Illinois Vehicle Collision Data Analysis

W200 Instructor: **Gunnar Kleeman**

April 12th, 2021

Kumar Narayanan, Estrella Ndrianasy, Viswanathan Thiagarajan

Context

Car collisions are a common occurrence that happens every day across the USA, and in general all over the world. The USA, as a major country, and other countries too, have adopted surface transportation using cars as a major mode, if not the primary mode, of transportation. As we move away from major cities cars and personal automobiles become the only mode of transportation. Given that suburbs and rural areas still account for a large population the importance of cars and the dynamics associated with driving aspects become crucial.

The reasons for these collisions can be varied from driver errors to external factors like weather and road conditions. The more we understand factors that lead to accidents the more we can do to avoid these incidents and make our lives safer. Even if we agree that the fatality rates may have dropped with all the advances in technology it's important to bear in mind that an accident has other implications - delayed arrival, ripple effects in the form of taking the car to the garage for repairs, issues associated with towing, working with insurance companies, and in many cases, the financial burden due to higher premiums.

Our project seeks to understand the collision statistics and gain insights into the external factors and the reasons that may be causing these collisions. For this project, we consider the crash data compiled by the state of Illinois, USA, to analyze the impact of various factors.

Source Data

Main dataset: The Crash Report 2014 dataset has the collisions in Illinois in 2014 with the details of the route, date, injuries, fatalities, weather condition, surface condition, lighting condition, county, city, traffic control feature, number of vehicles involved, number of drivers involved, vehicle maneuver, driver condition, injury type, whether collision happened at an intersection, work zone, or when speed exceeded or cell phone was used.

File: CrashReport2014.csv

Source: Kaggle (https://www.kaggle.com/qcarver/crashes-2014-csv)

Dataset has 292019 rows and 80 columns in CSV format.

Supplemental Dataset: Illinois Counties Population from the Decennial Censuses from April 2010. This dataset includes information on the county-level population of Illinois in 2010 which is the earliest dataset available corresponding to the 2014 collision data.

File: CountyPopulationIL.csv

Source: https://www.dph.illinois.gov/data-statistics/vital-statistics/illinois-population-data

Dataset has 102 rows and 8 columns transformed into a CSV format.

Questions

Through the analysis of the car crash dataset we intend to answer:

Influence of conditions, such as weather, surface, light:

- What is the impact on accidents due to surface conditions?
- What impact does the weather have on accidents?
- Are there more accidents during certain periods of the day when it's not so bright?

Accidents at various times (month, day of the week, etc.):

- What months saw the most number of accidents?
- Are there variations in the accident number by days of the week?

Accidents by city/town, routes:

- Which counties/cities are worse than others in terms of injuries and fatalities and any inference to the location of these counties?
- How's the spread of the accidents in the state? What is the geo heatmap of all the collisions?
- Do certain routes (Interstates and Highways) have a trend in types of accidents and can we infer any specific details by narrowing down to the most common collisions?

Other factors and cross-correlation:

- What other factors may have influenced accidents - the presence of intersections with no traffic control devices, and driver errors (judgment/cell phone usage/excessive speed)?
- What type of collisions causes a higher rate of fatalities and if the number of injuries and fatalities increase with the number of vehicles that were involved in the collision?

Initial Exploration and data preparation

We performed an initial analysis on the dataset to check the different columns and their potential uses. Since there was no description or additional information associated with the columns, we reviewed each column by its name and made notes of which ones to be used based on the

clarity of the column name in the dataset. Some columns had obvious names like route, year, month, day, hour, coll_type (collision type), weather, lighting, surf_cond (surface condition), and other easily understandable columns. There were a total of 20 columns out of the 80 which we could not easily understand what they refer to or the significance of them. These 20 columns were excluded from the analysis like ‘objectid’, ‘SFE’, ‘xcoord’, ‘ycoord’ etc. The total list of fields used in the analysis is in the appendix.

The total rows in the dataset were **292,019**. Each row corresponds to one recorded collision event. Some key columns used in the analysis had NaN values. Below are the details

Column	# NaN values	Notes
ROUTE	178,491	This column lists Interstate numbers and US Highway numbers. We assumed that if there is a NaN value, then the collision did not happen on a major highway or an interstate.
WEATHER	7,874	No weather conditions for 2.6% of the records. These were ignored in our analysis
LIGHTING	6,800	No lighting conditions for 2.3% of the records. These were ignored in our analysis
SURF_COND	10,554	No surface conditions for 3.6% of the records. These were ignored in our analysis
TRAF_CNTRL	4,414	No traffic control details for 1.5% of the records. These were ignored in our analysis
REC_TYPE	117	No injury type for less than 0.05% of the records. These were ignored in our analysis
CITY	17	No city name for less than 0.01% of the records. These were ignored in our analysis

The latitude and longitude for each of the cities and unincorporated areas were derived using Python packages “geopy” and “Nominatim”. The XCOORD and YCOORD in the CVS didn’t have a reference index. Using these values to derive latitude and longitude without knowing the reference (such as EPSG number) would lead to a lot of guesses and may not give us the right answer. The location “Unincorporated” appears in several rows. To localize the reference, we

used the associated county number and township of the corresponding row to calculate the latitude and longitude.

A new “Date” column was created using the three columns MONTH, DAY, and YEAR.

A new “Critical” column was created aggregating the total number of injuries and fatalities in each row

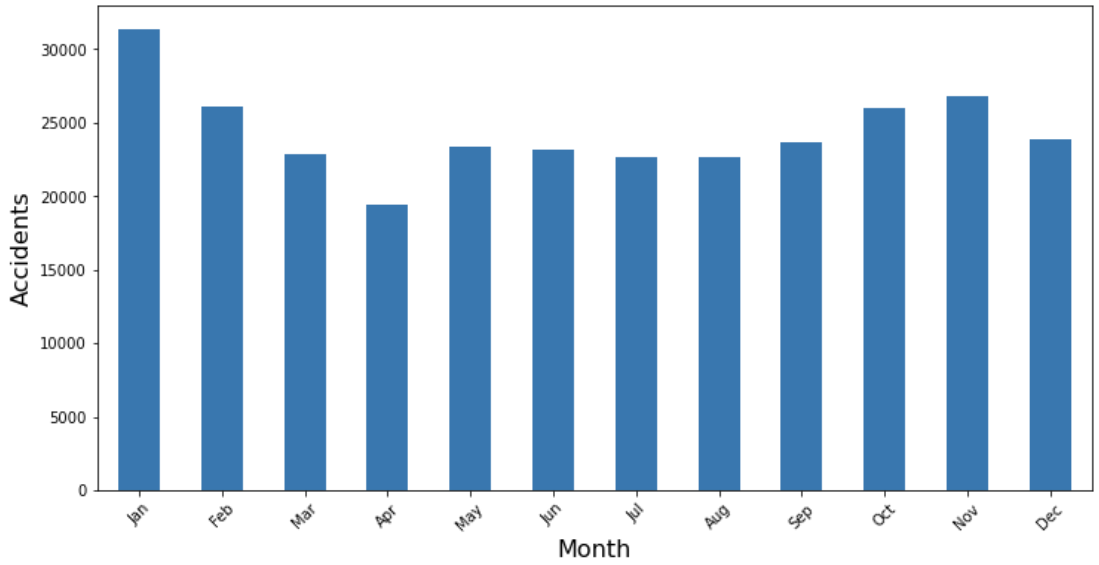
Some route names ended with a “*” and these were cleaned up for easy grouping and presentation.

Exploratory Analysis

We started the analysis of the data and created a monthly accident summary chart to see if we can spot any trends or seasonal correlation.

Acccidents by Month

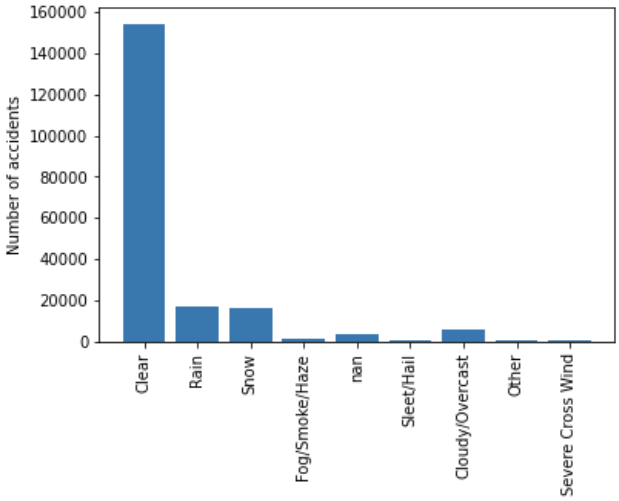
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
31406	26068	22910	19413	23394	23215	22615	22633	23673	25981	26834	23877

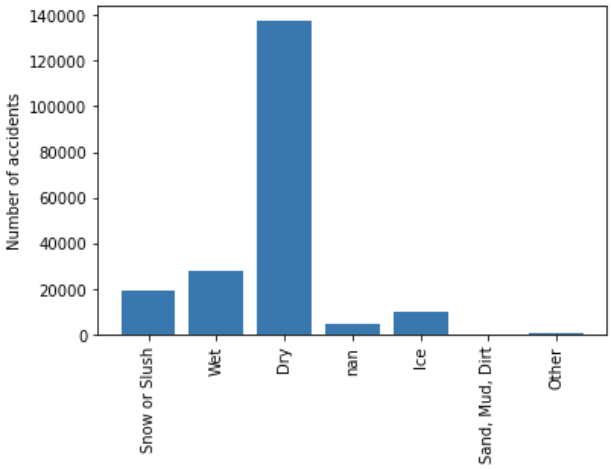


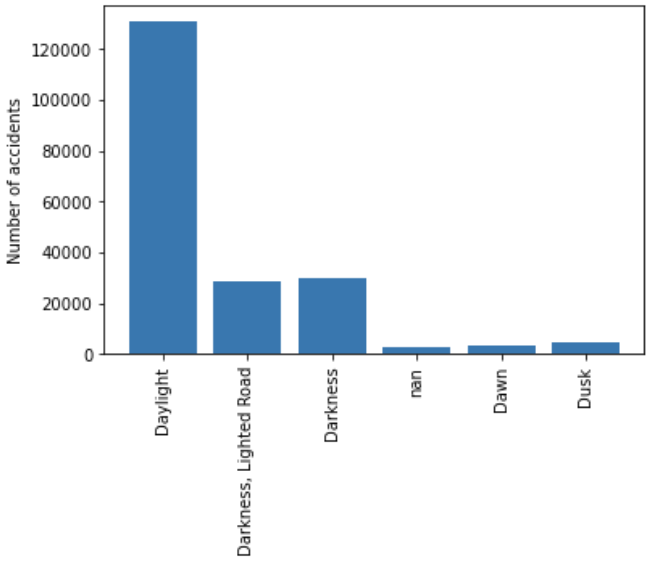
We observed that there was no seasonal correlation. The month of January saw the most accidents and the month of April saw the least. As later data would show the amount of snow and slushy conditions had some influence on the number of accidents.

We also looked at the influence of other factors such as weather, road conditions, and lighting.

Interestingly, a lot of the accidents seemed to have happened when the weather was clear, road conditions were dry, and when there was plenty of daylight.

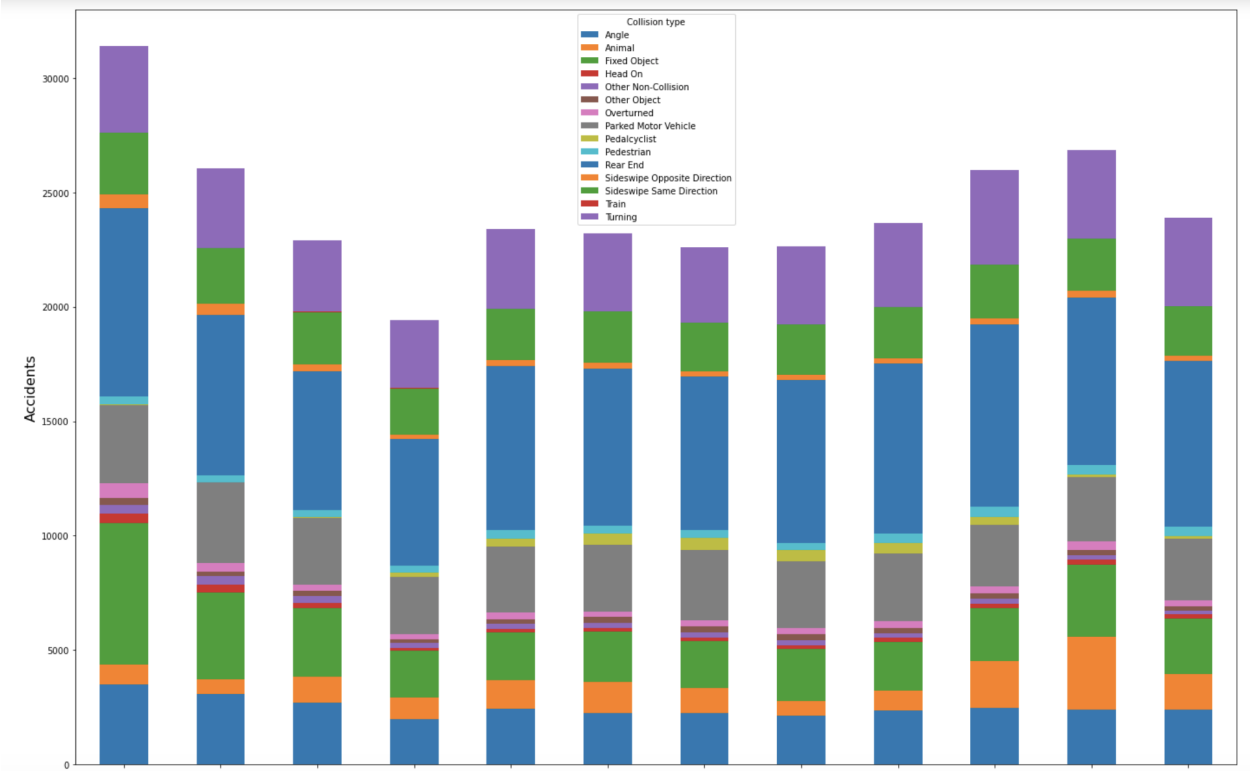




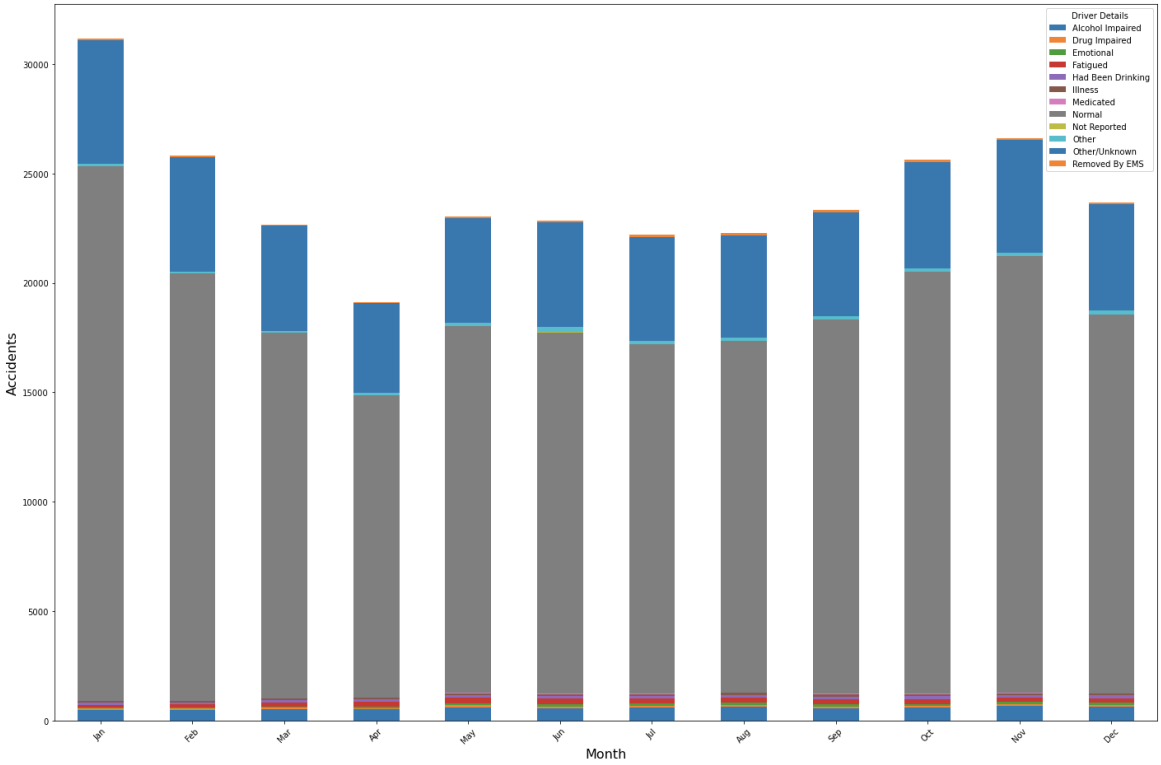


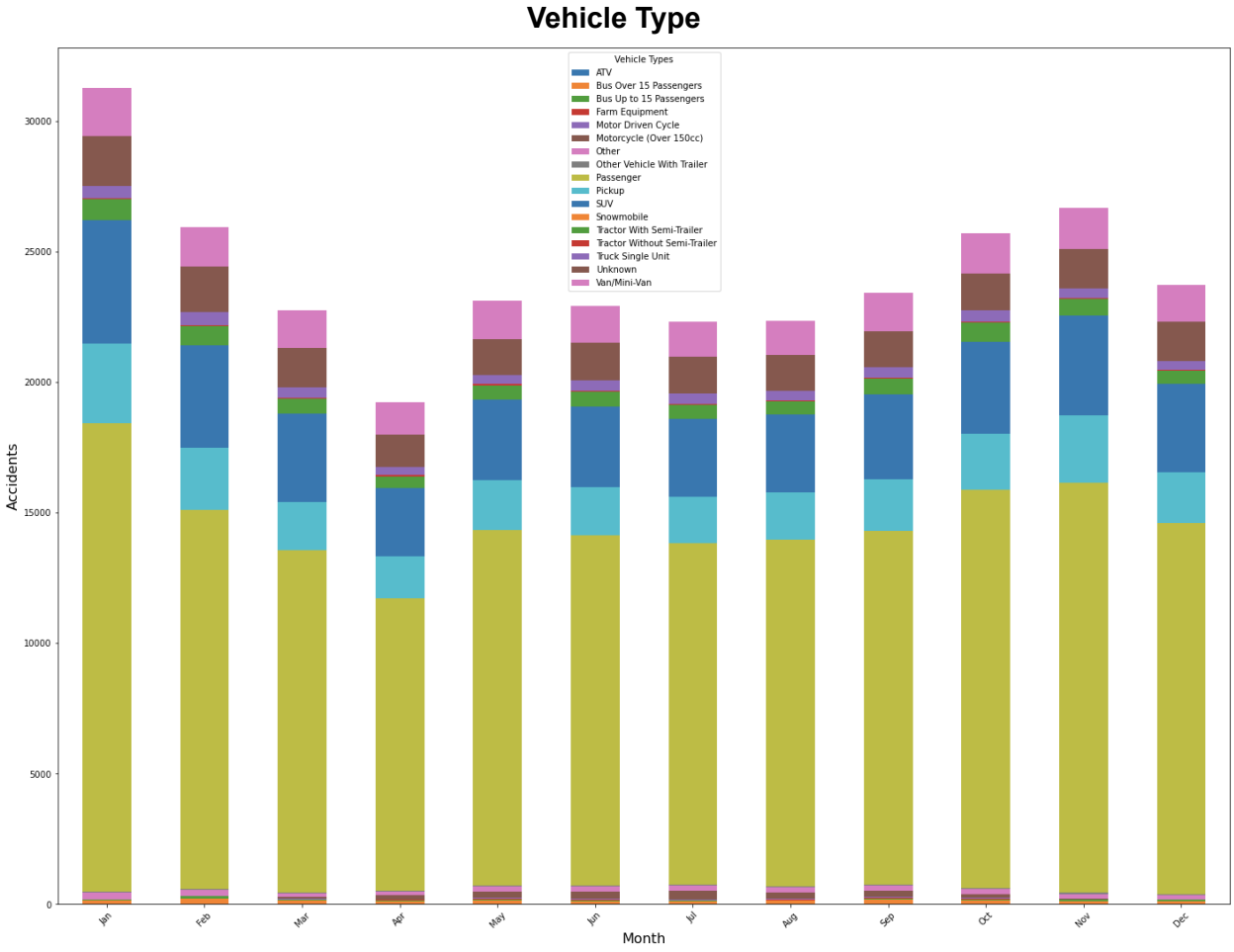
We also looked at the types of collisions, condition of the driver, and vehicle type involved.

Collision Type



Driver Condition

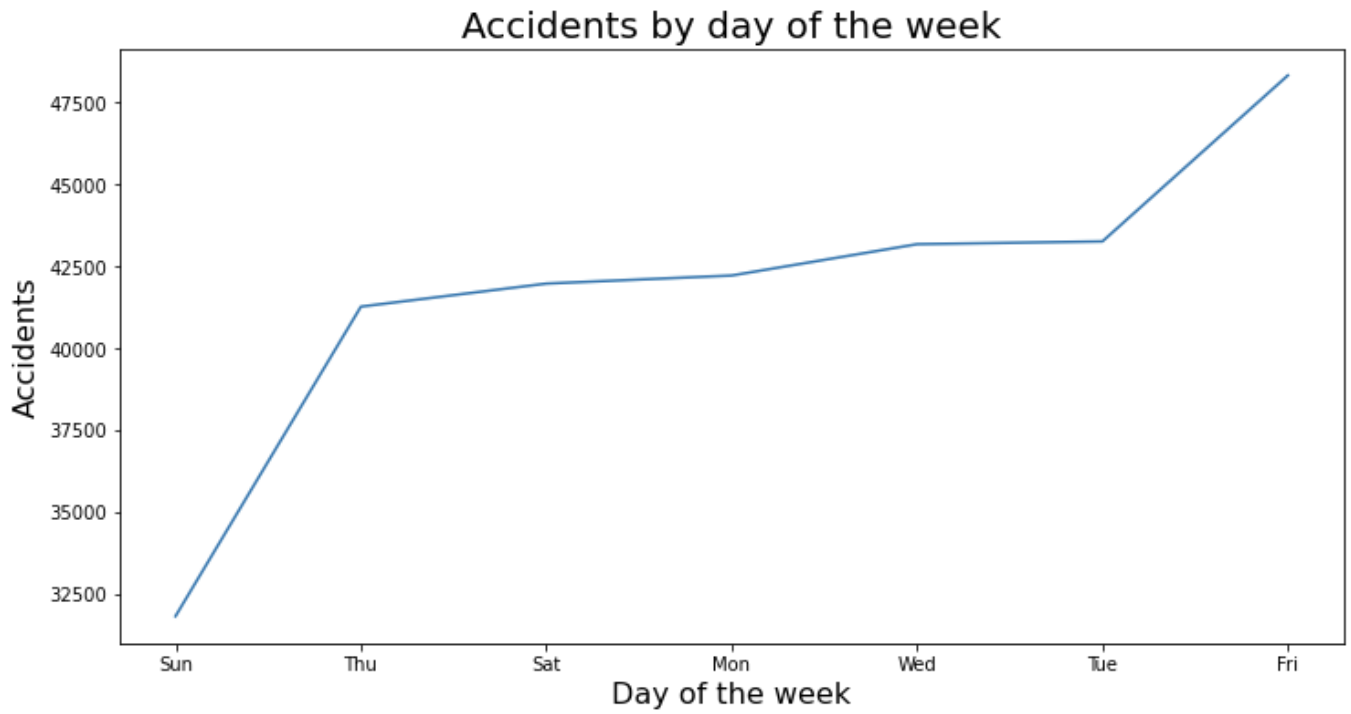




The observation was that most collisions were rear-ended collisions, the condition of the driver was normal, and the vehicle type was common passenger vehicle (sedans, minivans, etc.)

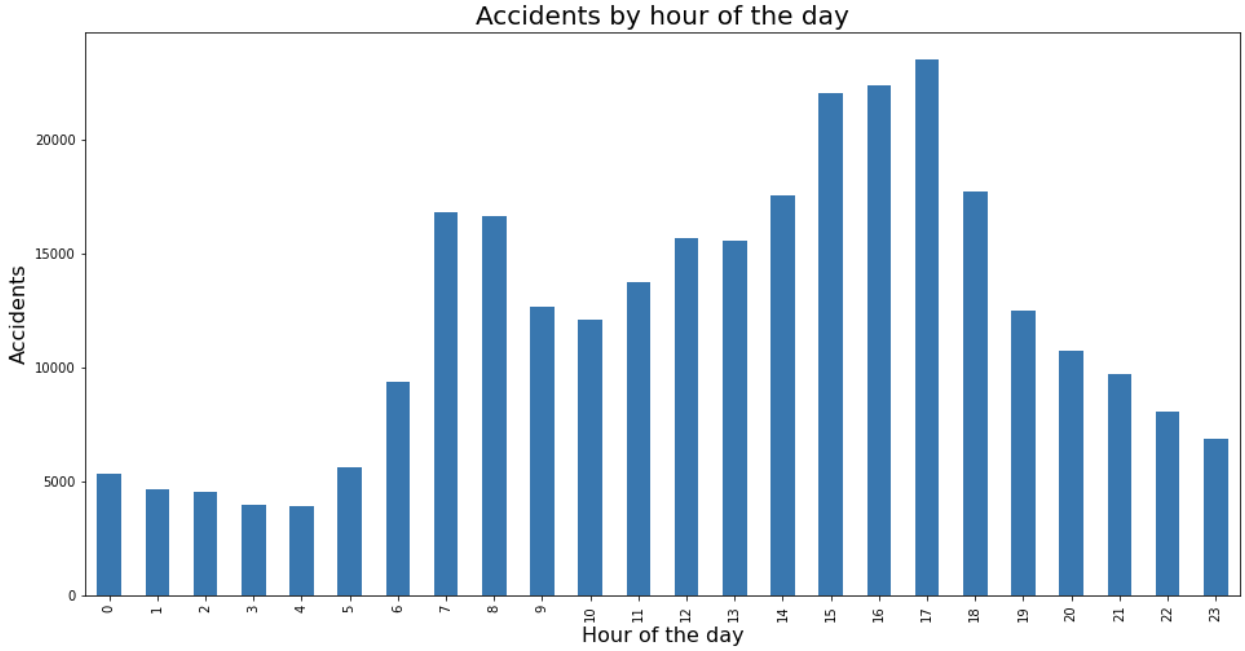
The First Insight:

We observed that a bulk of the accidents happened when there was enough light, under clear weather conditions, with passenger cars, and with drivers under normal conditions. This forced us to look at the day of the week for a trend.



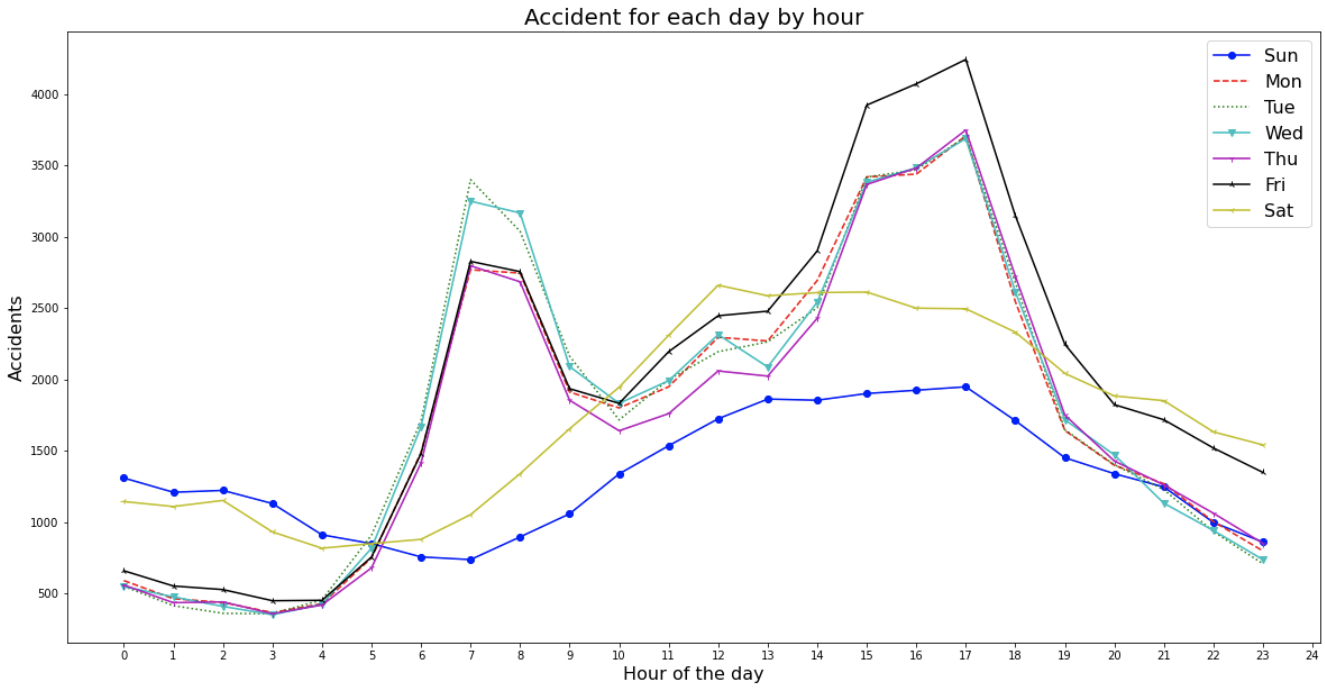
We observed that the accidents were low on Sunday, and climbed steeply on Friday. This was observed every month.

We then drilled down to the hour of the day trend. We saw the following trend.



We observed a consistent increase in the hours between 7AM and 9AM, and also between 3PM and 6PM.

This further forced us to look at the number of accidents for each day of the weak, on an hourly basis.



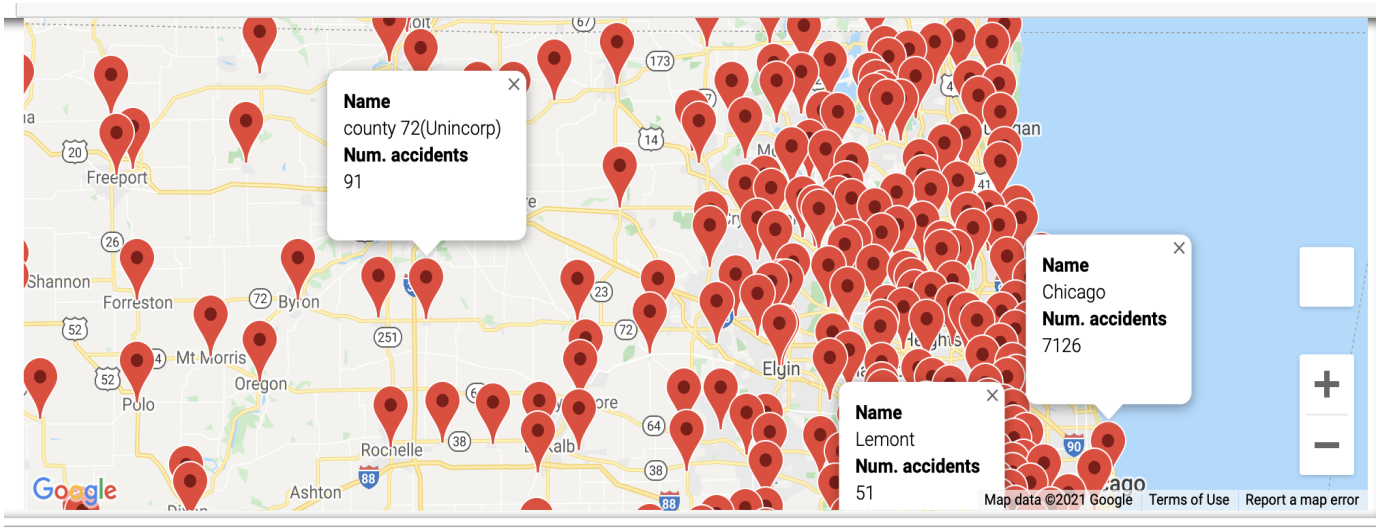
The trend revealed something clearly - commute hours to and from work is the major contributor. This trend is clearly visible during the working days, and comes down significantly on weekend days.

Is this the reason insurance companies focus on commute details for giving a quote?

Geo spread of the collision locations

We then looked at the spread of accidents across the state of Illinois. We initially wanted to a traditional heatmap; instead we chose to use Google APIs to plot on an interactive map the accident locations, with a pop-up for details.

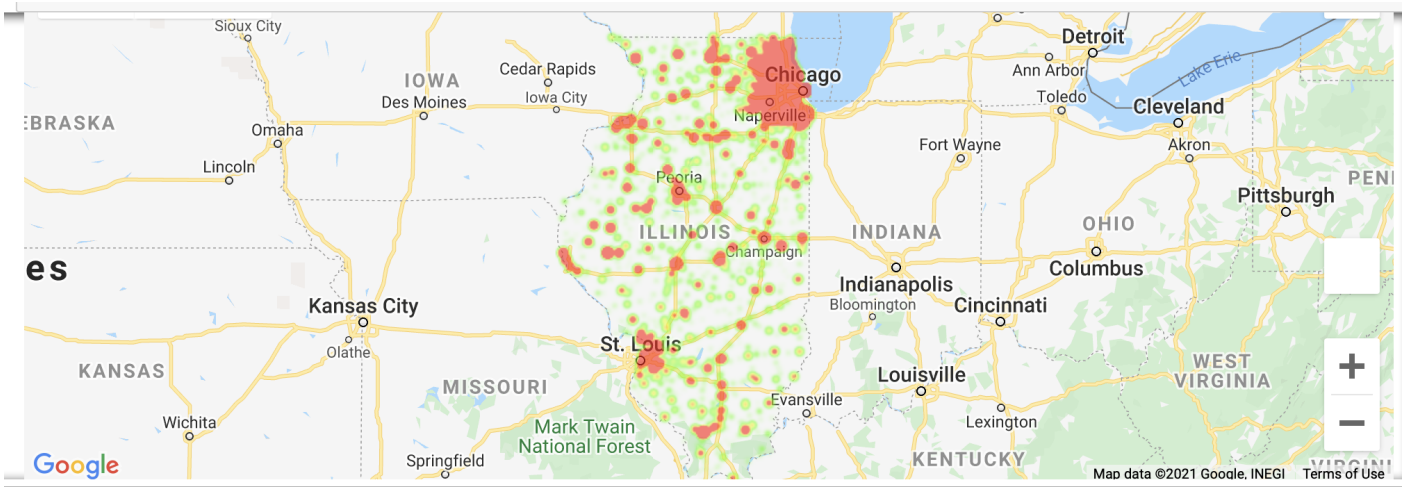
Note: The Google APIs create ‘layers” to overlay the accident locations and details. It’s not a static map, and needs to be seen with JUPTYER NB.



We observed that while Chicago accounted for most accidents the accidents are clustered around suburbs of Chicago, and thins out as we move away from Chicago, toward rural and unincorporated areas.

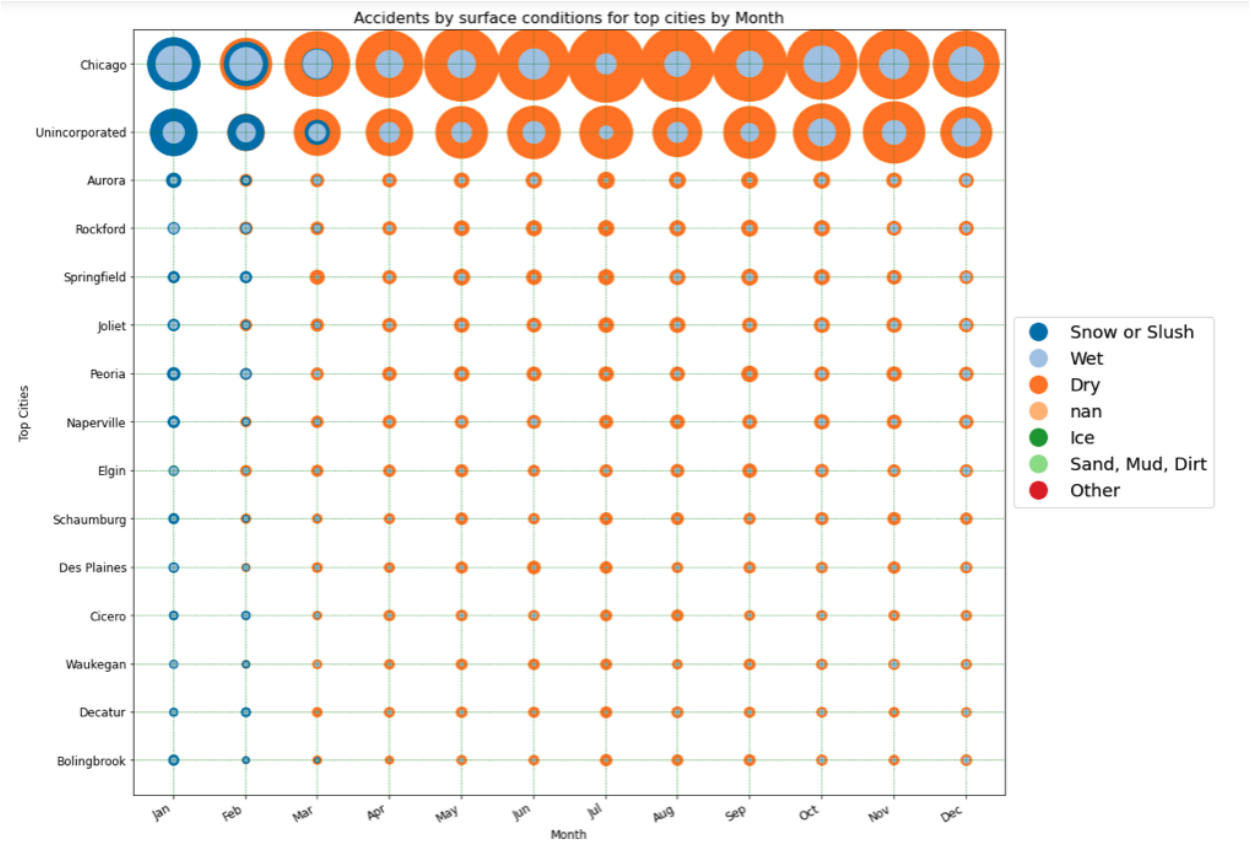
Heat map of the accidents by locations

A heat map across the state of Illinois was then created, again using Google API and GMAPS package.

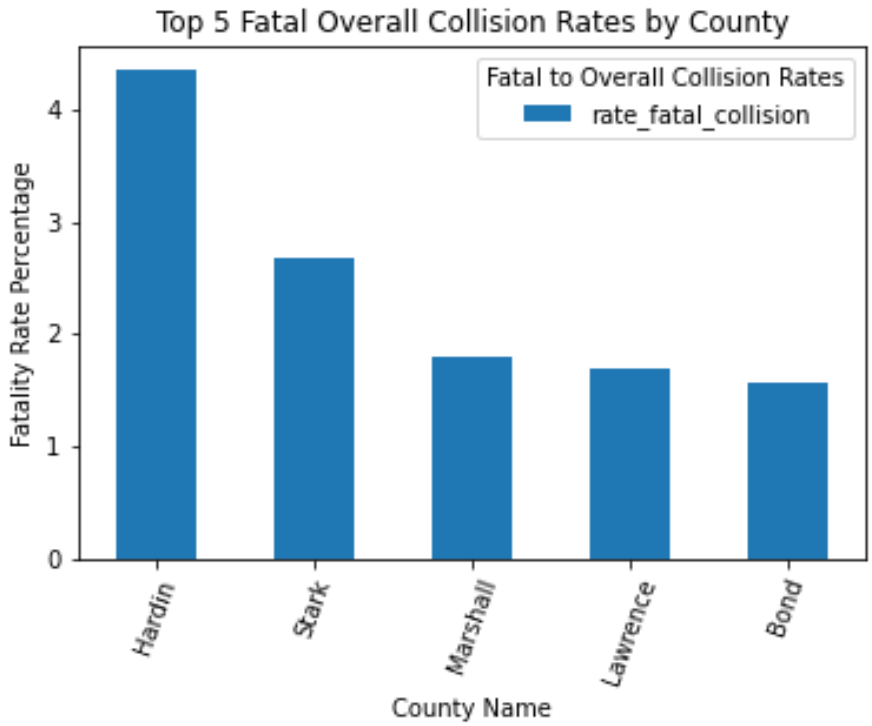


It can be seen from this map that in the state of IL, a good number of accidents happen around the Chicago area. While the previous graph gave an ability to drop pins on accidents locations and pop up details, this visualization allows us to get a feel for the entire state without being crowded by too many details.

Comparison of top cities on accidents and the causes.

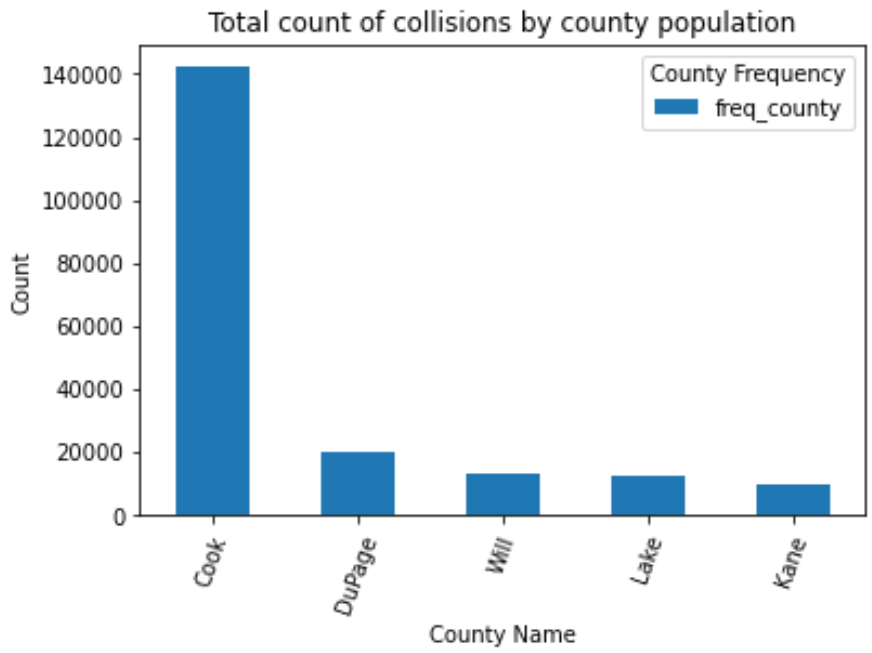


Clearly, the bulk of the accidents are in and around Chicago. In the months of January and February Snow or Slush has a significant role to play. However, accidents in other months are happening in Dry conditions.



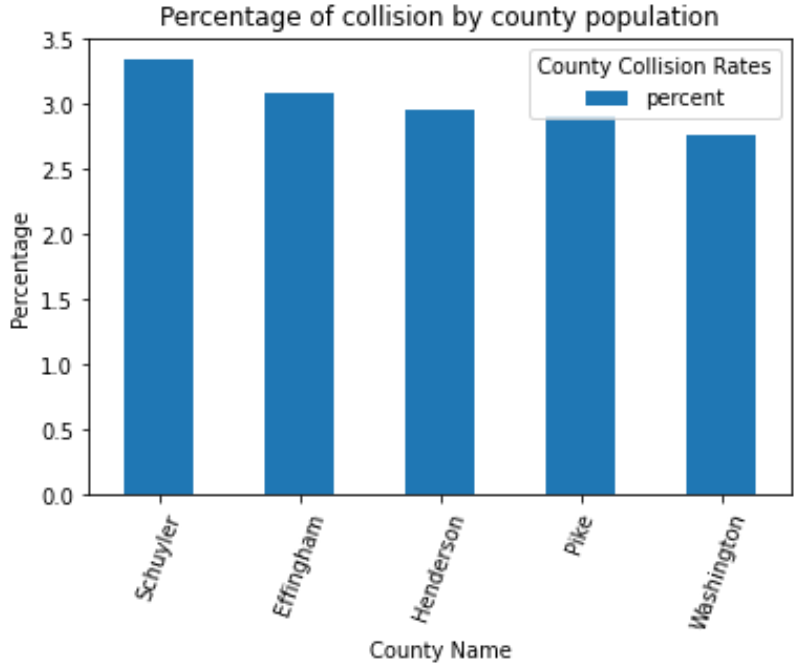
This graph shows the fatality rate in all collisions by counties. Stark, Marshall, and Bond are urban counties whereas the top county with collision fatalities, Hardin, is rural along with Lawrence per U.S Health Resources and Services Administration definition. Although the top county for collision fatalities is rural, it seems urban areas see much more collision fatalities than rural ones.

What are the top 5 counties by number of collisions?



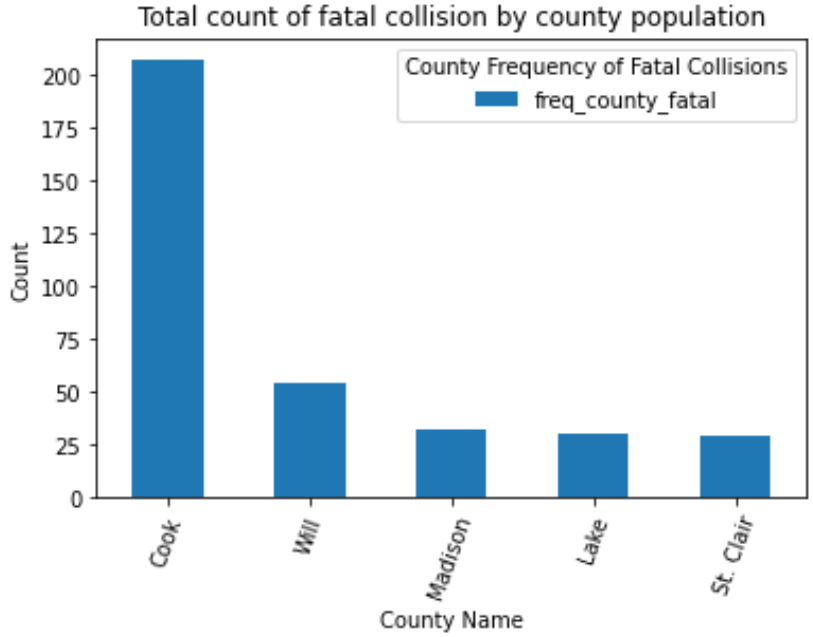
Cook county which has Chicago as a county seat had the highest number of collisions in 2014. The Cook county accidents accounted for almost half of all collisions in the dataset. Cook, DuPage, Will, Lake, Kane counties are all urban/non-rural counties allowing them to have better access to medical care and enhanced road infrastructures.

What are the top 5 counties with a higher rate of collisions with respect to the population?



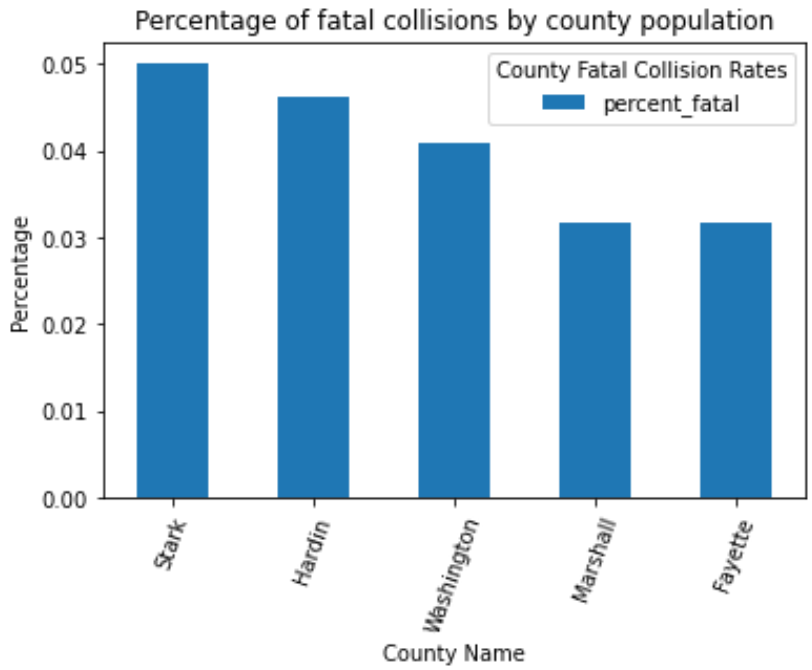
When assessed at the population level, rural counties disproportionately see a higher rate of collision per inhabitant. This may be due to lacking road infrastructures and a higher propensity to drive due to the lack of public transportation.

What are the top 5 counties by the absolute number of deadly collisions?



Cook county leads in the count of total fatal collisions. This may be due to its more dense population. The other counties with the most collision fatalities are also all found in urban areas lending credibility to population density driving collisions leading to a higher incidence of road death.

Deadly counties - What are the top 5 counties with higher rates of fatalities with respect to the population?



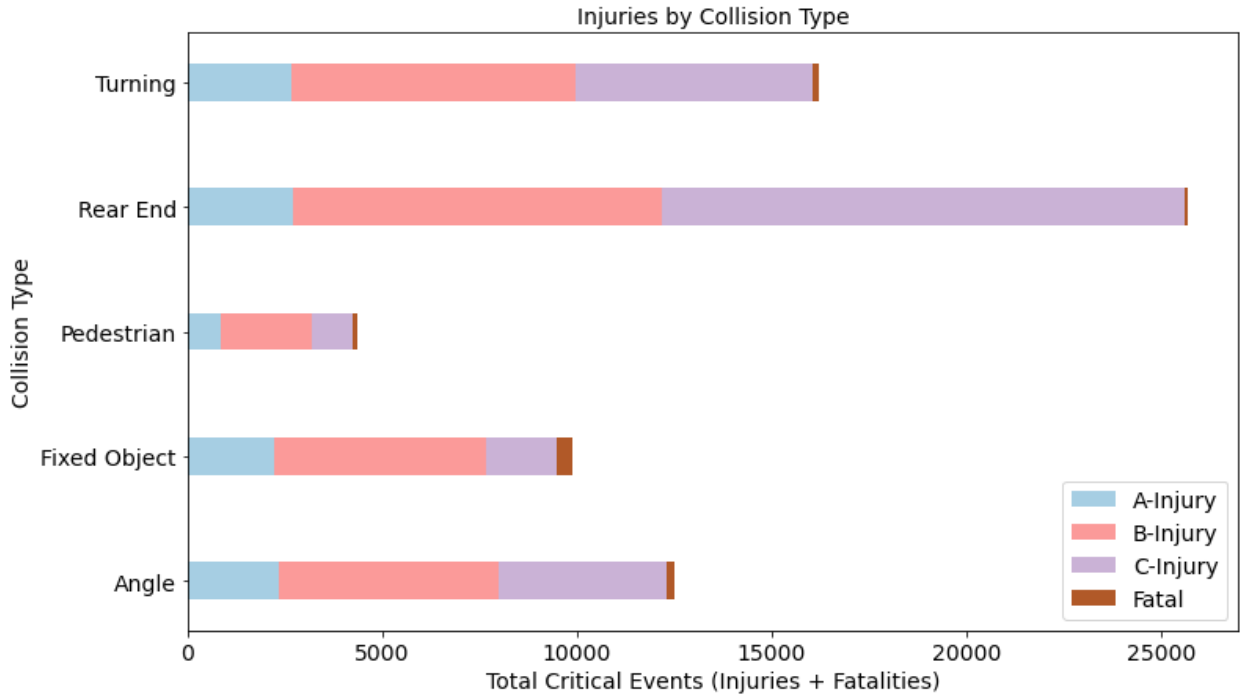
Similarly to the rate of fatalities compared to overall collisions, the population breakdown points to rural areas leading the state of Illinois in fatalities per inhabitant. Hardin, Washington, and Fayette are rural counties among the top 5 of all collision fatalities per capita. However, urban counties such as Stark and Marshall are significant in driving the collision fatalities rate at spot number one and four of the top five respectively.

Summary of the different top 5 collisions/fatalities by county analysis:
Urban areas see a higher count of fatalities to overall collision rate as a whole, although rural areas such as Hardin, have the single highest death toll in total and per inhabitant. In general, rural areas tend to experience more collisions' related fatalities per population than urban areas per inhabitant which may be due to a higher road utilization and subpar access to medical care.

Which type of collisions causes a higher number of injuries and fatalities?

To better understand the type of collisions that cause the different levels of injuries, we analyzed the injury type (REC_TYPE) and the collision type (COLL_TYPE) columns in the dataset. All collisions with only Property Damage (PD) were excluded in this analysis.

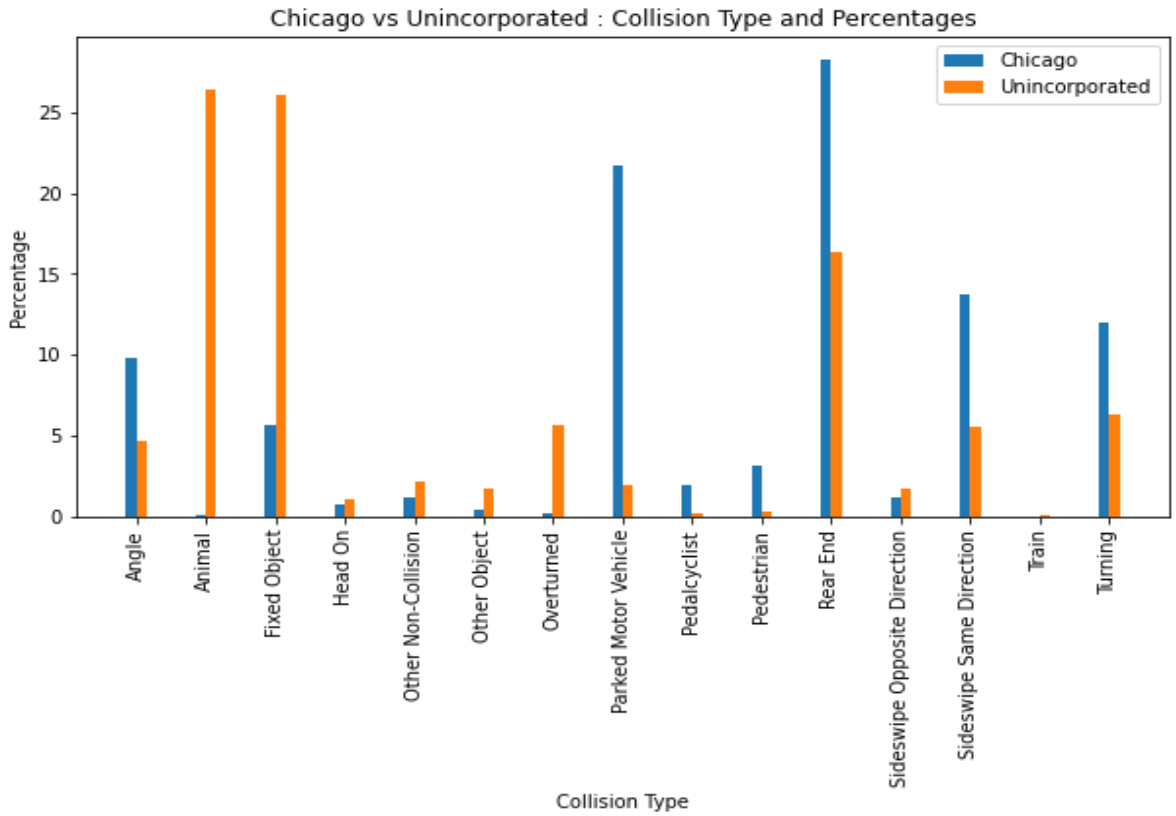
Injury Type	Injury Level
C-Injury	Minor
B-Injury	Moderate
A-Injury	Major or Serious
Fatal	Fatality



We see from the plot that most of the collisions that resulted in an injury or fatality were due to rear-end collisions. This type of collision happens when a driver is unable to safely stop the vehicle and hits the rear end of another vehicle that is traveling in the front. Most rear-end collisions are attributable to driver error.

Observation: Rear-end collisions caused the most number of injuries and these may be attributable to driver judgment errors. Whereas collisions with a fixed object caused the most number of fatalities. A fixed object collision may/may not be attributable to a driver error as it can also be due to external factors such as maneuvering to avoid collision with another driver who lost control.

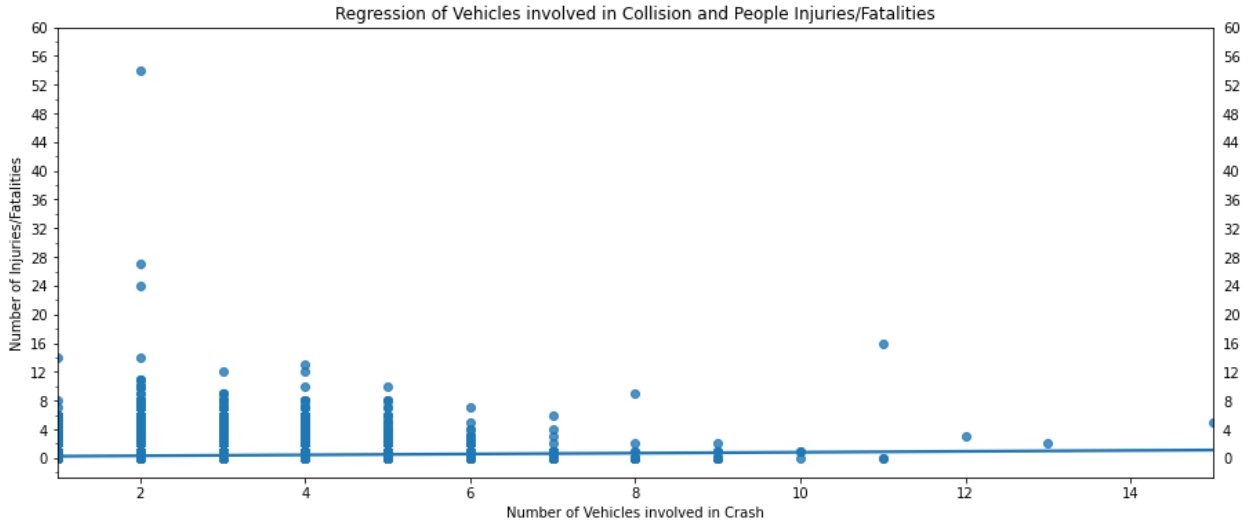
Is a difference in the types of collisions in the sparsely populated unincorporated areas and the densely populated Chicago area?



Observation: This side-by-side comparison of the collision type reinforces our understanding that parked vehicles and read-end collisions are common in densely populated urban areas like Chicago. Unincorporated areas see a majority of collisions with animals or fixed objects.

Is there a correlation between the number of vehicles involved in the crash and the total number of injuries and fatalities? Can the injuries and fatalities be predicted based on the number of vehicles involved in a crash?

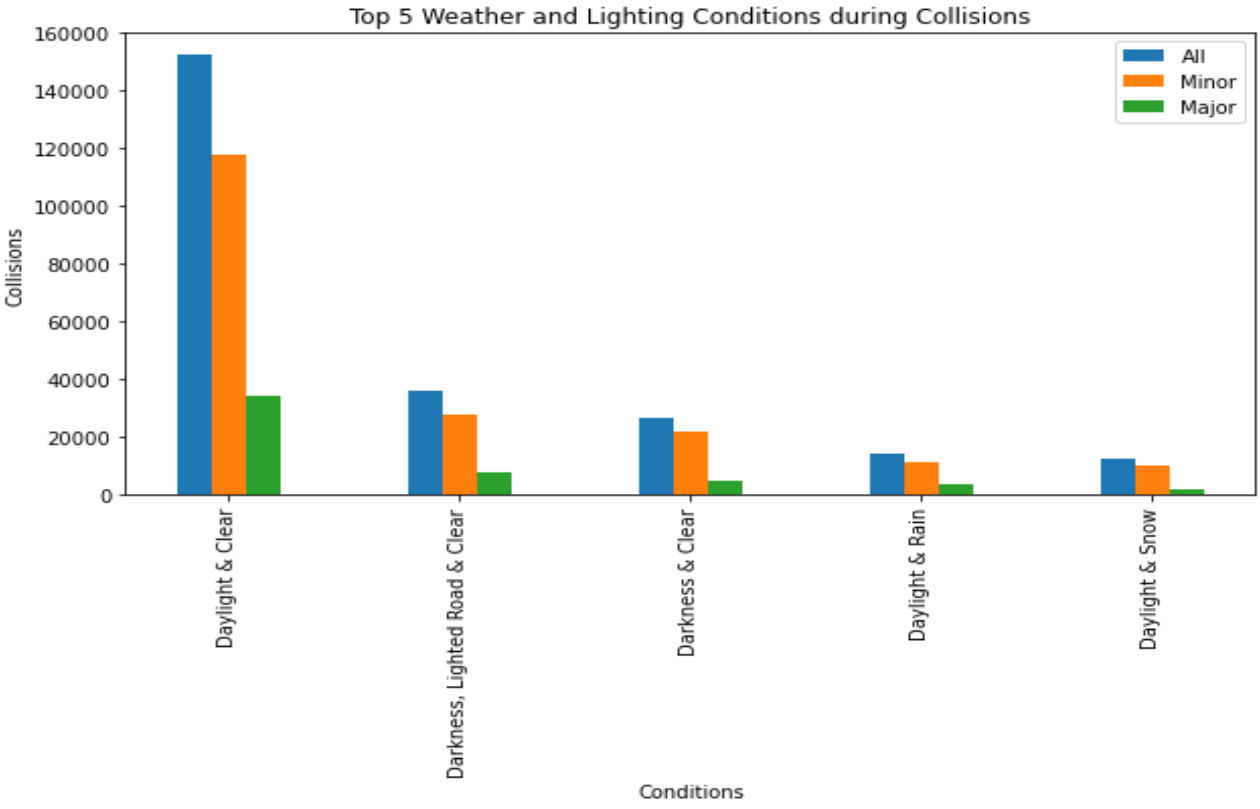
We analyzed the dataset to answer the above question by creating a regression plot between the number of vehicles and the count of injuries and fatalities. We considered only the reports which had at least 1 injury or a fatality due to the collision.



Observation: Based on the slope of the regression line, we anticipate that we will not be able to accurately predict the number of injuries or fatalities from the number of vehicles involved in the collisions.

Did any external factors like weather & lighting condition influence a higher number of collisions?

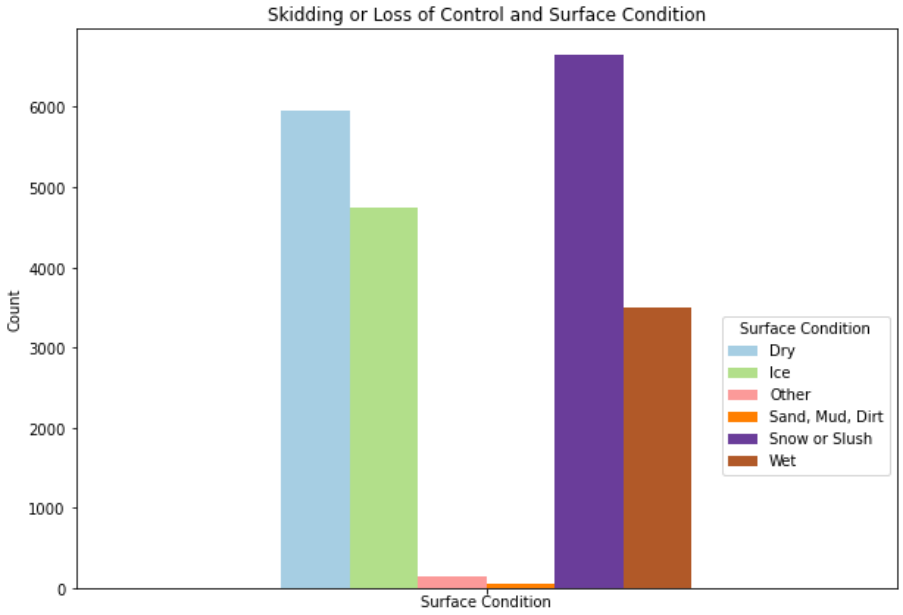
The collisions were split into major (collisions with injuries & Fatalities) and minor (collisions with only property damage) categories and analyzed across different weather and lighting condition combinations.



Observation: A majority of the collisions occurred during daylight and in clear weather conditions. Based on this observation, we think that weather and lighting conditions did not play a role in the majority of the collisions.

When a driver skidded or lost control of a vehicle, did the underlying surface condition play a role in such collisions?

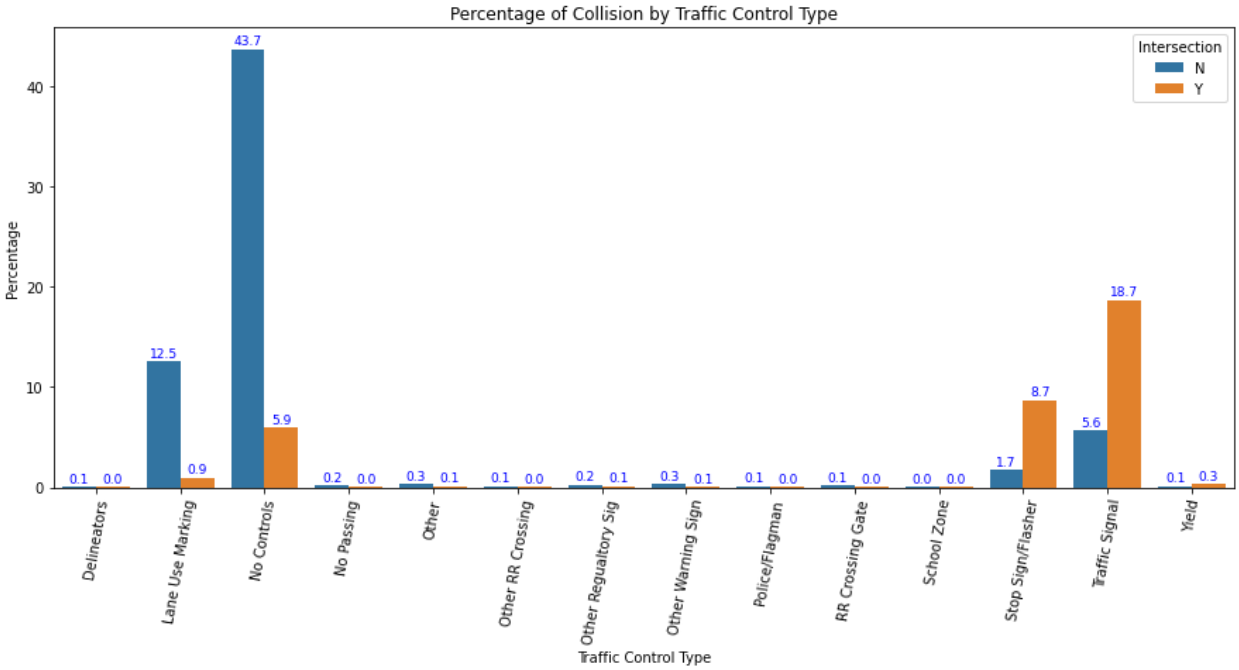
All the collisions that had the primary vehicle maneuver (VEH1_MANUV) as skidded or lost control were analyzed across surface conditions.



Observation: A majority of the drivers lost control or skidded in dry conditions after the highest condition of snow or slush. This may mean that driver errors played a much more important role than external surface conditions.

Did lack of traffic controls play a role in the collisions and whether there were a greater number of collisions at intersections with no traffic control?

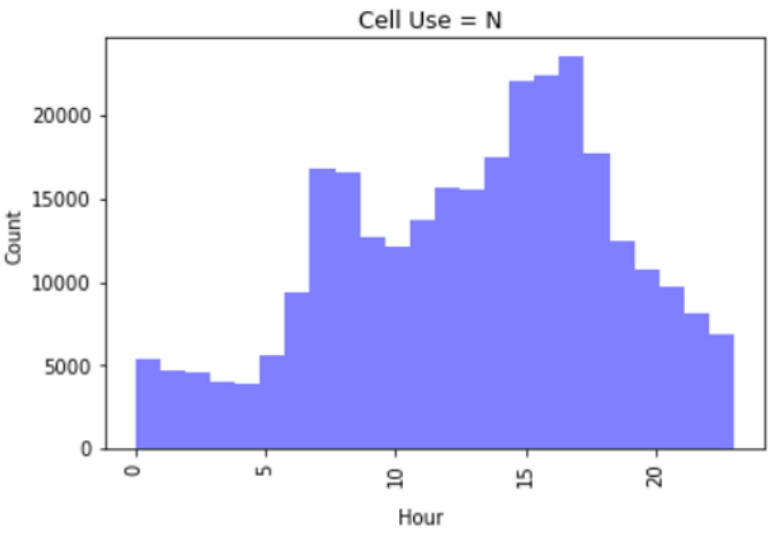
The dataset was analyzed to plot the percentage of collisions at different traffic controls and whether there was an intersection at the collision site.

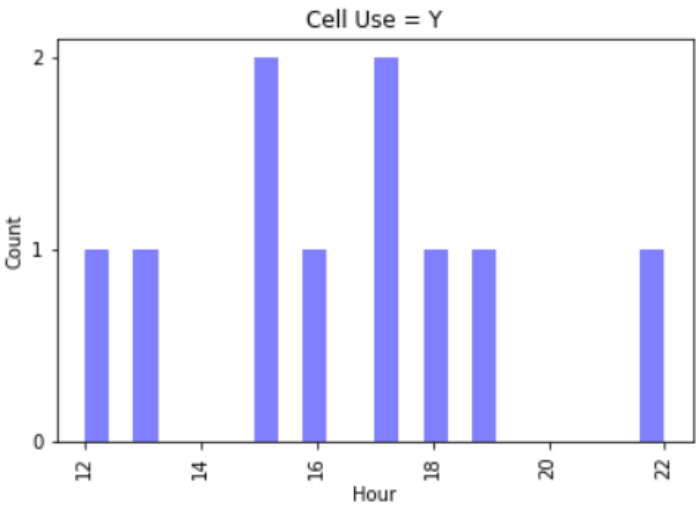


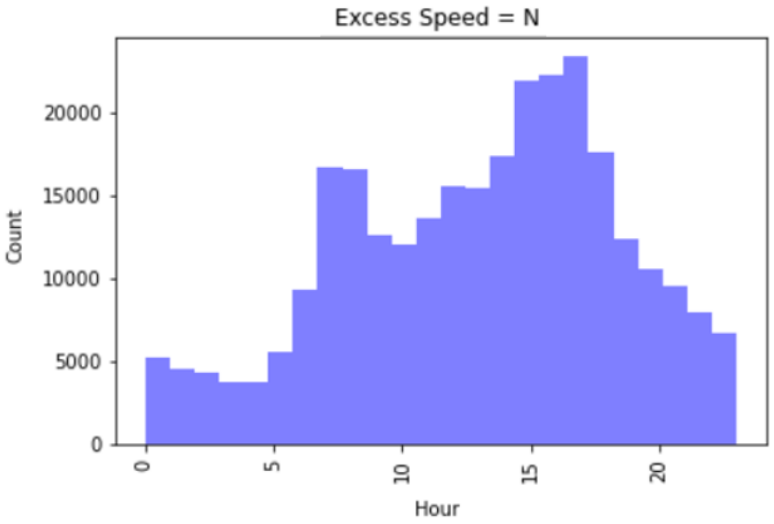
Observation: As expected, a majority (~50%) of the collisions happened at locations with no traffic control devices. An interesting point to note from the plot is that there were 3 times more collisions at intersections with a traffic signal (18.7%) than collisions at intersections with no controls (5.9%).

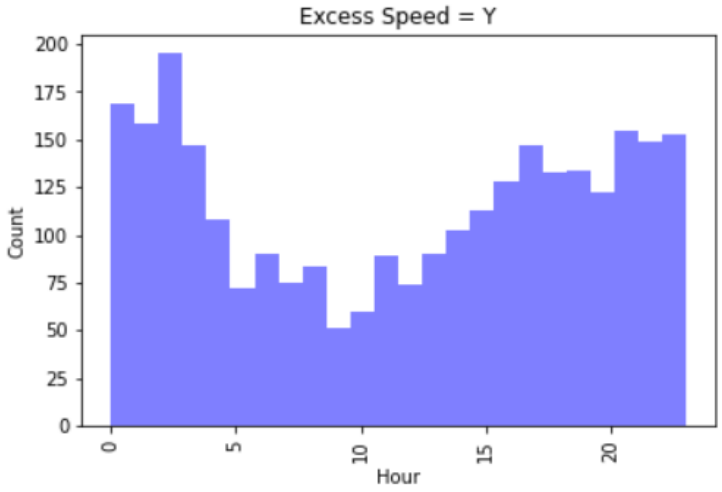
Were there any driver factors like using a cell phone or exceeding speed limits that correlated with a higher number of collisions?

The dataset was analyzed and plots were created to show the number of collisions with instances of cellphone usage and excessive speed across the different hours of the day.





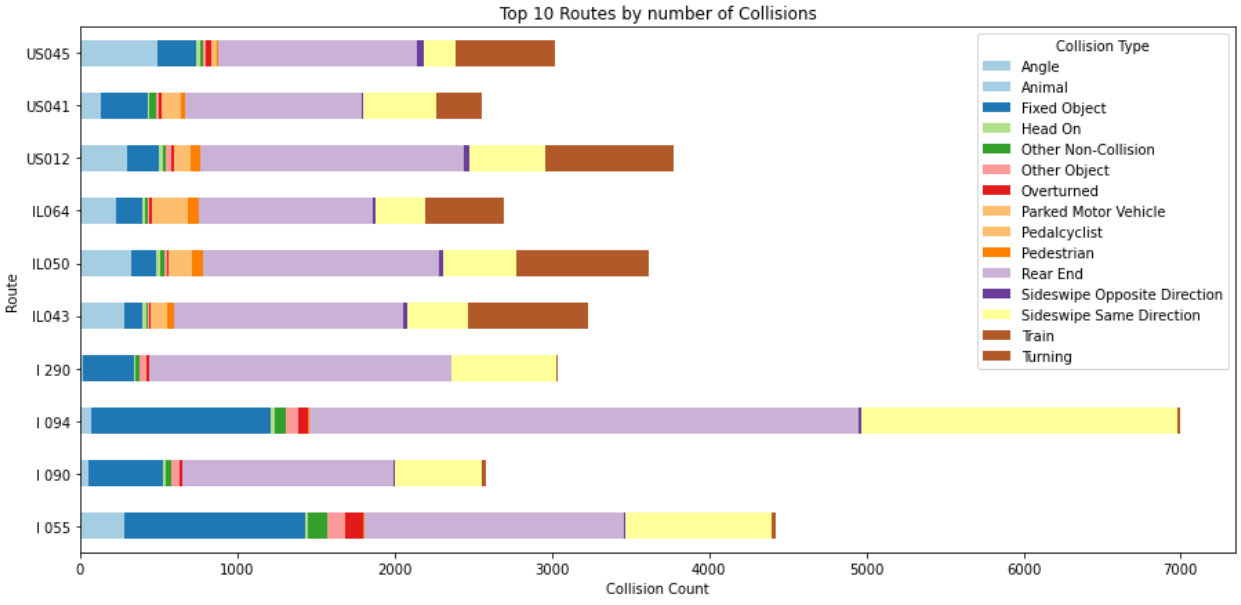




Observations: Both cell phone usage and excessive speed reports had a relatively lower number of collisions when compared to the total number of incidents without these driver factors. We think that the driver factors (cell phone usage & Excessive speed) may not have caused a high number of collisions based on this crash dataset.

Do certain routes (Interstates and Highways) have a trend in types of accidents and can we infer any specific details by narrowing down to the most common collisions?

A plot was created to show the different types of collisions on the top 10 highways & interstates by the count of collisions.



Observation: I094 had the most collisions on a highway/interstate but the least due to turning vehicles. US045, US041, US012, IL064, IL050, and IL043 had comparatively lower collisions but a higher number of turning vehicle collisions. These six highways may need monitoring in place to check what is causing a higher number of turning vehicle collisions on them.

Conclusion

This dataset provided a rich analysis of the collision statistics in Illinois and helped gain insights into the factors and the reasons that may be causing these collisions. Each of the above analyses and the corresponding plots provided answers to the questions that we initially intended to find answers to.

The analysis shows that the commute hours on working days of the week has a significant impact on the number of accidents. While most accidents happen over Dry roads, Clear weather, with sufficient lights and involved passenger vehicles the spike in the month of January may be attributed to the snow/slush additive effect. This observation was reinforced by the section that analyzed vehicle maneuvering.

It's also clear that the populated areas see more accidents. We see a clustering effect around cities and towns.

We could see that the presence of turns and intersections added to accidents. Additional monitoring may be implemented to reduce accidents. Intersections without traffic lights contributed to a vast majority of collisions at the intersections.

Somewhat counterintuitive but the weather and lighting conditions didn't play a huge role in increasing accidents.

Similarly, cell phone usage and/or speed limit exceeding didn't have a significant impact on the accidents numbers.

Analysis of collision type reinforces our understanding that parked vehicles and read-end collisions are common in densely populated urban areas like Chicago. Unincorporated areas see a majority of collisions with animals or fixed objects.

Appendix

Crash Report 2014 Dataset: Various columns of this data set are:

OBJECTID, geodb_oid, ROUTE, YEAR, MONTH, DAY, HOUR, DAY_O_WEEK, NUM_VEH, INJURIES, FATALITIES, COLL_TYPE, WEATHER, LIGHTING, SURF_COND, RD_DEFECT, RD_FEATURE, TRAF_CNTRL, COUNTY, TOWNSHIP, TS_ROUTE, MILE, CITY, DRIVER_1, VEH1_TYPE, VEH1_SPECL, VEH1_DIR, VEH1_MANUV, VEH1_EVNT1, VEH1_LOC1, VEH1_EVNT2, VEH1_LOC2, VEH1_EVNT3, VEH1_LOC3, DRIVER_2, VEH2_TYPE, VEH2_SPECL, VEH2_DIR, VEH2_MANUV, VEH2_EVNT1, VEH2_LOC1, VEH2_EVNT2, VEH2_LOC2, VEH2_EVNT3, VEH2_LOC3, DRIVER_3, VEH3_TYPE, VEH3_SPECL, VEH3_DIR, VEH3_MANUV, VEH3_EVNT1, VEH3_LOC1, VEH3_EVNT2, VEH3_LOC2, VEH3_EVNT3, VEH3_LOC3, DRIVER_4, VEH4_TYPE, VEH4_SPECL, VEH4_DIR, VEH4_MANUV, VEH4_EVNT1, VEH4_LOC1, VEH4_EVNT2, VEH4_LOC2, VEH4_EVNT3, VEH4_LOC3, DUP_CD, REC_TYPE, XCOORD, YCOORD, INTERSEC, SFE, AGENCY_NUM, RUNDATE, WorkZone, WorkZoneTy, WorkersPre, ExceedSpee, CellPhoneU

County Population Dataset
‘County’, ‘April 1, 1980 Census’, ‘April 1, 1990Census’, ‘April 1, 2000 Census’,, ‘April 1, 2010 Census’, ‘Percent Change 1980 to 1990’, ‘Percent Change 1990 to 2000, ‘Percent Change 2000 to 2010’

The study only uses the “County” and “April 1, 2010 Census” columns to approximate the county-level population count for the year of 2014 in Illinois.