

**US CAR ACCIDENTS DURING 2016-2021**

Muna Makhkamova

9/20/2022

**EXECUTIVE SUMMARY**

The dataset for this project represents the information on car accidents in the USA from February 2016 to December 2021. The dataset covers 48 states of USA and District of Columbia and contains the characteristics of car accidents such as time, date, location, severity of accidents, temperature, weather condition and etc. Currently, there are about 2.8 million accident records in this dataset.

The project will focus on analyzing the car accident patterns and impact of weather conditions on the amount of car accidents.

**DATA SOURCES**

The main dataset for this project is open source and was downloaded from [Kaggle](https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents).

They were collected using multiple APIs that provide streaming traffic incident (or event) data. These APIs broadcast traffic data captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

In addition to the main dataset, data on the number of licensed drivers per state was also downloaded from the official website of [Federal Highway Administration](https://www.fhwa.dot.gov/policyinformation/statistics/2020/).

**LIMITATION AND ETHICS**

* The main dataset represents only a sample of car accidents in the USA, but the creator of the dataset has claimed that this sample was collected correctly and can be a good population representative;
* The main dataset does not cover all 50 states (Hawaii and Alaska were excluded);
* The main dataset does not contain any information on the drivers involved in the car accidents, so this analysis will not involve personal characteristics;
* The main dataset does not contain any information on the drivers involved in the car accidents, so this analysis will not involve personal characteristics;
* The severity of accident was estimated by the data provider based on the impact on traffic flow, rather than the incident itself;
* The weather conditions in the main dataset contained 128 various weather conditions which for the sake of this analysis were categorized into 4 levels based on the risks for the drivers. The categorization was made based on common sense and weather risk levels for some countries due to the inability to find unified weather categories with a list of all of these weather conditions;
* The data on licensed drivers doesn’t cover the year 2021.

**DATA CLEANING AND DATA CONSISTENCY CHECKS**

* State abbreviations were replaced by full state names
* Following columns were created: Region, Year, Month, Hour, Weather, Licensed\_Drivers
* Following columns were dropped: End\_Time, End\_Lat, End-Lng, Wind\_Chill(F), Description, Number, Street, Side, Zipcode, Country, Timezone, Airport\_Code, Weather\_Timestamp, Civil\_Twilight, Nautical\_Twilight, Astronomical\_Twilight,
* Data was checked for missing values, 172786 values in Precipitation were filled based on Weather Conditions
* Data was checked for consistency
* Data was checked for duplicate values

**DATA PROFILE**

The number of rows and columns in cleaned dataset 2304260 x 37 (original dataset contained (2845342x47).

**COLUMN DETAILS**

| Column | Description | Data type | Time Variant |
| --- | --- | --- | --- |
| ID | This is a unique identifier of the accident record. | Qualitative, ordinal | no |
| Severity | Shows the severity of the accident, a number between 1 and 4, where 1 indicates the least impact on traffic (i.e., short delay, distance and duration as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay, distance and duration).  Severity is primarily defined based on the impact on traffic flow, rather than the incident itself | Qualitative, ordinal | no |
| Start\_Time | Shows start time of the accident in local time zone. | Quantitative, continuous | no |
| Start\_Lat | Shows latitude in GPS coordinate of the start point. | Quantitative, continuous | no |
| Start\_Lng | Shows longitude in GPS coordinate of the start point. | Quantitative, continuous | no |
| Distance(mi) | The length of the road extent affected by the accident. | Quantitative, continuous | no |
| City | Shows the city in address field. | Qualitative, nominal | no |
| County | Shows the county in address field. | Qualitative, nominal | no |
| State | Shows the state in address field. | Qualitative, nominal | no |
| Temperature(F) | Shows the temperature (in Fahrenheit) from the closest airport-based weather station | Quantitative, continuous | no |
| Wind\_Chill(F) | Shows the wind chill (in Fahrenheit). | Quantitative, continuous | no |
| Humidity(%) | Shows the humidity (in percentage). | Quantitative, continuous | no |
| Pressure(in) | Shows the air pressure (in inches). | Quantitative, continuous | no |
| Visibility(mi) | Shows visibility (in miles). | Quantitative, continuous | no |
| Wind\_Direction | Shows wind direction. | Qualitative, nominal | no |
| Wind\_Speed(mph) | Shows wind speed (in miles per hour). | Quantitative, continuous | no |
| Precipitation(in) | Shows precipitation amount in inches, if there is any. | Quantitative, continuous | no |
| Weather\_Condition | Shows the weather condition (rain, snow, thunderstorm, fog, etc.) | Qualitative, nominal | no |
| Amenity | Indicates presence of [amenity](https://wiki.openstreetmap.org/wiki/Key:amenity) in a nearby location. Refers to particular places such as restaurant, library, college, bar, etc. | Qualitative, binary | no |
| Bump | Indicates presence of speed bump or hump in a nearby location. | Qualitative, binary | no |
| Crossing | Indicates presence of [crossing](https://wiki.openstreetmap.org/wiki/Key:crossing) in a nearby location. | Qualitative, binary | no |
| Give\_Way | Indicates presence of [give\_way](https://wiki.openstreetmap.org/wiki/Tag:highway%3Dgive_way) in a nearby location. | Qualitative, binary | no |
| Junction | Indicates presence of [junction](https://wiki.openstreetmap.org/wiki/Key:junction) in a nearby location. Refers to any highway ramp, exit, or entrance. | Qualitative, binary | no |
| No\_Exit | Indicates presence of [no\_exit](https://wiki.openstreetmap.org/wiki/Key:noexit) in a nearby location. | Qualitative, binary | no |
| Railway | Indicates presence of [railway](https://wiki.openstreetmap.org/wiki/Key:railway) in a nearby location. | Qualitative, binary | no |
| Roundabout | Indicates presence of [roundabout](https://wiki.openstreetmap.org/wiki/Tag:junction%3Droundabout) in a nearby location. | Qualitative, binary | no |
| Station | Indicates presence of [station](https://wiki.openstreetmap.org/wiki/Key:station) in a nearby location. | Qualitative, binary | no |
| Stop | Indicates presence of [stop](https://wiki.openstreetmap.org/wiki/Key:stop) in a nearby location. | Qualitative, binary | no |
| Traffic\_Calming | Indicates presence of [traffic\_calming](https://wiki.openstreetmap.org/wiki/Key:traffic_calming) in a nearby location. | Qualitative, binary | no |
| Traffic\_Signal | Indicates presence of [traffic\_signal](https://wiki.openstreetmap.org/wiki/Tag:highway%3Dtraffic_signals) in a nearby loction. | Qualitative, binary | no |
| Turning\_Loop | Indicates presence of [turning\_loop](https://wiki.openstreetmap.org/wiki/Tag:highway%3Dturning_loop) in a nearby location. | Qualitative, binary | no |
| Sunrise\_Sunset | Shows the period of day (i.e. day or night) based on sunrise/sunset. | Qualitative, binary | no |
| Region | Shows which US region the state corresponds to . | Qualitative, nominal | no |
| Year | Shows the year of the accident. | Qualitative, ordinal | no |
| Month | Shows the year of the accident. | Qualitative, ordinal | no |
| Hour | Shows the year of the accident. | Qualitative, ordinal | no |
| Weather | Shows the category of weather condition (Light, Moderate, Heavy, Extreme) | Qualitative, nominal | no |
| Licensed\_Drivers | Shows the total amount of licensed drivers in each state during the year of the accident | Quantitative, discrete | no |

**QUESTIONS TO EXPLORE**

* **Did the amount/severity of car accidents change over time?**
* **Does the month, day of the week or hour of day have any impact on the number/severity of accidents?**
* **Does the amount and proportion of car accidents to licensed drivers vary across US states?**
* **Are there any hot locations (locations where most of the accidents happen) in US states?**
* **Does the weather condition affect the amount/severity of car accidents?**
* **Is there any correlation between temperature and amount/severity of car accidents?**

**DELIVERABLES**

**Visual results of this analysis are available** [**here**](https://public.tableau.com/app/profile/munavvar.makhkamova/viz/USCarAccidents2016-2021/Story1)**.**

**References:**

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.