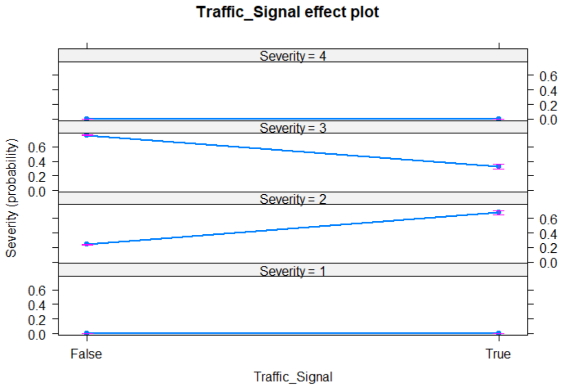
Number in graphic, Split arriving at node, less severe training count, more severe training count, classification, less severe training percentage, more severe training percentage, * indicates terminal node:

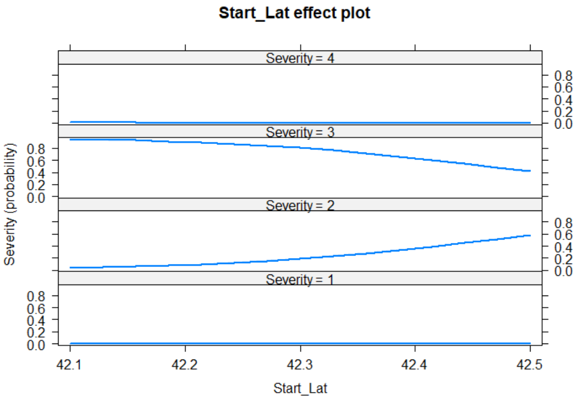
```
1) root 50153 21104 less severe (0.5792076 0.4207924)
  2) Start_Lat>=42.78239   24131  5062 less severe (0.7902283 0.2097717)
    4) Start_Lat< 43.24714   23080  4330 less severe (0.8123917 0.1876083) *
    5) Start_Lat>=43.24714   1051   319 more severe (0.3035205 0.6964795) *
  3) Start_Lat< 42.78239   26022  9980 more severe (0.3835216 0.6164784)
    6) Traffic_Signal=True   1959   677 less severe (0.6544155 0.3455845) *
    7) Traffic_Signal=False   24063  8698 more severe (0.3614678 0.6385322)
     14) Junction=True   3651  1763 less severe (0.5171186 0.4828814) *
     15) Junction=False   20412  6810 more severe (0.3336273 0.6663727)
       30) Start_Lng< -83.20295   10519  4012 more severe (0.3814051 0.6185949)
         60) Start_Lng>=-83.22021   1148   314 less severe (0.7264808 0.2735192) *
         61) Start_Lng< -83.22021   9371  3178 more severe (0.3391314 0.6608686) *
       31) Start_Lng>=-83.20295   9893  2798 more severe (0.2828262 0.7171738) *
```

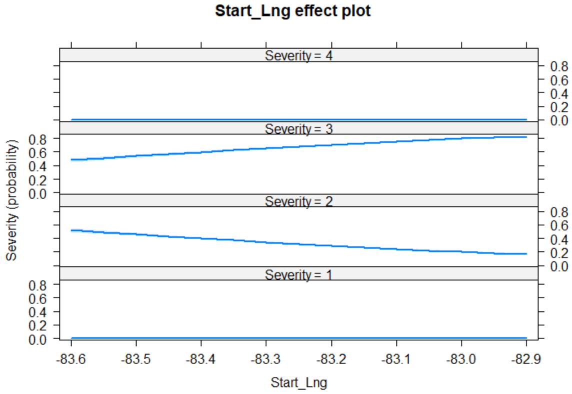
Ordinal Logistic Regression Appendix



```
                           Value  Std. Error       t value       p value
Start_Lat            -5.19957432 0.160253076 -3.244602e+01 6.164290e-231
Start_Lng             1.30106249 0.081767266  1.591178e+01  5.250204e-57
Distance.mi.         -0.41296446 0.045195114 -9.137370e+00  6.398980e-20
SideR                 0.48465523 0.091721114  5.284009e+00  1.263867e-07
Pressure.in.          0.31504395 0.047223849  6.671289e+00  2.535660e-11
Visibility.mi.       -0.02488582 0.006662129 -3.735415e+00  1.874057e-04
Wind_Speed.mph.       0.01082171 0.003755299  2.881717e+00  3.955145e-03
CrossingTrue         -2.07747098 0.222919301 -9.319386e+00  1.170156e-20
StationTrue          -1.62606253 0.297355234 -5.468417e+00  4.540716e-08
Traffic_SignalTrue   -1.30697152 0.073769281 -1.771702e+01  3.099229e-70
1|2                -327.46540491 0.001746116 -1.875393e+05  0.000000e+00
2|3                -319.31026949 0.455128620 -7.015825e+02  0.000000e+00
3|4                -315.18288454 0.458137022 -6.879664e+02  0.000000e+00
```







After splitting the data into training and testing, 70% and 30% respectively, the data showed that 97% and 95% of the train data and test data, respectively, consisted of Severity levels 2 and 3. After running several iterations of the regression model with different predictors, the predictors Starting Latitude, Starting Longitude, Distance mi, Side, Pressure in., Visibility mi, Wind Speed mph, Crossing, Station, and Traffic Signal have the most significant effect on the Severity of accidents.