

Background

I am going to explore a large data set of traffic crashes to learn about what factors are connected with injuries. I will use data from the city of Chicago's open data portal. (This activity is derived from a blog post by Julia Silge). Along with this I am going to do some mutating of the original data set to answer some questions at the bottom.

This data set has a lot of different variables that are unnecessary for the Exploratory Data Analysis that we are going to do today. Therefore we are going to manipulate it into a few smaller data sets so it is easier to work through and get graphs out of. One important aspect of the data set that we don't have but want is a variable called `injuries` which indicates if the crash involved injuries or not. So we are going to first create that variable. We are then going to create an unknown category for missing `report_types` and decide any other variables to keep in our new data set.

```
## # A tibble: 228,423 x 8
##   injuries report~1 latit~2 longi~3 crash_date           crash~4 weath~5 crash~6
##   <fct>     <chr>    <dbl>    <dbl> <dttm>          <int> <chr>    <ord>
## 1 injuries ON SCENE    41.8   -87.6 2022-02-28 10:21:00      2 CLEAR Monday
## 2 injuries ON SCENE    41.9   -87.7 2022-04-27 09:30:00      4 CLEAR Wednes-
## 3 none      ON SCENE    41.9   -87.7 2023-03-05 01:13:00      1 CLEAR Sunday
## 4 none      ON SCENE    41.8   -87.7 2022-11-30 08:50:00      4 CLEAR Wednes-
## 5 none      NOT ON ~    41.8   -87.7 2022-03-04 10:30:00      6 CLEAR Friday
## 6 none      ON SCENE    41.7   -87.6 2022-09-20 15:47:00      3 CLEAR Tuesday
## 7 none      ON SCENE    41.7   -87.7 2023-07-03 21:09:00      2 CLEAR Monday
## 8 none      ON SCENE    41.7   -87.6 2022-05-10 21:18:00      3 CLEAR Tuesday
## 9 injuries ON SCENE    41.9   -87.6 2023-02-22 15:43:00      4 RAIN  Wednes-
## 10 injuries ON SCENE    41.9   -87.6 2022-07-21 13:26:00      5 CLEAR Thursday
## # ... with 228,413 more rows, and abbreviated variable names 1: report_type,
## #   2: latitude, 3: longitude, 4: crash_day_of_week, 5: weather_condition,
## #   6: crash_day_name

## `summarise()` has grouped output by 'week_day'. You can override using the
## `.` argument.

## # A tibble: 14 x 3
##   week_day  injuries number_of_crahses
##   <ord>     <fct>        <int>
## 1 Sunday    injuries       826
## 2 Sunday    none          4684
## 3 Monday    injuries       777
## 4 Monday    none          5011
## 5 Tuesday   injuries       850
## 6 Tuesday   none          5083
## 7 Wednesday injuries       799
## 8 Wednesday none          4741
## 9 Thursday  injuries       843
## 10 Thursday none          5111
## 11 Friday   injuries       925
## 12 Friday   none          6139
## 13 Saturday injuries       928
## 14 Saturday none          5658
```

Exploratory Data Analysis

Now that we have taken away some of the clutter that was in our original data set, by making new ones, we are going to answer a few questions.

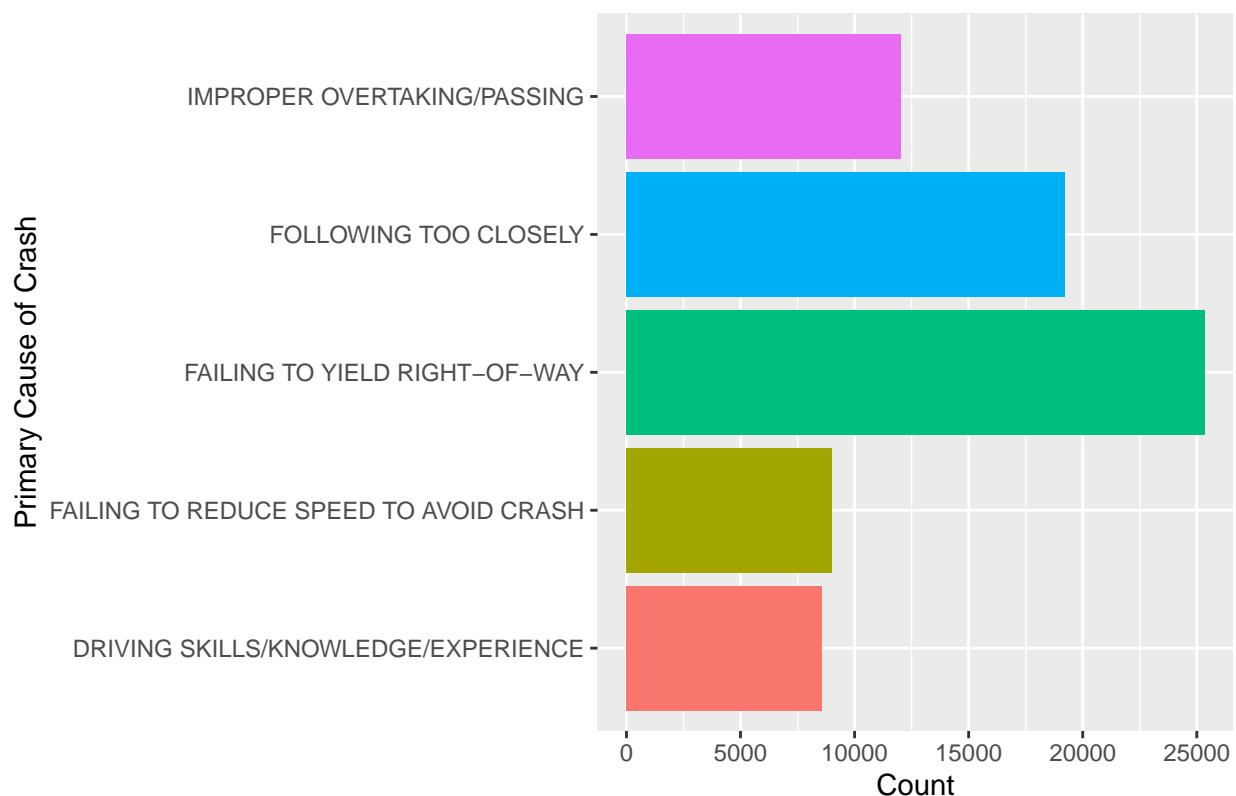
1. Take a look at crashes by latitude and longitude, colored by injuries. What do you notice?



One thing I notice is that the outline that forms looks like the city limits of Chicago. Along with this there doesn't seem to be much else that we can take away from this graph. It is a little jumbled and doesn't tell a clear story.

2. What are the most common contributing factors to a crash?

Top Five Primary Crash Contributory Causes

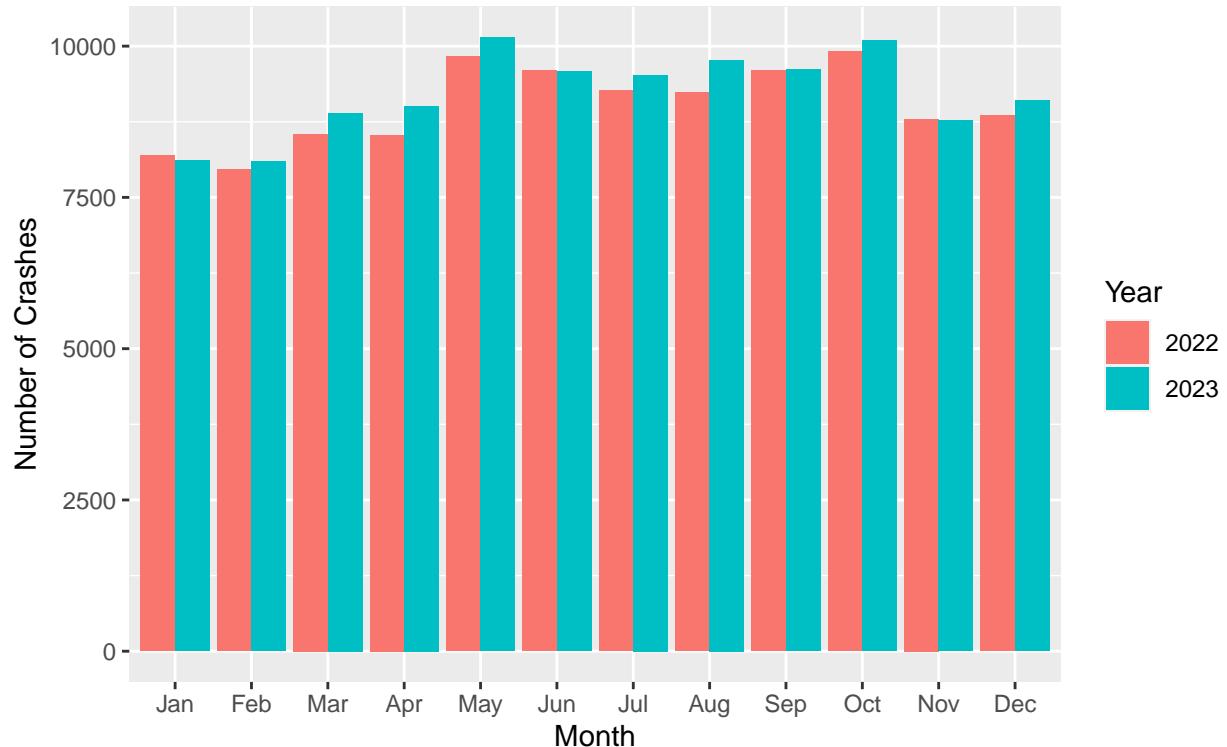


Here we are able to extract the top five primary factors for a car crash. These are; 1) Failing to yield right of way, 2) Following too closely, 3) Improper overtaking/passing, 4) Failing to reduce speed to avoid crash, 5) driving skills/knowledge/experience.

3. How do crashes vary month by month? Compare crashes by month in 2022 to 2023.

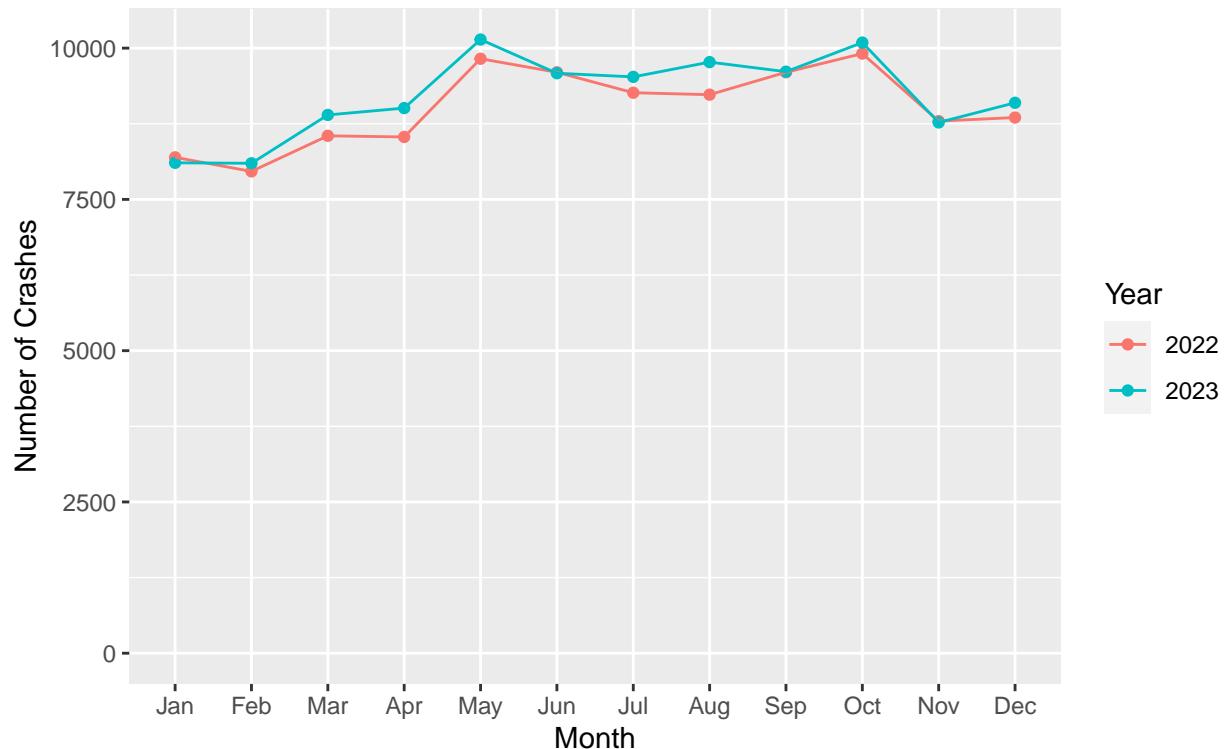
Crashes by Month

2022 and 2023



Crashes by Month

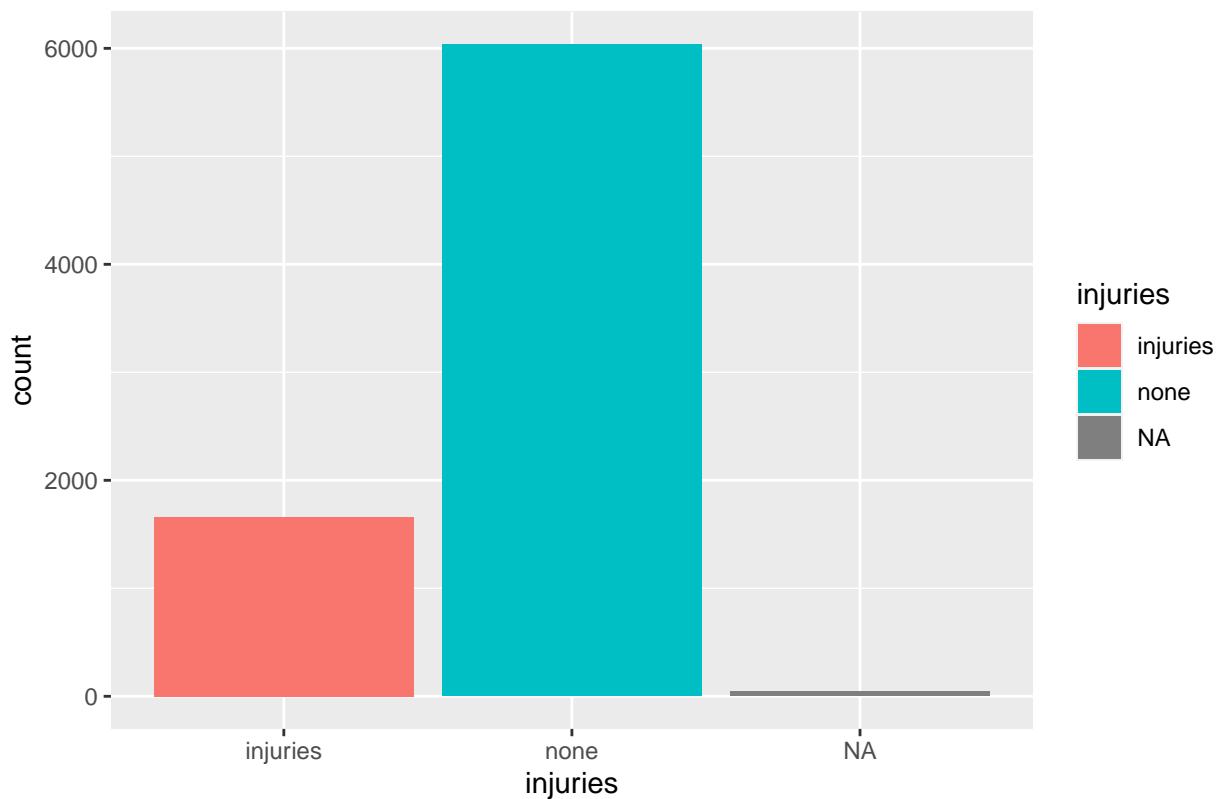
2022 and 2023



When looking at all the data, both May and October are the two months that have the most crashes. As you can see by looking at these graphs, there seems to be a very similar distribution of crashes throughout the months.

4. Are crashes more likely to cause injuries when it is rainy and dark? Use the variables `weather_condition` and `lighting_condition` to explore.

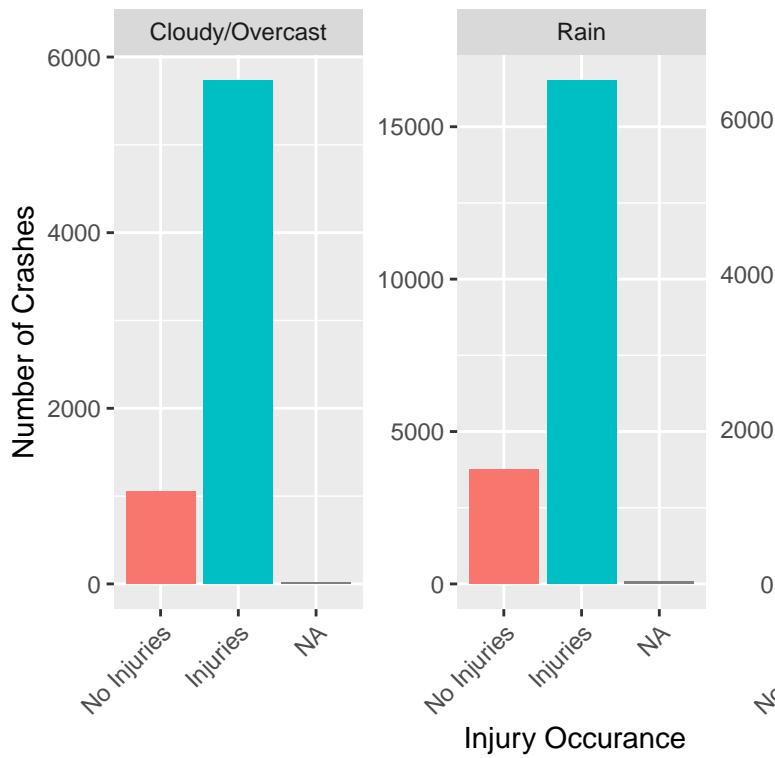
Crashes in Rain and Darkeness by Injury Status



When looking at injuries that occurred under the conditions when it was rainy and dark, with or without lighted roads, injuries do not occur more often than not. As you can see in the graph above, of the many crashes that occurred under these conditions injuries happened in roughly 1 out of every 5 crashes.

5. Choose a question you want to explore, and create an appropriate visual.

Injuries by Condition Cloudy, Rain, & Snow



What conditions cause the highest percentage of injuries?

By looking at these even though we can see the y-axis are set to different scales, the bars that depict when there were no injuries are all at a equal height. Therefore we can then look at the proportions of each graph to determine which condition has the highest probability to have an injury in an accident. Out of these three conditions we can see that Rain has the highest probability, followed by Cloudy or Overcast and then lastly, Snow.