Accident Rate and Severity – US Traffic Accident Rates

Andrea Fox

Summer 2021

https://github.com/anfox86/US-Traffic-Accident-Rates

Introduction

I decided that looking into US traffic accident rates would be an interesting problem because I live in the Midwest and every time, we get any type of precipitation it feels like everyone forgets how to drive. This inevitably leads to numerous accidents. There is also the argument in Kansas City that Kansas drivers are way worse than Missouri drivers, which the data does not support even though my husband is adamant on this fact.

At first this project was created to predict the state with the highest accident rate so that a plan could be created to help remedy the issues such as an education campaign on safe driving or maybe it would mean more trucks out after a snow fall clearing the streets. The more I dove into the data the less I was inclined to try prediction. I posed the question to the course section, and Sam Loyd mentioned it might make more sense to look at severity and have the whole project be a more story telling piece than a prediction piece. The data used in this project came from Kaggle and there are 47 variables, which can be found in the Appendix under table 1.

Business problem/Hypothesis

The issue I set out to solve in this project is to find out what was causing the traffic accident rates to increase. With more knowledge on what was potentially causing higher rates, I could work with state or even federal transportation officials to work on a plan to help alleviate the problem. This could

be a national education campaign like the do not text and drive campaign or it could mean more money is spent to clear roads during a snow event or something like Kansas City is doing and creating a toll lane that would bypass most rush hour traffic experienced on the major highways. There are numerous methods that can be deployed to alleviate the problem, I just need to identify the biggest issue(s). The second problem I wanted to look at was whether severity was affected by anything specific like weather or obstacles. “Nearly 1.25 million people die in road crashes each year, on average, 3,287 deaths a day (Rawat, 2020).” With rates this high, it is definitely worth looking into severity of accidents to potentially find any patterns that may at least lower the more severe accident rates.

Clean-up

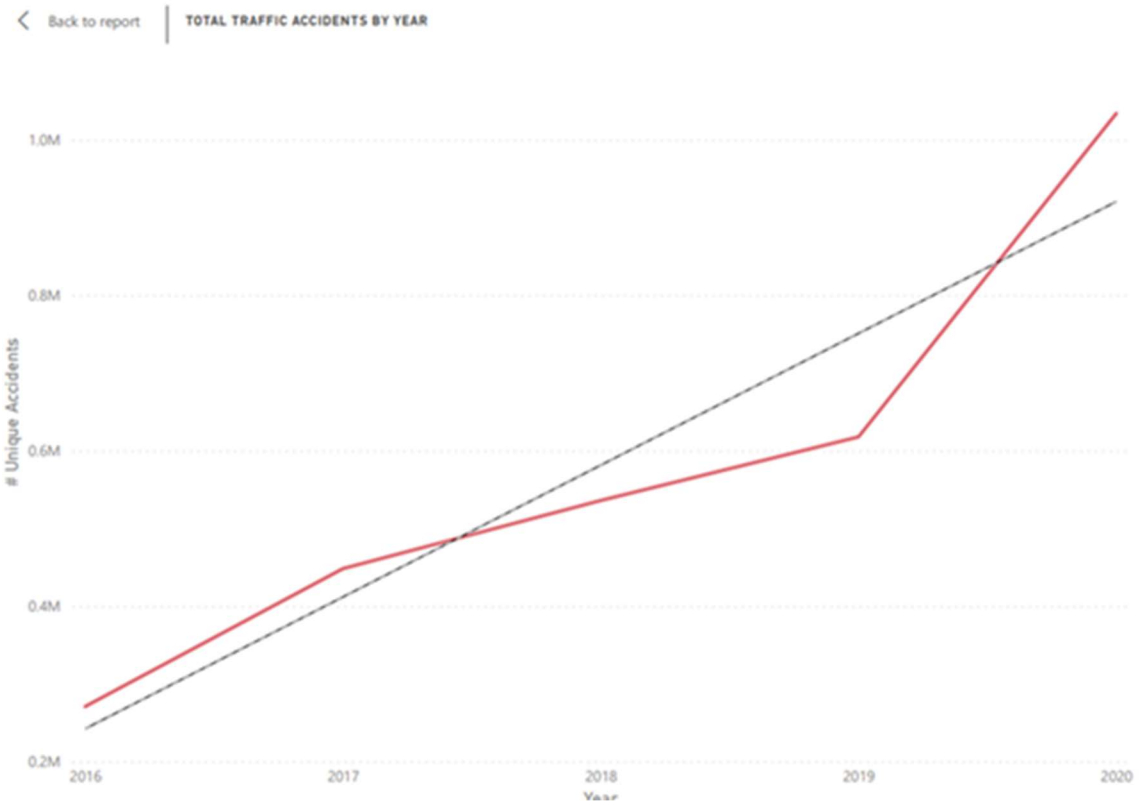
The first step after downloading the data was to read it into a Jupyter notebook so I could begin to do some cleaning. After printing a list of columns, I determined there were many columns that I had no intention of looking into included in the dataset. I removed those columns then started to look at null values. This was the first blow in my plans as there were several thousand null values in the weather-related columns. I reached out to the class on this as well since I know there are benefits to keeping null values as well as changing them to 0. Overall, I decided to change my null values to 0 since the rest of the weather data was specific (example, temperature of 76 degrees F). I then changed the True/False values in the obstacle related columns into integers so they would be easier to work with when it came to modeling. Again, Sam Loyd is an amazing resource as I spent most of a Saturday trying to do this for the entire dataset. Ultimately, I could not get it to work by applying the integer type to the dataset, I had to do it column by column, which was tedious but effective. Then I had to go find out what each of the different twilight phases that were variables in the dataset meant. I found a website that explained the different twilight phases which went from where there was still quite a bit of light out with the sun going down to complete darkness and the sun had fully set (Bikos & Kher, n.d.). From there I had about 110 null values for each twilight phase, so I updated those to 0 since the values were either Day or Night.

The last piece of cleaning was removing the columns that had date/times in them as I really only cared about year and month from the start time. I created two new columns of Year and Month and then dropped the other date/time columns.

Analysis

So far most of my analysis has come from creating visualizations. The first thing I wanted to look at was the accident rates by year, so I created a simple line chart that would give me an idea of what the rate looks like from 2016 to 2020.

Figure 1 – Total Traffic Accidents by Year

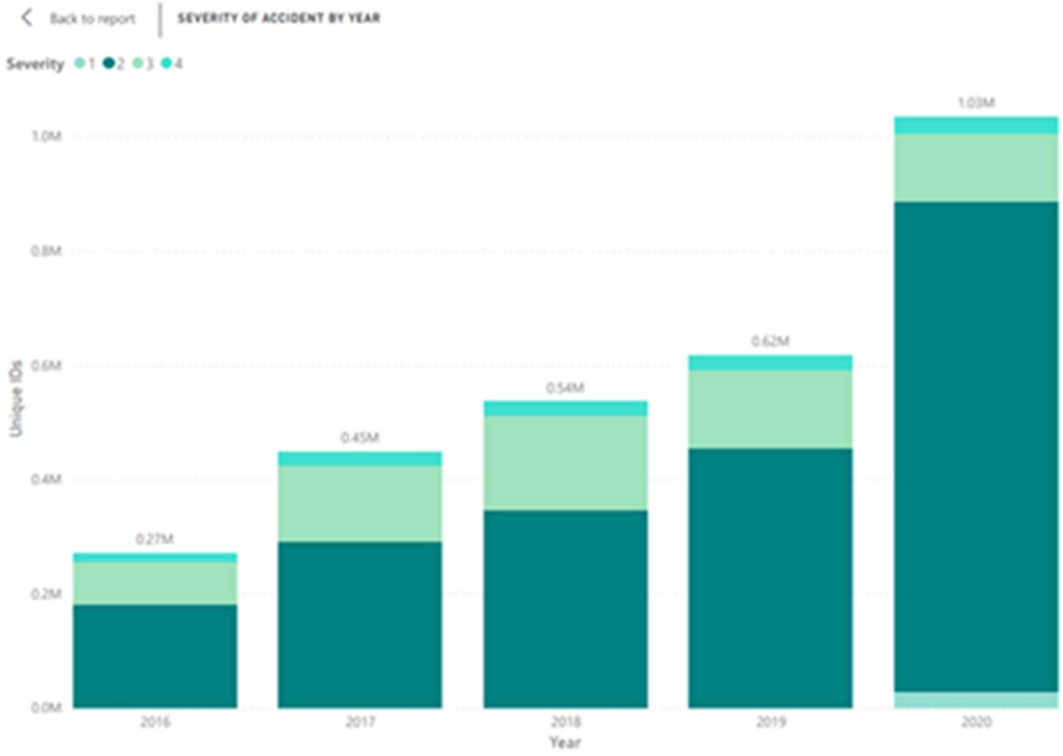


What I did not expect looking at the trend was that it steadily increased especially in 2020. My initial thoughts were 2020 would be quite low due to the pandemic, but the accident rate might be higher because many people traveled by vehicle instead of flying. I then broke the data into a column chart that

would look at total accidents by month and then used year as my values, so the bars were stacked. I expected the winter months to be high due to potential snow and/or ice events, but January and February did not look high in comparison to other months. The months with the highest rate of accidents overall were October, November, and December. This could be due to holidays and people are traveling at a greater rate than they normally would.

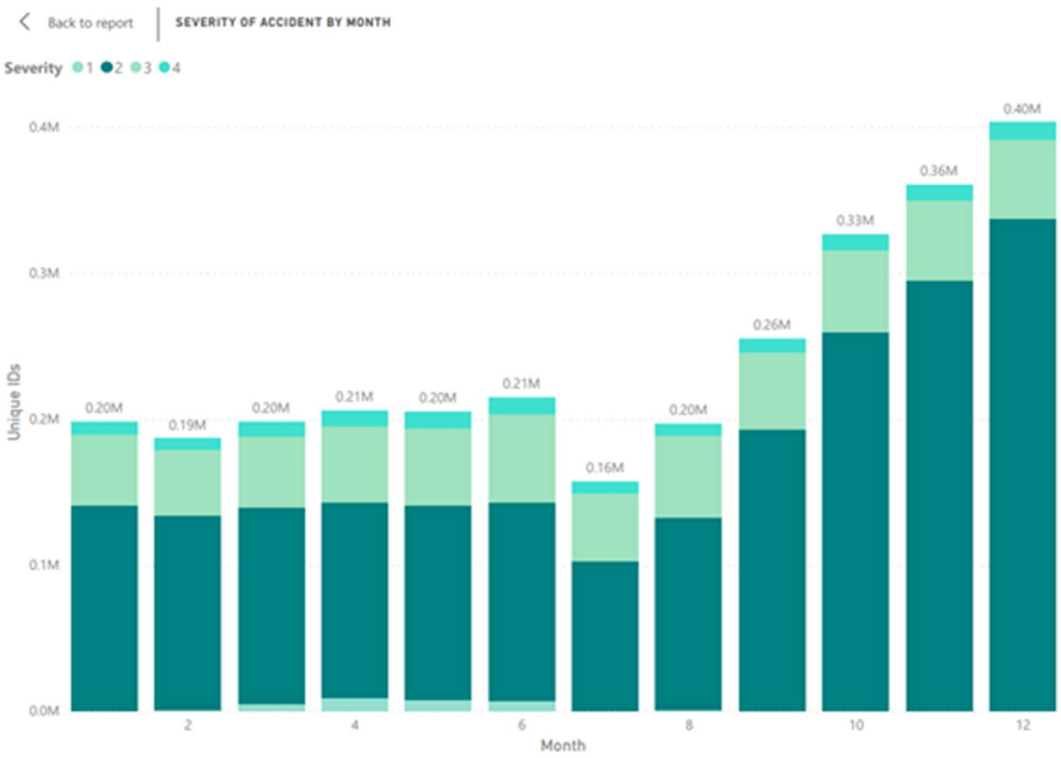
Next, I wanted to evaluate severity by both month and year. The dataset I used from Kaggle had all the column descriptions. Severity "shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic and 4 indicates a significant impact on traffic (Moosavi, 2019)." It is my assumption that the more severe an accident is causing an impact to traffic patterns the more likely there are injuries or fatalities. In figure 2, I have severity broken down by year as well as level severity.

Figure 2 – Severity of Accident by Year



Here you can see that 2020 had the highest number of accidents and that 2 was the most common severity rate for that year, as well as all other years. In Figure 3, I break down severity level for all years by month. Again, we see that October, November, and December have a much higher rate of accidents than any other months.

Figure 3 – Severity of Accident by Month



I wanted to point out though that severity 4 also increased across the months with December having the highest number at 11,915. Still under the assumption that these rates and severity rates are higher at this time of year due to holiday travel. The last visualization/analysis I did as an overview was broke down each state with their totals by years in a table format. It was a cleaner way to just see totals by state and year. This can be found in Figure 1 in appendix.

I then moved into analysis for weather related incidents. I created 4 different horizontal bar graphs that looked at total numbers (Unique IDs) by what felt the most influential weather type which

include: visibility, precipitation, wind speed, and weather condition. I also created another table that broke down state with totals in each of these different weather types which can be found in the appendix as figure 2. On this I did not expect California to lead in visibility or wind speed, however, that lead was not significantly far ahead of other states. Florida led in accidents by precipitation and Minnesota led by weather condition. Both make sense as Florida can have heavy rainfall especially during hurricane season, and Minnesota experiences heavy snowfall during the winter months. I did not look at severity levels here as I was more focused on totals overall by state along with the weather condition. I did create a separate Jupyter notebook that looked at correlation between severity and all the different weather-related variables. None of the correlations had a major impact as they were either positively or negatively 0 to 0.3, which according to Bruce Ratner, those are weak correlations (Ratner, 2009). With the number of records, I expected weather-related traffic accidents to have more significance than it did. Most of the totals are quite low compared to the overall number of records which are in the millions.

After evaluating weather, I then investigated the different obstacles for lack of a better word. California and Florida were both at the top of the list in several of the obstacle related accidents. Junction had the highest overall numbers but looking at the rest of the data California had a high rate of accidents at junctions at 103K compared to the next highest at 17k for Florida. Here is an outlier that I may decide to remove since it has a huge gap between California and the other states. If I do not count the outlier in junction, traffic signals seemed to have the highest overall rates of accidents. Florida, Texas, and California had the highest numbers of accidents at traffic signals. This could mean that traffic signals may not be clearly marked, or people are trying to get through without slowing down if the light is yellow. In Missouri, if we have a traffic signal coming up that is not easy to see (over a hill, etc) there is usually a warning sign with a light that turns on when the traffic signal is at red. This could be a potential fix for some of these large numbers. Stop signs and railways also had high numbers, but California again

had a large number of accidents at railroads compared to other states. Moving forward I may remove California from this chart as well as it skews the data since CA has 7.8K accidents at a railway compared to the next highest of IL at 2.0k. I also ran correlation numbers for severity in comparison to all the obstacles in the dataset, which I did not use all the obstacle variables in the visualizations. Again, there were weak correlations between all the different obstacles and severity. I may need to do some more digging here as I would think some of the more severe accidents would happen around traffic signals and/or stop signs due to the nature of the accident (example, someone runs a red light while another person is crossing it would cause a T-bone type of accident). I have not split the data or gotten into the modeling as the visuals took several days/hours to perfect and dig into.

The last dashboard I created to do analysis was accidents by time of day. Here I created two horizontal bar charts that looked at number of accidents by state for all 3 of the different twilight phases as well as total accidents by state for time of day. Again, all my assumptions proved to be incorrect in relation to time of day. I expected there to be a higher rate at nighttime in comparison to daytime. Mostly because of vision issues, but daytime also makes sense if I consider rush hour in the morning and evening.

Conclusion

I am not done with my modeling or digging into the data, so I expect some of what I have summed up in my methods to change. So far, most of my earlier assumptions have proven to be incorrect. I do plan on spending more time looking into severity in comparison to traffic obstacles this week as I think there could be more to find in this space. I did not look into this while creating visuals, so that is my next step along with splitting my data and looking into modeling.

10 Questions/Q&A

1. Did you consider population sizes when evaluating data?

2. Why do you think accident rates increased in 2020?
3. Did you immediately suspect California was going to be at the top of the list in number of accidents considering LA traffic?
4. What do you hope to find when looking into the different severity levels?
5. Will you look at rate of fatalities in comparison to traffic accidents?
6. Will you investigate distracted driving as a cause for higher rates of traffic accidents?
7. By looking at the numbers, how do you propose the different states make changes to alleviate the problem?
8. Since weather did not factor much into accident rates will you keep investigating this area?
9. You mentioned rush hours in the time-of-day accidents, will that be investigated more?
10. What has surprised you most in your investigation?

References

Bikos, K., & Kher, A. (n.d.). *Twilight, Dawn, and Dusk*. Retrieved from Time and Date: https://www.timeanddate.com/astronomy/different-types-twilight.html

Moosavi, S. (2019, May 20). *US Accidents (3 million records -- updated)*. Retrieved from Kaggle: https://www.kaggle.com/sobhanmoosavi/us-accidents

Ratner, B. (2009, May 18). *The correlation coefficient: Its values range between +1/-1 , or do they?* Retrieved from Springer Link: https://link.springer.com/article/10.1057/jt.2009.5

Rawat, S. (2020, February 21). *USA Accidents Data Analysis*. Retrieved from Towards Data Science: https://towardsdatascience.com/usa-accidents-data-analysis-d130843cde02

Appendix

Table 1 – Variables from Dataset (Moosavi, 2019)

Variable	Description
ID	Unique identifier of accident record
Severity	Severity of accident
Start_Time	Time of beginning of accident.
End_Time	End time of accident.
Start_Lat	Latitude of starting point.
Start_Lng	Longitude of starting point.
End_Lat	Latitude of ending point.
End_Lng	Longitude of ending point
Distance(mi)	Length of the road affected by accident
Description	Brief description of accident.
Number	Street number in address
Street	Street from address record
Side	Side of street (right/left)
City	City from address record
State	State from address record
Zipcode	Zipcode from address record.
Country	Country from address record
Timezone	Timezone accident-based in.
Airport_Code	Closes airport weather station.

Weather_Timestamp	Timestamp of weather observation.
Temperature(F)	Temperature at time of accident.
Wind_Chill(F)	Wind chill at time of accident.
Humidity(%)	Humidity at time of accident.
Pressure(in)	Air pressure at time of accident.
Visibility(mi)	Visibility in miles at time of accident.
Wind_Direction	Direction of the wind at time of accident.
Wind_Speed(mph)	Speed of wind at time of accident.
Precipitation(in)	Precipitation amount at time of accident.
Weather_Condition	Weather condition at time of accident.
Amenity	Amenity nearby accident location.
Bump	Presence of bump at accident location.
Crossing	Presence of crossing at accident location.
Give_Way	Presence of a give way near accident location.
Junction	Presence of junction near accident location.
No_Exit	Presence of no exit near accident location.
Railway	Presence of railway near accident location.
Roundabout	Presence of roundabout near accident location.
Station	Presence of station near accident location.
Stop	Presence of stop sign near accident location.
Traffic_Calming	Indicates if traffic was calming at accident location.
Traffic_Signal	Presence of traffic signal at accident location.

Turning_Loop	Presence of turning loop at accident location.
Sunrise_Sunset	Time of day (day/night).
Civil_Twilight	Time of day based on civil twilight.
Nautical_Twilight	Time of day based on nautical twilight.
Astronomical_Twilight	Time of day based on astronomical twilight.

Figure 1 – Total Accidents by Year for Each State (broken into two images)

Back to report | TOTAL OF ACCIDENTS BY YEAR FOR EACH STATE

State	2016	2017	2018	2019	2020	Total
AL	135	1813	7544	10148	13650	**33290**
AR	26	95	491	925	3189	**4726**
AZ	2770	7976	13138	13633	24190	**61707**
CA	87207	95651	98862	168526	280498	**730744**
CO	1324	4702	8981	11920	10353	**37280**
CT	3120	4579	4278	3588	8489	**24054**
DC	306	536	736	940	2683	**5201**
DE	246	640	1024	908	1807	**4625**
FL	23945	38775	40019	35407	125154	**263300**
GA	8277	13635	15624	14559	17441	**69536**
IA	1386	1961	1353	1717	3075	**9492**
ID	419	346	389	540	2471	**4165**
IL	12455	18238	15988	13298	17647	**77626**
IN	1610	4483	7784	4822	6294	**24993**
KS	762	1453	1718	1232	2279	**7444**
KY	1067	2098	4571	3737	3823	**15296**
LA	849	6500	10891	9892	21971	**50103**
MA	3006	5005	5192	4988	6930	**25121**
MD	4857	8469	7415	8051	23963	**52755**
ME	135	585	687	622	237	**2266**
MI	9898	16593	14872	12645	13065	**67073**
MN	2567	7005	11207	20852	38081	**79712**
MO	3265	5725	5918	5425	8341	**28674**
MS	402	673	1475	1238	1631	**5419**
MT	62	133	130	156	2852	**3333**
NC	6147	21901	30563	24482	39704	**122797**
ND	3	1	18	15	426	**463**
NE	3081	3273	3034	2568	1722	**13678**
NH	281	1607	1775	1616	971	**6250**
NJ	6078	9430	9720	8435	16551	**50214**
Total	**270881**	**448268**	**535921**	**617712**	**1033828**	**2906610**

NJ	6078	9430	9720	8435	16551	**50214**
NM	153	724	1168	926	1002	**3973**
NV	1357	1828	1456	1653	1518	**7812**
NY	13742	22674	22543	25051	42166	**126176**
OH	7621	9249	11522	7111	12333	**47836**
OK	570	5323	10380	10201	8631	**35105**
OR	1005	3823	19225	38300	35999	**98352**
PA	12483	15411	12190	10863	38798	**89745**
RI	1034	1813	2031	1865	1571	**8314**
SC	2058	14996	28933	30144	44331	**120462**
SD	8	11	11	27	160	**217**
TN	1688	6831	13376	11458	22142	**55495**
TX	29803	49743	47246	42755	57093	**226640**
UT	968	3297	7882	20938	13812	**46897**
VA	4546	12630	16437	14991	41126	**89730**
VT	17	113	127	170	128	**555**
WA	6107	11919	11977	9702	9750	**49455**
WI	1549	3253	3278	4125	2464	**14669**
WV	417	653	600	453	1298	**3421**
WY	69	96	142	94	18	**419**
Total	**270881**	**448268**	**535921**	**617712**	**1033828**	**2906610**

Figure 2 – Overview dashboard of visualizations

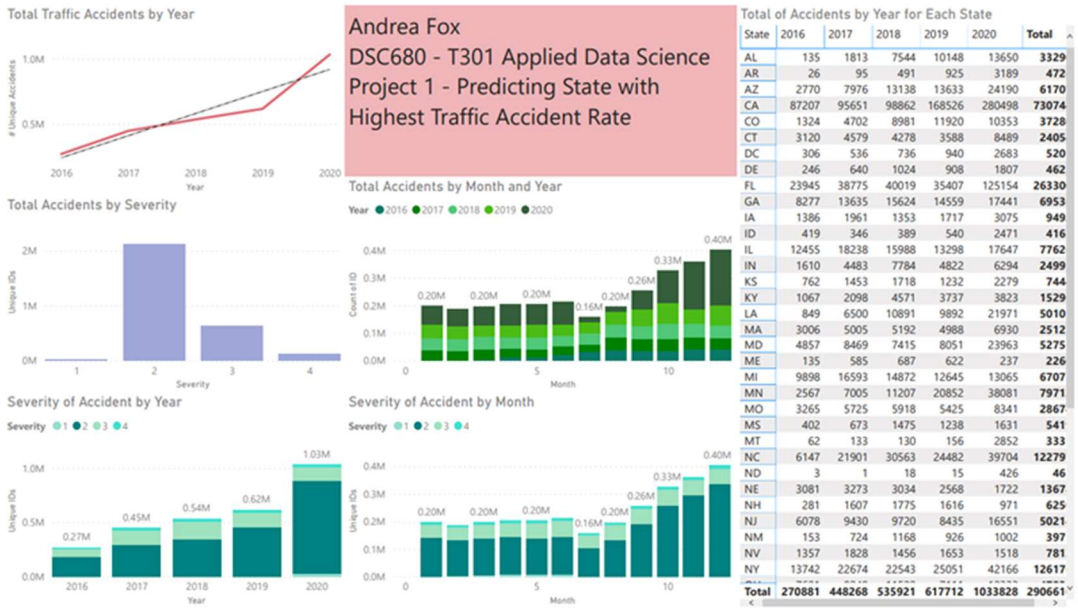


Figure 3 – Accident by Weather dashboard

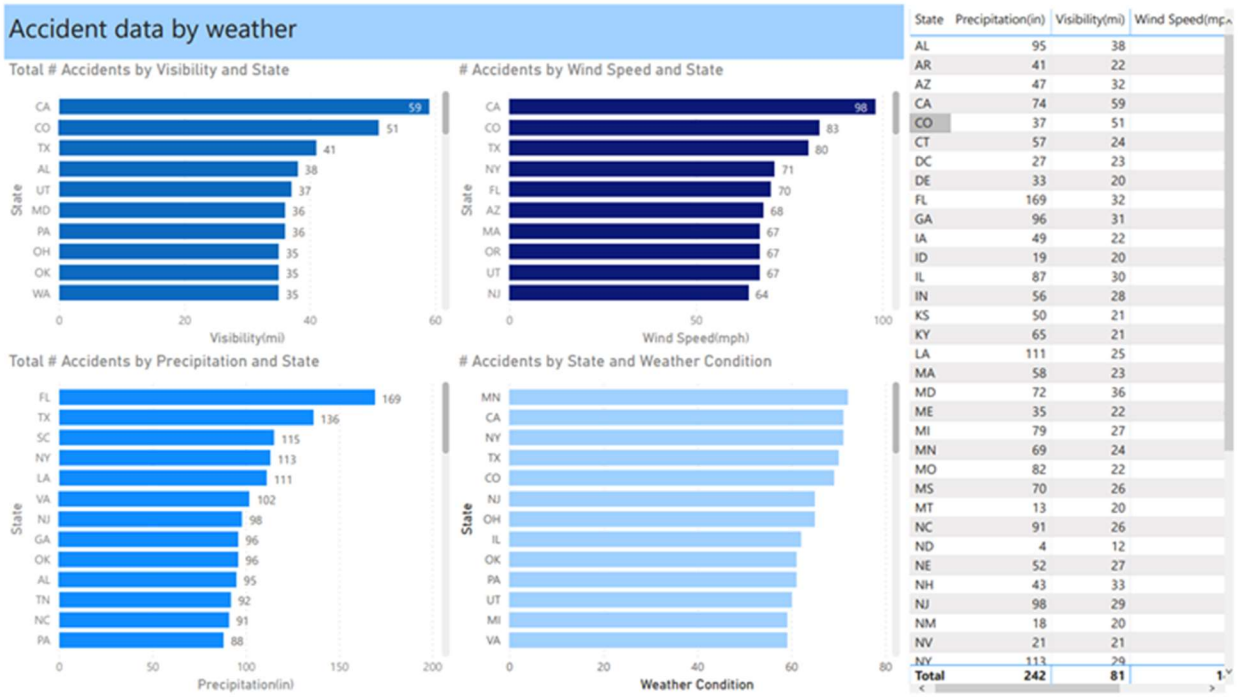


Figure 4 – Accident data by Road Obstacles dashboard

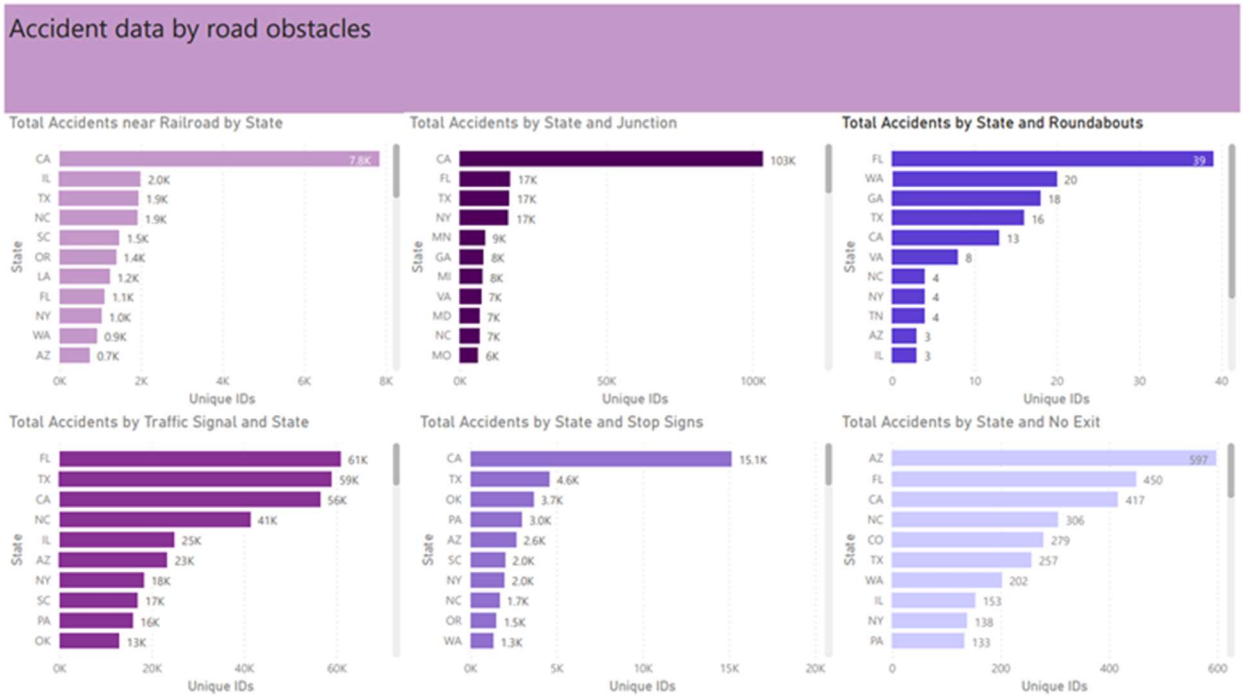


Figure 5 – Accident Data by Time-of-Day Dashboard