W200 Project 2
Illinois Vehicle Collision Data Analysis

April 12th, 2021

Team 1
Kumar Narayanan
Viswanathan Thiagarajan
Estrella Ndrianasy

Dataset Summary & Questions

- Analysis of **CrashReport2014.csv** file
  - Data source: Kaggle (https://www.kaggle.com/qcarver/crashes-2014-csv)
  - Data has 292019 rows and 80 columns

    Key columns like Route, Number of Vehicles, Injuries, Collision Type, Weather conditions, City, Collision type, Cell phone usage etc.

- What are the key questions that we set out to answer?
  - What months saw the most number of accidents?
  - Are there variations in the accidents number by days of the week?
  - How's the spread of the accident in the state? What is the geo heatmap?
  - What factors influenced accidents - road condition, weather, brightness level, number of vehicles, presence of intersections?
  - Is there a trend in the type of collisions on certain routes?
  - Is there a larger proportion of collisions due to cell phone usage or excess speed?
  - Do the number of injuries and fatalities increase with the number of vehicles that were involved in the collision?

Analysis Steps

- Cursory analysis of the file
  - Are there columns that needed clean up?
  - Could we ignore some columns and not impact the trend and analytics?

- Important changes made
  - Date split into 3 columns - made sense to combine it into a single date column
  - There were unexpected values in ‘city’ name column; needed to be handled.
    - Fortunately, it was statistically insignificant
  - XY coordinates lacked reference; derived lat-long through other means
  - Had to account for “Unincorporated” areas by factoring the associated county
  - Added a column to indicate the total number of critical collisions with injuries or fatalities
  - Cleaned up the route numbers which had a “*” at the end in some of the values

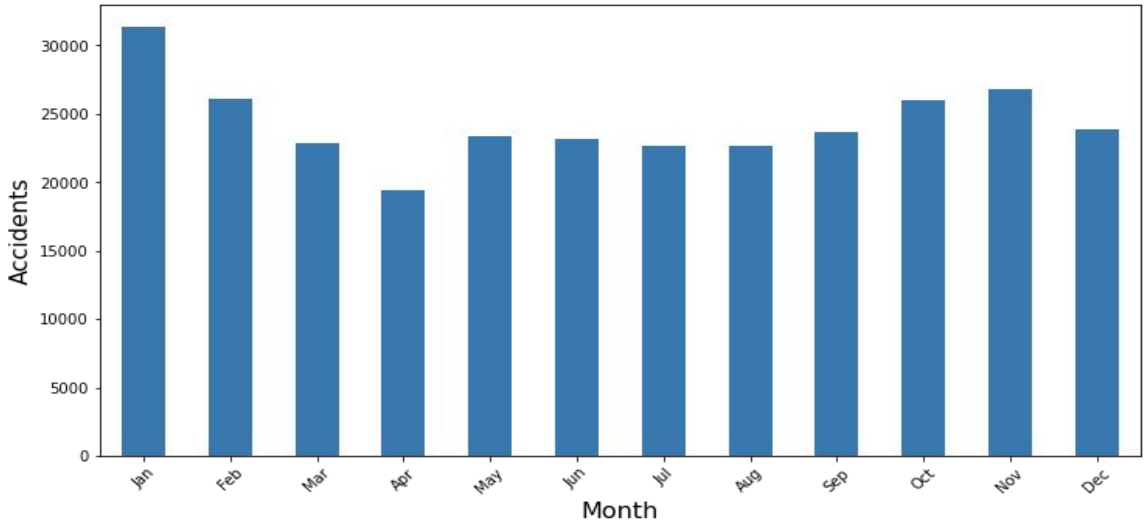
Assumptions

- The special value NaN
  - Chose to ignore this; a few city names (< 20) and weather column (< 1000)
  - The XY coordinates unusable due to lack of reference
  - Some trends considered only a month’s data; else can take hours to compute
    - Future work to run the logic in a loop for each month
  - Around 20 columns were ignored in the study for which there was no reference like objectid, geodb_oid, SFE, rundate, agency , dup_cd etc.

Monthly Accident Summary Chart

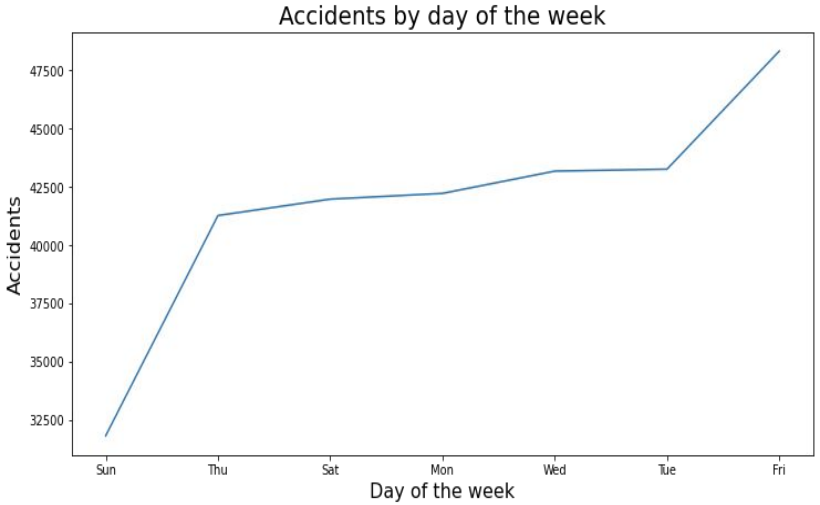
Acccidents by Month

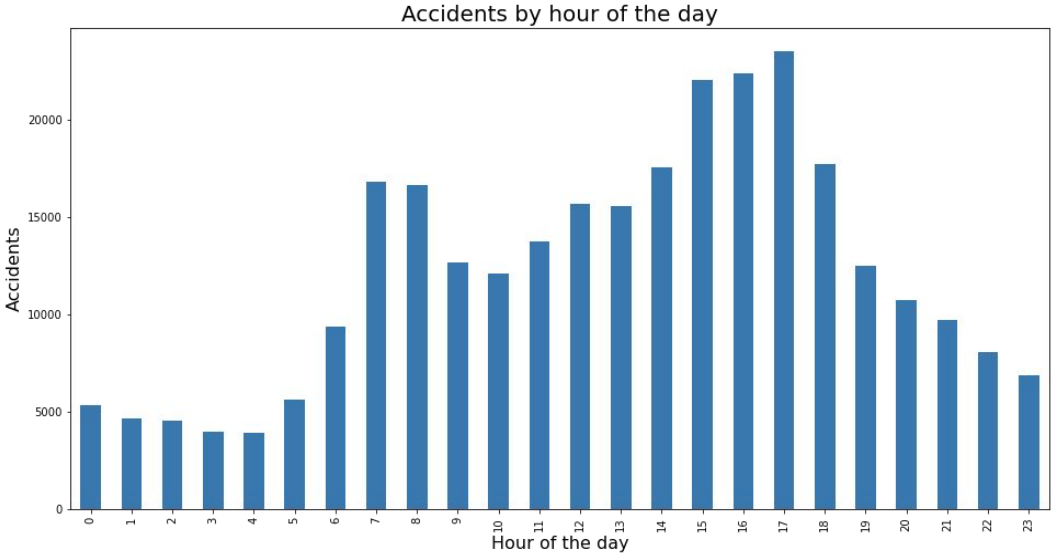
Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
31406	26068	22910	19413	23394	23215	22615	22633	23673	25981	26834	23877



First level view of all the accidents by month - does it show seasonal correlation?

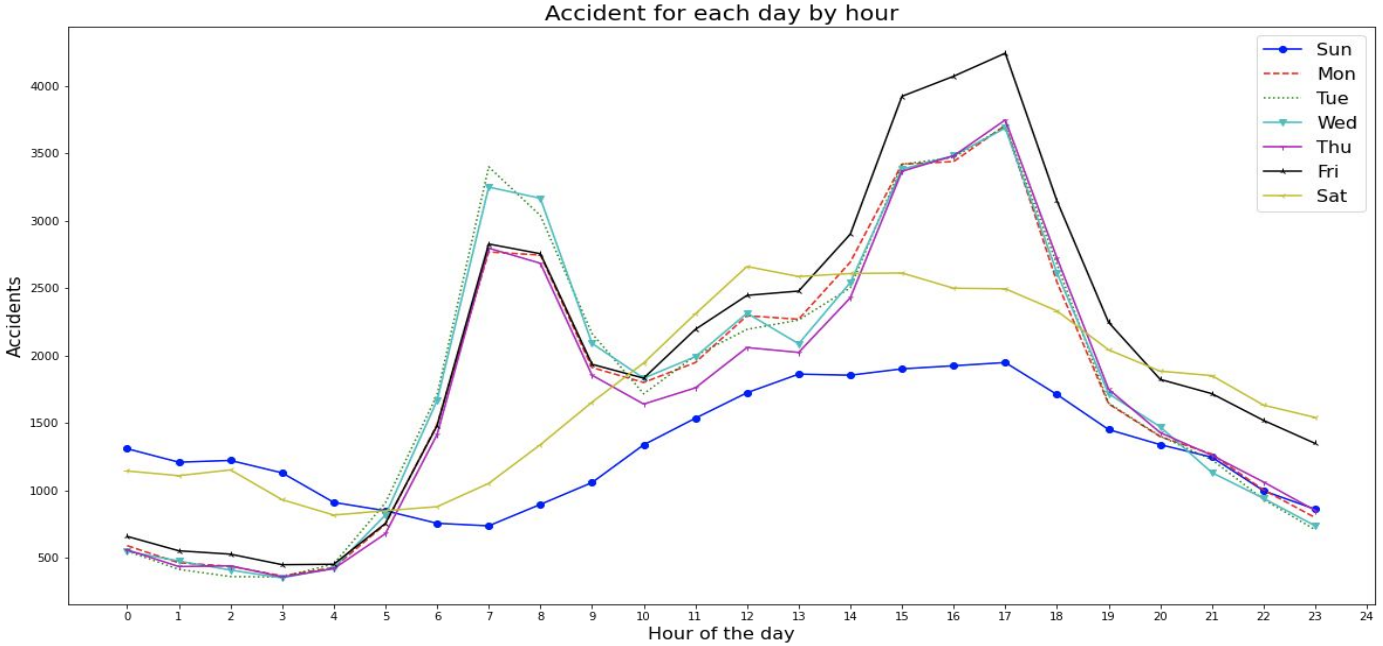
Days of week/Hours of the day





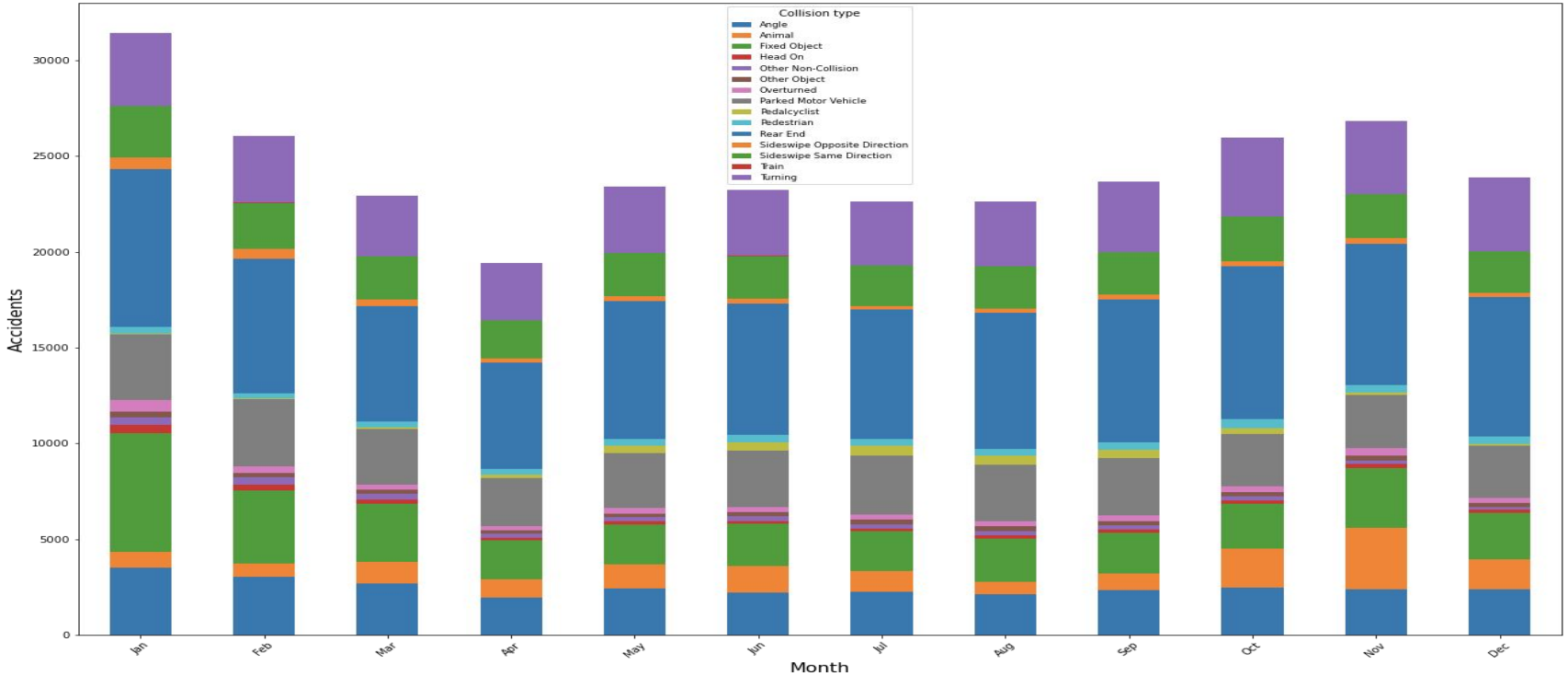
What is the correlation to the day of the week, and hours of the day? Sun. lowest, Fri. highest, quite steady in-between

Commute Hour Impact



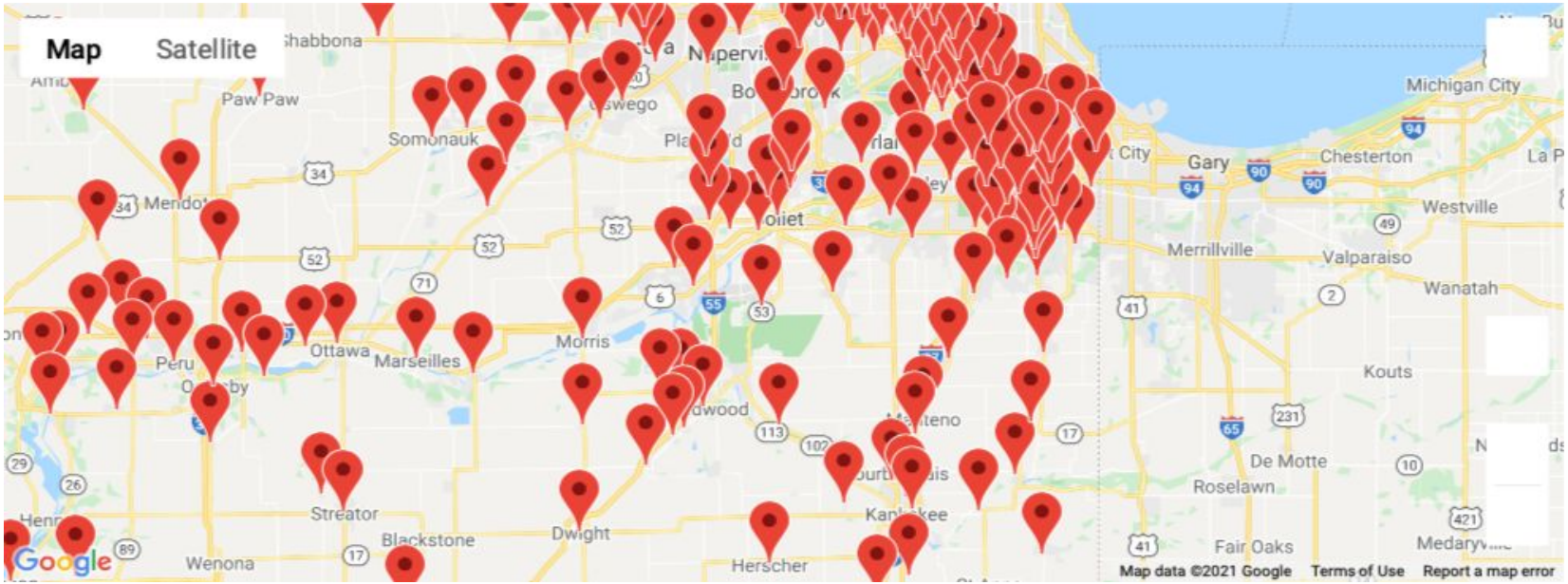
Further drill down - impact of commute hour traffic. Side bar: Is that why insurance companies ask for commute data?

Collision Type Stacked by Month



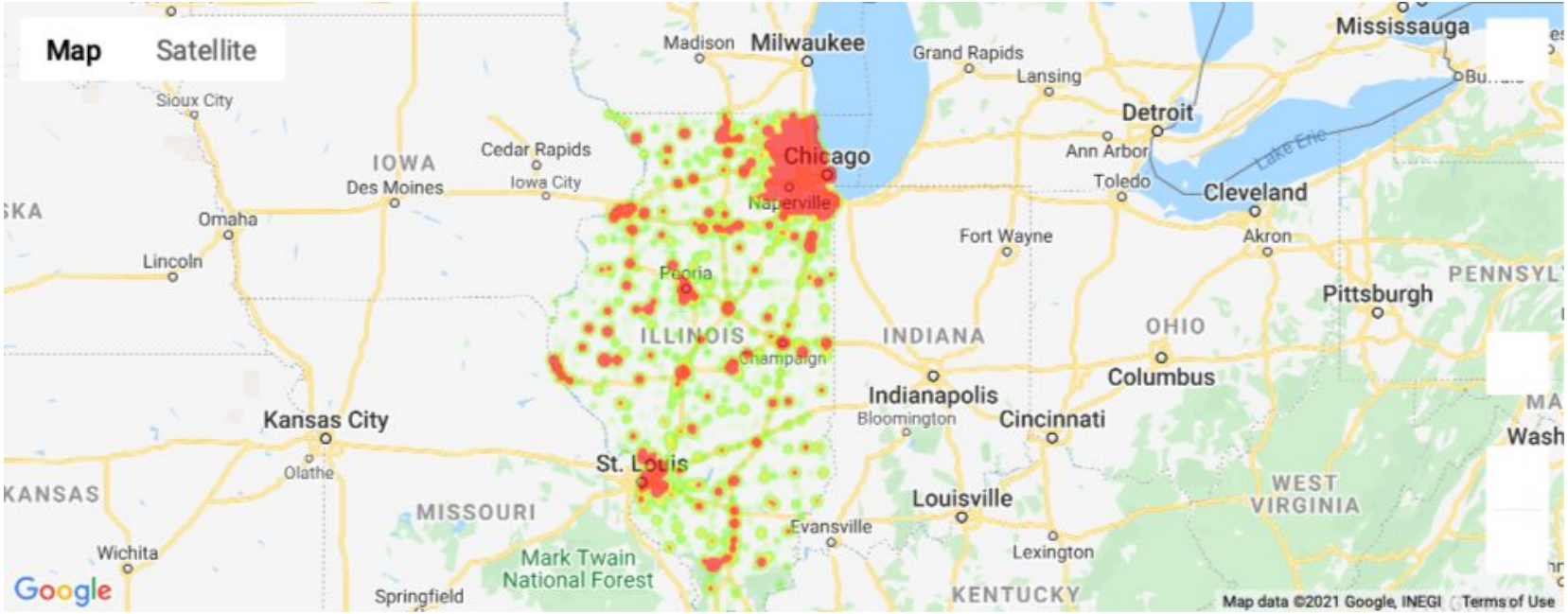
Rear-end collision is the most common type of collision

Geo Spread



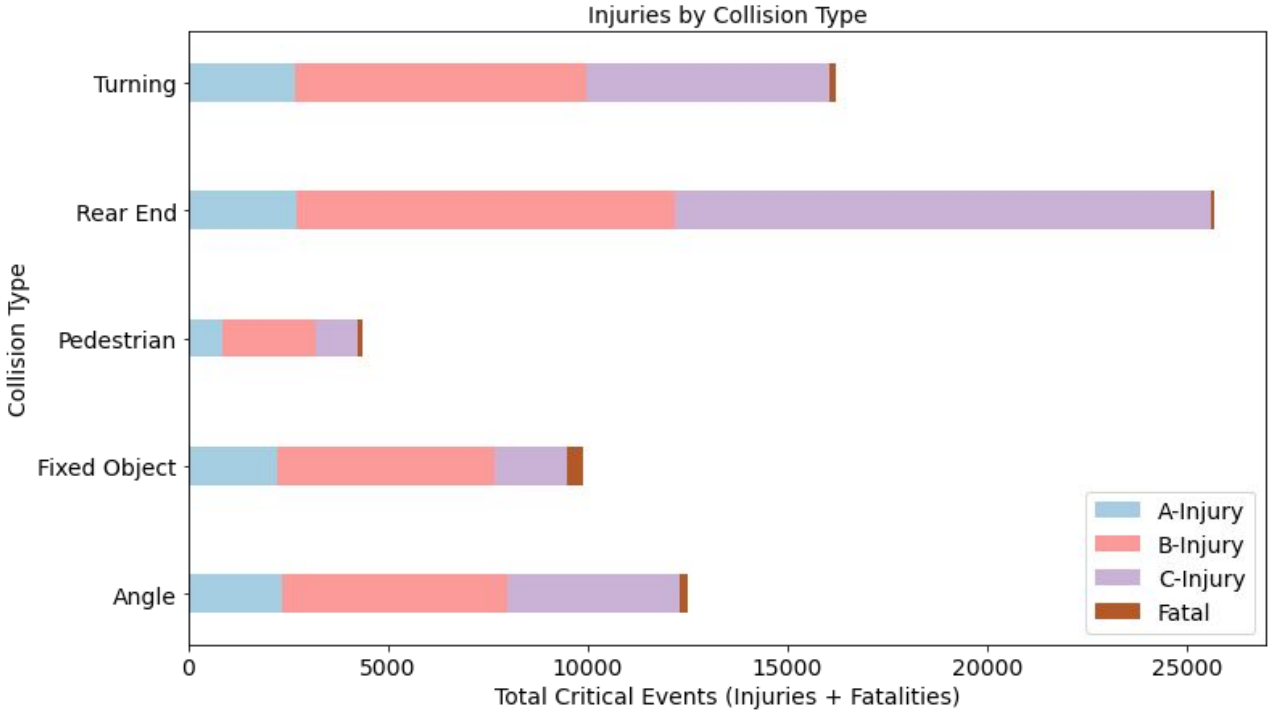
How did the accident spread out in the state of IL? See the cluster around Chicago and major towns

Heat Map



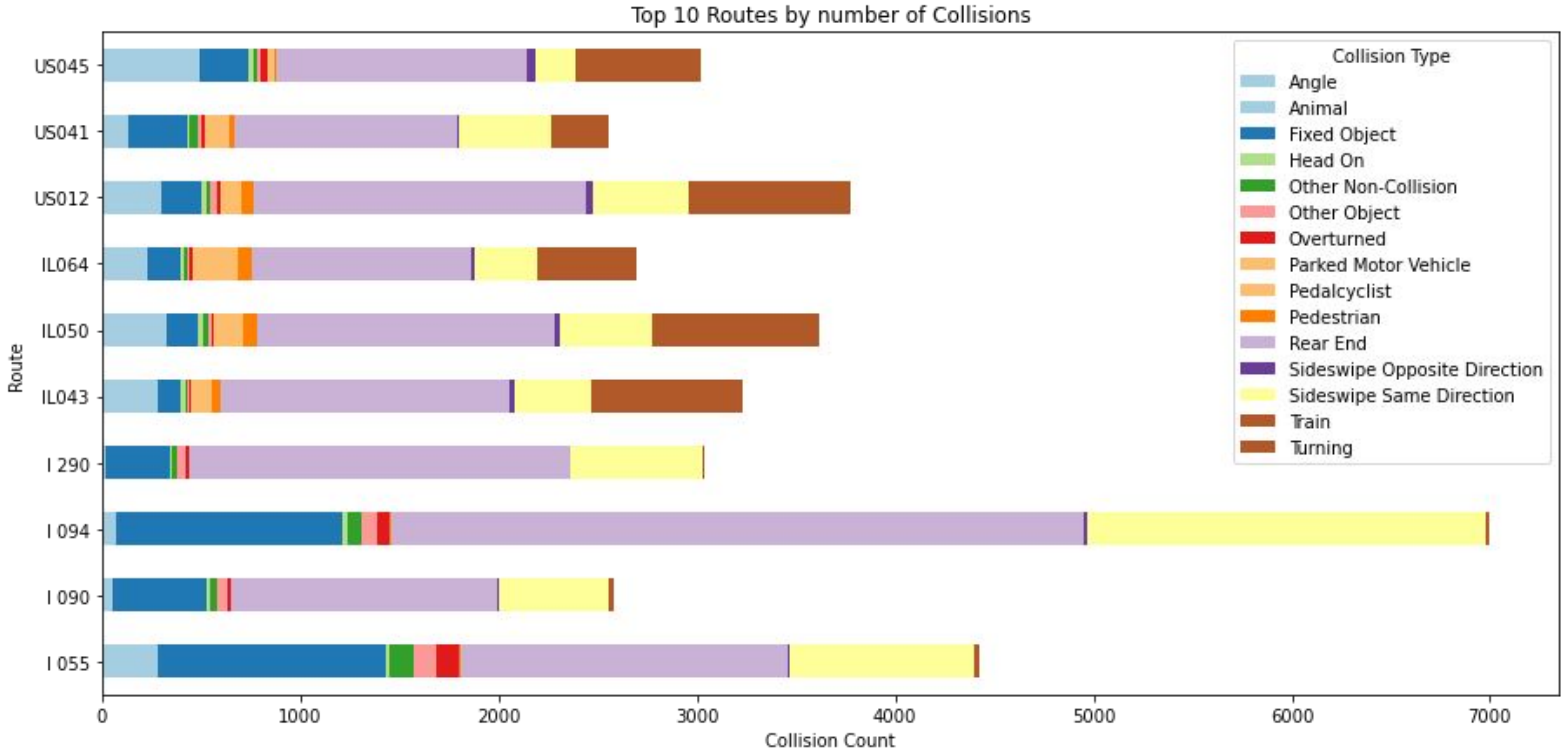
Heat map is easier to see where the accidents are? The previous map gives additional details

Collision Type and Injuries



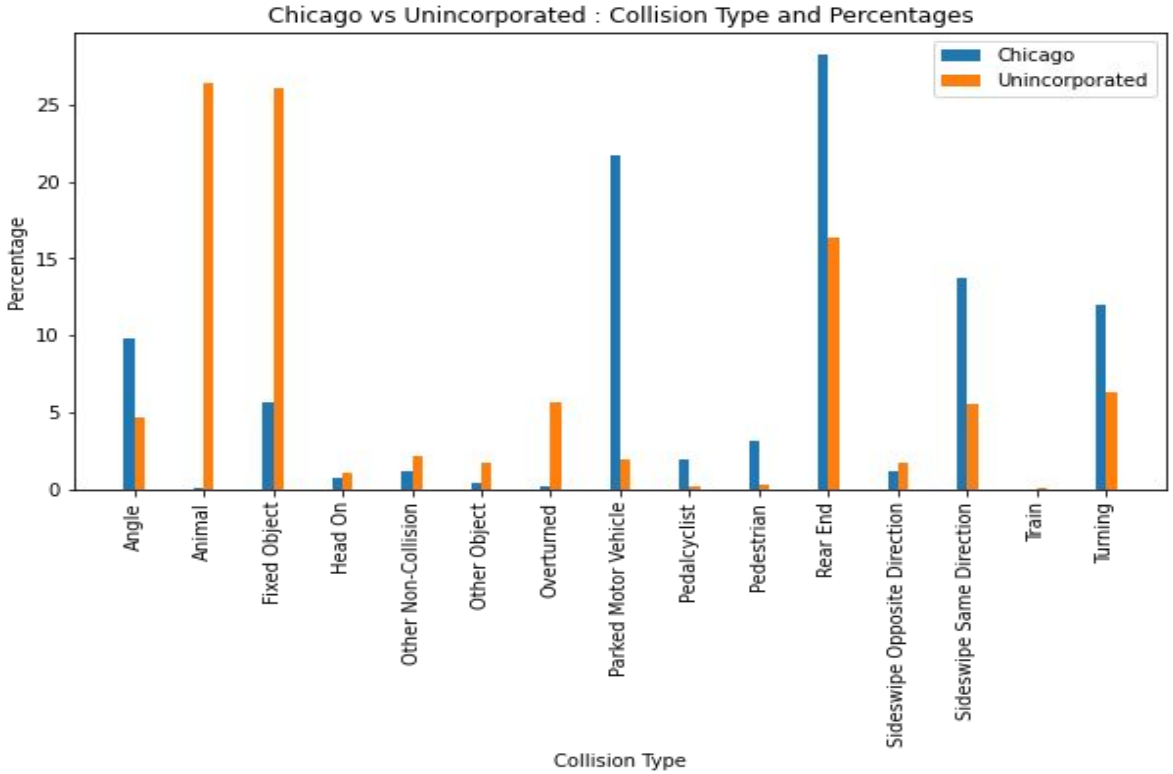
Most collisions were rear end collisions but it had the lower fatalities. Fixed object collisions were having higher fatalities even with less number of collisions

Collision type in the top 10 routes by count

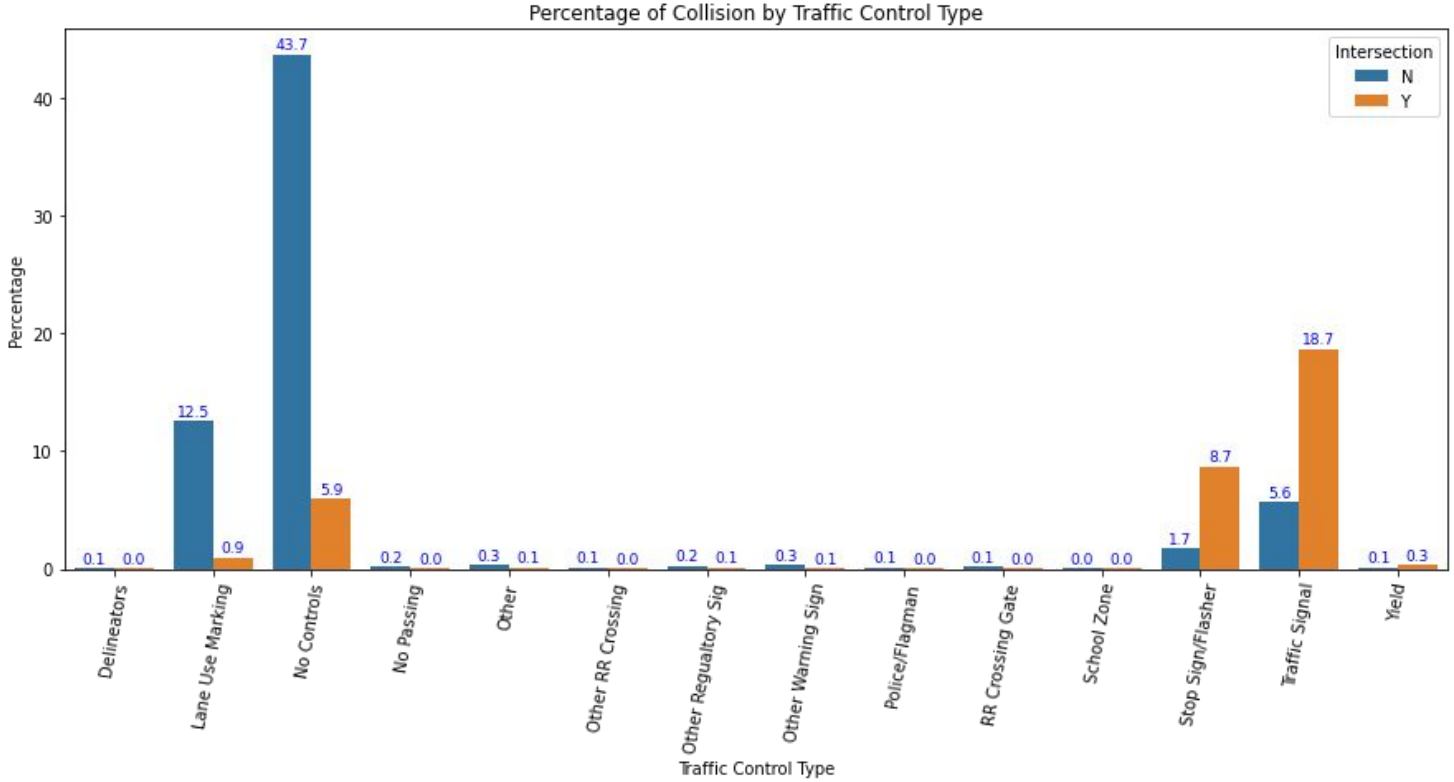


I094 had the most collisions but least due to turning vehicles. US045, US041, US012, IL064, IL050, and IL043 had comparatively lower collisions but higher number of turning vehicle collisions

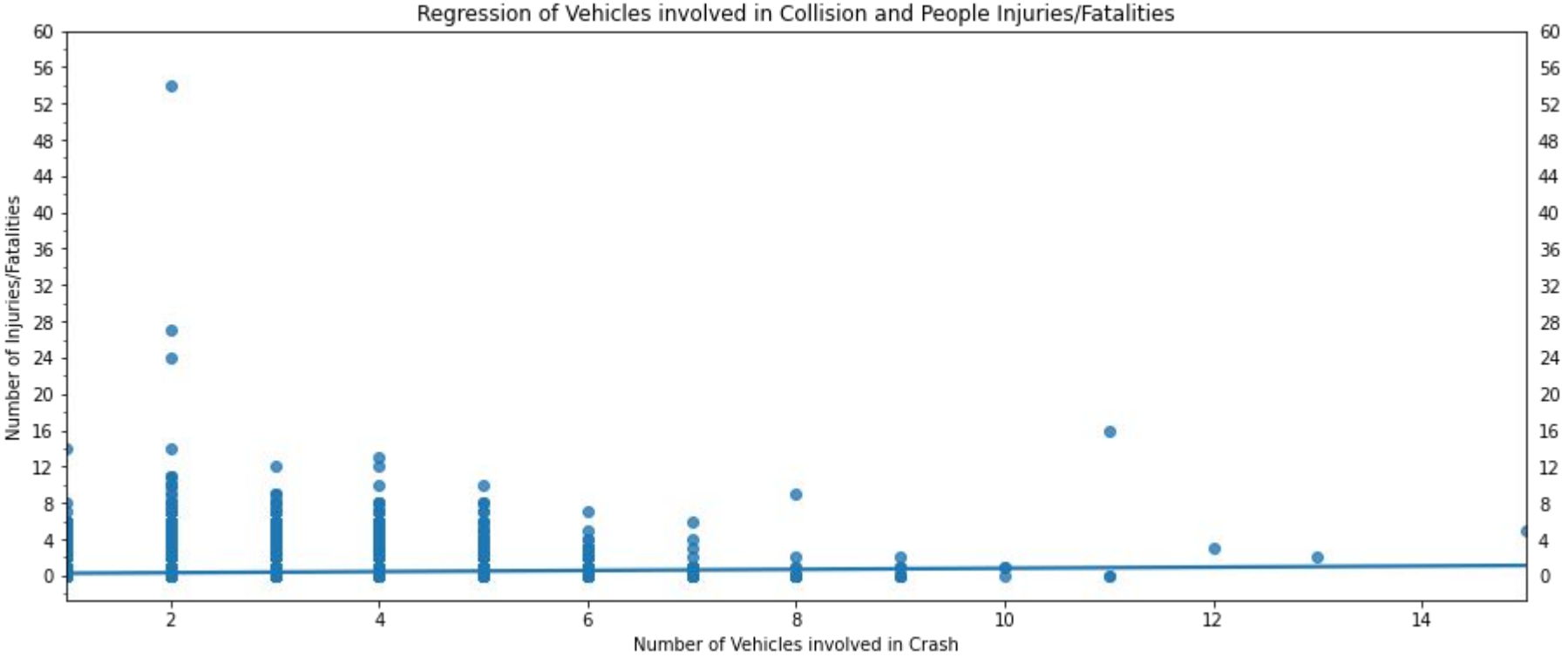
Collision types in Urban vs Unincorporated areas



Percentage of Collisions at different Traffic controls

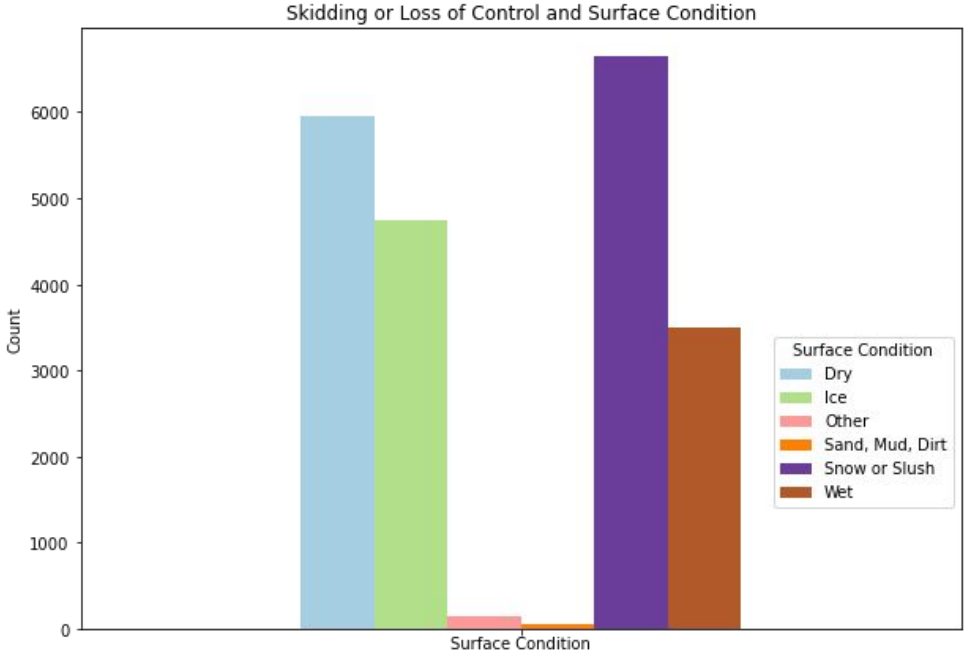


Regression of # Vehicles in Collision and # Injuries



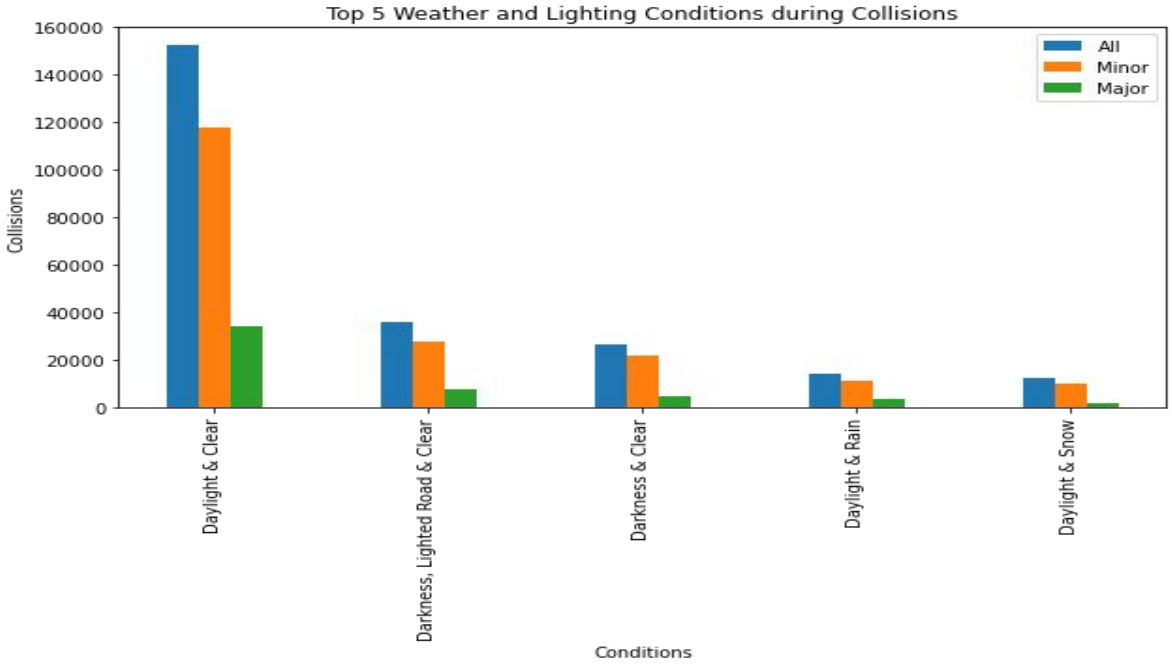
The number of injuries or fatalities did not increase proportionally as the number of vehicles involved in the crash increased

Surface conditions when the vehicle driver lost control



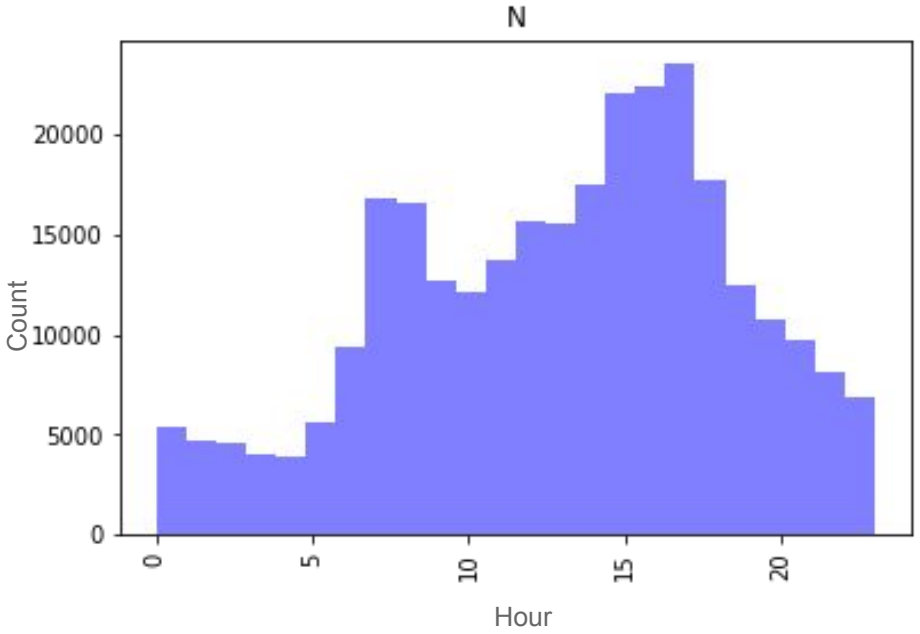
Interesting to note that when a driver skidded or lost control the surface condition of "Dry" was in the top two conditions

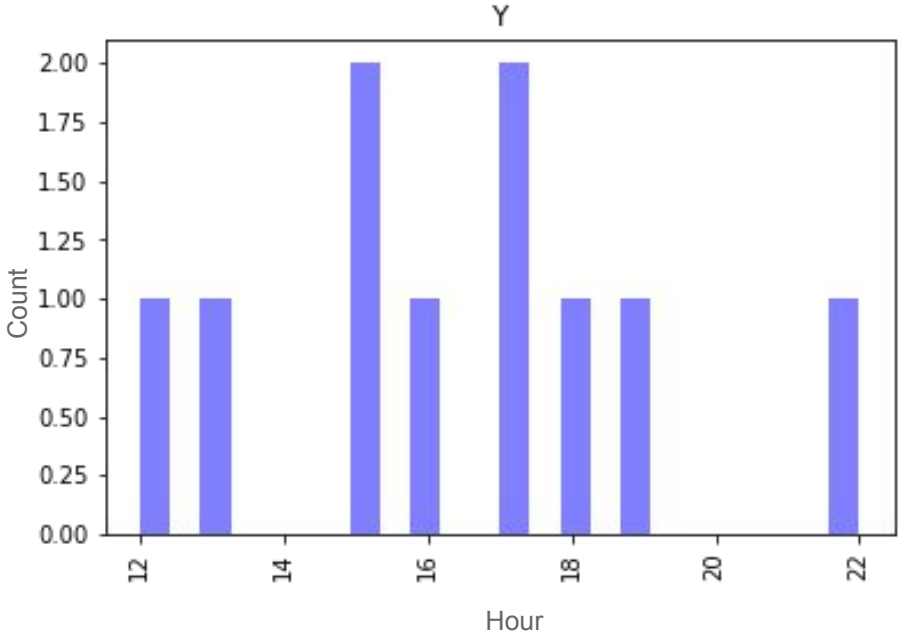
Weather Conditions during Collisions



Most of the collisions (both minor and major) happened during daylight and clear weather conditions

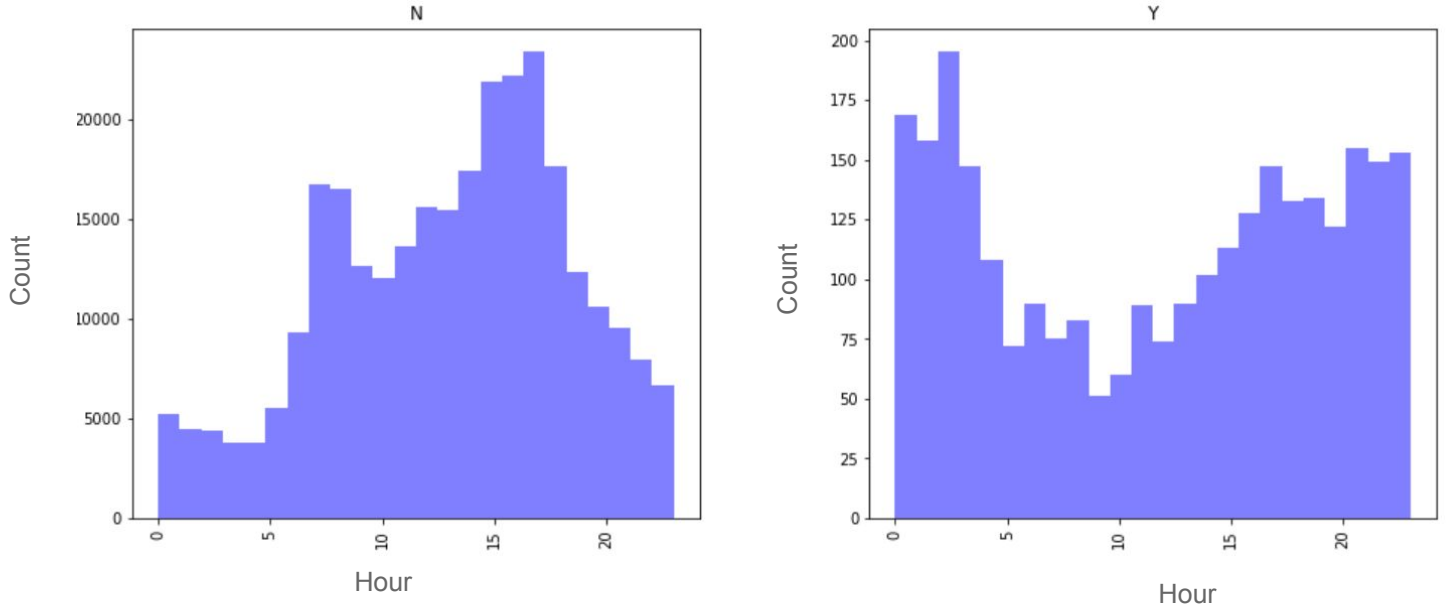
Collisions by hour of the day - Cellphone usage (Yes/No)





Most collisions occurred when there was no cell phone usage. Collisions with cell phone usage were rare as there were only 10 total collisions out of 292K.

Excess speed in collisions (Yes/No)



The proportion of collisions due to exceeding speed limit is relatively lower when compared to collisions that happened within posted speed limits.

Thank you

References

- **CrashReport2014.csv** file having collisions that were reported in Illinois in the year 2014
  - Data source: Kaggle (https://www.kaggle.com/qcarver/crashes-2014-csv)