**Car Accidents of the United States**

## **Faisal Fuad Nursyahid**

## National Taiwan University of Science and Technology

faisal.fuad.n@gmail.com

## **蔡承泰**

## National Taiwan University of Science and Technology

## littlemic123@gmail.com

## **張雅雯**

## National Taiwan University of Science and Technology

m10801303@gapps.ntust.edu.tw

**Abstract**

This paper aims to analyze and obtain information from the US car accident dataset. Before conducting the analysis, we first conduct data pre-processing such as feature selection and missing data cleaning. After that, we did a random sampling using the Cochran’s formula. After the data has been done pre-processing and random sampling, we do the analysis using various built-in python module such as pandas, matplotlib, and seaborn. The final result of our analysis is a various graphical, chart, and maps visualizations that represent our data.

**Keywords:** pandas, matplotlib, seaborn and maps visualizations

# Introduction

Traffic accident are one of the highest causes of death in the world[[1]](#endnote-1), including in the United States (US). According to[[2]](#endnote-2) Association for Safe International Road Travel (ASIRT) 38,000 people die every year in car accident on U.S roadway, or about 12.4 fatality rate per 100,000 inhabitants. This certainly is a terrible number. Therefore we want to learn more about this terrible fact using the US car accident dataset that we got on the Kaggle website.[[3]](#endnote-3)

In this paper we want to utilize python built in module such as: pandas, matplotlib, and seaborn to dig up information and make visualization in this dataset. The result is that multiple visualization chart and map to show to the audience the facts about US car accident.

### Dataset

This US car accident dataset consist of 49 columns and 2974335 rows collected by using various streaming accident data APIs, such as: Internet, Traffic Message Channel (TMC), and weather station. This data collected from February 2016 to December 2019. In this data included the time, duration, severity, weather, and location of the accident. The introduction of these columns is in [Figure 1](#fig1) and [Table 1](#t1).



Figure 1: The brief look of the dataset.

|  |  |
| --- | --- |
| **Column Name** | **Description** |
| Source | Indicates source of the accident report (i.e. the API which reported the accident.). |
| TMC | A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event. |
| 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 as a result of the accident) and 4 indicates a significant impact on traffic (i.e., long delay). |
| Start\_Time | Shows start time of the accident in local time zone. |
| End\_Time | Shows end time of the accident in local time zone. |
| Start\_Lat | Shows latitude in GPS coordinate of the start point. |
| Start\_Lng | Shows longitude in GPS coordinate of the start point. |
| End\_Lat | Shows latitude in GPS coordinate of the end point. |
| End\_Lng | Shows longitude in GPS coordinate of the end point. |
| Distance(mi) | The length of the road extent affected by the accident. |
| Description | Shows natural language description of the accident. |
| Number | Shows the street number in address field. |
| Street | Shows the street name in address field. |
| Side | Shows the relative side of the street (Right/Left) in address field. |
| City | Shows the city in address field. |
| County | Shows the county in address field. |
| State | Shows the state in address field. |
| Zipcode | Shows the zipcode in address field. |
| Timezone | Shows timezone based on the location of the accident (eastern, central, etc.). |
| Airport\_Code | Denotes an airport-based weather station which is the closest one to location of the accident. |
| Weather\_Timestamp | Shows the time-stamp of weather observation record (in local time). |
| Temperature(F) | Shows the temperature (in Fahrenheit). |
| Wind\_Chill(F) | Shows the wind chill (in Fahrenheit). |
| Humidity(%) | Shows the humidity (in percentage). |
| Pressure(in) | Shows the air pressure (in inches). |
| Visibility(mi) | Shows visibility (in miles). |
| Wind  \_Speed(mph) | Shows wind speed (in miles per hour). |
| Wind\_Direction | Shows wind direction. |
| Precipitation(in) | Shows precipitation amount in inches, if there is any. |
| Weather  \_Condition | Shows the weather condition (rain, snow, thunderstorm, fog, etc.). |
| Amenity | A Point-Of-Interest (POI) annotation which indicates presence of amenity in a nearby location. |
| Bump | A POI annotation which indicates presence of speed bump or hump in a nearby location. |
| Crossing | A POI annotation which indicates presence of crossing in a nearby location. |
| Give\_Way | A POI annotation which indicates presence of give\_way sign in a nearby location. |
| Junction | A POI annotation which indicates presence of junction in a nearby location. |
| No\_Exit | A POI annotation which indicates presence of no\_exit sign in a nearby location. |
| Railway | A POI annotation which indicates presence of railway in a nearby location. |
| Roundabout | A POI annotation which indicates presence of roundabout in a nearby location. |
| Station | A POI annotation which indicates presence of station (bus, train, etc.) in a nearby location. |
| Stop | A POI annotation which indicates presence of stop sign in a nearby location. |
| Traffic\_Calming | A POI annotation which indicates presence of traffic\_calming means in a nearby location. |
| Traffic\_Signal | A POI annotation which indicates presence of traffic\_signal in a nearby location. |
| Turning\_Loop | A POI annotation which indicates presence of turning\_loop in a nearby location. |
| Sunrise\_Sunset | Shows the period of day (i.e. day or night) based on sunrise/sunset. |
| Civil\_Twilight | Shows the period of day (i.e. day or night) based on civil twilight. |
| Nautical  \_Twilight | Shows the period of day (i.e. day or night) based on nautical twilight. |
| Astronomical  \_Twilight | Shows the period of day (i.e. day or night) based on astronomical twilight. |

Table 1: The introduction of the columns.

# Methods

* 1. **Pre-processing Data**

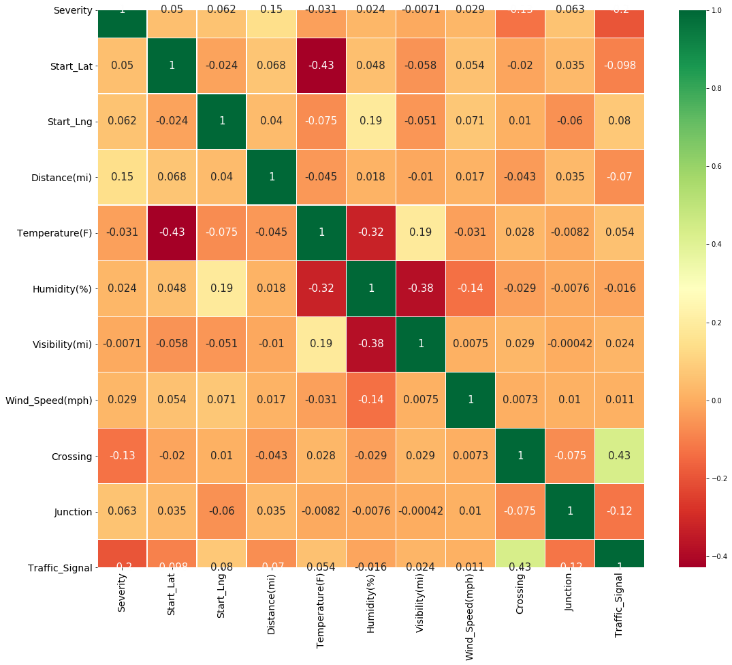
Due to a lot of redundant information, missing values, and other information that we do not want, therefore we have to do pre-processing before we conduct analysis for this dataset. These are some of the pre-processing method that we do, such as:

* + 1. **Feature Selection**

As previously noted, this US car accident dataset has 49 columns, and most of these columns contain irrelevant information and missing values. Therefore our first step to do is to remove columns with minimum 40% missing values as recommended from this site.[[4]](#endnote-4) Upon searching we found that there are 4 columns with missing values above 40%, i.e. End\_Lat, End\_Lng, Number, Precipitation(in), Wind\_Chill(F).

Then, the next step is we remove the columns that only contain one unique value, or two unique values but are dominated by one other value (99%). These columns includes: Roundabout, Traffic\_Calming, Bump, Railway, Give\_Way, Turning\_Loop, Station, No\_Exit, Stop.

The next thing we do is we remove columns that has same meaning, such as Nautical\_Twilight and Astronomical\_Twilight that has close meaning to Civil\_Twilight. The final step is we remove the columns that we don’t use because they contain information that cannot be interpreted. Such as, Airport\_Code, Country, Street, ID, Pressure(in), Wind\_Direction, Weather\_Timestamp, and TMC. In the end, from the initial 49 columns we only used 24 columns which we found useful.

* + 1. **Missing Value**

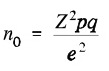
There are 3 types of missing value[[5]](#endnote-5):

* Missing at Random (MAR), missingness of the data can be predicted by other feature in the dataset
* Missing completely at Random (MCAR), the certain value that missing has nothing to do with its hypothetical value and with the other values of other variable.
* Missing not at Random (MNAR), the missingness is related to the factors we don’t know.

Based on the above explanation, our data is included in missing completely at Random (MCAR).So, then we drop all rows that have missing values. From this dropping method, we still have 84.35% data from the original data.

* 1. **Sampling Size**

After several methods of removing useless information, we still find that our data is still large enough to be analyzed. Therefore we do the sampling method using the Cochran’s formula[[6]](#endnote-6):



Z=Z value

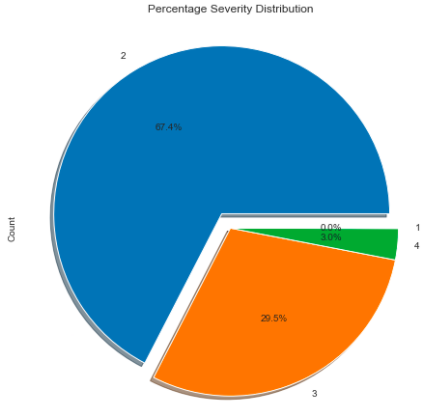
p=proportion of the population which has the attribute in question

q=1-q

e=the margin of error

We set a confidence level of 95%, with confidence interval of 0.1-0.01%. So we get our final data of 28.5% from the original data.

# Data/Results

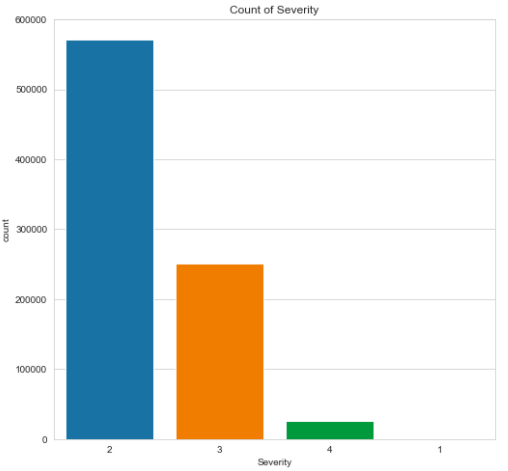
We use dataframe.shape, dataframe.colum-ns, .to-list(), dataframe.isnull(), .values(), .sum(), and dataframe.nunique() to summary the data first. Then plot the heatmap to observe the correlation (See [Figure 2](#f2)). We can see the correlation between traffic signal and crossing is high(R=0.43), it make sense because traffic signal and crossing usually cooperate in the intersection. The correlation between visibility and humidity is also higher than others(R=-0.38). This make sense, too. Since the high humidity may lead to fog or rain which will influence the visibility.

Then we aim to severity and draw pie chart and bar chart (See [Figure 3](#f3) and [Figure 4](#f4)). The pie chart shows there are 67.5% accidents belong to level 2, 29.5% accidents belong to level 3, 3.0% accidents belong to level 4 and no accident belong to level 1. If we see the bar chart, we can apparently see that most accidents belong to level 2

Figure 2: The heatmap of the dataset. It shows the correlation among several variables.

(About 600 thousands records). Despite of severity, we also draw pie chart and bar chart with weather condition (See [Figure 5](#f5) and [Figure 6](#f6)). However, because there are so many weather condition are included in this dataset, we only choose 10 top weather condition for accidents. We can see that most accidents actually happened with clear weather condition from both [Figure 5](#f5) and [Figure 6](#f6). There are more than 200 thousands accidents happen under clear condition. Maybe drivers tend to drive cautiously when the weather is bad while drive less careful when the weather is stable. By the way, notice that the percentage here does not represent the proportion of overall, it only represent the proportion of these 10 top weather condition. After knowing the information about weather and severity, we are wondering that, though bad weather doesn’t cause so much accidents as we guessed, will bad weather condition lead to a more serious accident? We plot a line graph and get the answer (See [Figure 7](#f7)). The line graph indicate that the average severity of light snow and rain are higher than other condition. The finding shows that though less accidents happen under the bad weather condition, however once the accidents happen, they usually more serious than those happen under a good weather condition.

Figure 3: The pie chart of severity. It shows the distribution among severity categori



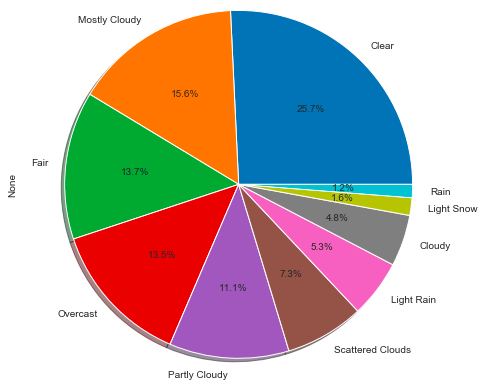


Figure 4: The bar chart of severity. Most accidents belong to level 2. No accident belong to level 1.

Figure 5: The pie chart of weather condition. It shows the distribution among 10 top weather condition.

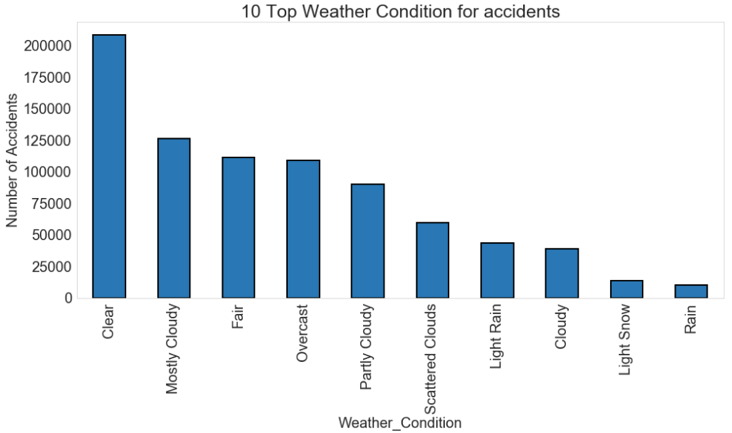


Figure 6: The bar chart of weather condition. Most accidents happened under a clear weather condition.

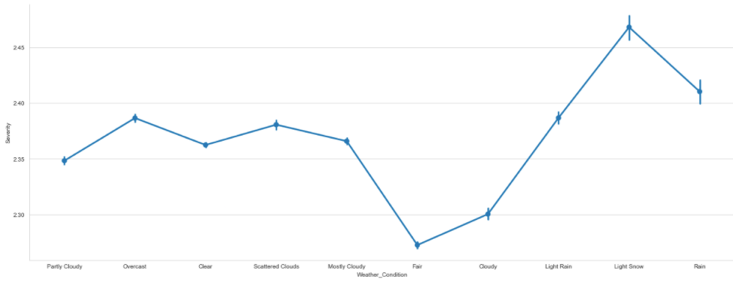


Figure 7: The line chart of severity and weather condition. It shows the relationship between severity and weather condition.

Aside from the perspective of weather condition, we targeted at the time series as well. (See [Figure 8](#f8) , [Figure 9](#f9) and [Figure 10](#f10)) [Figure 8](#f8) shows the accidents count for every month in 2019. From the picture ,we noticed that the possibility of traffic accidents during autumn and winter is relatively higher that other seasons. The major reasons we had inferred are due to poor visibility and worse weather condition. However , in our dataset, we had dropped the column of visibility due to the missing and error of these data. Also, all of the obvious variation within different months come from the second degree of severity. Furthermore, if we shrink time length to weekly data, we could see that, the possibility of the accident counts happening on the first three workdays is way higher than the other days. Holidays remains nearly half of the weekdays’ accidents.(See [Figure 9](#f9)) According to our inference, this kind of phenomenon was resulted from the high work and transportation activities during this period, people tended to be relatively unable to concentrate on commuting as well as the logistics industry. The road traffic was accompanied by low concentration of people and sudden massive.

For the daily data, we also checked severity for each day in 2019, and finding that the accidents seemed to mostly happen in day rather than night even though the traffic sight at night was way worse than day. And among all the case within the day, the severity 2 still takes the major of those 4 types of degree.(See [Figure 10](#f10))

In addition to time attribute , we also conducted analysis on

the geographic attribute. In [Figure 11](#f11) we adopted the choropleth map of Python, discovering the rate of traffic accidents in several states is high. From the graph we can see that California owns the highest traffic accident rate. Texas, Florida, New York, South Carolina as well as North Carolina were catching up. The bar beside the USA map indicates the total accident case. However, this kind of analysis could not fully address our observation toward the major accident-happening states. The reason was that all of the states just mentioned were virtually the states with high population. It is very intuitive that higher population leads to higher accident rate, it would be much better if we use density as the observation index. In order to make up for this deficiency, we additionally included population as further consideration. In [Figure 12](#f12), we can see the total accident for each state corresponding to its population. Without the interference of large population, we discovered the state “South Carolina ”owned the highest traffic accident rate. Oregon and California were the second and third.

# 4 Discussion/Conclusions

Accidents are virtually unpredictable. However, via immense and large enough historical data, we still can find out where the accident happens with high probability. In our analysis, we’ve gone through several progress to examine and find out the insight from the data, including the sampling and cleaning of the original data, the analysis toward the basic data structure, the relationship between severity and weather condition as well as traffic object. In conclusion, most of the accidents from the dataset belongs to second degree of severity. As the weather condition gets more rain or snow involving, the accident severity will positively increase.

If we focus on time analysis result for this dataset, we can tell that the U.S. federal government should spend more effort on the traffic regulation and education promotion especially during Monday to Wednesday as well as the autumn and winter seasons. Both the regular commuters and logistics industry should pay more attention during these periods. In addition, by setting up the choropleth map in Python, it can illustrate that California is the state with most traffic accident cases ,which is nearly 178,436 cases. But in order to make the result more previse without being interfered by the states’ population, we especially tested the total accident case per population for each state of America. The local government of South Carolina and cities in the west coasts may have to formulate more complete and rigorous traffic law system to face the bad situation right now .

The future outlook of this study is to find out more specific features as new analytical features such as the accident vehicle types, perpetrators’ past driving path, age, driving time as well as the items associated with severity like injury numbers. Also, the traffic object features should be imported more by the data authority. The agency releasing this dataset can track more detail on the traffic object proximity. Furthermore, the zip-code column of this dataset can be optimized and even divided into smaller district area, which may enable researchers to make more precise analysis toward target area.

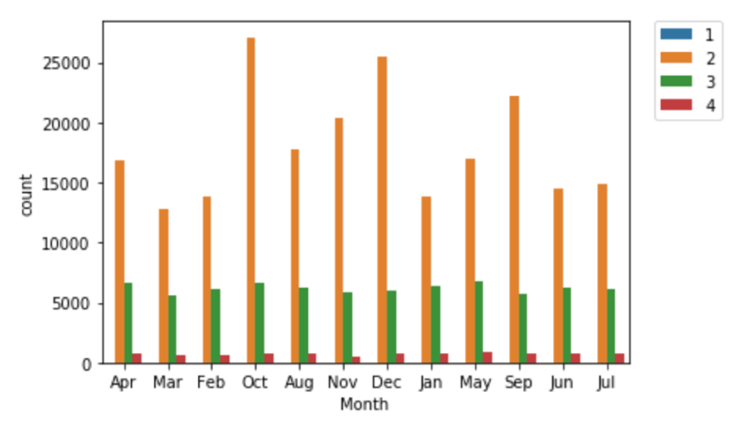


Figure 8: The Bar chart of traffic accident severity counts for every month in 2019. It tells that autumn and winter have the most accident cases for severity 2.

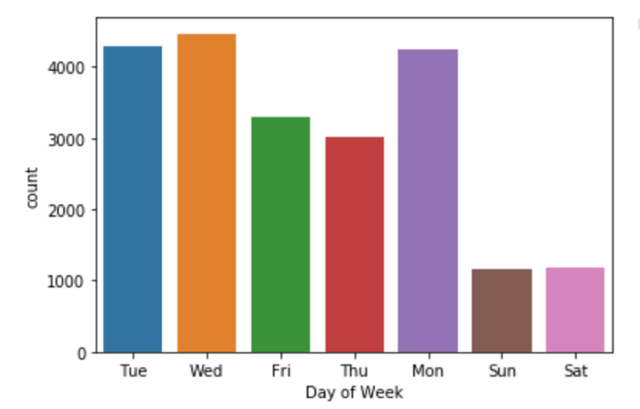


Figure 9: The bar chart of traffic accident numbers during every day of week. Monday until Wednesday holds the top.

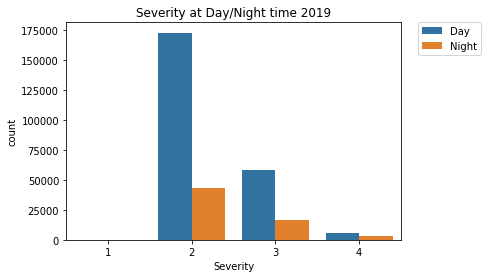


Figure 10: This bar chart indicates that the traffic accidents happened in day more often than night in 2019.

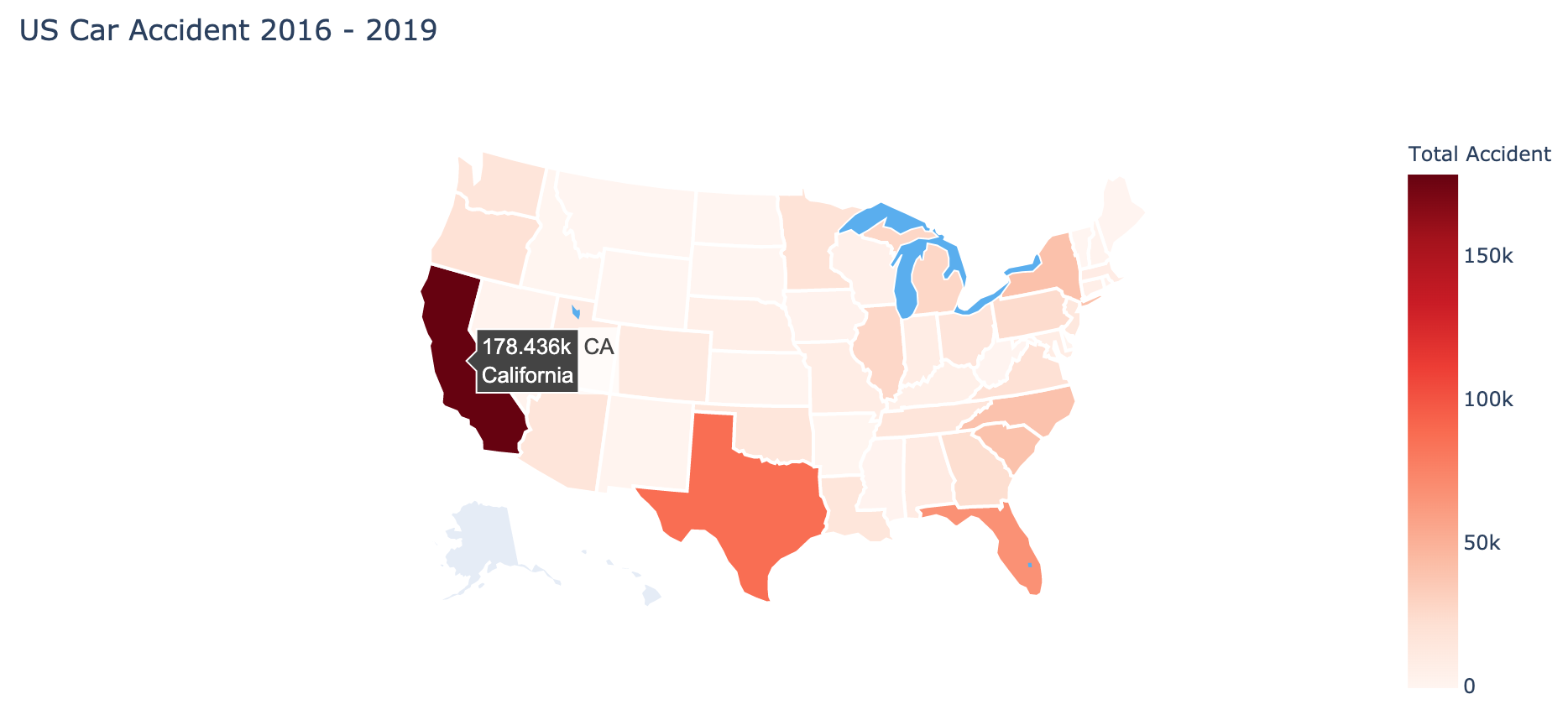


Figure 11: The choropleth map of US Car accident cases during 2016 ~ 2019 without considering population for each state. California, Texas as well as Florida take the top 3.

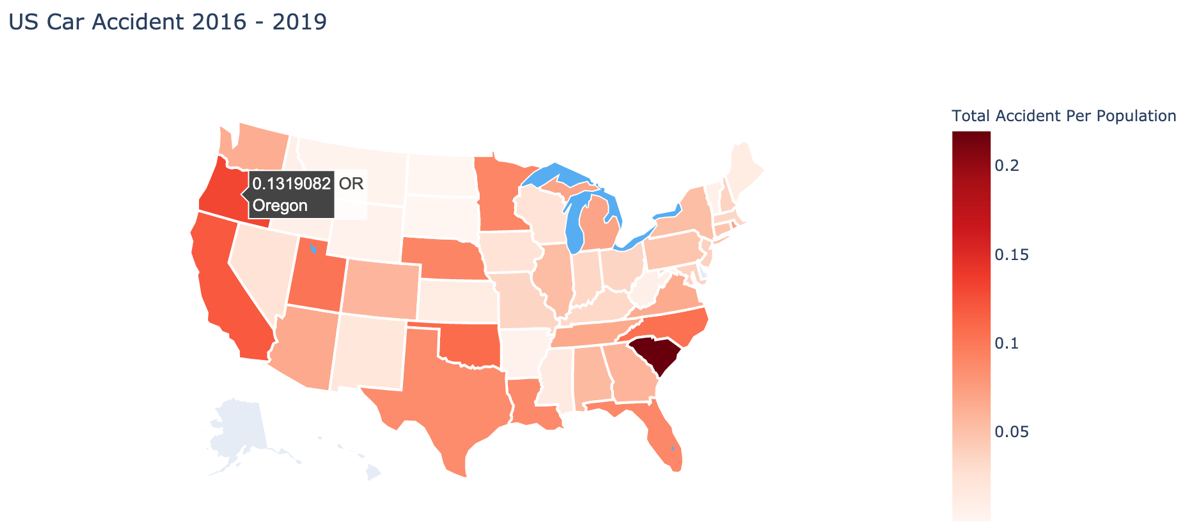


Figure 12: The choropleth map of US Car accident cases during 2016 ~ 2019 which considers the population. South Carolina owns the most accidents during this case.

# References

1. <https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death> [↑](#endnote-ref-1)
2. <https://www.asirt.org/safe-travel/road-safety-facts/#:~:text=Annual%20United%20States%20Road%20Crash,enough%20to%20require%20medical%20attention.> [↑](#endnote-ref-2)
3. <https://www.kaggle.com/sobhanmoosavi/us-accidents> [↑](#endnote-ref-3)
4. <https://www.sciencedirect.com/science/article/pii/S0895435618308710> [↑](#endnote-ref-4)
5. <https://medium.com/@danberdov/types-of-missing-data-902120fa4248> [↑](#endnote-ref-5)
6. <https://www.statisticshowto.com/probability-and-statistics/find-sample-size/> [↑](#endnote-ref-6)