Which State has the Highest Accident Rate?

# **Andrea Fox Summer 2021 https://github.com/anfox86/Predicting-the-US-State-with-Highest-Traffic-Accident-Rate**

# Which Domain?

My main dataset is coming from Kaggle that was listed as a task. The following are references I have looked into and found may be useful in my analysis:

1. <https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02> - Using to compare analysis. This article used the same dataset, so I’m curious how they approached the analysis but plan to complete using my own analysis.
2. <https://www.kdnuggets.com/2015/05/top-10-data-mining-algorithms-explained.html/3> - While I think I have a general idea of the algorithm I want to use as this is a prediction project, I wanted to read through each of these to make sure I’m not missing any other options.
3. <https://www.hindawi.com/journals/jat/2018/3869106/#data> – Journal article about traffic analysis using deep learning. Not sure on the dataset used for this, but again I like having a reference and see what others have done.
4. <https://www.hindawi.com/journals/jat/2020/1257627/> - Same author as journal article above, but this talks more about predicting traffic accident severity. I don’t plan on focusing on severity of accident but have not completely ruled that out.
5. <https://meshbesher.com/news-and-updates/after-a-car-crash/top-5-causes-of-car-accidents/> - Law firm website that walks through the top 5 causes of traffic accidents. This will play a huge factor in my predictions.
6. <https://www.cdc.gov/injury/wisqars/> - Has multiple pages for injury vs fatal accidents. Again, I don’t plan on looking at fatalities, but I don’t want to rule this out if it plays a part in my analysis. For example, what if one state has more traffic accidents while another has a higher fatal traffic accident rate. I think looking at both could be beneficial.
7. <https://www.driverknowledge.com/car-accident-statistics/> - Accident cause statistics that will help me in analysis. Right now I have a few factors that I plan on considering in my analysis but will look through this to make sure I’m not missing anything important.
8. <https://www.idrivesafely.com/defensive-driving/trending/which-state-had-most-dangerous-roads-2020> - This website has states ranked by traffic accident ranks. I want to compare what I find to this to see how far off I might be. This website does only cover 2020 though. I’d be curious while looking at the larger timeline my dataset contains whether these rankings hold true or if there are huge shifts between years.
9. <https://newscenter.lbl.gov/2019/11/04/machine-learning-help-predict-traffic-headaches/> - Machine learning that help predict traffic headaches. I’m not looking at traffic patterns necessarily, but I am under the assumption that areas that have a high traffic rate, such as Los Angeles, may contribute to the overall rankings among states.
10. <https://medium.com/geoai/using-machine-learning-to-predict-car-accident-risk-4d92c91a7d57> - This article talks about predicting car accident risk. I think once I find my rankings on state with highest traffic accident rate, this could be beneficial in devising a solution or an approach to solve the problem.

# Which Data?

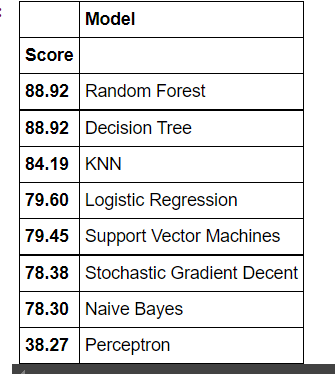
For this project, I will be utilizing the US Accidents (3 million records) dataset from Kaggle as my main form of data. This can be found here - <https://www.kaggle.com/sobhanmoosavi/us-accidents>. I plan to use this dataset because it not only gives location of the accident, but it also provides weather conditions, time of day, and location type (crossing, roundabout, etc.). This data was collected for 49 states in the United States and covers the February 2016 to December 2020 timeframe, so there is plenty of data to analyze for trends (Moosavi, Samavatian, Parthasarathy, and Ramnath, 2019).

# Research Questions? Benefits? Why analyze these data?

After cleaning up my dataset and handling my outliers if any exist, I think my first approach will be to look at traffic accident rate overall total by state. I think having that baseline will help guide my future analysis. After I do this, I plan on breaking down my analysis into three main areas that deal with cause: population density, weather trends, and distractions. When I first mentioned to my husband what I planned on doing for my project he immediately guessed California would have the highest rate because of LA traffic. However, I know living in the Midwest accident rate goes up during major weather events such as thunderstorms and snow. I also need to factor in the distraction potential as that seems to be an issue that keeps increasing with younger drivers. All of these causes I think will affect the overall prediction rate because if one state has more weather events while another has a higher population will they be even? Will one come out higher than another? There could also be potential factors that I am not considering yet as I saw in my dataset there are areas of accident like railroad crossing or roundabout. Do accidents tend to happen more often around these areas? I am sure traffic lights play into this some with people running red lights. As for now these are my focus areas, but with a deeper analysis of the data as well as looking into my references more deeply that could easily change.

# What Method?

I think initially I just want to see totals of accident rates and then accident rates by the different causes. I plan on doing this by getting a sum and/or by doing some quick visuals. After that, Naïve Bayes looks to be my best option when it comes to prediction, but I included a few references that go over several different options. I may do something like my Titanic project, where I get some code down and then look at the potential score of the different algorithm types.



I am not sure if this will be feasible, but I do not think it will hurt to try it and see what happens. I may do some more research into prediction as I have not done it much outside of some past exercises and the term project from the Predictive Analytics course. There we were not able to predict hospital readmissions with any kind of accuracy so I may also review that project and see what I could use now and where we may have gone wrong.

# Potential Issues?

The obvious here is time. I plan on spending quite a bit of time in cleaning and analysis, but life happens, and things come up, so I am listing time as a potential risk. My coding skills are still raw, so this could potentially hold me back on doing some things that I might do if I had more skill or experience. I do plan on using Google as much as possible as well as my classmates. The last issue I have is my dataset is quite large compared to others I have worked with in the past, so not sure whether that will throw my schedule off.

# Concluding Remarks

Overall, my intention with this project is to hopefully predict which state(s) have the highest rate of traffic accidents. Ideally, if I were working on this for a business, I could then pinpoint the reasons why the accident rates are so high in those states and then suggest potential ways to fix or alleviate the problem.

# References

* Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “[A Countrywide Traffic Accident Dataset](https://arxiv.org/abs/1906.05409).”, 2019.