Crash Risk Prediction Model using Data Science

This manuscript ([permalink]) was automatically generated from [uiceds/cee-492-term-project-fall-2022-swifties@fb57a64] on October 31, 2022.

Authors

- **Lara Diab**
  0000-0001-8489-2015 · diablara
  Illinois Center for Transportation (ICT); Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

- **Renan Santos Maia**
  0000-0002-0877-4006 · renanssmaia
  Illinois Center for Transportation (ICT); Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

- **Farid Saud**
  XXXX-XXXX-XXXX-XXXX · fsaudm
  Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

- **Johann J Cardenas Huaman**
  0000-0002-4695-7639 · Johann-Cardenas · transporter_pe
  Illinois Center for Transportation (ICT); Department of Civil and Environmental Engineering, University of Illinois Urbana-Champaign · Funded by The Grainger College of Engineering

Project selection and Introduction

Road accidents are responsible for a significant number of injuries reported every year. According to the World Health Organization (WHO), approximately 1.3 million people die each year as a result of road traffic crashes (as of June, 2022). In addition, road traffic crashes cost countries 3% of their gross domestic product (Safarpour et al, 2020; WHO, 2022). Consequently, understanding what influences these accidents on roads is of utmost importance. However, it is not easy to decide which exact conditions lead to these accidents. Different road, climate, vehicle and driver conditions affect the likelihood of a driver to be in a fatal/serious accident.

The ability of predicting in an accurate way the potential occurence of car crashes is a valuable contribution for road safety. In an approach frequently used in the literature, crash records' data are used for the development of crash prediction models, so that agencies can allocate investments to priority areas of the roadway network. However, given that the budget for infrastructure improvements is limited, adopting countermeasures for all facilities that crashes are potentially occuring is not financially feasible. Therefore, informing drivers about the potential safety risks is a way to proactively compensate the aforementioned limitations. Moreover, with the development of connected and autonomous vehicles, this information can be provided in a more optimized way, contributing for vehicles' route decision, as well as for real-time alerts that can lead drivers to take the necessary precautions to operate more safely (Yu et al, 2021).

Project Objective and Plan Proposal

The objective of this work is to use the Illinois Department of Transportation (IDOT) extensive crash data to be analyzed and, finally, be used for a crash risk prediction model based on main categorical data that can be real-time updated (such as the weather/lighting/pavement conditions). Ideally, it could be used by navigation systems to allert drivers to adopt more cautious behavior as soon as they enter higher-risk sections.

The plan to be carried out will follow the basic steps described as follows:

1. Read the IDOT's crash data CSV files available as an open data source.
2. Clean the data by deleting unwanted columns, handling missing data, and removing irrelevant observations.
3. Tidy the data by organizing the variables into columns and the observations into rows.
4. Analyze and visualize the data by finding correlations both analytically and graphically.
5. Model a prediction algorithm for crash risk according to categorical variables.

Description of the Data Set

IDOT has generated datasets with statewide crash locations produced by the Crash Information Section of the Illinois Department of Transportation (IDOT). The accident data has been collected throughout the years using Application Programming Interfaces (APIs) that provided streaming traffic incident data. There are about 300,000 accident records per year in these datasets, and each record contains attributes that include conditions like (among others that are not listed or described here because these are not relevant for this study):

1. Time and date (day, month, year)
2. Coordinates (x,y)
3. Type of collision
4. A quantitative description of fatalities and injuries

5. Crash severity classification based on their impact on traffic
6. The road surface condition (“Dry”, “Wet”, “Snow or slush”, “Ice”, or “Sand/Dirt/Mud”)
7. Road defects (“Debris on roadway”, “Rut/Holes”, “Unkown”, or “No defects”)
8. Lightning conditions (rated in a scale from 1 to 9)
9. Geometric characteristics of the road section
10. Work Zone (“construction”, “maintenance”, “utility”, “unknown”, or “N/A”)
11. Possible causes of the accident.

The datasets for different years are available for download as .CSV files at the IDOT’s website:

https://gis-idot.opendata.arcgis.com/search?groupIds=6d2862031a6d47c7a8c211e38e423e05

Exploratory Data Analysis

Open-source crash data is published by the Illinois Department of Transportation (IDOT) yearly. Each crash report was found to have extensive entries with up to 85 attributes, which include several independent variables to describe each occurence. Each dataset is organized with observations filled out according to the IDOT Traffic Crash Report SR 1050 Instruction Manual (2019). The datasets for each year are available online in .CSV format at the IDOT website, and they contain observations arranged in rows and attributes in columns. The datasets from 2017, 2018, and 2019 were included in this Exploratory Data Analysis. The datasets from 2020 and 2021 were discarded in this analysis given the COVID-19 pandemic outbreak, which altered drastically the dynamics of traffic worldwide, and thus crash-related data (Yasin, Grivna & Abu-Zidan, 2021). Regarding the dataset size, each one had originally over 300,000 rows (944,328 in total, combined). The Exploratory Data Analysis will be carried out following the steps described in the next sections.

Reading the Data

The datasets were imported to Visual Studio Code using the CSV library. It was found that the latest report contained 5 additional attributes that could not be used since they were missing in previous reports. It was verified that any other attribute was arranged in the same way for each file, thus it was decided to discard this information. After deleting these attributes, all 03 datasets were merged into a unified file, which contains 80 variables (columns) and 944,328 observations (rows). This final database serves as the baseline to start the cleaning process.

Cleaning Process

It was found that several independent variables would not provide fruitful information due to missing, unknown or incomplete data. First, this observation was obtained by visual inspection, and later by analyzing the number of different and unique values present in each attribute. Thus, the datasets were processed to filter out irrelevant or incomplete variables. For instance, information pertaining the location (latitude & longitude, or X & Y coordinates) have not been taken into account. A map plot was initially produced to see the distribution of the data though. Columns containing codes describing the city, county or ID of the location where the crash took place have also been excluded. Columns involving duplicate information (e.g. two columns describing the same independent variable with a label and a number), and traffic structures were also removed. For few other independent variables, information that could potentially be useful was found to be significantly incomplete. For instance, this was the case of attributes such as the number of lanes and the type of intersection. As a consequence, these variables were not included on the clean dataset. Finally, additional cleaning was carried out for independent variables with a high number of description labels. For example, the "Railroad Crossing Number" variable had up to 100 different values which would have not been handy information for the end-user. After filtering out all the attributes that would not be utilized for this analysis, the number of independent variables went down from 80 to 21.

When it comes to crash reports, several inconsistencies are considerably frequent. In the literature, for example, it is mentioned that "investigation of traffic safety by means of crash records is a reactive approach, where researchers need to deal with imprecise, incomplete, inconsistent, and, sometimes, inexistent records", and thats why the acquisition of historical series to provide minimal consistency to the analysis of crashes to reduce misinterpretations and misleading conclusions is crucial (Hauer & Hakkert, 1989; Chin & Quek, 1997; Farmer, 2003). This justifies the need of dedicating a considerable amount of time, after filtering the columns (variables) of interest, to the cleaning process of the rows (observations). For each column, the observations labeled as "blank", "unknown", and "other" were matter of discussion among the group on how these inconsistencies would be handled. For all variables, the “blank” observations were immediately removed from the dataset.

Analysis and Visualization

Bar Plot Crashes

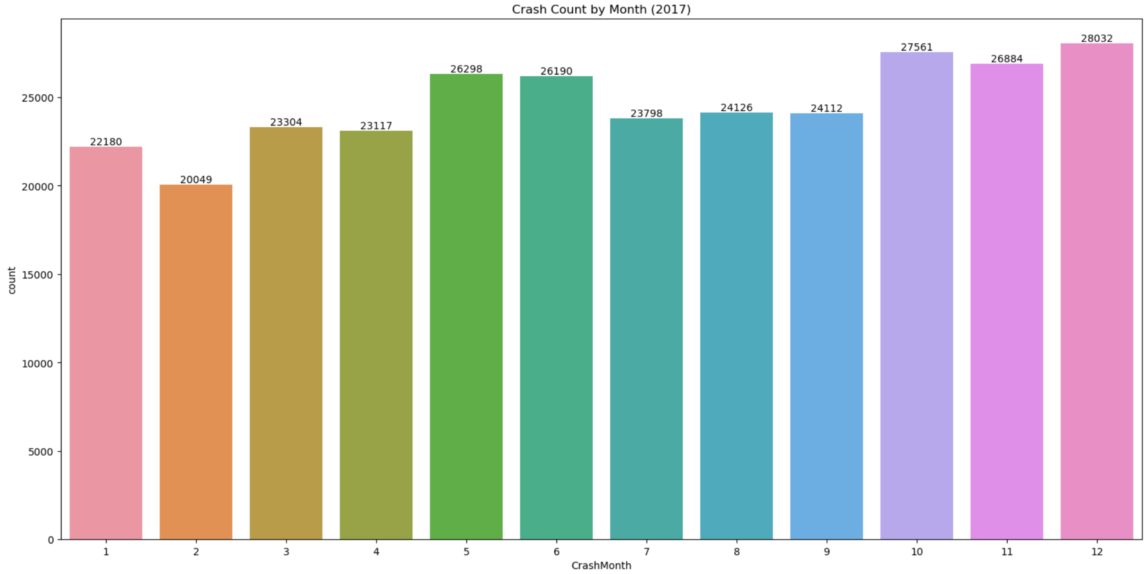

Figure 1: Bar Plot Historical Crashes. Crash reports from 2017 in the state of Illinois, USA.

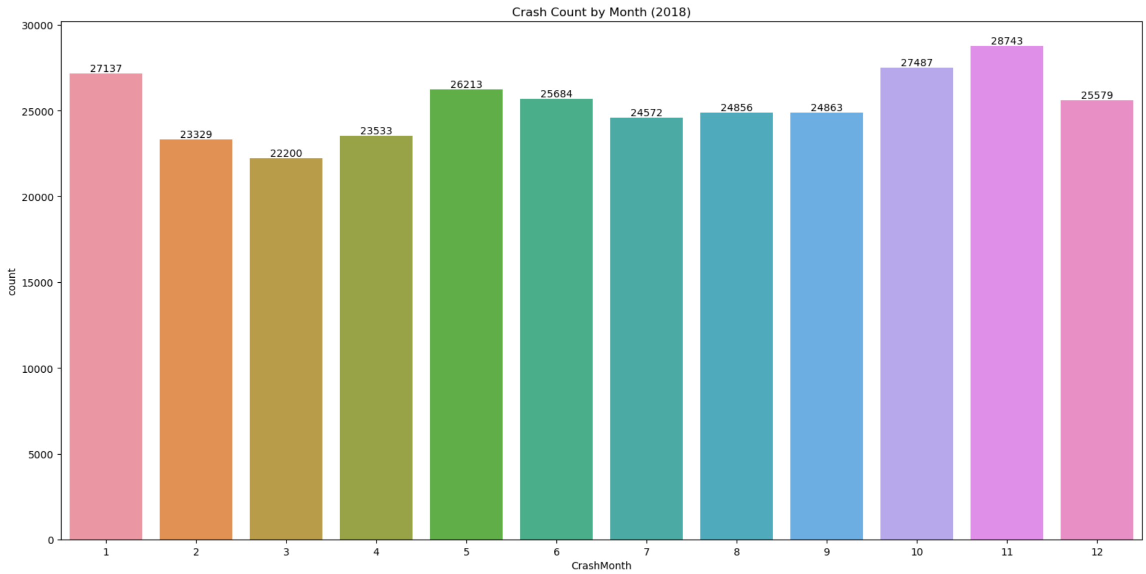

Figure 2: Bar Plot Historical Crashes. Crash reports from 2018 in the state of Illinois, USA.

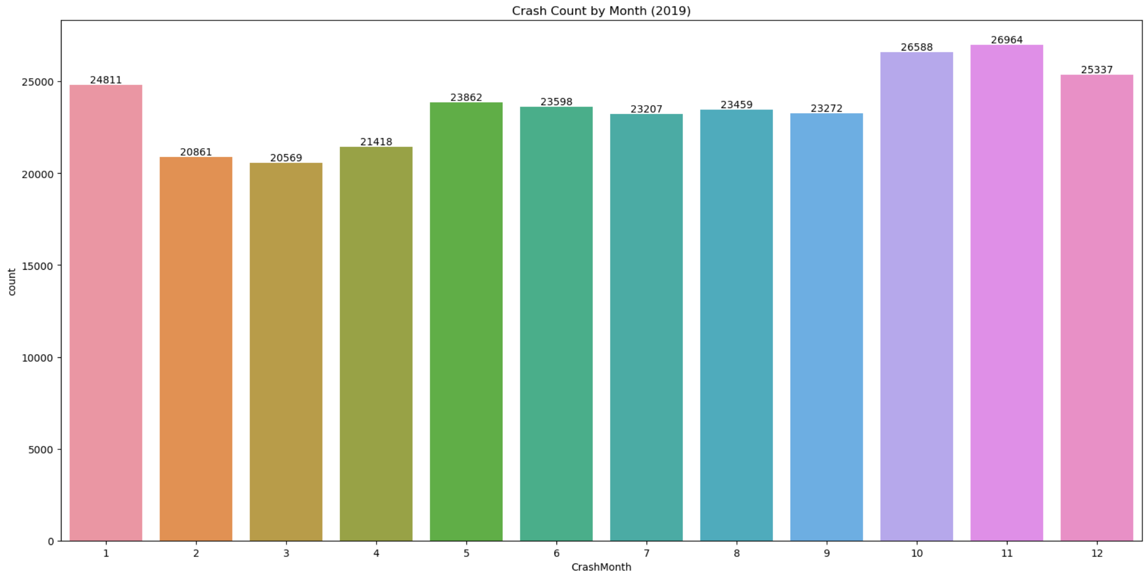

Figure 3: Bar Plot Historical Crashes. Crash reports from 2019 in the state of Illinois, USA.

Map

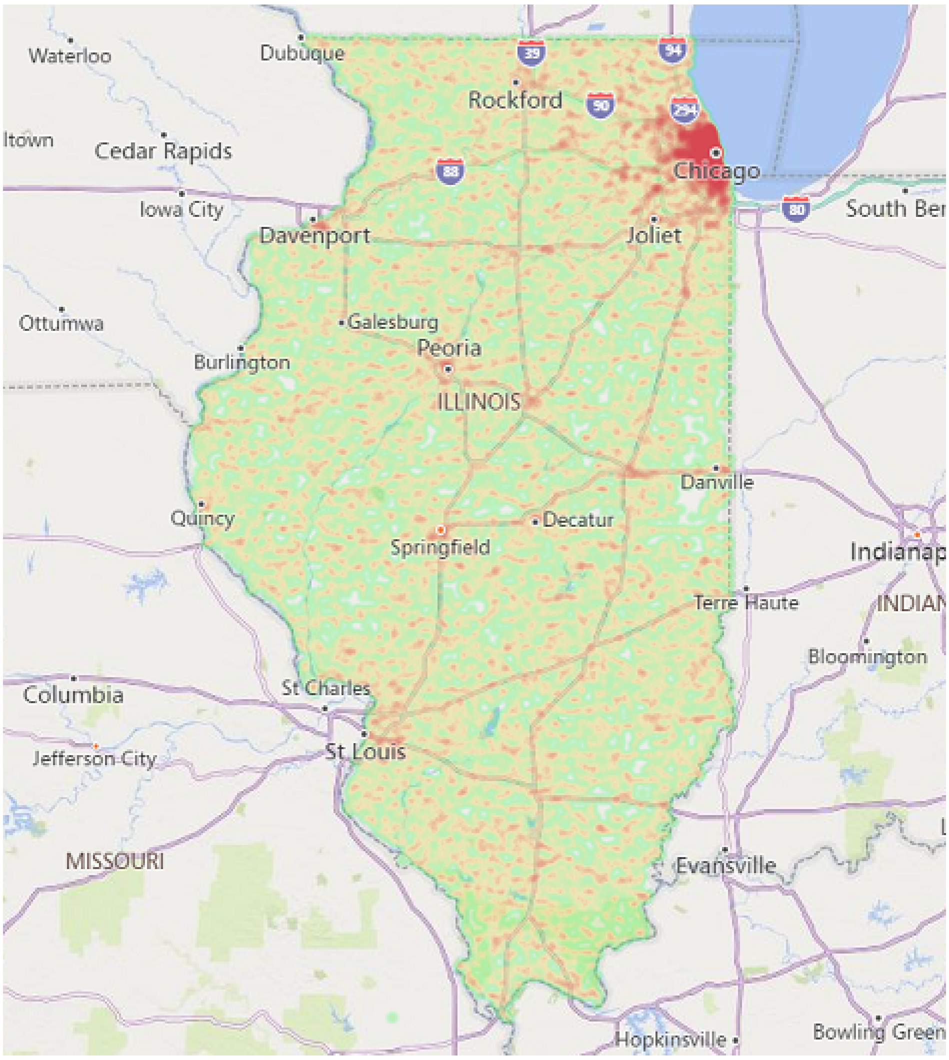

Figure 4: Distribution of crash occurences. Crash reports from 2017,2018,2019 in the state of Illinois, USA.

Pie Chart | Lightning Conditions

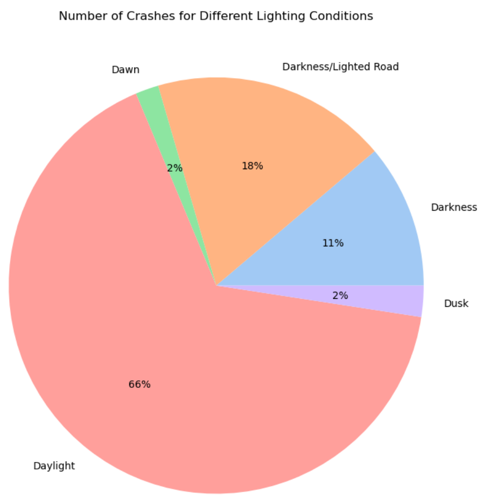

Figure 5: Distribution of accidents by Lightning Condition. From 2017 to 2019.

Bar Chart | Road Defects

Figure 6: Distribution of accidents by Road Defects.

Figure 6: Distribution of accidents by Road Defects.

Pie Chart |Road Surface Condition

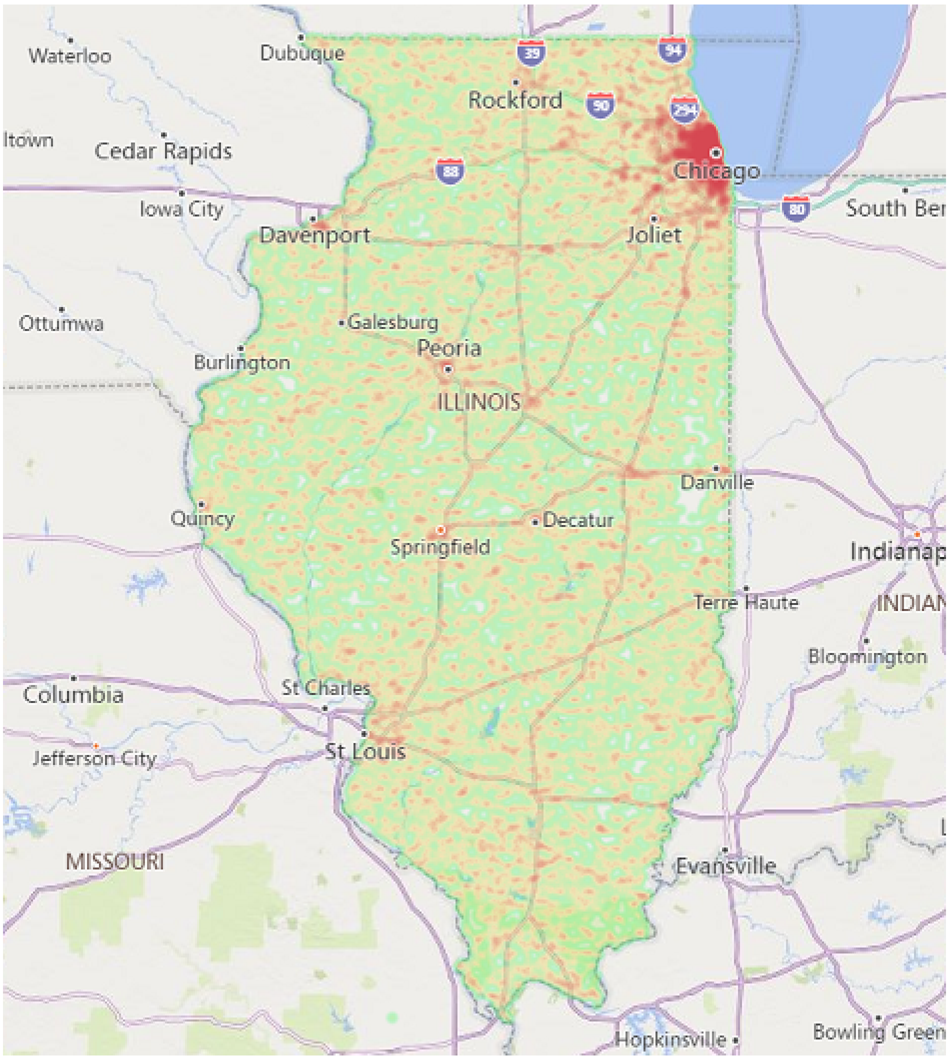

Figure 7: Distribution of accidents by Road Surface Condition. From 2017 to 2019.

Pie Chart | Road Defect Condition

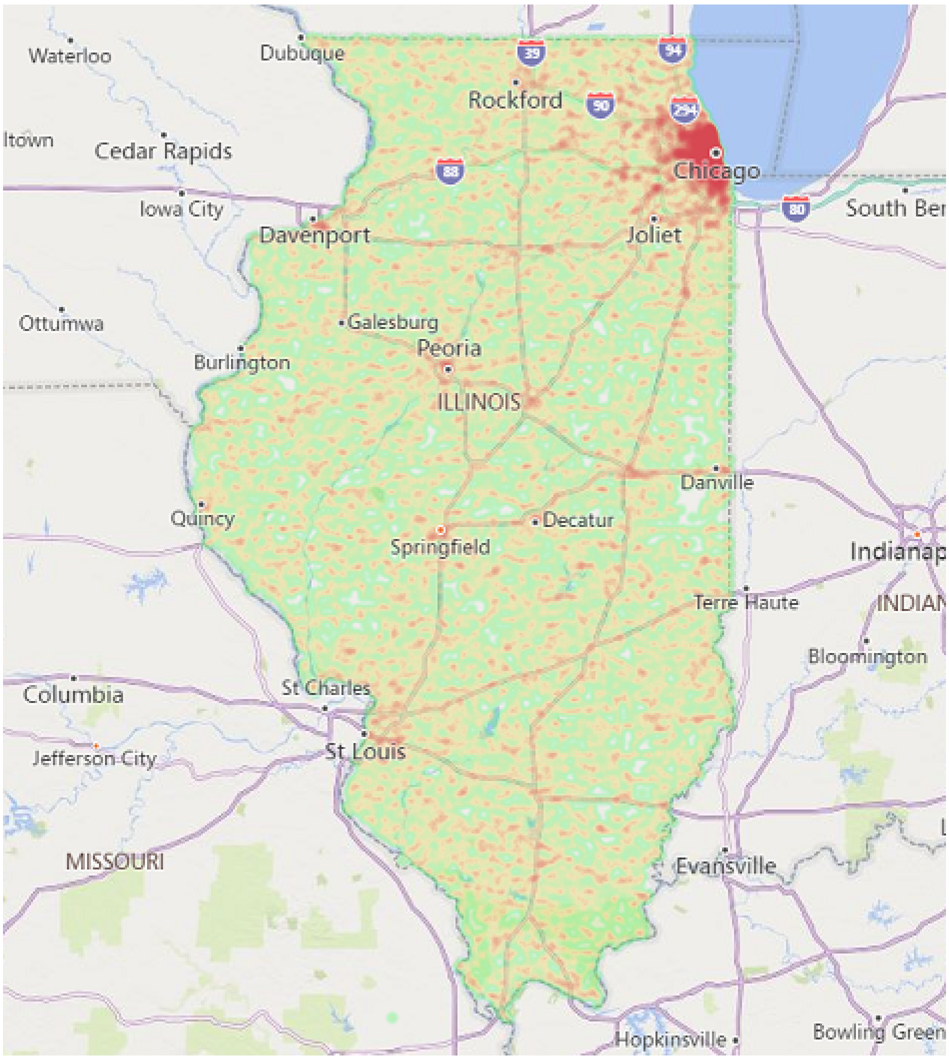

Figure 8: Distribution of accidents by Road Defect Condition. From 2017 to 2019.

Bar Plot

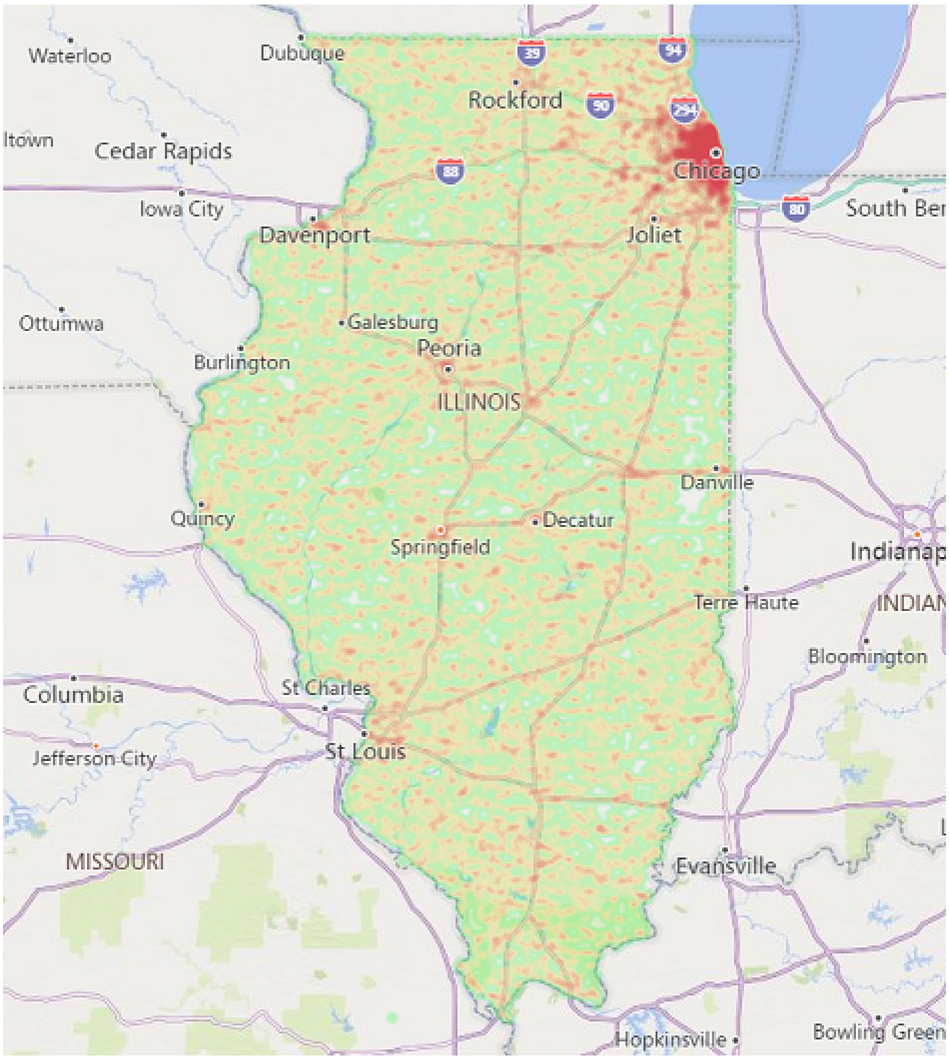

Figure 9: Distribution of accidents over time. From 2017 to 2019.

Correlation Plot

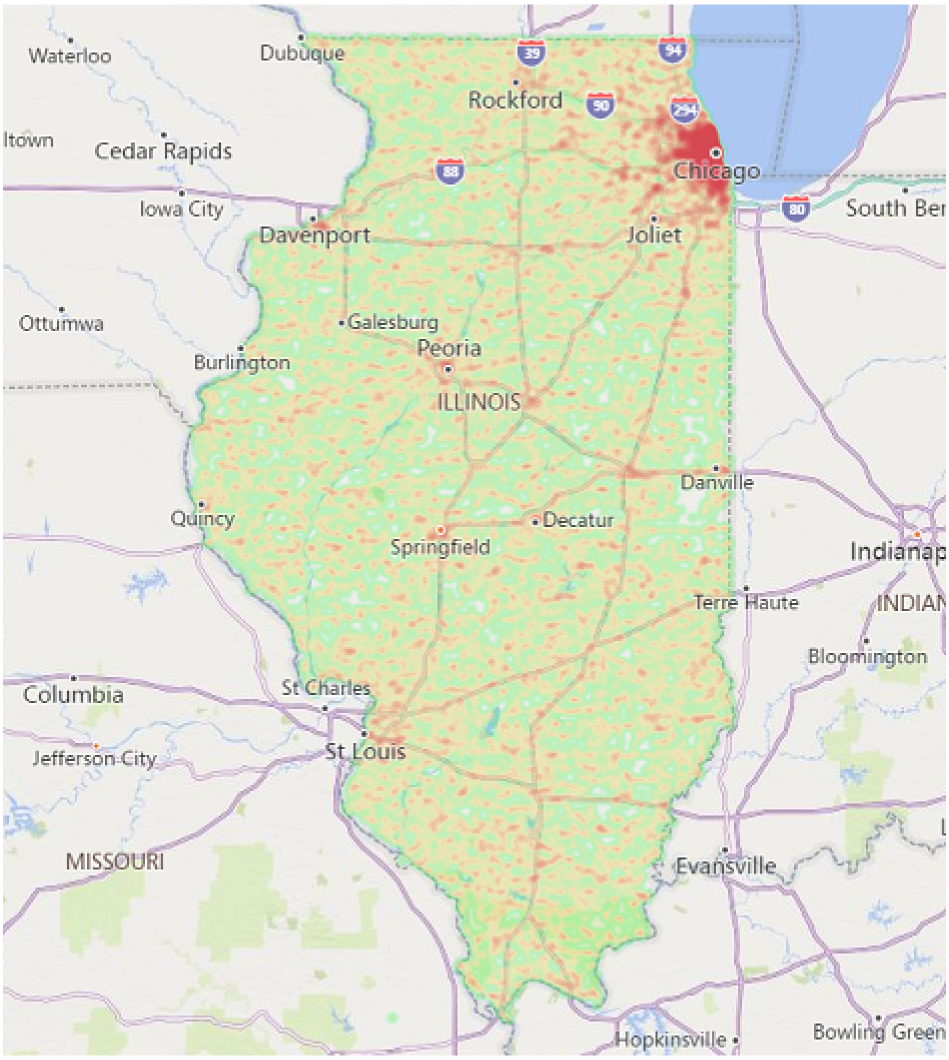

Figure 10: Correlation of Independent Variables. Only the top 5 independent variables have been cosidered.

Trends

Since not all the variables have to be present for an accident to occur, it can be seen that most accidents happen in the absence of adverse conditions. However, we should take into account that this reflects the fact that adverse conditions are exceptions, and accidents happen on a daily basis with other factors as underlying reasons such as human behavior. However, adverse conditions do increase the likelihood of accidents and it can be observed an increase in the overall number of occurrences in specific hours (evening), days (weekends), and months (winter).The road type is also found to be directly correlated with the maximum speed limit, and as a consequence is tied to the number of accidents per day.

Potential Issues

As long as the number of entries containing a value for an independent variable overcome by large any other value for the same independent variable, we may experience problems related to "imbalanced data" due to the uneven distribution of observations. Similarly, it can be seen that most of the independent variables are "classifications", and therefore their entries don't provide meaningul numerical values to be analized or correlated. For some of them we could replace the text values by boolean variables, but for some others a rating system may be needed if a numerical interpretation is required.

It can be noticed that most of the attributes are subjective observations trying to describe the potential causes of an accident, and may be dependent on the observer itself. However, the casualties are a meaningful numerical observation that will be thoroughly used througout this report.

Predictive Model

Open-source crash data is published by the Illinois Department of Transportation (IDOT) yearly. Each crash report was found to have extensive entries with up to 85 attributes, which include several independent variables to describe each occurence. Each dataset is organized with observations filled out according to the IDOT Traffic Crash Report SR 1050 Instruction Manual (2019). The datasets for each year are available online in .CSV format at the IDOT website, and they contain observations arranged in rows and attributes in columns. The datasets from 2017, 2018, and 2019 were included in this Exploratory Data Analysis. The datasets from 2020 and 2021 were discarded in this analysis given the COVID-19 pandemic outbreak, which altered drastically the dynamics of traffic worldwide, and thus crash-related data (Yasin, Grivna & Abu-Zidan, 2021). Regarding the dataset size, each one had originally over 300,000 rows (944,328 in total, combined). The Exploratory Data Analysis will be carried out following the steps described in the next sections.

References

Illinois Department of Transportation | IDOTAdmin https://gis-idot.opendata.arcgis.com/search?groupIds=6d2862031a6d47c7a8c211e38e423e05

Illinois Department of Transportation | Traffic Crash Report SR 1050 Instruction Manual https://idot.illinois.gov/home/resources/Manuals/Manuals-and-Guides

Safarpour, H., Khorasani-Zavareh, D., & Mohammadi, R. (2020). The common road safety approaches: A scoping review and thematic analysis. Chinese journal of traumatology, 23(02), 113-121. https://doi.org/10.1016/j.cjtee.2020.02.005

World Health Organization (WHO). (2022). Road traffic injuries World Health Organization Regional Office for the Eastern Mediterranean. Geneva, Switzerland. https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

Yasin, Y. J., Grivna, M., & Abu-Zidan, F. M. (2021). Global impact of COVID-19 pandemic on road traffic collisions. World journal of emergency surgery, 16(1), 1-14.

Yu, R., Han, L., & Zhang, H. (2021). Trajectory data based freeway high-risk events prediction and its influencing factors analyses. Accident Analysis & Prevention, 154, 106085. https://doi.org/10.1016/j.aap.2021.106085.

HAUER, E.; HAKKERT, A. S. The Extent and Implications of Incomplete Accident Reporting. Transportation Research Record: Journal of the Transportation Research Board, n.1185, p.1-10, 1989.

CHIN, H. C.; QUEK, S. T. Measurement of Traffic Conflicts. Safety Science, Vol. 26(3), p. 169-185, 1997. DOI: https://doi.org/10.1016/S0925-7535(97)00041-6.

FARMER, C. M. Reliability of Police-Reported Information for Determining Crash and Injury Severity. Traffic Injury Prevention, 2003, n.4, p.38-44, 2003. DOI: https://doi.org/10.1080/15389580309855.