Iowa Crash Data

DS 202 Final Project Report

Aaron Jencks, Hunter Rose, Eli Musgrove & Tanner Boyle

Background

Our dataset comes from the Iowa Department of Transportation in the form of 2 .csv files. Our first dataset describes information involved in the crash such as weather, date, street names, location, cause, etc… The second dataset contains information about each of the vehicles involved and other factors in the accident such as visual range, drug test results, road conditions, etc… These datasets contain over 600,000 and over 1,000,000 rows respectively. The datasets cover years from 2009 to 2020 and were recently updated in March 2020.

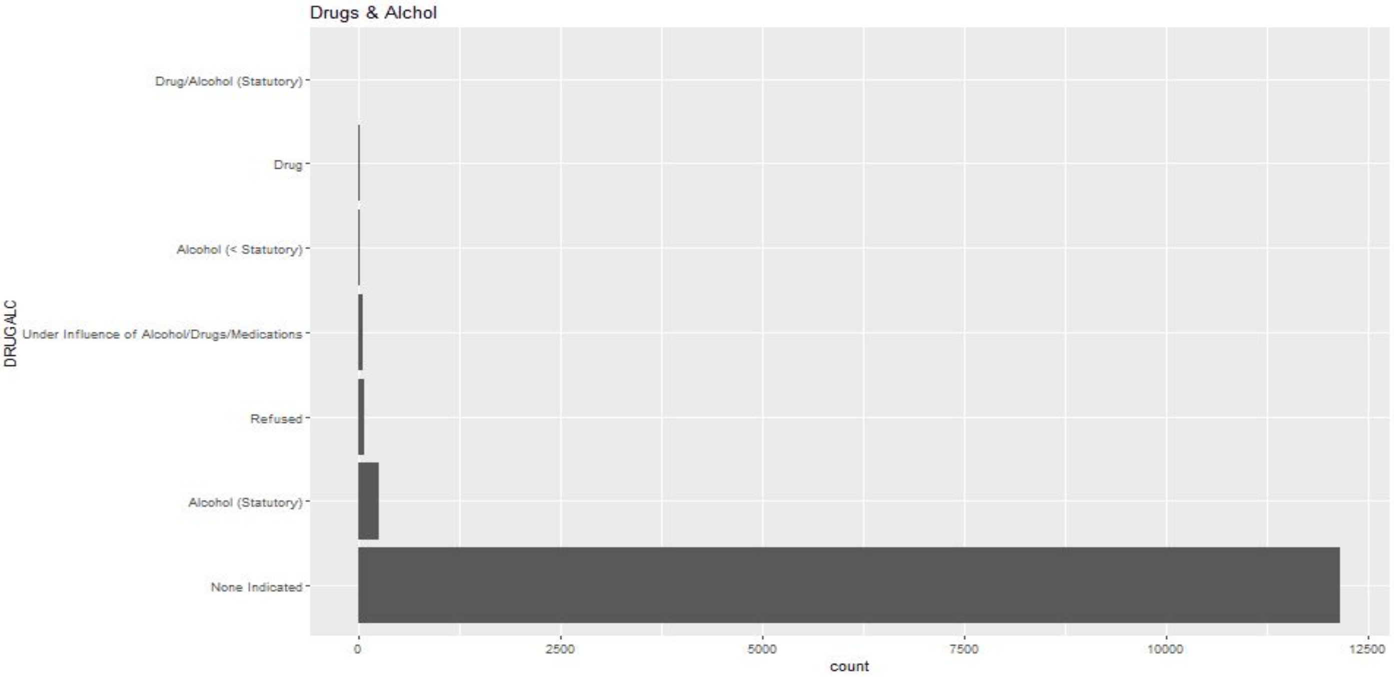
Obtaining data & cleaning

Obtaining the data was pretty simple, the Iowa DOT has all of their information online for free, though, there is little actual documentation of any sort. As far as cleaning goes, the dataset was pretty massive in the beginning, in order to thin it down, we decided to only look at data from 4 main cities: Cedar Rapids, Iowa City, Ames, and Des Moines, later we expanded that to 4 counties, Linn, Jones, Story, and Polk respectively. Once that was done, there were still outliers that had to be thinned out, according to a document I found, about 1 in 60 data entries in the DOT is erroneous, so we had to combine both the county number and latitude/longitude bounding boxes for each county. For further cleaning, we also had to change several of the datatypes, some of the ints came in as chars, and ALL of the string data was read in as factors.

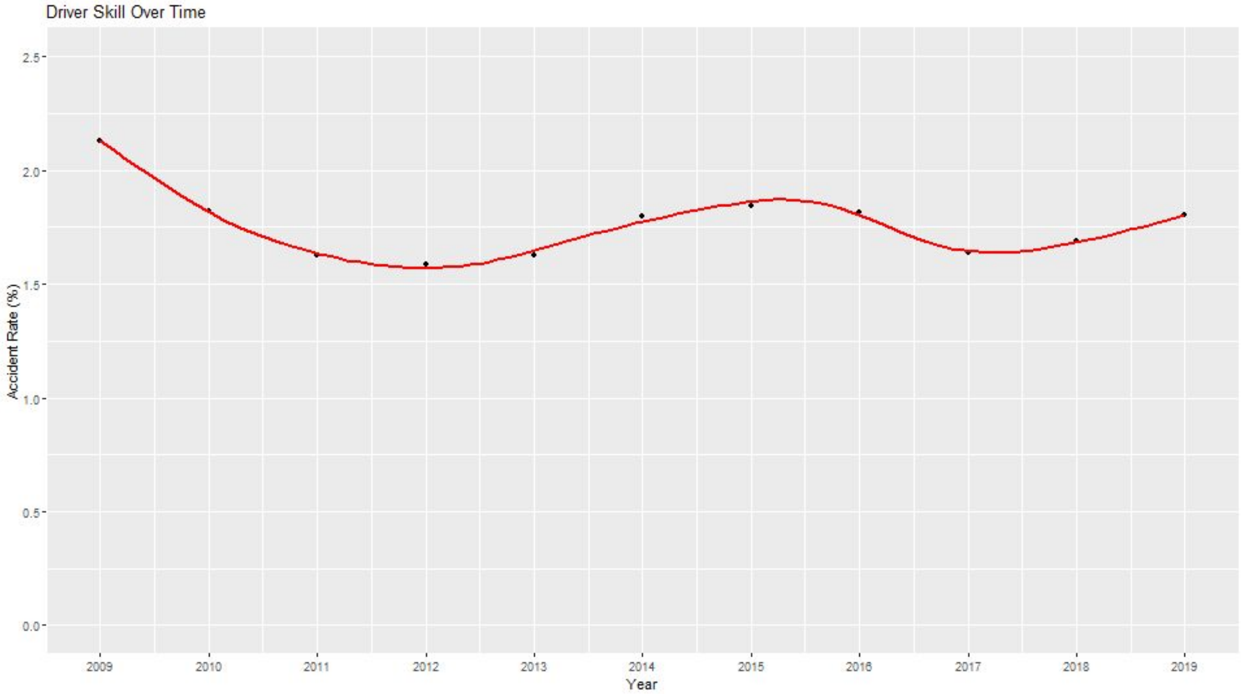
Exploratory Analysis

For our exploratory analysis, we all agreed on finding the most dangerous intersections within our counties to be used for our infographic. After that we chose different things to look at for each city so that we had a variety of data. Some things were commonly chosen, these were substance-related accidents and age of drivers. The reason we wanted some uniqueness for each city, is because we had such large datasets with a lot of interesting information and didn't want it to be overlooked. This analysis will include a few of the graphs that we have created, the rest can be found on our GitHub repository along with the respective code. The graphs being referenced in parentheses are in chronological order of appearance.

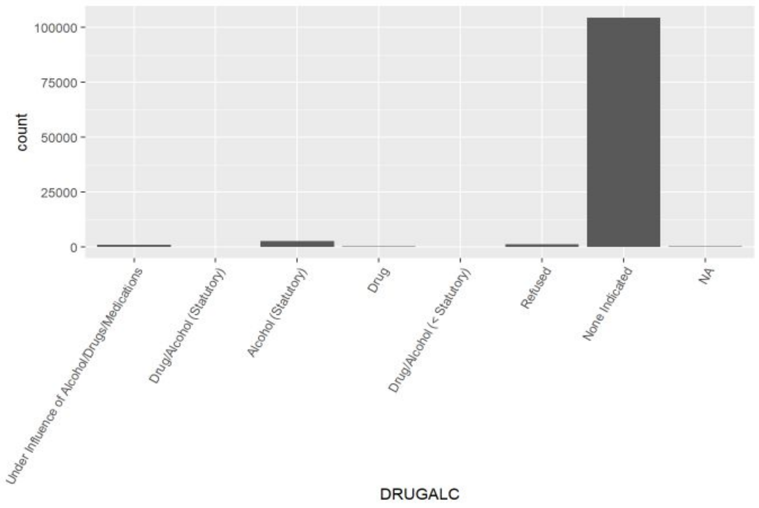
Starting off with Ames, after sub-setting the data based on longitude and latitude, there were 12,567 accidents. When looking at the substance-related accidents, over 10,000 of them didn't have any substances involved (graph 1). One of the unique things that was done with the Ames data, was calculating driver skill over time (graph 2). This was done by taking the number of accidents in a year divided by the Ames population creating the percentage of Ames getting in an accident. This was plotted with accident rate in the y and year in the x, and from 2009-2019 the accident rate decreased by roughly .4%. This was somewhat surprising as Ames population has increased over those 10 years. Eli also looked into the driver's age for each accident, finding that the age group with the most crashes is 17-23. He also looked at the



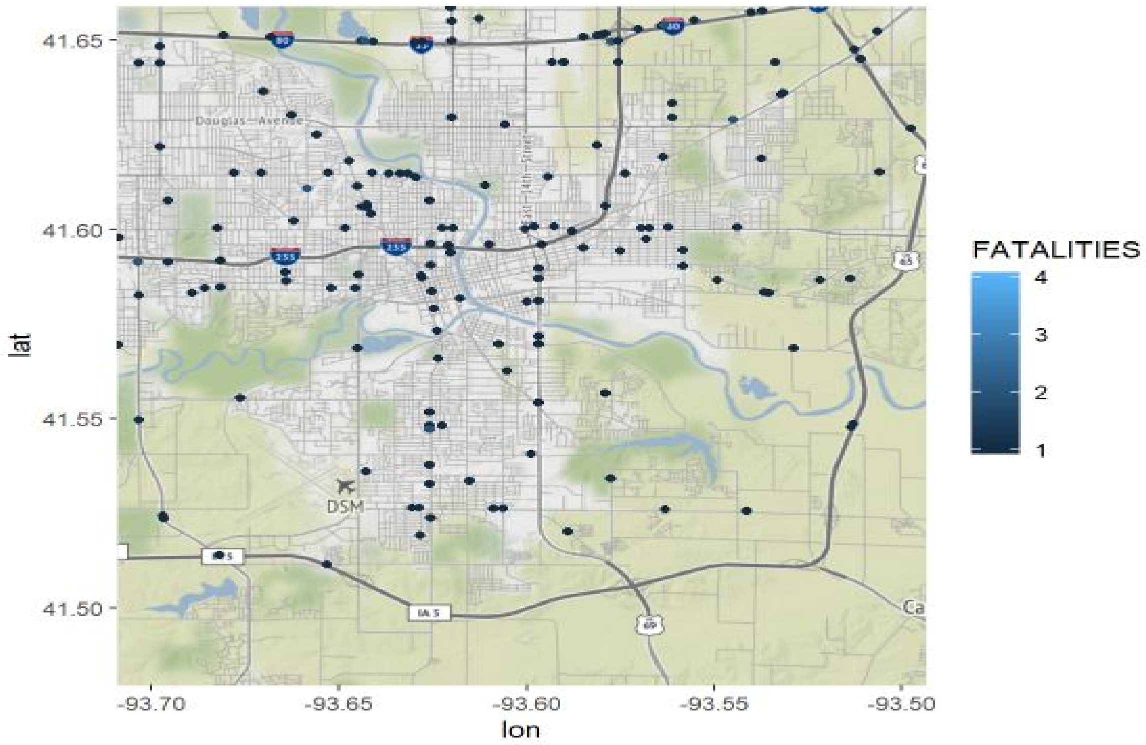
number of accidents at each speed limit and showed the spread of surface conditions for each crash at each speed limit. The age and speed limit/surface conditions graphs can be found in our GitHub repository.



Moving on to Des Moines, after subsetting the data to Polk county, there was over 140,000 crashes. Looking at the substance related accidents, Des Moines had over 100,000 crashes that didn’t involve any substances (graph 3). There were around 4500 crashes that did involve alcohol/drugs or the driver refused to be tested.

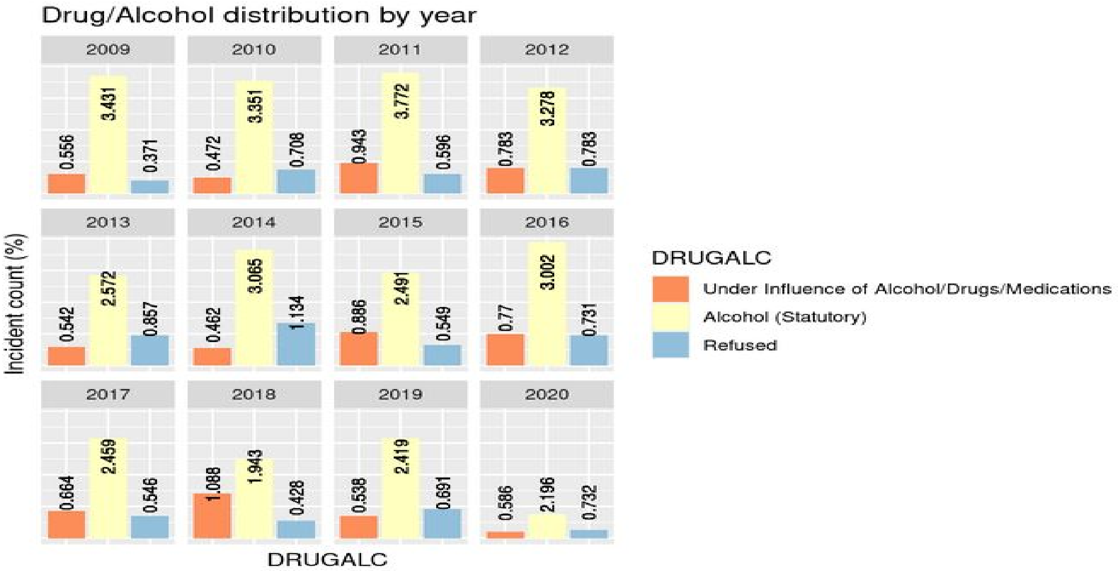


One of the unique things that was done with the Des Moines data, was trying to plotting the fatalities that happened in the crashes to see if there were any areas that had more crashes that lead to fatalities than others (graph 4).

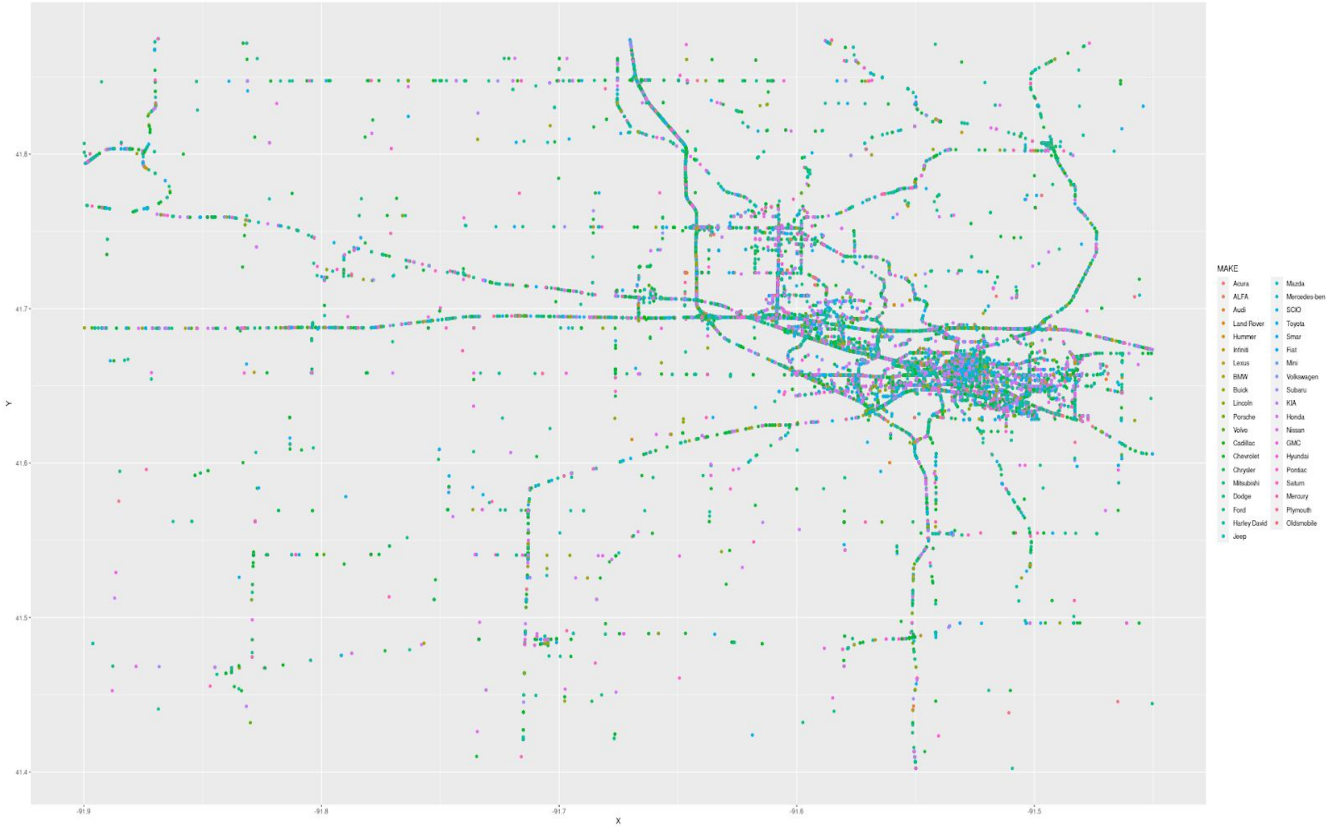


We can see that the interstate on/off ramp by I-80 and I-235 did have a small cluster of fatalities, along with the slight curve on Hickman by the Des Moines River. Some of the other things that were looked at were age/gender in each crash, which had similar results as Ames. As far as gender, there were slightly more males than females involved in the crashes. Weather and surface conditions were looked at as well, with the majority of crashes happening on clear or cloudy days and dry roads.

Then we have Iowa City, or more specifically Jones County. This subset had over 25,000 accidents. When it comes to the drug/alcohol related crashes in this county, around 1,000 had substances involved (graph 5).



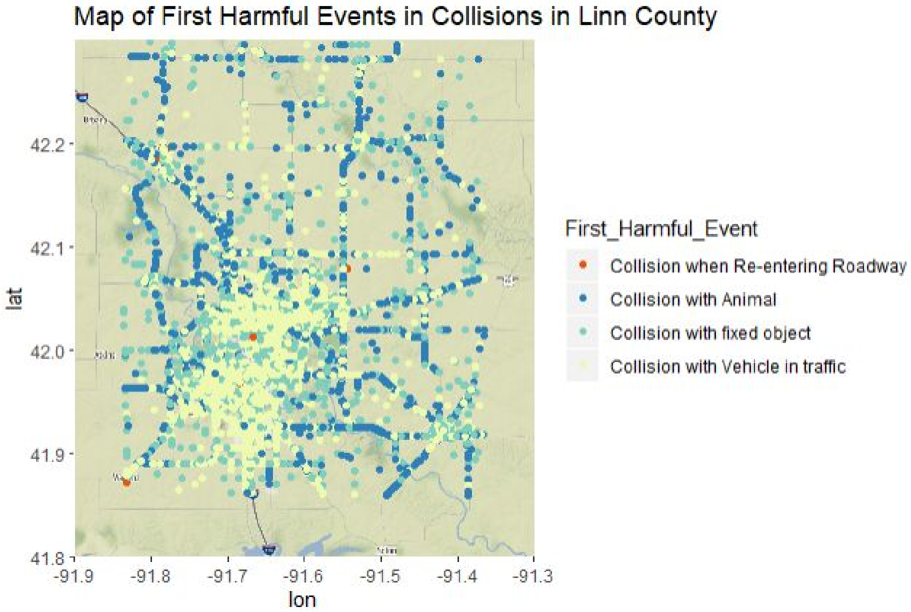
One of the unique things that was done with the Iowa City data was taking a look at the vehicle distribution to try and find some economic data mapped in Iowa City. After cleaning the vehicle make column thoroughly, Aaron sorted the list into more expensive on average to less expensive then plotted them using longitude and latitude (graph 6).



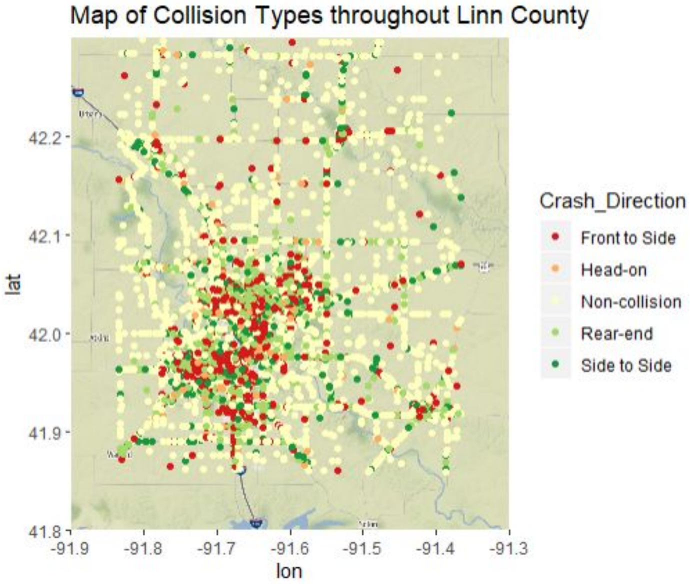
This was to try and see if we could identify where lower-income/higher-income people are driving and getting into crashes. Aaron also looked at the drug/alcohol related accidents in the most dangerous intersections, along with plotting where all the drug/alcohol related accidents happened using longitude and latitude. These other graphs can be found in the github repository.

Lastly, we have Cedar Rapids. Cedar Rapids is the second largest dataset we had, with over 37,000 records. One of the first thing that Tanner looked into was the most prevalent types of collisions that occured in Cedar

Rapids (graph 7).

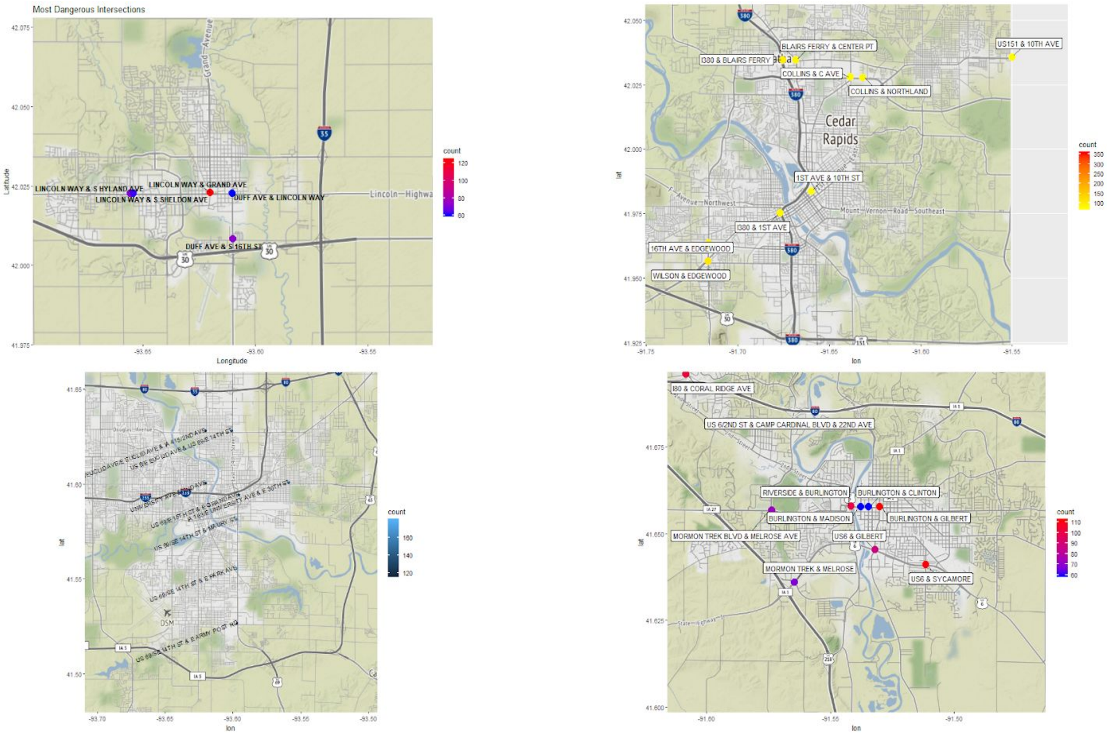

Map of First Harmful Events in Collisions in Linn County
42.2
42.1
42.0
41.9
41.8
lat
-91.9
-91.8
-91.7
-91.6
-91.5
-91.4
-91.3
lon
First_Harmful_Event
Collision when Re-entering Roadway
Collision with Animal
Collision with fixed object
Collision with Vehicle in traffic

We can see that the majority of crashes occurred in traffic with 2+ vehicles and that county roads/rural highways had the highest risk of animal collisions. Tanner then tried to see if there were any time effects when it came to prevalence of collisions. Looking first at days of the week, collisions were most common on Fridays and least common on Sundays with most other days holding a steady average. However, the more interesting effect came from plotting collisions by month of the year -- winter months typically had the highest number of collisions for every year that we looked at. This led Tanner to look at road conditions of collisions in the winter months; unfortunately, this turned up almost nothing as dry roads were more common than non-dry roads in accidents during winter months. Lastly, Tanner also looked at the type of collisions, and found that you have a higher risk of front-to-side or side-to-side crashes in downtown Cedar Rapids and a higher risk of non-collision crashes (veering into ditches, etc) on rural highways and county roads (graph 8).



All other figures can be found in our github repository.

Featured Infographic



- For our featured infographic, we wanted to provide something actionable for anyone interested in our project. While it's good to know which types of accidents are prevalent and which car models are being crashed where, we didn't think that it was going to be anything that help people out. Being that the majority of Iowa State students are likely from Iowa, we decided to look at the most dangerous intersections from some of the biggest cities in Iowa. To do that, we utilized a few different packages that allowed us to pull Google Maps data and better visualize our city maps. Then we were able to tile the four plots into one single infographic using gridExtra's function "grid.arrange()". We felt that this was a strong takeaway from our presentation as, even if you aren't from Iowa, any Iowa State student can benefit from knowing the most dangerous parts of Ames.

Steps for data wrangling and Visualization

In order to wrangle the data, we had to do quite a bit of work. To further wrangle the data, we often had to summarize, and join several pieces together. Most often, I would end up grouping the dataset, then summarizing into counts, then sewing that back into the original dataset by matching the DOT case numbers, or by euclidian distance of the lat/long numbers.

There was also wrangling when it came to handling mispellings, or multiple spellings, etc. One such instance being the LITERAL column, which is supposed to handle the exact address/intersection that the incident occured at. There are so many ways to spell out an address, that what we had to end up doing, is we combined the lat/long into a euclidian distance, then grouped the dataset by that distance, then summarized with count = n(), and then joined that dataset back into the original, adding a new column in the process. Once that was done, we ordered the columns by the count column, found all of the unique values, then took the top 10 street names, as the most dangerous streets in the city/county. Tanner did something similar to this when searching for the most harmful incidents as well.

Aaron also did a graphic modelling car makes and their distribution throughout Jones County. To do this, he needed to find unique identifiers for the car names originally there were over 300 identifiers for car makes, so he corrected the spelling of more than 75, then filtered out the foreign, or unknown makes.

Hunter and Tanner did a graphic modelling the position points on top of a map, to do this, he learned how to use a 3rd party package, and then was able to map latitude and longitude points to an actual map.

Answers to questions raised

We each raised individual questions but also shared a few points to analyze in every city. We all analyzed the effect of drug and alcohol on accidents in every city, and found that our four cities had similar rates of accidents caused by alcohol at around 2-3%. Ames and Iowa City are both big college towns where there is a lot of partying, but this didn’t seem to increase the rate compared to Des Moines and Cedar Rapids. Although this is surprising, this is a positive sign and may indicate that students don't drink and drive at very high rates to boost these numbers.

We also analyzed the most dangerous intersections in each of our respective cities and found potential roads to avoid. For Ames specifically, Lincoln Way contained several of the intersections in the top 10 most dangerous, which matched the age analysis that revealed that the most at-risk group of drivers are between 19-23 years old. This shows that the accidents on Lincoln Way are caused by college students rather than a pinch point where a lot of traffic has to funnel through.

Additionally, several group members analyzed the effect of weather in causing accidents and found that the vast majority of accidents occur in dry road conditions (even in the winter). This was surprising to all of us, but this may be due in part to our data being in the city where drivers are more careful and the danger in spinning out is lower.

Conclusion

- Unsurprisingly, we found that, in general, cities within Iowa are pretty similar. One of the biggest takeaways was that, while winter months typically led to increased collision numbers, that we couldn't correlate it to dangerous road conditions. Intuitively, it makes sense that snow/rain/sleet lead to crashes (just listen to traffic reports during the first snowstorm). While our analysis didn't find the cause of heighted collision numbers in the winter months, we are hopeful that if we had more time, that we would be able to find something.
- Another important conclusion was that Iowa City and Ames, both notorious college towns, tended to be similar to Cedar Rapids and Des Moines in terms of drunk driving related accidents. While we ideally would have found no accidents related to alcohol, it is good to know that students might not be living up to the reputation that they have been given (earned or not). One last thing to note is that alcohol related accidents were distributed fairly evenly throughout the counties that we looked at, when one might've expected them to be centered in the counties' respective downtowns.

- Finally, in our analysis of most dangerous intersections in each of our cities, we saw that there was typically at least one road that was a repeat offender. As Iowa State Students, we are all very familiar with Lincoln Way and Duff Avenue. In fact, many of us have probably found ourselves in more than one traffic jam in one or both of those roads. This analysis confirms that it is likely better to avoid them than to grit your teeth and bear them -- finding data to back up these thoughts was very satisfying for us.

Personal contribution section

- About 1 paragraph each

Aaron:

I did the county of Jones, and I also supplied the source code for finding the most dangerous streets in the city, I helped with data cleaning, and wrangling, along with data visualization, giving code snippets to help the others with their work. I wrote the Data cleaning and Data wrangling parts of the report, and did the Iowa City section of the presentation.

Hunter:

I decided to choose Polk county. Everything related to Des Moines in the presentation was done by me. I cleaned the Polk county data to a point that it would let me upload to github. As far as data visualization, I helped with using the Open Street Map package to plot our data on detailed maps of our respective cities. I also wrote the Exploratory analysis part of the report.

Eli:

My analysis is focused on Ames. I did individual analysis on driver skill, road and weather conditions resulting in accidents in various weather conditions, analyzed alcohol as a cause of accidents, etc… I also spent time editing and producing our video and the background/questions raised sections of the report.

Tanner:

I helped find us our datasets and ended up choosing to cover Cedar Rapids because that is where I was from. As we all did, I took care of all of the data exploration and wrangling for my city. In addition to the alcohol/weather questions that we all answered, I looked at a few different questions that ended up requiring a decent amount of extra cleaning/wrangling. Finally, I put together the featured infographic and then wrote about it and the conclusion for our report.