The slower you drive, the more likely you are to get in accident in Chicago

By: Mehul Khajuria

(2nd Trend) Alt title 1: **You are more likely to get in a hit-and-run in Chicago than a regular crash**
(3rd Trend) Alt title 2: Your chances of getting into an accident in Chicago is at 3 p.m.
CC: https://www.tripadvisor.com/Attractions-g35805-Activities-Chicago_Illinois.html

---

Introduction

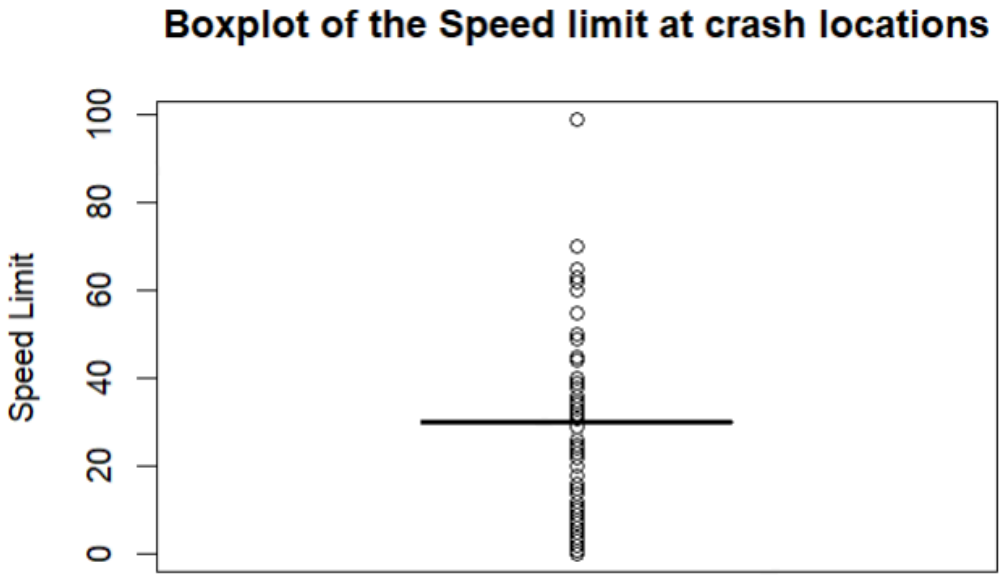
When I first learned to drive, my Dad would always make sure I took it as slow as possible. If the speed limit was 30 then I needed to go 20 and if I gained an extra mile on that speedometer I would be in trouble. It ended up becoming a habit till this day where I'd regularly check the speedometer if I saw a speed limit sign. All of this goes without saying that my Dad would blow past that speed limit easily. But all these memories came to mind when I was looking at a particular data set from Kraggle. It's called "Chicago Traffic Crashes" and it comes from the Chicago Police Department, if you'd like to check it out then you can find it here: **https://bit.ly/42yj9Og**. It's filled with a lot of interesting information and all categorized with 48 columns! Just some background, each row is a recorded crash and subsequent rows provide information about the conditions about and around it like weather or location. But it wasn't exactly the crashing part that reminded me of my first months driving, although there were some close calls, it was something interesting with the speed limits.

First Trend

When I checked the summary of the data, I saw something weird in the "POSTED_SPEED_LIMIT" column where the 1st Quartile, 3rd Quartile, and median were all the same number: 30.0. Below is the summary of this column I got from R.

```
POSTED_SPEED_LIMIT
Min.   : 0.0
1st Qu.:30.0
Median :30.0
Mean   :28.4
3rd Qu.:30.0
Max.   :99.0
```

It would be easy to write of this oddity in the data if the recorded speed limit got cut off at 30 but it spans all the way from 0 to 99.There are two main mysteries I want to focus on, why all these crashes were occurring at such low speed limits and what was with the triple 30. But to show you how interesting this is all is, take a look at the distribution of data on a box plot.

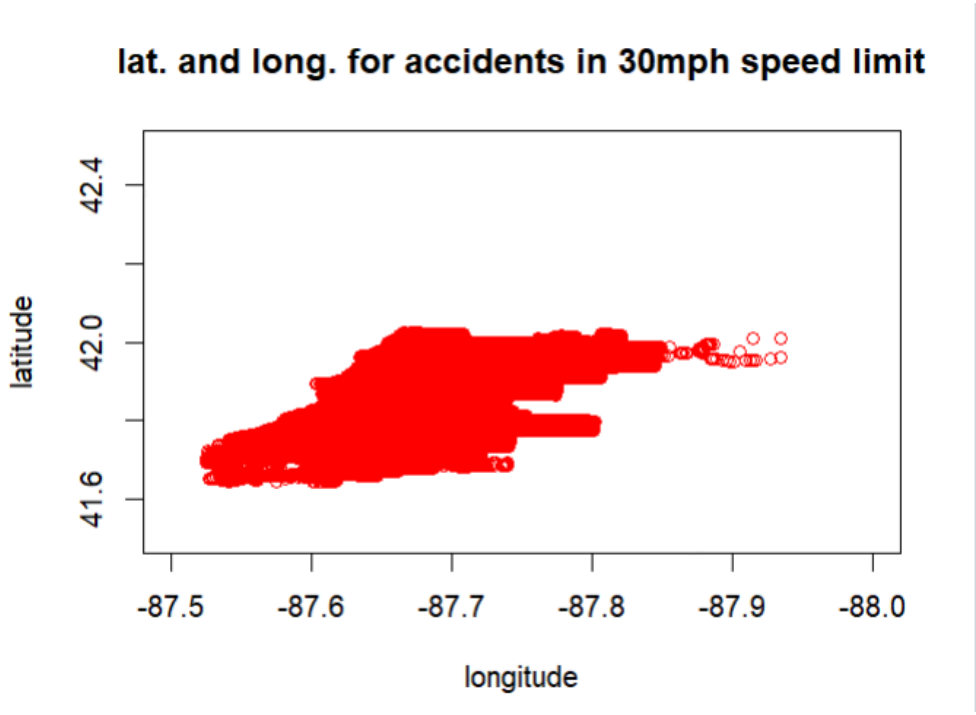


The box is evidently gone, it's just a line. After inspecting it further, it seems that the lower the speed limit is, the more likely you are to get in a crash in Chicago. That statement seems like the makings of a good click bait title (might be the reason why you clicked on this article) but I'd like to do more research and possibly provide some reasoning as to why this is happening.

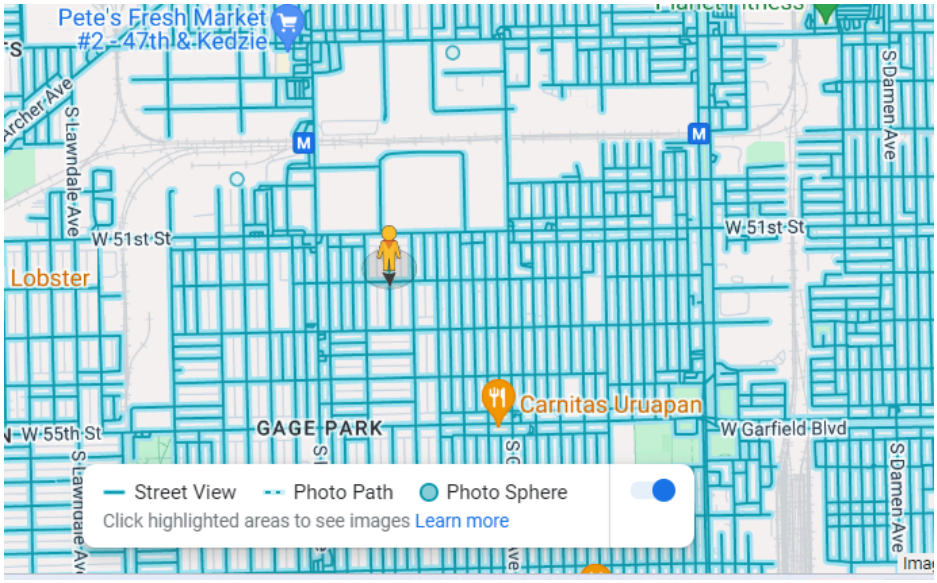
Going back to my anecdote of my first months driving a car, I spent some weeks with a driving school. In some of these classes, we would go into neighborhoods that had winding roads and narrow space to drive through. I had trouble navigating through these spaces and it became evident that speed isn't the only factor in getting into a car crash. I think that location could be a major culprit behind these accidents at these low speed limits, like what if the roadways have

many turns, are too narrow, or even have many pedestrian crosswalks. That's why I think that the location of where the 30 mph speed limits are can help see if how the road is setup can be causing the accident. All of this is to pose the question: are Chicago's roads causing accidents to become more frequent?

To attempt to answer this, I will use the longitude and latitude from this data set filter for the ones where the speed limit was either equal to or less than 30 mph in order to see where these types of accidents were occuring. From that, I created a scatter plot to see where most of these accidents occur. If there are 'hot-spots' where there is high concentration, we can see if the roads are set up in a way that can increase the chances of accidents. Below is the scatter plot.



After making it more zoomed in, the coverage is still pretty large as there is a .8 latitude and .5 longitude coverage and it's concentrated in one area. But the scatter plot is most dense between 41.7 - 42.0 latitude and -87.65 - -87.75 longitude. I picked -87.7 longitude 41.8 latitude as a point in this area to get an idea of where this area is covering and checked it on google maps. Sky-view shows grid-like road patterns with so many intersections that could lead to accidents. See that sky view below:



Moreover, street-view shows neighborhoods with very narrow streets with cars parked on each side of the road making it harder for cars to pass through. Walking down, you can see pedestrian crosswalks that can also allow for more traffic and subsequent accidents. It is also important to note that this is a neighborhood and a quick google search of 'chicago speed limits in neighborhoods' would tell you that the speed limits would be 30 mph. As shown by street view, these are big neighborhoods so it could be assumed that the reason we saw triple 30 was because these accidents are in residential neighborhoods where the default speed limit is set at 30. See this street view below:

This all shows that these accidents at lower speed limits can be the result of bad road layouts and lack of space. The location of these accidents also explains the mystery of the triple 30, the

standard speed limits in neighborhoods at 30 mph. If you'd like to check this street out for yourself then here is the reference link: bit.ly/3SyLHTc.

Second Trend

Although we have found one possible reason, I'd like one more that is more data oriented to maybe find a good predictor for when an accident is most likely to occur. I want to specifically look into the hit-and-runs column at this speed limit because I feel like this data might be subject to selection bias and that may have caused the data to seem to converge at a low speed limit of 30. But to check if this was the case, I needed to look further. Below is a table of the number of accidents that were a normal crash (left) and the number of hit-and-runs (right) from the REPORT_TYPE column.

INJURY AND / OR TOW DUE TO CRASH	NO INJURY / DRIVE AWAY
211779	583177

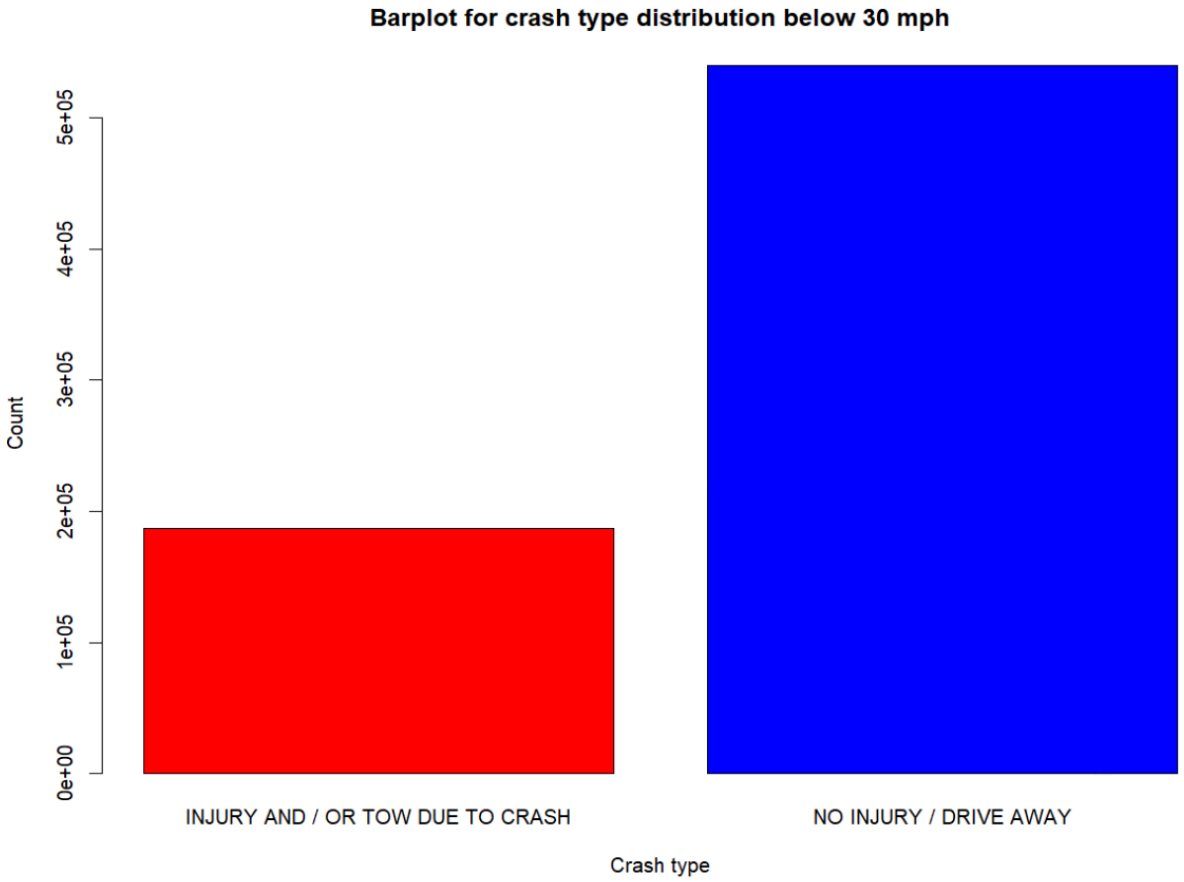
Not only does it show us that most of the crashes are hit-and-runs, but it develops the possibility of selection bias. If there is a high amount of hit-and-runs, there is a chance that not all hit and runs get recorded. Especially the ones on higher speed limits may not be reported because in some cases there was no recorded information like a license plate to find the person who caused the accident or it is even possible they got away quickly because of the high speed limit. This could mean that data points at these higher speed limits could be missing and that is why our data seems to concentrate around the 30 mph speed limit.

But before proving anything, I need to outline a possible method that can test this hypothesis further. I first make two sections of accidents of the data that are above and below the 30 mph speed limit and only focus on the REPORT_TYPE column. If the number of hit-and-run accidents above the 30 mph speed limit are less than the normal 'towed' crashes and the opposite is true for below then there must be selection bias present as accidents of hit-and-runs at higher speed limits can not be reported. The first table below is of accidents less than or equal to the 30 mph speed limit (crashPartOne) and the 2nd table is above that (crashPartTwo).

```
> print(crashPartOne)
CRASH_TYPE
INJURY AND / OR TOW DUE TO CRASH           NO INJURY / DRIVE AWAY
                          187227                           540257
> print(crashPartTwo)
CRASH_TYPE
INJURY AND / OR TOW DUE TO CRASH           NO INJURY / DRIVE AWAY
                           24552                            42920
```

It did not end up being that case. The table for above 30 mph speed limit for hit-and-runs was still greater and the same was apparent for the table for less/equal to the 30 mph mark. There might be selection bias occurring but the data does not point to such a conclusion. But there are some things we can extract from these two tables, one more trend that can shed more light on this crash data.

From past understanding and what we can see from the first table (CrashPartOne), all crash types had the highest count before the 30 mph limit. We can justify this based on our initial explanation on how these roads are set up. But one interesting trend emerges, the number of hit-and-runs (or the ‘no injury / drive away' as it is recorded in the column) is way higher than the number of crashes. See bar graph of this trend below:



It is way higher than the normal crash and tow. But the solution to this mystery also lies in our past research. That is why I can also say with high certainty that the reason why the number of hit-and-runs is so high is because (referring to the picture of street-view google maps part) there are a lot of cars parked on BOTH sides of these narrow roads. This makes it easy for some to hit a parked car and just drive away. As a result the owner could later report it as a hit-and-run, adding to our large surplus of hit-and-run records.

Predictor for 1st and 2nd Trend

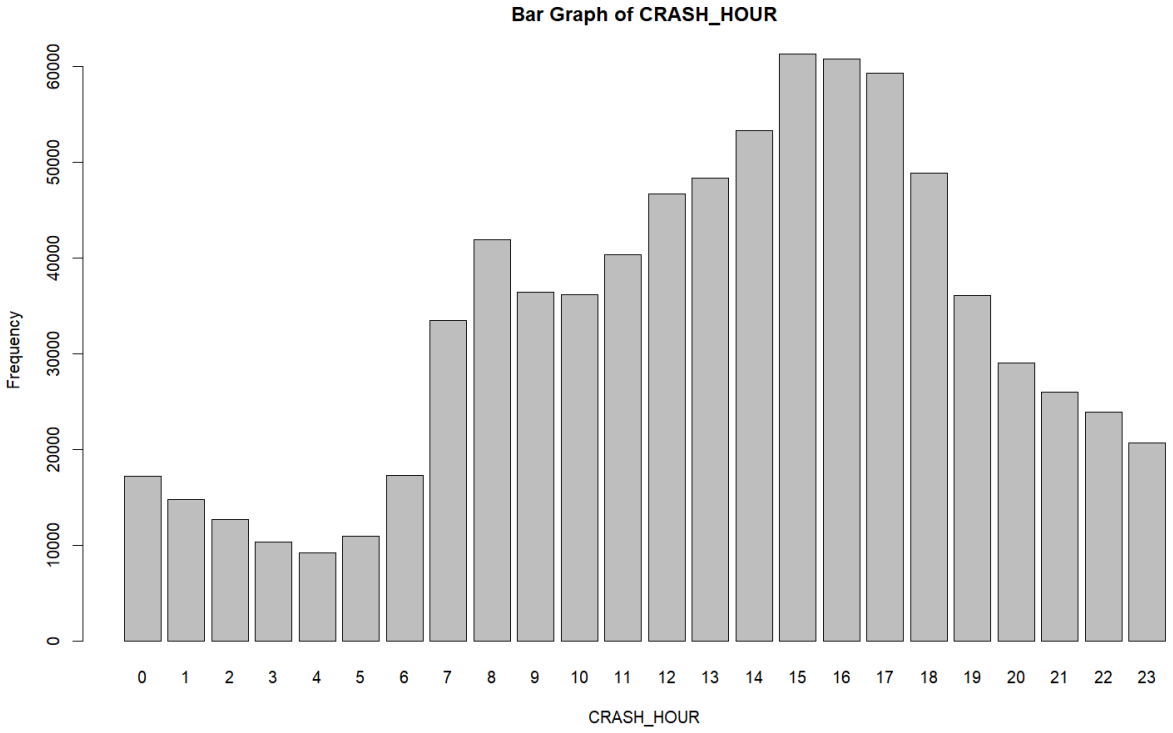
Before I conclude this section, I think we can actually make a predictor from this as well. You can find the code for this and everything mentioned in this article in the reference.R file. Just to summarize the code, it uses a simple rule-based method to determine if a crash is hit-and-run or just a normal tow. Our first rule makes sure the crash in the bounds we found from the concentration of accidents and defines the latitude/longitude. We also have a second rule where

the recorded speed limit must be less than or equal to 30 mph because of what we learned before about the location of these crashes. If it fits these two metrics, it is most likely a hit-and-run otherwise a normal tow. This predictor was 70% effective, providing some credibility to our previous conclusion about the data. See the exact accuracy below:

```
[1] 0.7335947
```

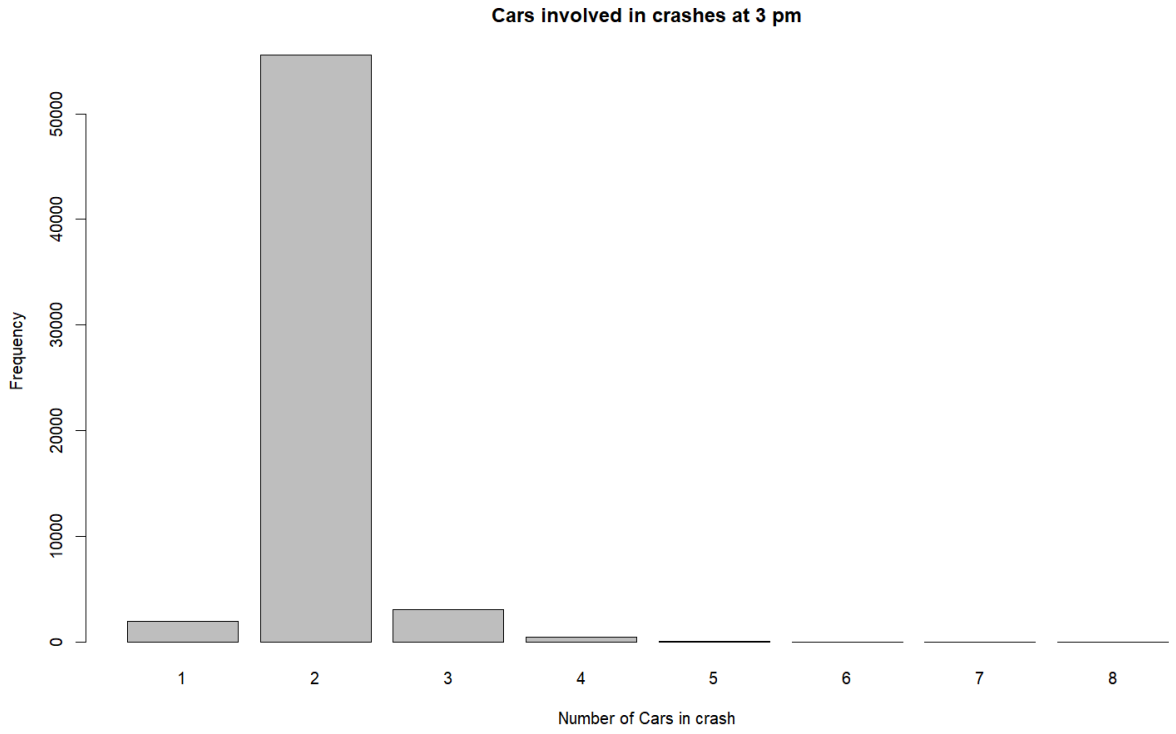
Third Trend

Now I will move beyond these main trends and show you one last trend I found worth mentioning. I found the crash hours column to be interesting, something to preface is that it is not presented like a XX:XX:XX format but by the hours where it goes from 0 to 23, where 0 is first hour and 23 is last hour so it's a bit weird to think about. The website had little previews into the data and I may have accidentally seen this column and it piqued my interests. I double checked what I saw below in a bar plot:



The trend I saw was also here, there was an increase towards a specified hour and that one hour had way higher number of accidents compared to other hours. I then proceeded to compare it to the NUM_UNITS column (shows number of cars involved in crash) to see if my possible reason is true: if there are more cars involved in the crash then there might a lot of cars driving at this hour so that is why that hour has the most accidents.

I took the highest hour of 15 (3 p.m.) from the bar plot above and filtered the NUM_UNITS column according to that hour. From that I made a barplot you can see below:



What was interesting was that there was more than one car involved in a crash so there is a chance that around 3 p.m. there is the most traffic in Chicago so that is why most of the crashes took place then. But I wanted to go back to the table I made for this data and compare it to a table for the entire column of NUM_UNITS, you can find the 1st table below the filtered one for 3 p.m. and the 2nd one is the entire column:

1	2	3	4	5	6	7	8
2011	55571	3113	458	98	19	8	3

1	2	3	4	5	6	7	8	9	10	11	12	13
43915	695536	44003	8497	2052	587	204	89	36	16	7	5	1

14	15	16	18
2	1	1	4

It does not distinctly show that 3 p.m. had the most cars involved in a crash because compared to the entire column these numbers are way smaller. There are also car crashes above 8 cars in the column that never appeared in the filtered data so it kind of diminishes the chances that traffic could be the source of 3 p.m. being a reason. I would still say that it might have some influence on why 3pm has the highest number of crashes just from experience, but without a column focused on traffic it would be hard to prove.

Conclusion

With that last insight, our exploration into this data had ended. That warps up the most interesting trends I found from the Chicago Traffic Crashes data set. There might be a lot more things left to find in there, especially with 48 columns, but I just wanted to show what I found myself, especially the interesting trend where the slower you drive the more likely you are to get in a hit-and-run. In reality, that statement avoided the more nuances of the data like location of where these accidents took place. Hopefully someone may find this analysis beneficial but remember data trends always change so this analysis may not work in the future. Thanks for reading.