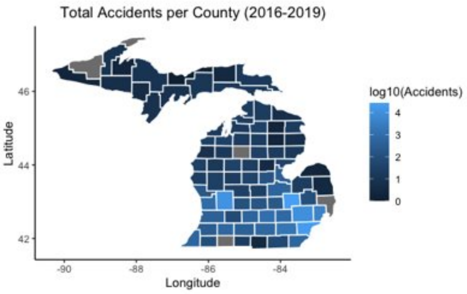
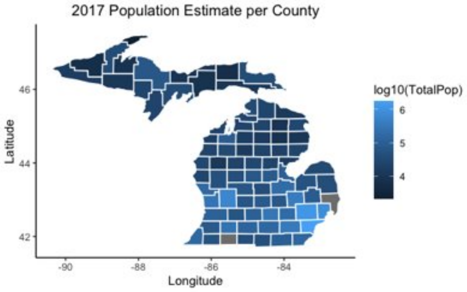
Michigan Accident Data Analysis Midterm Report

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

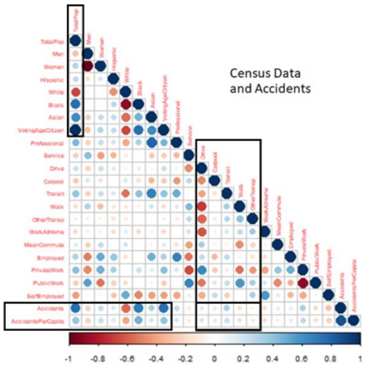




Introduction: This project uses two datasets (1) the accident dataset[a] containing information about accident severity, location, weather conditions, time, and features of the roadway near the accident. The data was captured by the (U.S. and Michigan departments of Transporation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks). (2) Michigan 2017 census prediction data[b] is comprised of Michigan county specific information including population, sex, ethnicity, income, employment, and types of transportation used. These datasets combined will be used to address the following specific aims:

County (City)	Total Accidents	Population	Accidents Per Capita
Genesse (Flint)	27,072	410,881	6.59%
Wayne (Detroit)	22,214	1,763,822	1.26%
Kent (Grand Rapids)	12,989	636,376	2.04%

- Determine if there is correction between accident severity and other predictors (e.g. population, weather, time of year, etc.) Find patterns in accident data.
- Determine the cause of accidents in specific locations.
- Predict the time of day, location, and severity of an accident.
- Build a visualization to allow intuitive interpretations of accident, census, and traffic data.
- Identify population factors correlated with accidents or accident severity.
- Determine the effect of road construction on the number and severity of accidents.



Work Done: (1) Michigan accident and predicted census data were selected from United States datasets to build a Michigan-specific dataset for analysis. Outliers were removed to prevent skewing analysis and the formatting of the datasets were standardized to enrich analysis on data derived from the datasets. (2) Data visualization is used to determine if there is a correlation between accidents per capita and population by county (census data). Beyond population, we looked at additional census data variables for correlations with accidents per capita (gender, race, voting age citizen, profession, and commute). To determine if there are patterns in the accident data, we plotted the volume of accidents as well as accidents indexed to population on each Michigan county. We identified specific city streets that accidents occurred frequently by mapping the "Description" frequencies which allowed us to associate specific variables (Points of Interest in the roadway, Days of the Week, and Times) and to understand if specific locations are having repetitive causes. (3) Accident and accident severity were explored as responses to the predictors from the accident, census, and traffic datasets to build a prediction model. County and time of day were used to predict monthly accident volume.

Teamwork activities: (1) GitHub repositories have been utilized by our team to upload, edit, and comment out coding activities to clearly delineate what has been done, what needs to be completed, and if problems are arising. (2) Group meetings transitioned quickly from in-person to virtual meetings as a response to the coronavirus emergency. These meetings are briefly summarized in the table below:

Michigan Accident Data Analysis Midterm Report

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

Team Meetings								
Date	**Time**	**Location**	**Format**	**Attendence**	**Questions**	**Conclusions**	**Agreements**	**Resolved Issues**
2/10/20	-	-	Virtual	Mazhad, Ameen, Jordan, David	Which dataset would we like to use for our group project?	Datasets need to be explored before an agreement can be made.	To individually look at the suggested datasets.	-
2/19/20	8 - 10 pm	UGL	In-Person	Mazhad, Ameen, Jordan, David	Which dataset did we decide to use?	Decided to use the US-Accident & Census datasets.	To use the accident and census datasets.	Which dataset to use.
					Using Github for code management.	Github will enable version control.	To use Github for version control	-
2/24/20	8 - 10 pm	UGL	In-Person	Mazhad, Ameen, Jordan, David	-	-	-	Github problems.
					What should we include on our project proposal?	Came up with questions to try to answer with the datasets.	To modify the project proposal as per our meeting.	-
2/25/20	-	-	Virtual	Mazhad, Ameen, Jordan, David	Should we invite George into our group?	Invited George to our group.	Modify project proposal to include George.	-
3/11/20	5:30 - 6:45 pm	-	Virtual	Mazhad, Ameen, Jordan, David, George	What progress has been made on the flight prediction mini-project?	Explored logistic regression and LDA, logistic regression gave better prediction results.	Finalize code and split the file into two.	-
					Can we add an ROC and AUC to the function?	Explored logistic regression cut-off.	Housekeep functions for submission	
					Can we use KNN for prediction?	Removed duplicates in the dataset.	Exploration of questions.	
					Can we use the entire dataset as the training set?	Cost function used as the dataset contains a small category to predict.	-	
					Model accuracy or secondary criteria?	Cross-validation to perform the validation of the prediction.		
3/18/20	5:30 - 6:45 pm	-	Virtual	Mazhad, Ameen, Jordan, David, George	What progress has been made for the midterm project submission?	-	-	Bullet points for midterm submission.
					What progress has been made for the flight prediction mini-project?			-
					How will changing a specific road affect accidents?	Use MDOT to find construction data.		
					Can we slice the data down to the top accident containing counties?	Yes we can.	Scope change to predict using counties with high accident rates.	
3/26/20	7-8 pm	-	Virtual	Mazhad, Ameen, Jordan, David, George	What do we include in our midterm report?	-	-	Midterm report
								Midterm presentation

Methods Used: (1) dplyr, tidyr, and tidyverse packages are used to pare down, manipulate, and combine datasets. (2) ggplot2, maps, mapproj, and ggmaps packages are used to create county map, city, and street plots of accident and population data. (3) lubridate, SnowballC, wordcloud, and wordcloud2 packages are used to identify specific city streets and street features to be investigated as predictors of accidents and accident severity. (4) The corrplot package was utilized to evaluate the correlation between accidents and variables included in the census data. (5) Linear regression analysis, logistic regression analysis, and k-nearest neighbors were evaluated to predict accidents by night and day in the counties the highest accident density.

Key Findings: **(1)** Interestingly, there are no strong correlations with commute (drive, carpool, transit, walk, other transportation). As expected, accidents were correlated with population and those variables that are correlated with population density (gender, race, voting age citizen). (2) Genesse county had the highest total accidents and highest accidents per capita, followed by Wayne and Kent counties. (3) There was a lower test MSE when no distinction between months were made compared with including each month as a predictor for accidents occurring during the day or night. There is a significant correlation between the time of day and frequency of accident occurrences, most accidents occur during the day during the times people commute to and from work. (4) Counties in the Upper Peninsula have the lowest frequency of recorded accidents. (5) Accidents with low severity are underrepresented in the dataset. (6) Using KNN to predict accidents had an accuracy of 60%.

Next Steps: (1) Use accident severity as the response variable with appropriate classification models. (2) Find dataset with Michigan road laws to compare this information with accident locations to identify if there are correlations between accidents with laws and traffic signals. (3) Use decision tree modeling to increase the accuracy of accident severity prediction.

Scope Changes: (1) Focus on counties with high accident rates to determine if it is possible to predict the monthly volume of accidents based on predictors like day/night, traffic, etc.

References: a) M. Sobhan, M. H. Samavatian, S. Parthasarathy, and R. Ramnath. 2019. b) M. Sobhan, M. H. Samavatian, S. Parthasarathy, R. Teodorescu, and Rajiv Ramnath. ACM, 2019. c) U.S. Census Bureau