Crash Risk Prediction Model using Data Science

This manuscript (permalink) was automatically generated from uiceds/cee-492-term-project-fall-2022-swifties@74f8005 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

In this section, a set of plots, charts and visuals are presented to present some of the findings of the exploratory data analysis. Certain conditions that were identified as more significant are displayed in the graphs below

Bar Plot Crashes

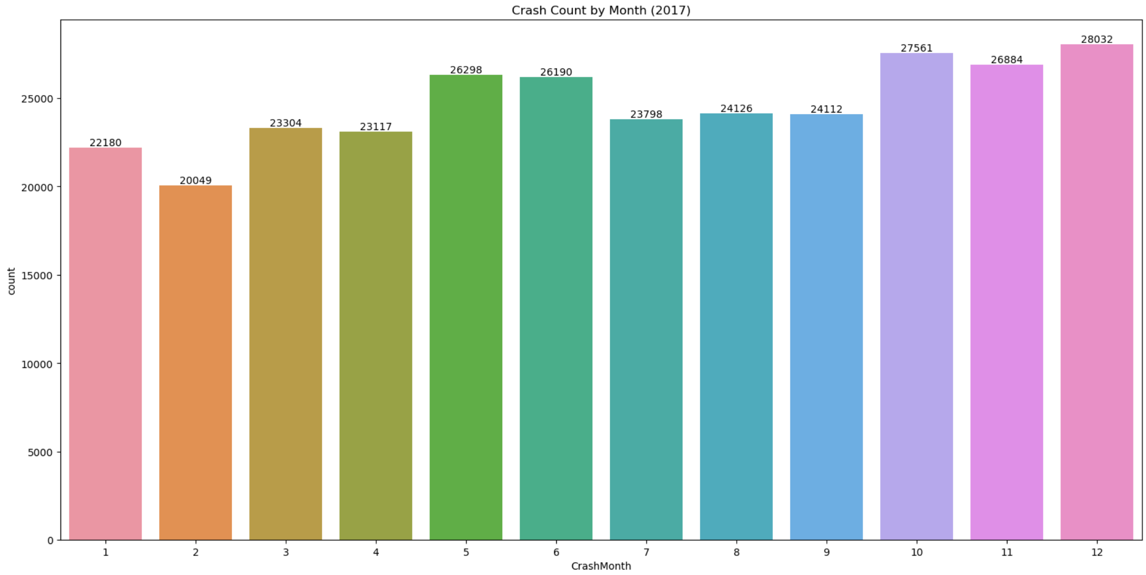


Figure 1: Bar Plot Historical Crashes. Crash reports from 2017 in the state of Illinois, USA.

For 2017, the month with the most incidents is December (“12”), with 28032 car crashes (out of 295651 for that year), and the month with the least car crashes is February (“2”), with 20049 incidents that month. These findings can represent the effect of weather conditions as well, given that during the very first and the later years of the year, the amount of accidents increases. These months correspond to winter season, and englobes certain holidays where people might be very active and potentially more car crashes might happen.

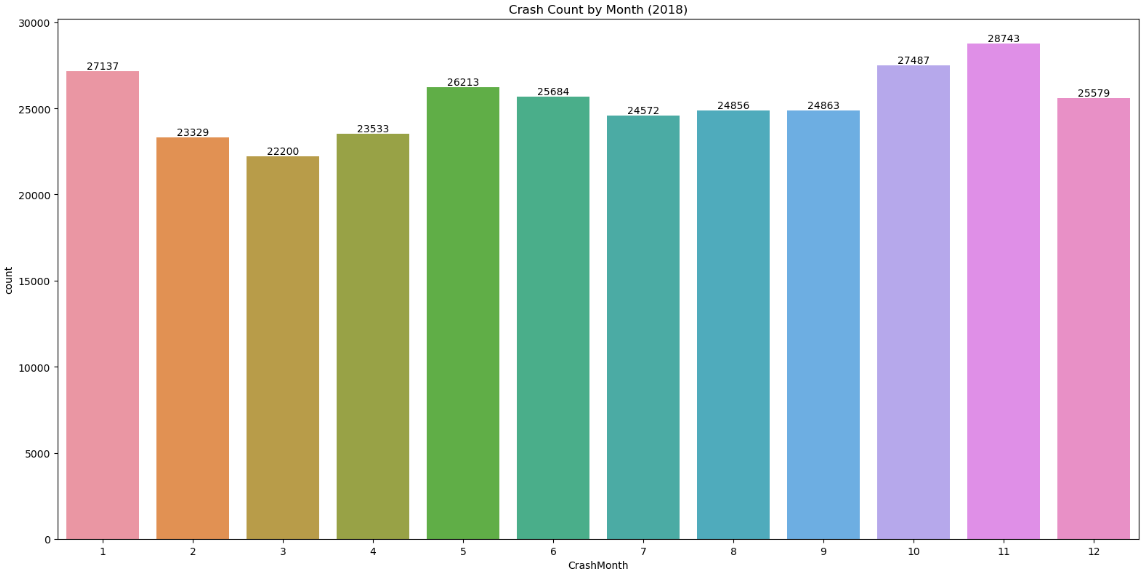


Figure 2: Bar Plot Historical Crashes. Crash reports from 2018 in the state of Illinois, USA.

Now, for 2018, different from the previous year, the month with the largest count of accidents is November (“11”), with 28743 recorded crashes (out of 304196). The month with the least amount of crashes is March (“3”), with 22200 crashes. For this year, just like for 2017, the months were peaks happen can be associated to the worse seasons for a driver.

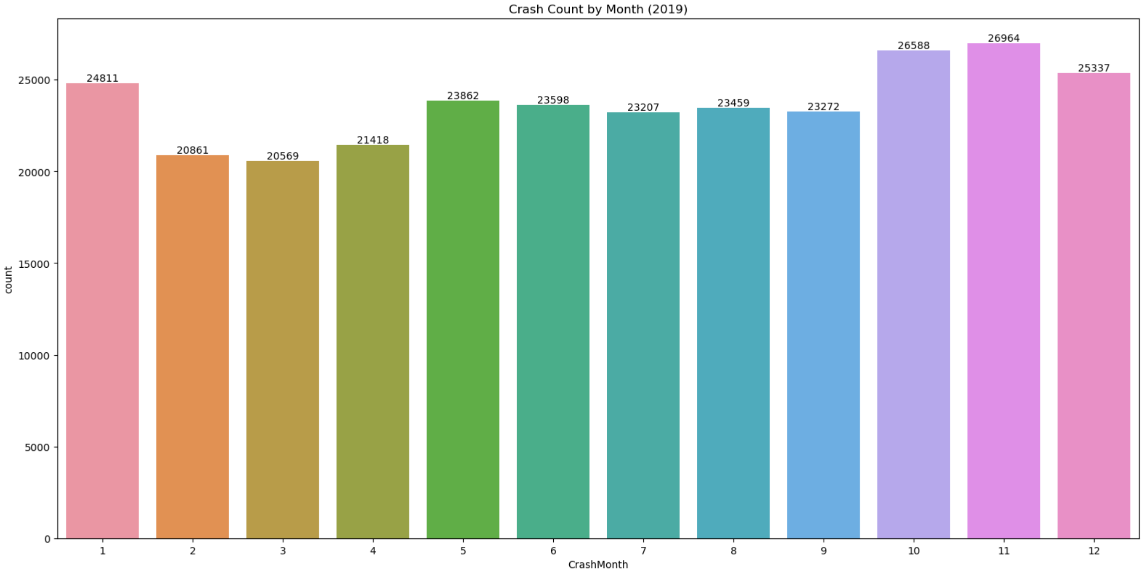


Figure 3: Bar Plot Historical Crashes. Crash reports from 2019 in the state of Illinois, USA.

For the year 2019, the month with most accidents is November (“11”), with 26964 recorded crashes (out of 283946), and the month with the least accidents is March (“3”), with 20569 incidents. in a similar way to years 2017 and 2018, the months with the highest count happen in the same season.

Map

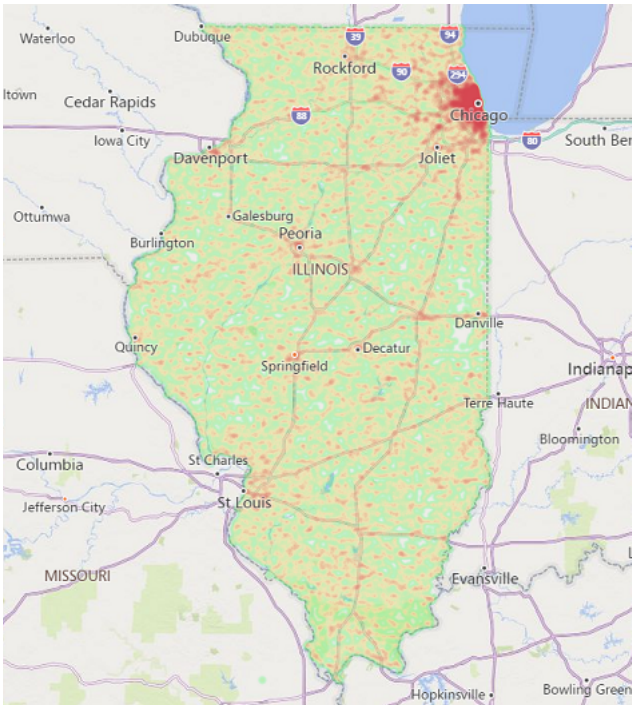


Figure 4: Distribution of crash occurrences. Crash reports from 2017,2018,2019 in the state of Illinois, USA.

Using the "X" and "Y" information present in the dataset, a visual distribution of the location of the recorded car crashes is illustrated in the map Figure. In the map of Illinois, the red dots represent a pair of "X" and "Y" coordinates, being the location where a car crash happened. The majority of car crashes appear to had happened in cities and towns. The best example would be Chicago, where all around the area, the number of occurrences, or red dots, is significantly higher, which may be presumable given that Chicago and the surrounding areas host a big fraction of the state's population. In a smaller scale, this is also visible that in other cities and towns, such as Springfield and Davenport. In other locations, the recorded car crashes are much more spaced out, with certain "hotspots" in some highways and roads.

Pie Chart | Lightning Conditions

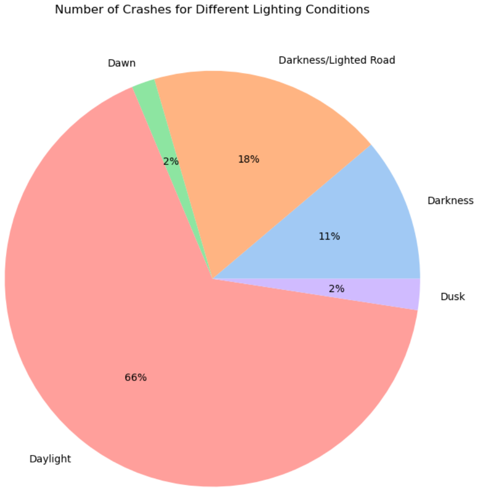


Figure 5: Distribution of accidents by Lightning Condition. From 2017 to 2019.

This figure displays the different percentages of the different lighting conditions presented in the dataset. The condition with the most car crashes associated to it is "Daylight", representing a 66% of the crashes in the dataset. During night time, "Darkness/Lighted Road" accounts for 18% of the crashes and "Darkness", 11%. Conditions "Dawn" and "Dusk" account for 2% each of the car crashes recorded.

Pie Chart |Road Surface Condition

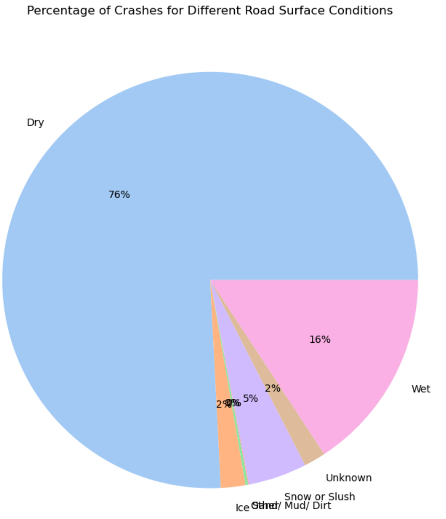
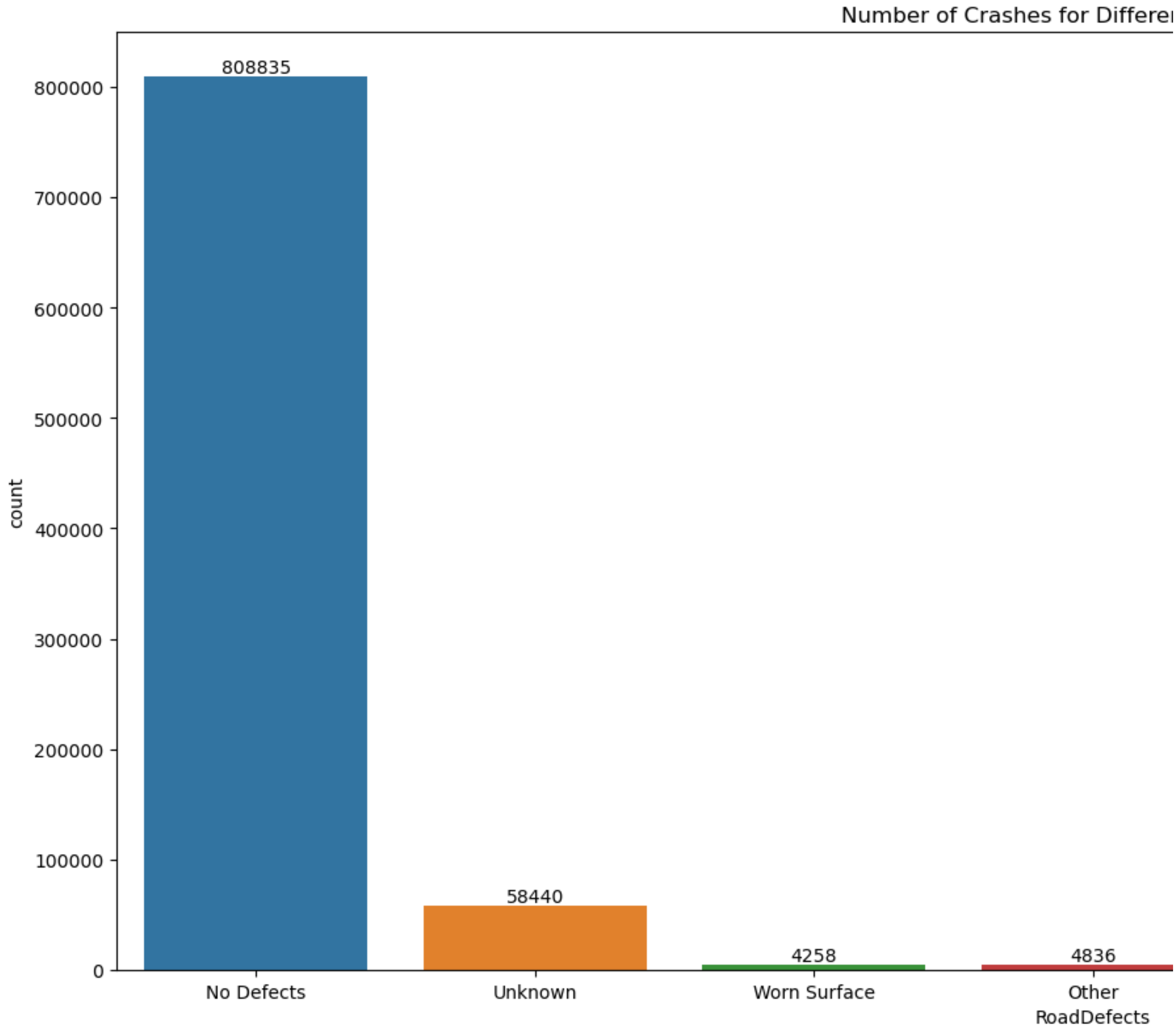


Figure 6: Distribution of accidents by Road Surface Condition. From 2017 to 2019.

The chart above summarizes the analysis performed on the data on the influence of the road surface condition on the amount of car crashes. It was found that 76% of the recorded crashes correspond to a "dry" road surface, which can be thought of as the least dangerous condition. For the not too favorable road surface conditions, 16% of the car crashes analyzed correspond to a "wet" road surface, 5% to "snow", 2% to "ice" and other "unknown" road surface conditions.

Bar Plot | Road Defects



{#fig:road_def}

This figure shows the number of car crashes associated with the different Road Defects. The horizontal axis labels the road defects from the data, and the vertical axis accounts for the number of observations. As seen in the bar chart, the vast majority of the incidents (808,835 out of 883,793 observations, which accounts for a 91.52%) happened where “no defects” were present in the location. The following condition of road defects is “unknown”, accounting for a 6.61% of accidents. Now, the percentage of observations where defects were reported are 0.5% for “worn surface” and “other road defects”, 0.14% for “shoulder” road defects, 0.49% for debris on roadways, and 0.21% for “ruts and holes”.

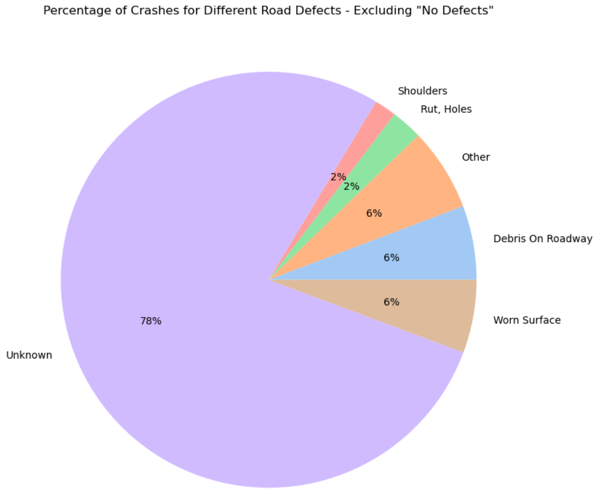


Figure 7: Distribution of accidents by Road Defects, excluding No Defects condition

Given that a great number of crashes happened without any road defects, it might be interested to account for the likelihood of road defects to be associated with a car crash. Figure X displays the different percentages of the road defects accounted for in the dataset, where “unknown” represents 78% of the data, followed by a 6% for “worn surfaces”, “debris on roadway” and “other” road defects, and 2% for “ruts and holes” and “shoulder” defects.

Bar Plot

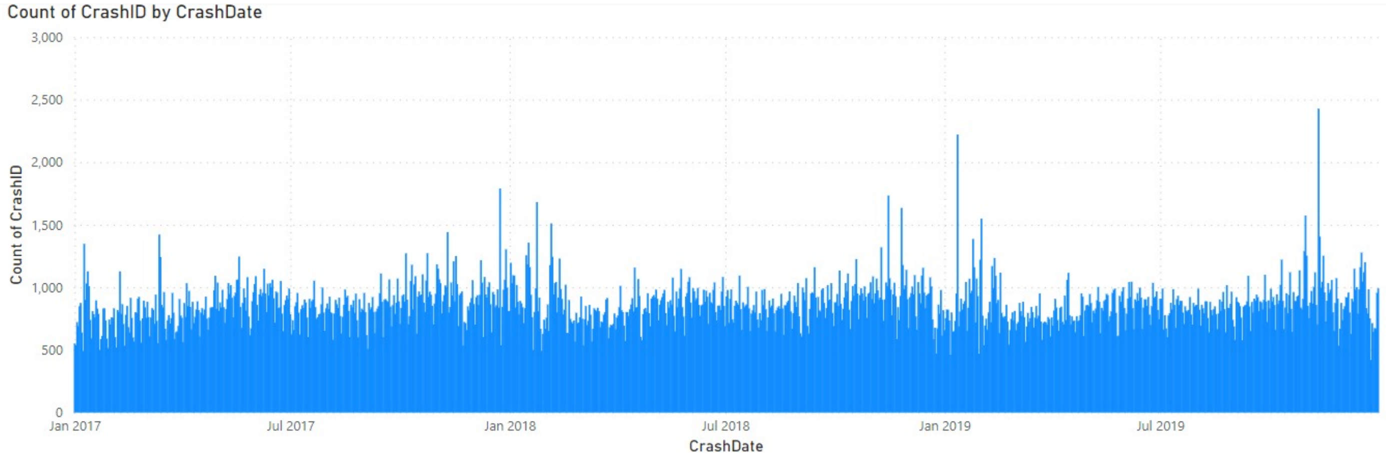


Figure 8: Distribution of accidents over time. From 2017 to 2019.

Correlation Plot

As mentioned earlier, one of the objectives of analyzing this data is understanding how different road and environment conditions would affect crashes and their severity. This can be obtained by finding the associations between the different variables (columns) in the dataset, meaning how is one variable affected by the other. However, most of the variables are of categorical type, i.e., variables that are identified based on names or labels given to them and not based on numbers. This makes the built-in correlation functions in Python or Julia not helpful. One very commonly used method to measure the correlation between two categorical variables is Cramer’s V statistic. Cramer’s V is based on a nominal variation of Pearson’s Chi-Square Test. Like correlation, the output takes values between 0 and 1 (inclusive), with 0 corresponding to no correlation between the variables and 1 corresponding to one variable being completely determined by the other. On the other hand, and unlike the usual correlation, there are no negative values. For this project, a function was created in Python that calculates the association between any 2 categorical columns using confusion matrix which can be obtained via built-in pandas method for categorical columns (pd.crosstab). For this data that has 24 columns, running this function for every pair of variables would take too much time and may not give many insights. Therefore, the function was used to find how the column “CrashSeverity” is correlated with every other variable. This column was chosen because finding how different conditions affect the severity of the crash is one of the most important outcomes of studying this dataset, and this would give an idea about the variables that have a significant impact on the crashes.

The output is described in the table below:

	Association with CrashSeverity
CrashYr	0.00807
CrashMonth	0.0268889
CrashDay	0.00470057
NumberOfVehicles	0.101171

	Association with CrashSeverity
DayOfWeekCode	0.0110936
CrashHour	0.0255177
CollisionTypeCode	0.259482
TotalFatals	0.707104
TotalInjured	0.705812
NoInjuries	0.355953
CrashSeverity	1
IntersectionRelated	0.141338
RoadwayFunctionalClassCode	0.0714974
WorkZoneRelated	0.00473353
TypeOfFirstCrash	0.259498
CityName	0.0738726
ClassOfTrafficWay	0.0608374
Cause1	0.165848
TrafficControlDevice	0.097696
TrafficControlDeviceCond	0.0515391
RoadSurfaceCond	0.0329927
RoadDefects	0.0424504
LightingCond	0.02852
WeatherCond	0.0307212

From the table above, it can be observed that the the factor that is the most correlated to the crash severity is the “TotalFatals” column which indicates the number of fatalities for each crash, with a correlation value of 0.707104. Similar observations can be made for the number of injuries. This makes sense because it is expected that the higher the severity of the crash, the higher the number of fatalities and injuries is expected to be. However, this is not very helpful for understanding how different conditions affect the severity of the crash. For this purpose, the columns that can be compared are: “IntersectionRelated”, “RoadwayFunctionClassCode”, “WorkZoneRelated”, “ClassOfTrafficWay”, “TrafficControlDevice”, “TrafficControlDeviceCond”, “RoadSurfaceCond”, “RoadDefects”, “LightingCond” and “WeatherCond”. Comapring these, it can be seen that presence of intersections has the highest correlation with the severity of the crash followed by the traffic control device. In addition, the characteristics of the workzone seem to have the least correlation with the severity of the crash. This observation can be useful to understand the dataset and get an idea about which variables are important for predictions.

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

Road crash prediction models are very useful tools in highway safety, given their potential for determining both the crash frequency occurrence and the degree severity of crashes (Abdulhafedh, 2017). While crash frequency refers to the number of predicted crashes for a given road under specific conditions, crash severity aims to correlate the casualties with contributing factors such as driven behavior, road conditions, and external factors (weather, lightning, etc). Identifying and analyzing the attributes influencing forecasting accuracy is of great importance in road crash prediction (Rashidi et al, 2022).

In a road crash dataset, the fatal crash samples, often constitute a very small proportion in comparison with non-fatal crash samples. Accurate prediction of fatal crashes, as a minority class, is one of the important challenges in such imbalanced sample distribution in most machine learning algorithms (Danesh et al, 2017). On top of that, several other factors such as the traffic flow or the average speed can greatly influence the prediction, so assumptions have to be made in order to develop a prediction model.

Given the nature of our database, the prediction model to be developed will focus on estimating crash severity based on our known attributes. For the reasons established before, crash frequency would require traffic data. Thus, trying to estimate it without this specific independent variable would lead to a completely innacurate model.

In the previous section, the Exploratory Data Analysis provided insightful information regarding the correlation of the independent variables, and a regression model will be the first approach for crash severity prediction. In this section, attention will be placed to understand the contribution of every independent variable to the overall result, to later start working on solving the data imbalance issue already identified.

The steps to be carried out can be summarized as follows:

1. Assign numerical values to the classification attributes to further study the influence of each factor in our target value.
2. Analyze the correlation between independent variables, and filter out those who are highly correlated.
3. Define an error metric and build a regression model.
4. Train the model to find the parameters that minize the error metric.
5. Divide the database into training data and test data.
6. Compare the predicted crash severity with both training data and test data, and assess the preliminary results.

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.