You have **2** free member-only stories left this month. Sign up and get an extra one for free.

USA Accidents Data Analysis

The data of countrywide traffic accidents from February 2016 to March 2019 is analyzed.

Shubhankar Rawat Follow
Feb 21 · 9 min read ★

INTRODUCTION

> Road accidents have become very common these days. Nearly 1.25 million people die in road crashes each year, on average, 3,287 deaths a day. Moreover, 20–50 million people are injured or disabled annually. Road traffic crashes rank as the 9th leading cause of death and accounts for 2.2% of all deaths globally. Road crashes cost USD 518 billion globally, costing individual countries from 1–2% of their annual GDP.
>
> In the USA, over 37,000 people die in road crashes each year, and 2.35 million are injured or disabled. Road crashes cost the U.S. $230.6 billion per year or an average of $820 per person. Road crashes are the single greatest annual cause of death of healthy U.S. citizens travelling abroad.
>
> (Source)

Looking at the severity of road accidents, I decided to analyze the accidents' data to discover something useful. And here I am, sharing my results.

DATA

The dataset is taken from **Kaggle**. You can find it **here**.
This is a **countrywide traffic accident dataset**, which covers **49 states** of the United States. The data is continuously being collected from **February 2016 to March 2019**,

using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events 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.

The dataset contains **2,243,939(2.24 million)** rows and **49** columns(Quite a large dataset). A point to be noted is that even though the dataset contains data for only three years, there are 2.24 million accidents already.

Feature Description

As discussed earlier, the dataset contains 49 features, and the following is their description.

S.No.	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	Source	Indicates source of the accident report (i.e. the API which reported the accident.).
3	TMC	A traffic accident may have a Traffic Message Channel (TMC) code which provides more detailed description of the event.
4	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).
5	Start_Time	Shows start time of the accident in local time zone.
6	End_Time	Shows end time of the accident in local time zone.
7	Start_Lat	Shows latitude in GPS coordinate of the start point.
8	Start_Lng	Shows longitude in GPS coordinate of the start point.
9	End_Lat	Shows latitude in GPS coordinate of the end point.
10	End_Lng	Shows longitude in GPS coordinate of the end point.
11	Distance(mi)	The length of the road extent affected by the accident.
12	Description	Shows natural language description of the accident.
13	Number	Shows the street number in address field.
14	Street	Shows the street name in address field.
15	Side	Shows the relative side of the street (Right/Left) in address field.
16	City	Shows the city in address field.
17	County	Shows the county in address field.
18	State	Shows the state in address field.
19	Zipcode	Shows the zipcode in address field.
20	Country	Shows the country in address field.
21	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
22	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
23	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).
24	Temperature(F)	Shows the temperature (in Fahrenheit).
25	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
26	Humidity(%)	Shows the humidity (in percentage).
27	Pressure(in)	Shows the air pressure (in inches).
28	Visibility(mi)	Shows visibility (in miles).
29	Wind_Direction	Shows wind direction.
30	Wind_Speed(mph)	Shows wind speed (in miles per hour).
31	Precipitation(in)	Shows precipitation amount in inches, if there is any.
32	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
33	Amenity	A POI annotation which indicates presence of amenity in a nearby location.
34	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
35	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
36	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
37	Junction	A POI annotation which indicates presence of junction in a nearby location.
38	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.

39	Railway	A POI annotation which indicates presence of railway in a nearby location.
40	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
41	Station	A POI annotation which indicates presence of station in a nearby location.
42	Stop	A POI annotation which indicates presence of stop in a nearby location.
43	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.
44	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.
45	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
46	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
47	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
48	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.
49	Astronomical_Twili	Shows the period of day (i.e. day or night) based on astronomical twilight.

Feature description of the dataset

THE APPROACH

Before getting into the analysis part, let's look at the null values present in the dataset.

The figure below shows only those features which have null values.

```
TMC                    516762
End_Lat               1727177
End_Lng               1727177
Description                 1
Number                1458402
City                       68
Zipcode                   646
Timezone                 2141
Airport_Code            23664
Weather_Timestamp       47170
Temperature(F)          62265
Wind_Chill(F)         1852370
Humidity(%)             64467
Pressure(in)            57280
Visibility(mi)          71360
Wind_Direction          47190
Wind_Speed(mph)        442954
Precipitation(in)     1979466
Weather_Condition       72004
Sunrise_Sunset             78
Civil_Twilight             78
Nautical_Twilight          78
Astronomical_Twilight      78
```

Features with null values

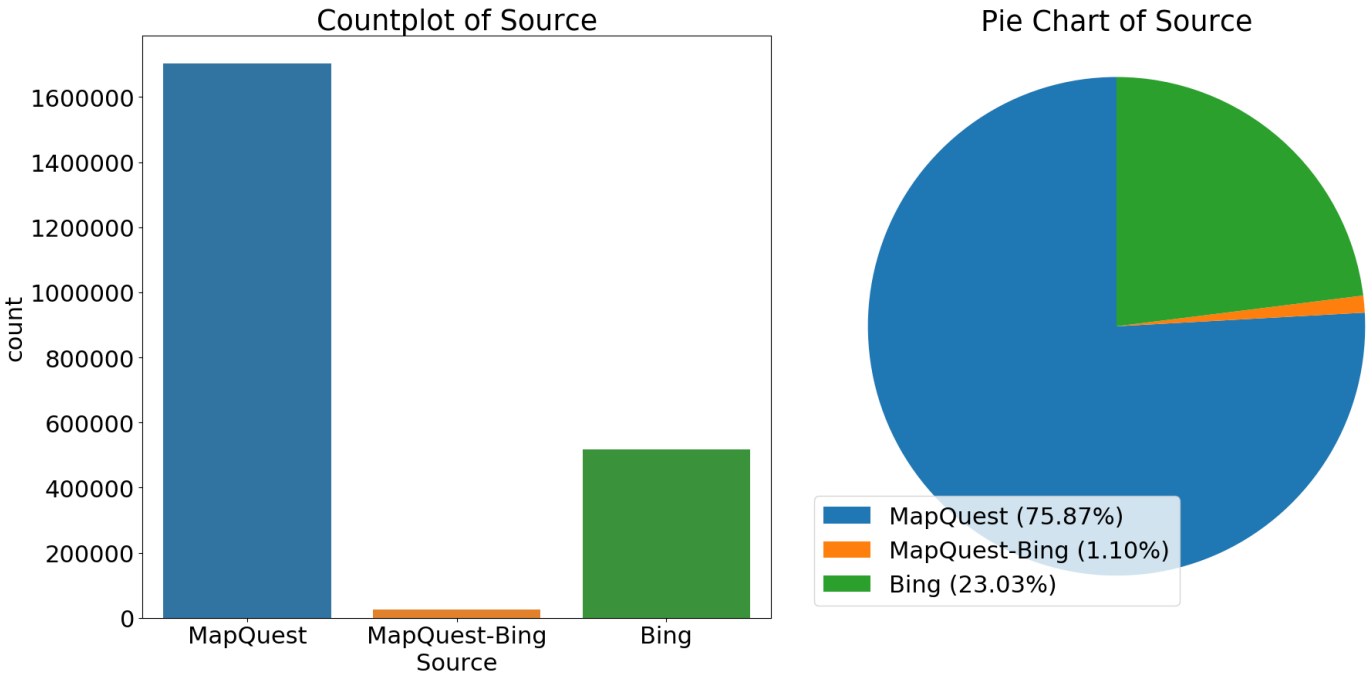
Exploratory Data Analysis

I will start by eliminating the unnecessary features first.

The feature *Country* contains only one entry — USA, which is quite apparent since we are dealing with the USA's dataset. Hence, I will be deleting this feature.

The feature *Turning_Loop* also contains one value — False. This means that there was no turning loop in the vicinity of any of the accidents. As this feature includes only one value, I'll be dropping this as well.

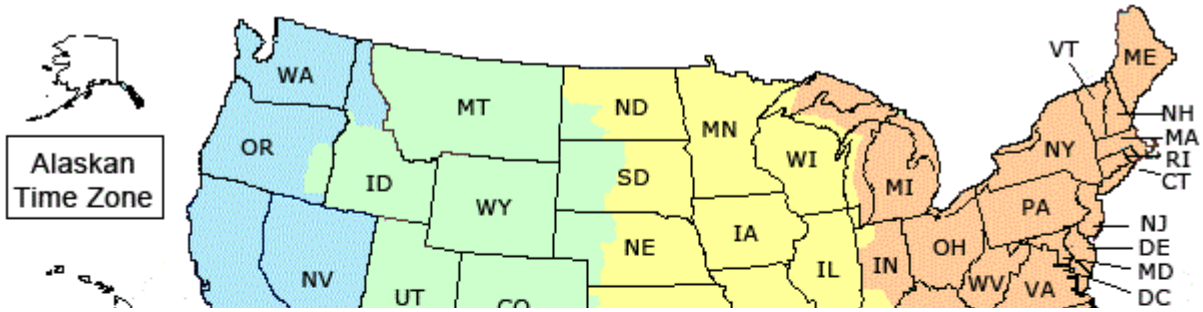
Let's look at the *Source* feature. It represents the API that reported the accident.

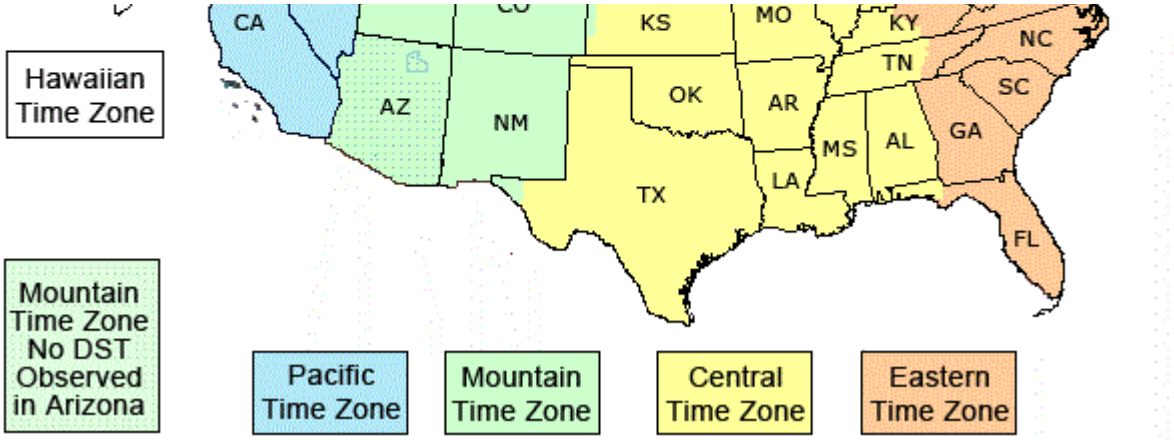


Number of accidents reported by each source

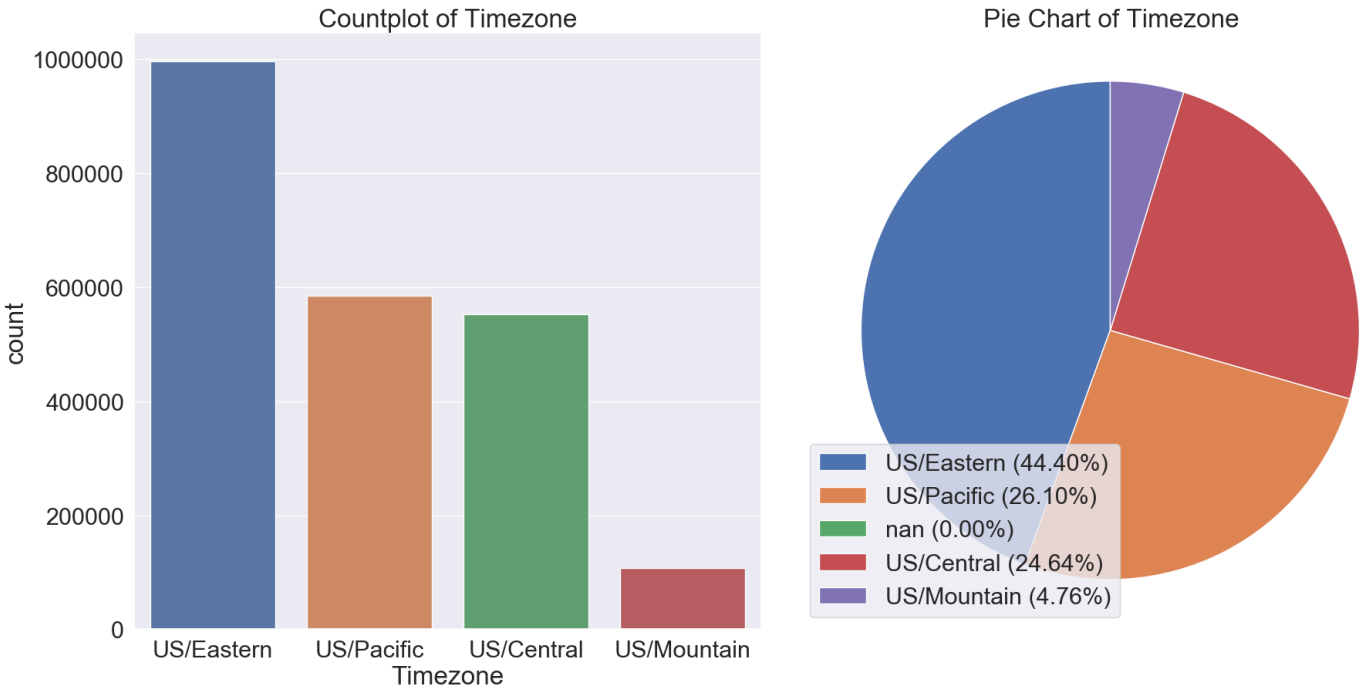
There are only three API sources that reported the accidents. It can be observed that most of the accidents(around 1,700,000) were reported by MapQuest, followed by Bing.

There are **nine** standard **time zones** in the US officially defined by federal law. The entire 50 states and the District of Columbia have **six** main time zones.





Out of these six significant timezones, we see(from the figure below) that there are only four timezones present in the dataset.



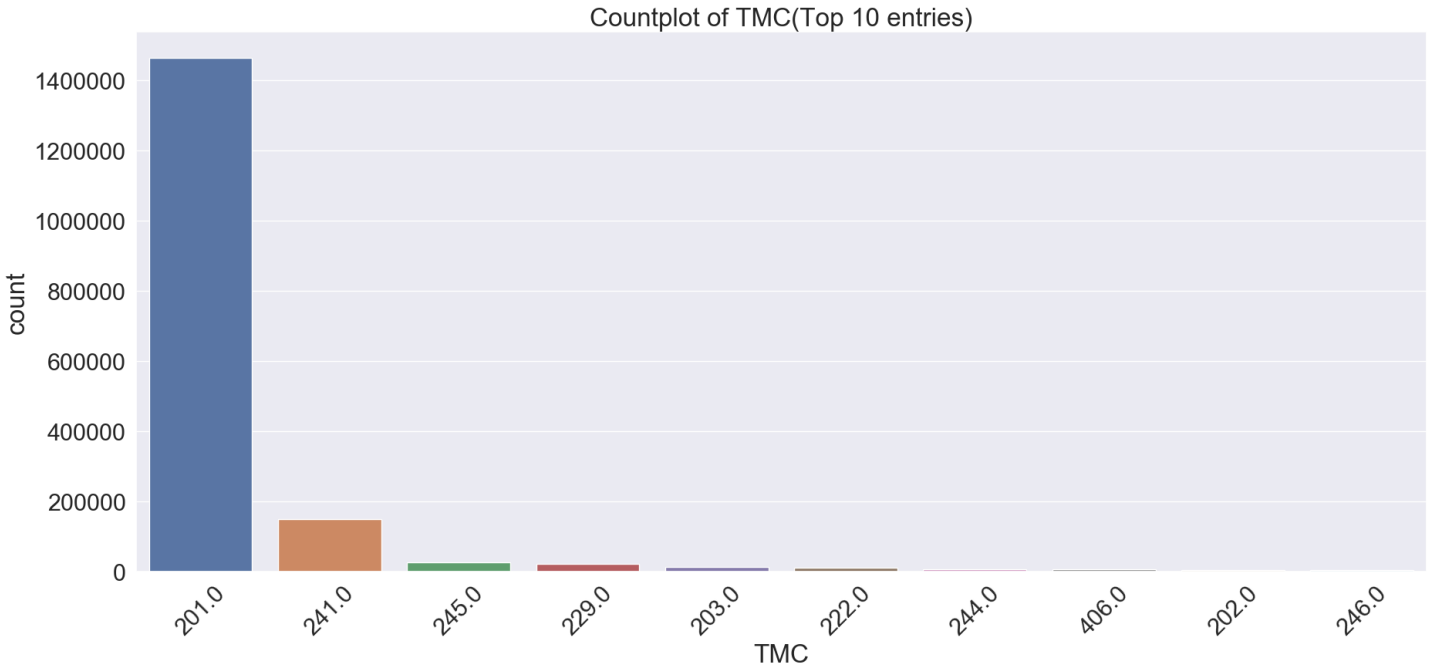
Most numbers of accidents took place in regions with timezone: Eastern Standard Time followed by Pacific Standard Time.

The dataset consists of an exciting feature: *TMC*.
A quick google search gives the following:

Traffic Message Channel (TMC) is a technology for delivering traffic and travel information to motor vehicle drivers. It is digitally coded using the ALERT C or TPEG protocol into RDS Type 8A groups carried via conventional FM radio broadcasts. It can

also be transmitted on Digital Audio Broadcasting or satellite radio. TMC allows silent delivery of dynamic information suitable for reproduction or display in the user's language without interrupting audio broadcast services.

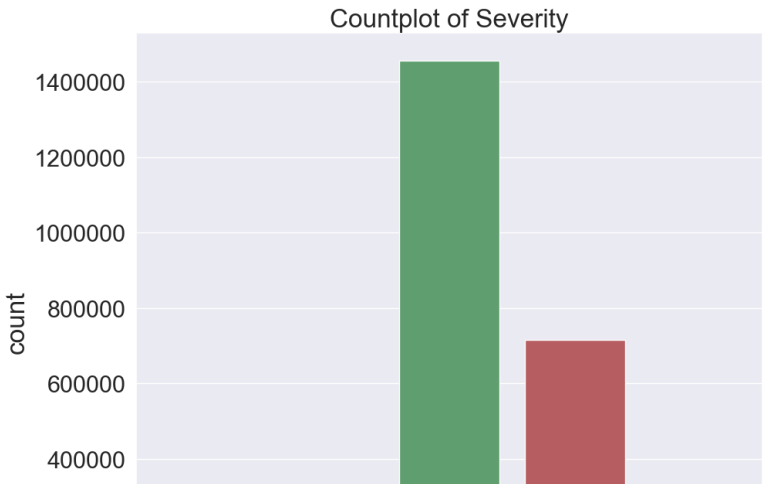
Now, let us plot the number of accidents with respect to the *TMC* feature.

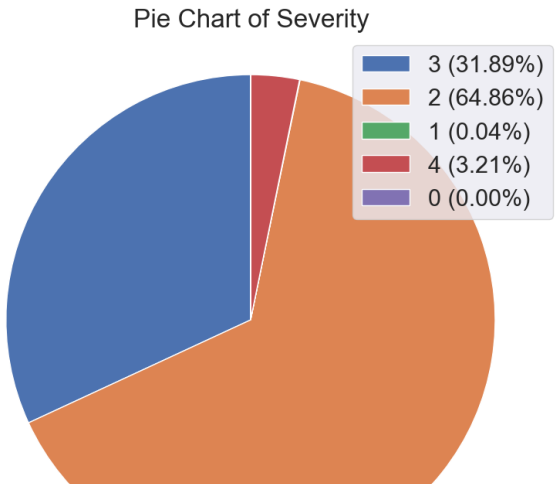


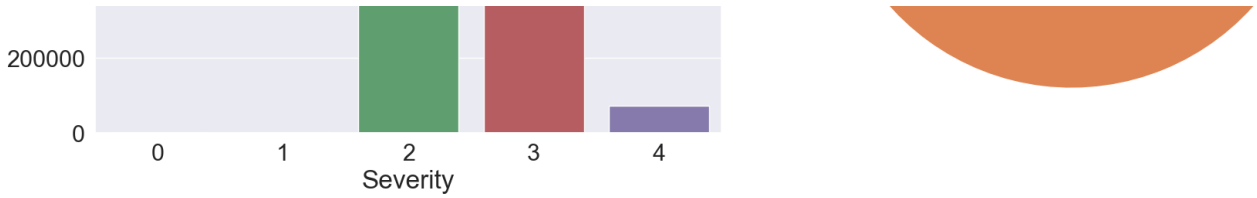
Number of accidents vs TMC

The plot depicts that most numbers of accidents have a TMC of 201. You can refer to the TMC code list for better understanding.

The most exciting feature is the *Severity*. It represents the severity of an accident.







Number of accidents for each severity

The plot depicts that mostly the accidents had severity equal to 2(average) followed by 3(above average), which is unfortunate. There are hardly any accidents with very low severity(0 and 1).

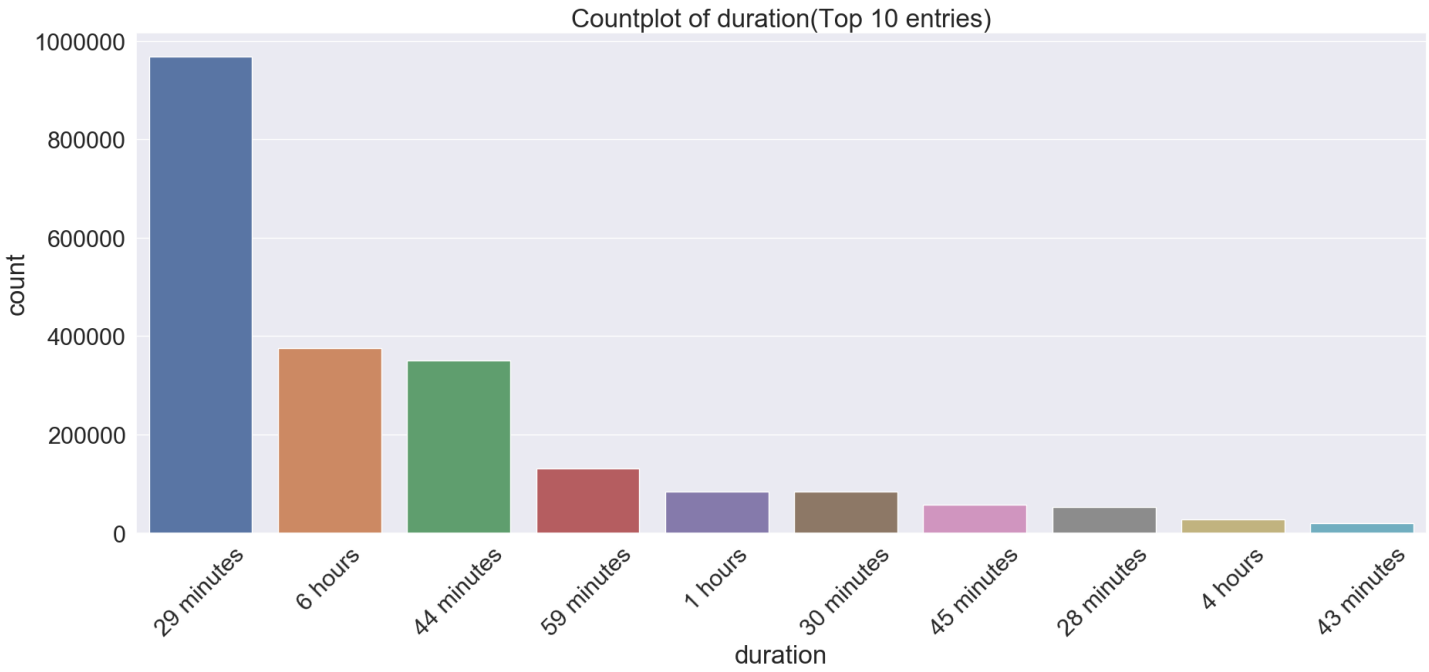
Let's look at the *Start_Time* and *End_Time* features:

Index	Start_Time	End_Time
0	08-02-2016 05:46	08-02-2016 11:00
1	08-02-2016 06:07	08-02-2016 06:37
2	08-02-2016 06:49	08-02-2016 07:19
3	08-02-2016 07:23	08-02-2016 07:53
4	08-02-2016 07:39	08-02-2016 08:09
5	08-02-2016 07:44	08-02-2016 08:14
6	08-02-2016 07:59	08-02-2016 08:29
7	08-02-2016 07:59	08-02-2016 08:29
8	08-02-2016 08:00	08-02-2016 08:30
9	08-02-2016 08:10	08-02-2016 08:40
10	08-02-2016 08:14	08-02-2016 08:44

Start_Time and End_Time features

The *Start_Time* and *End_Time* features depict the start and end time of an accident. To gain a better understanding, I have computed the duration of each accident.

It is interesting to see that the duration of accidents varies from minutes to years.

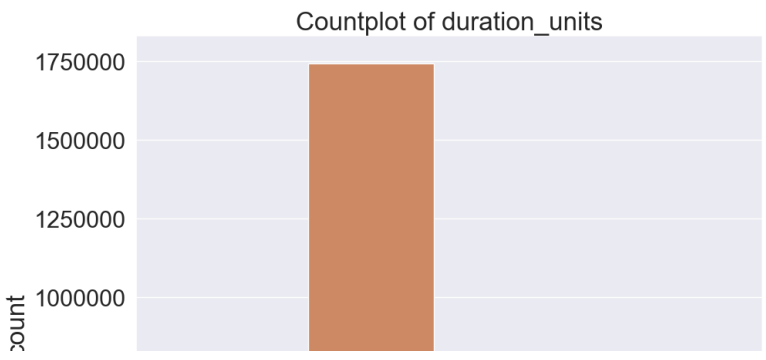


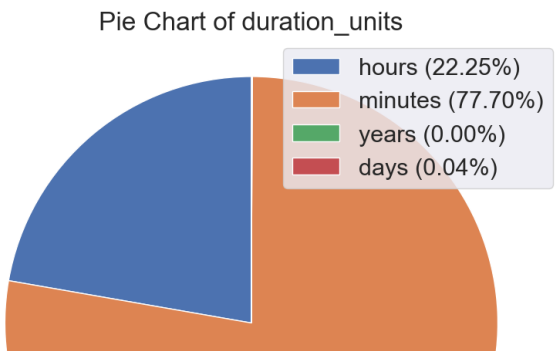
Number of accidents vs duration for top 10 values

The above plot is not as significant as the one that follows, but, I wanted to check the most common durations. Around 950,000(43%) accidents had a duration of 29 minutes, followed by 6 hours.

A point to be noted is that the dataset description tells that *Start_Time* and *End_Time* represent the starting and ending time of the accident. Although the duration(which is calculated by taking the difference between *Start_Time* and *End_Time*) for some accidents comes out to be in hours or even in months and years, it is not evident that an accident lasted for a few days or years. Maybe the two features also include the repair time as well, not much can be concluded.

A more significant way to look at the duration for each accident is to look at the duration unit.

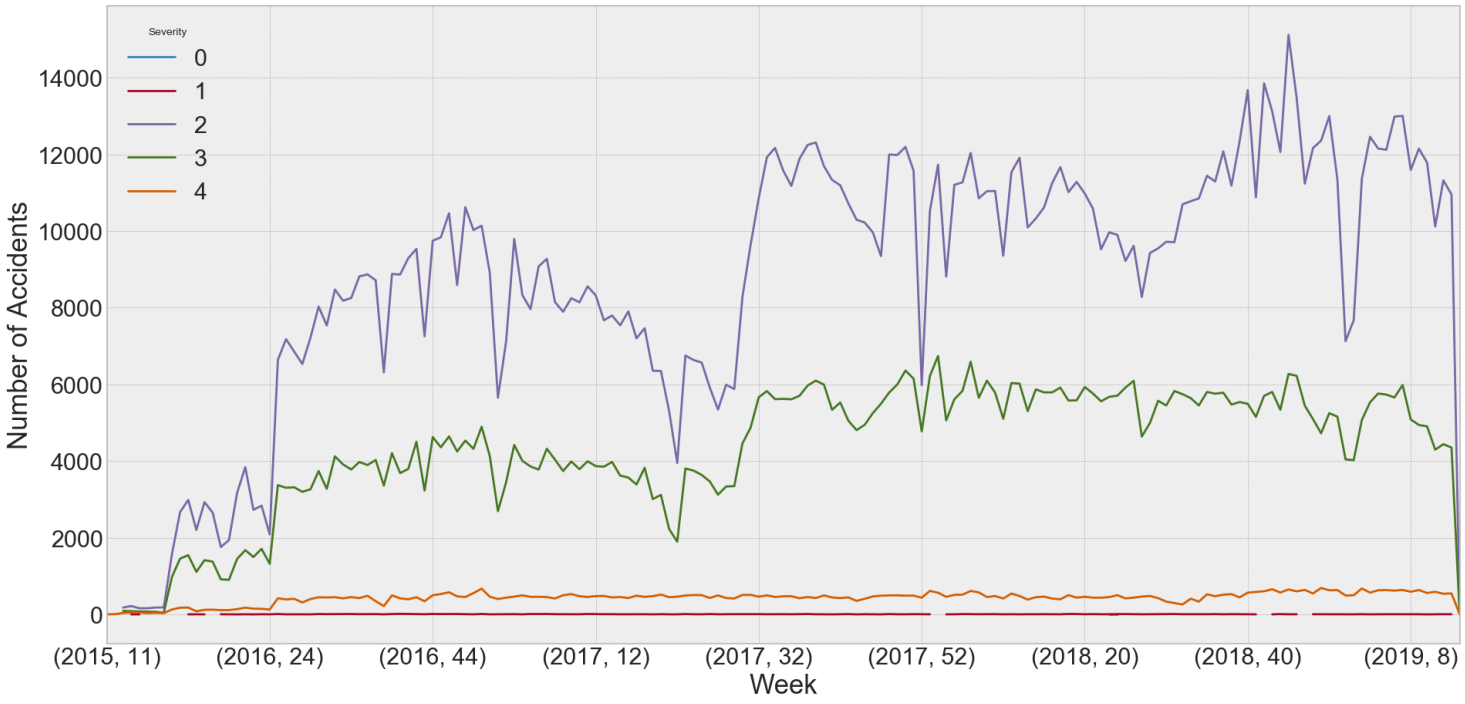




Number of accidents for each duration unit

The plot depicts that about 77% of the accidents have a duration in minutes, whereas about 22% of the accidents have a duration in hours. Only 52 accidents have a duration in years and 975 in days. This means that accidents that have a longer duration are rare in the USA. Also, accidents with small durations are much more frequent.

Now, let's see the trend of accidents for each severity over time.

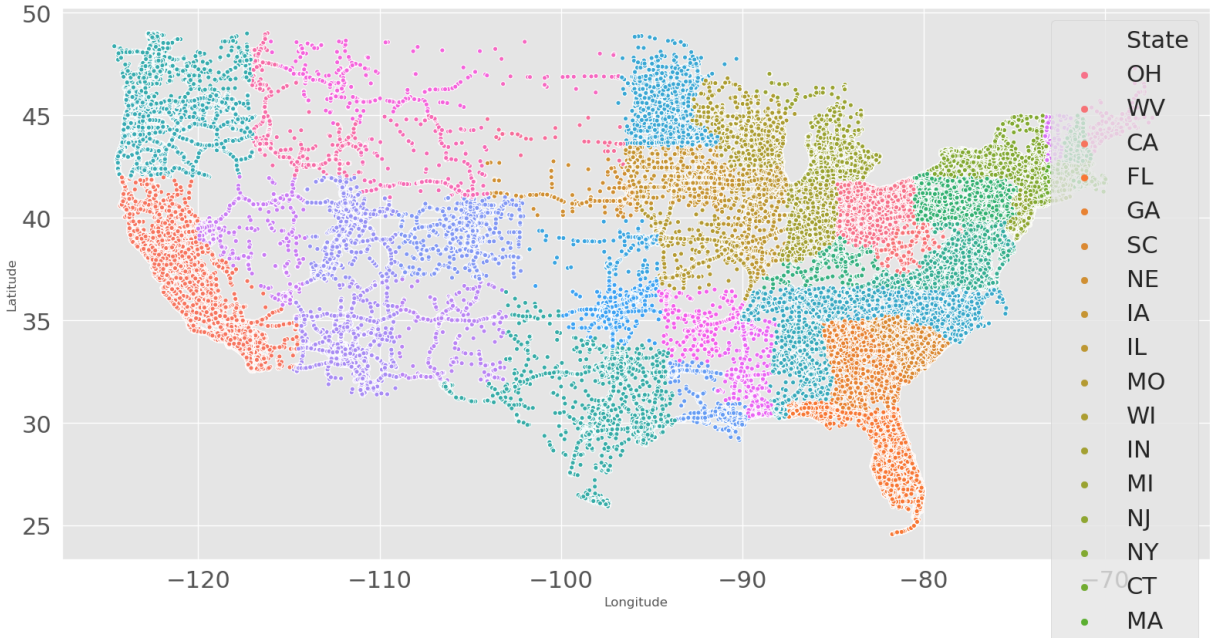


Number of accidents vs time for each severity

It can be observed that the number of accidents has increased over time for each severity. This is alarming and requires serious action. Even though there was a decrease in the number of accidents in 2017, around week 12, the number of accidents increased

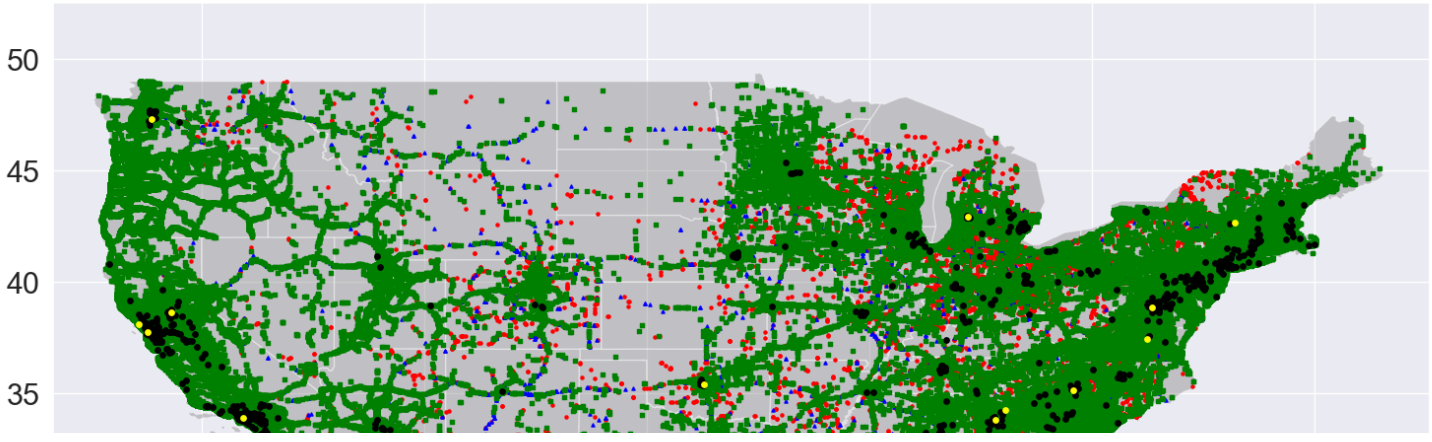
after that. Accidents with severity = 2 are more frequent and have increased the most, followed by accidents with severity = 3.

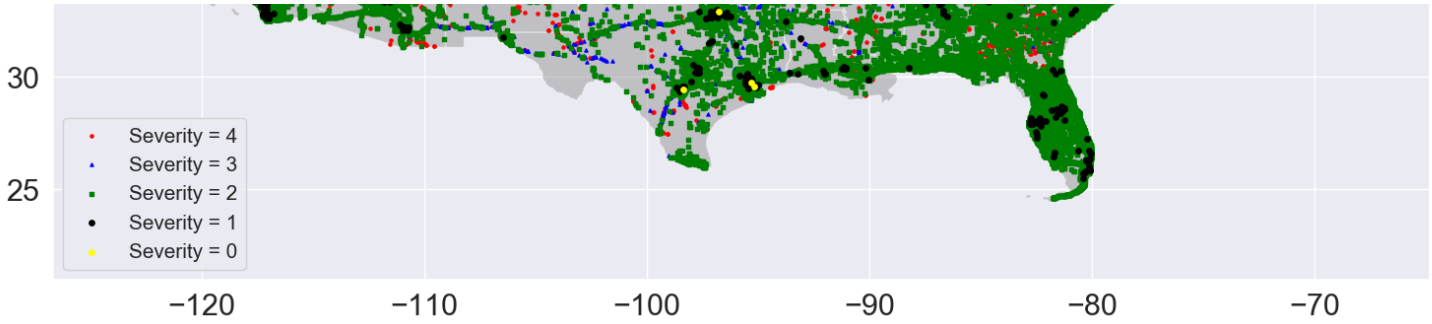
The *Start_Lat* and *Start_Lng* features are interesting since they can be plotted on a map, to get the exact location of the accident. First, I will draw a scatterplot between the two.



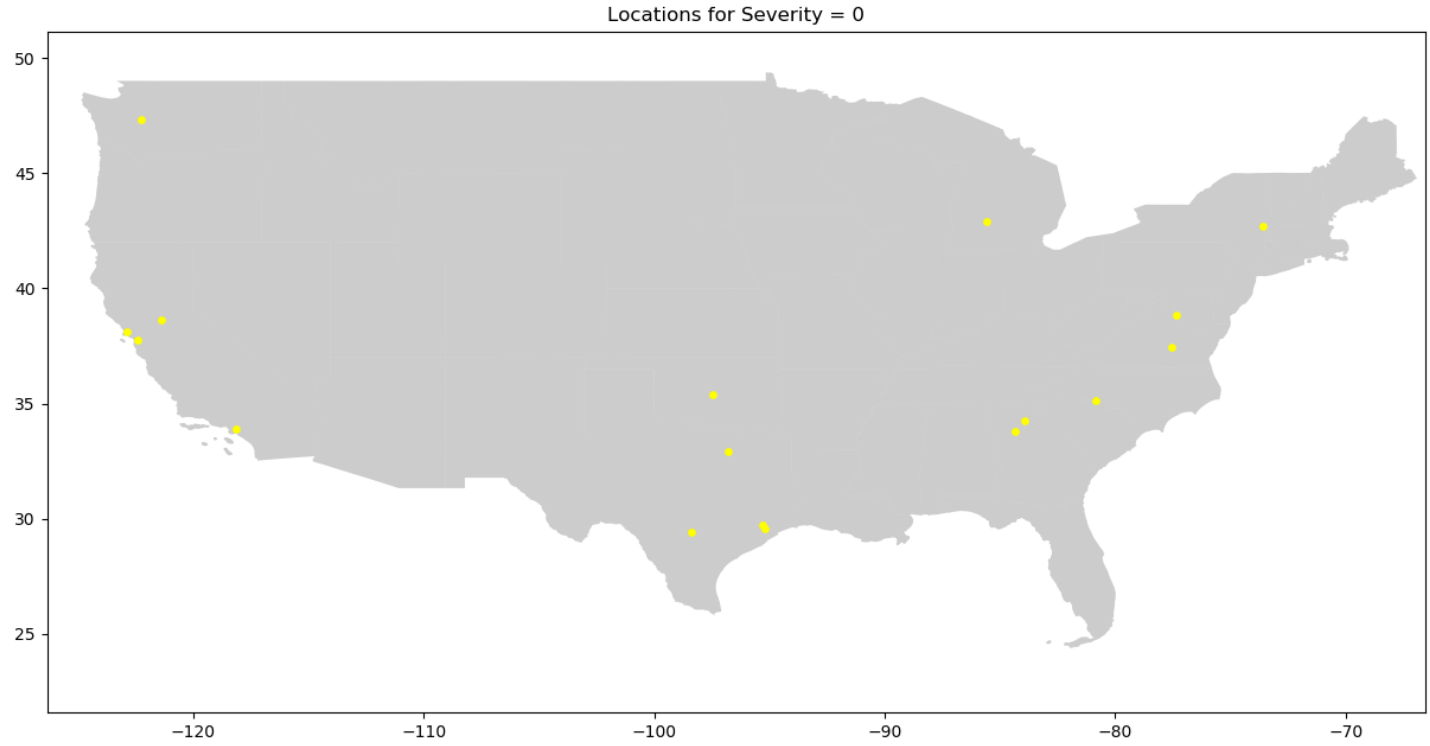
The scatterplot looks nice, but at the same time, it is alarming that almost every corner of the USA is covered, meaning that the accidents occurred over a large number of locations over the past few years.

To get a clear idea, I have plotted the accident's site using the coordinates given in the dataset on the USA map for each severity.

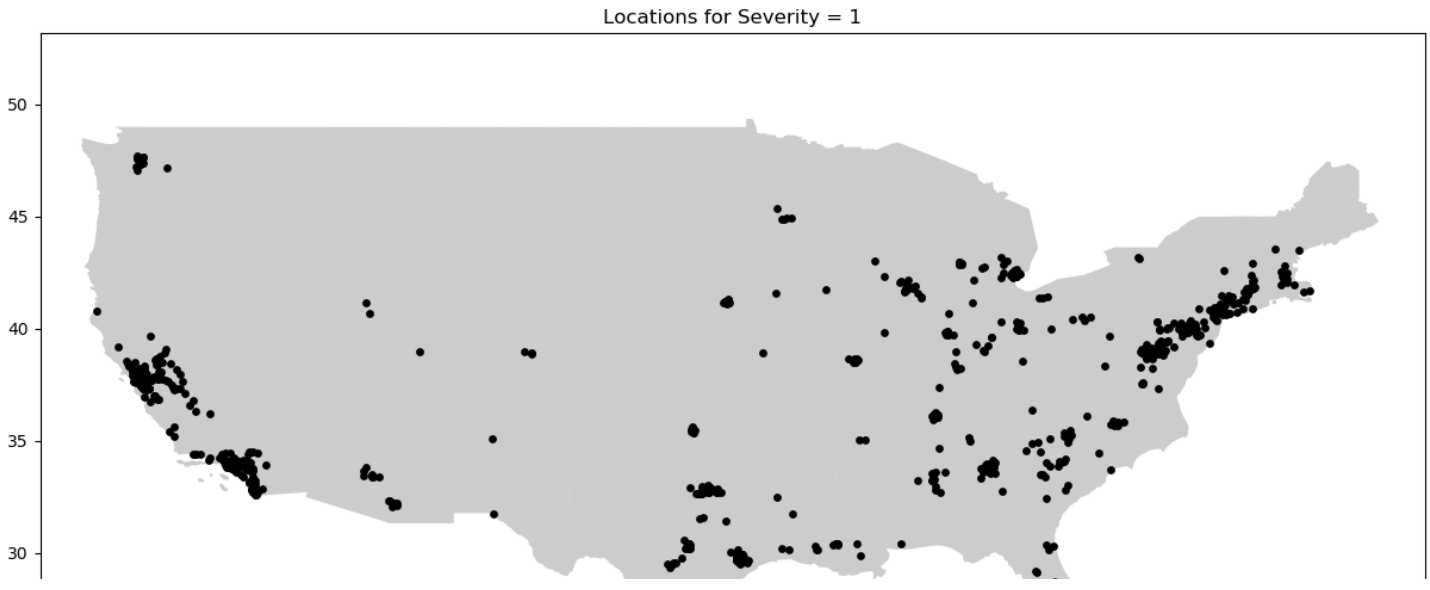




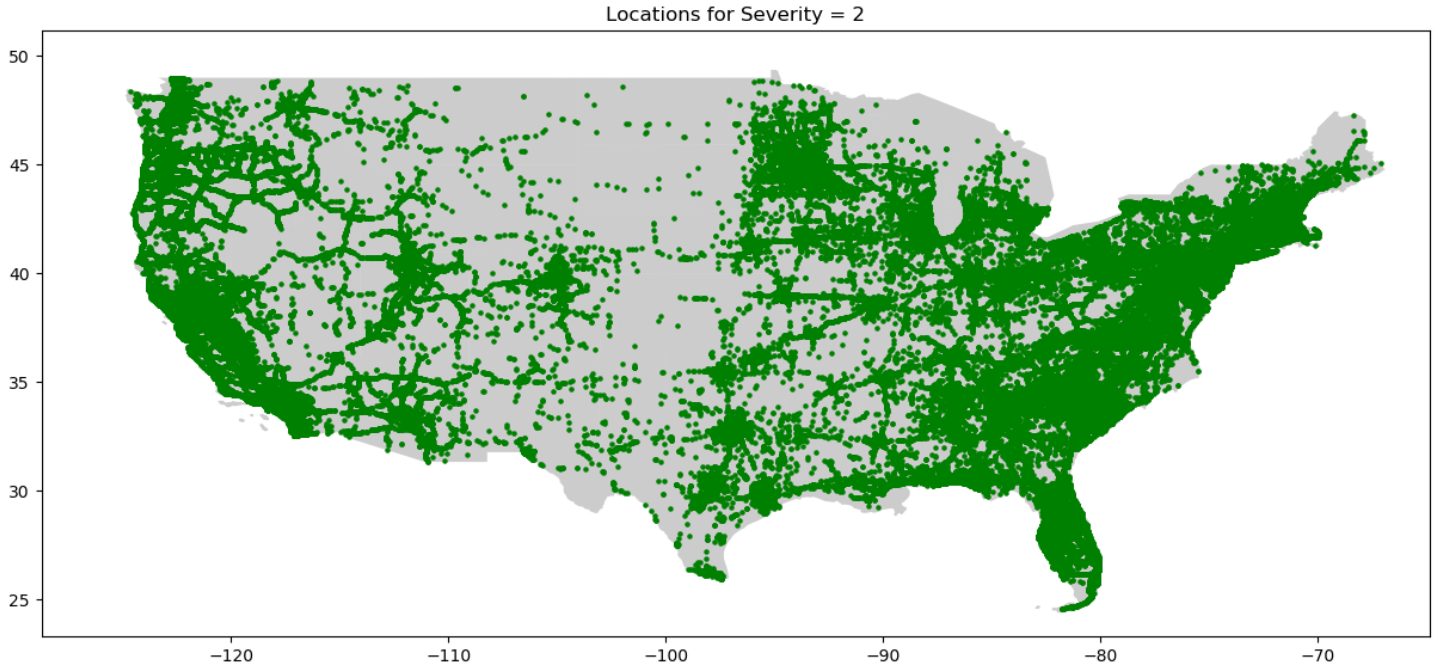
The above plot looks messy! So let's break it down.



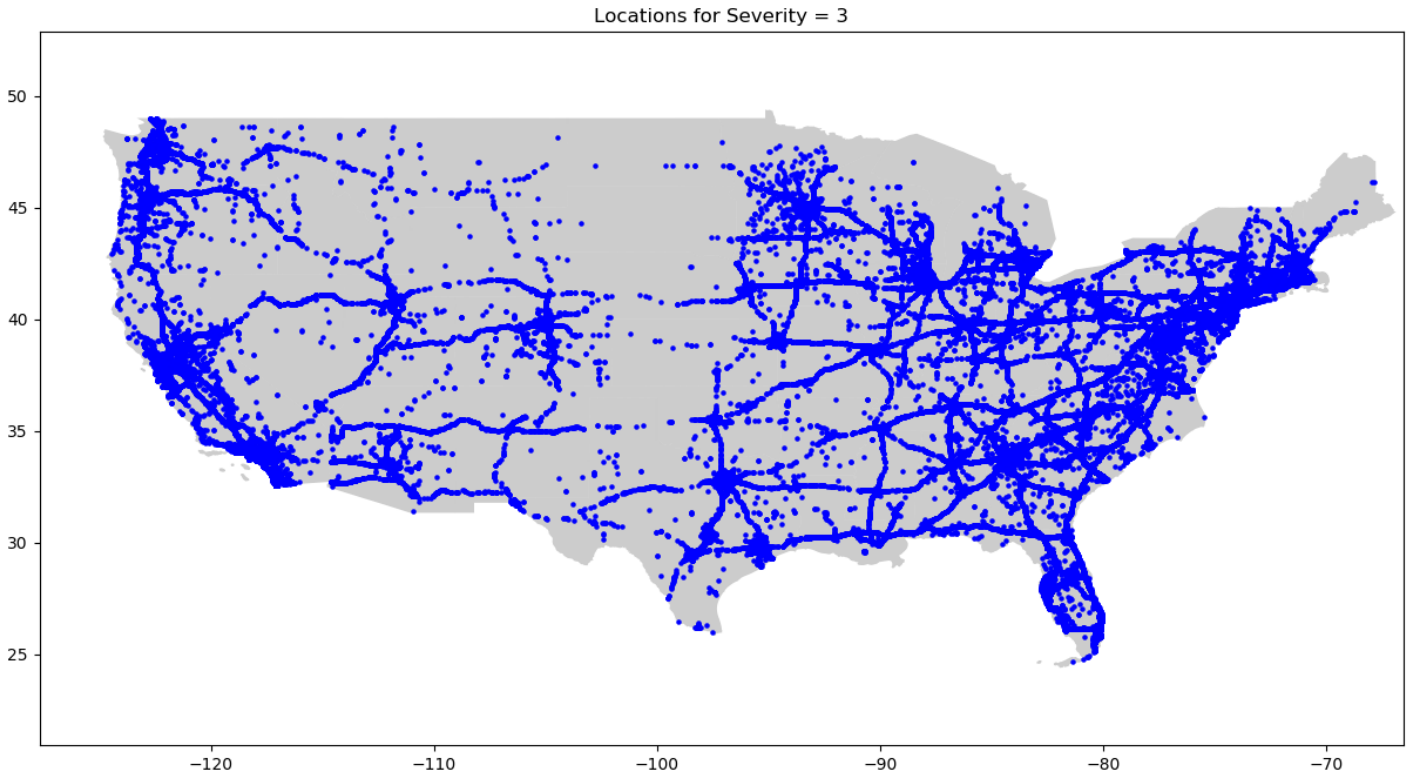
Accidents for severity = 0



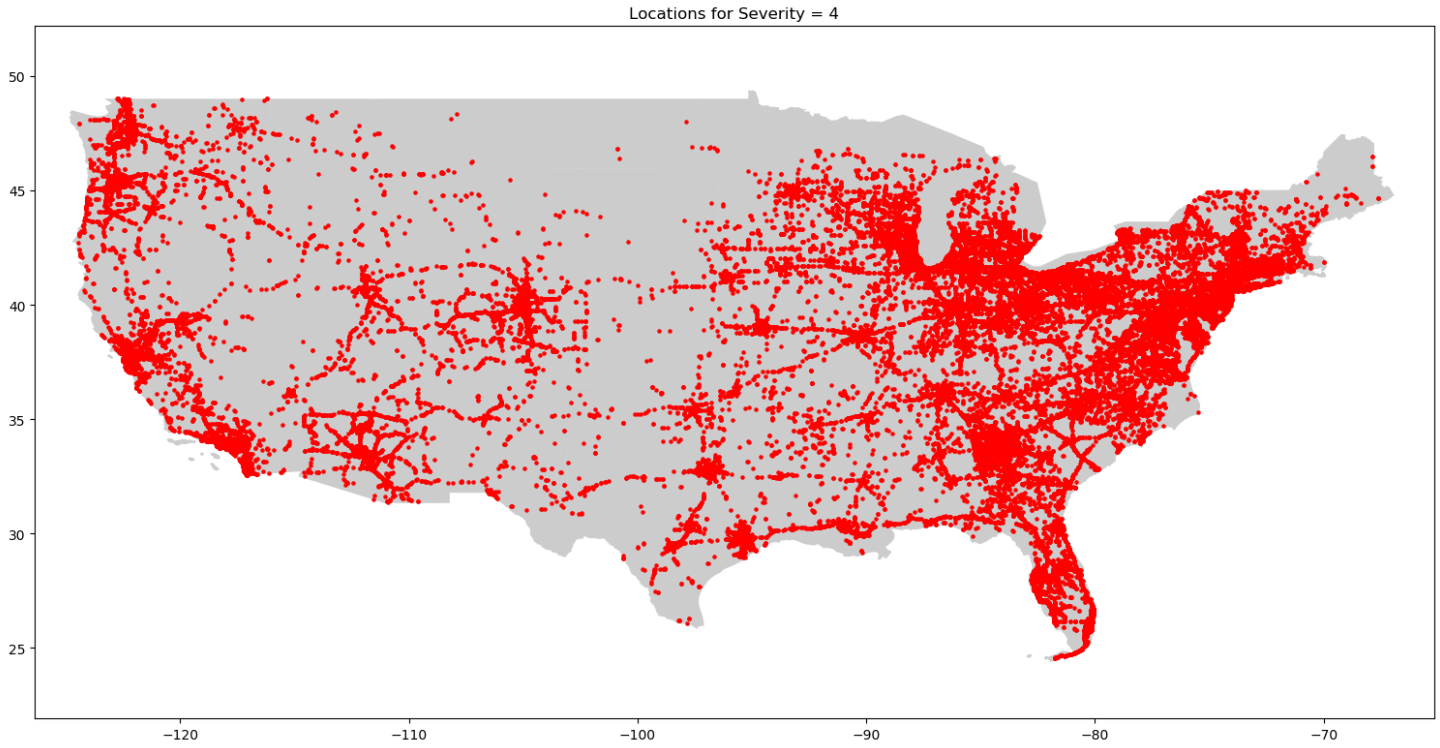
Accidents for severity = 1



Accidents' locaitons for severity = 2



Accidents' locations for severity = 3

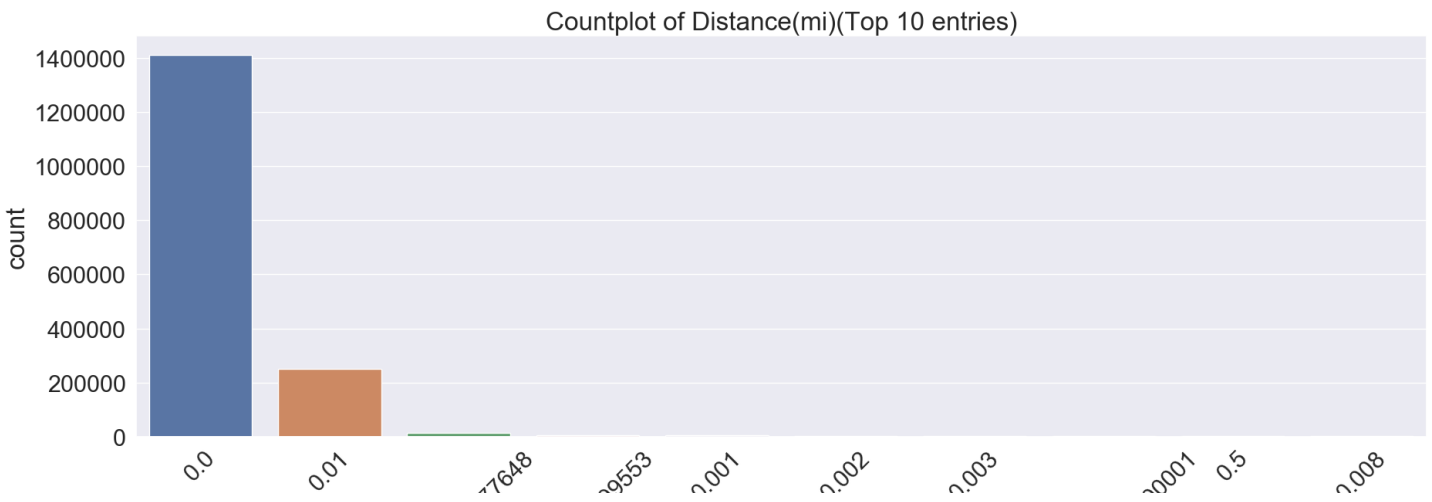


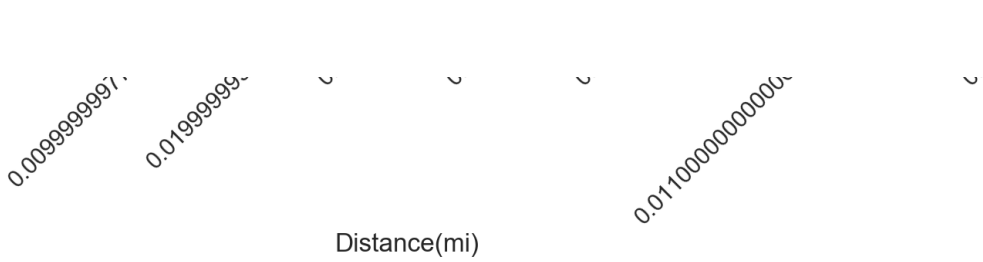
Accidents' locations for severity = 4

I have plotted the locations for each severity individually.

From the above plots, we can conclude that most numbers of accidents occurred in the Eastern and Western part of the USA, which accounts for the fact that most numbers of accidents took place in regions with timezone: Eastern Standard Time and Pacific Standard Time.

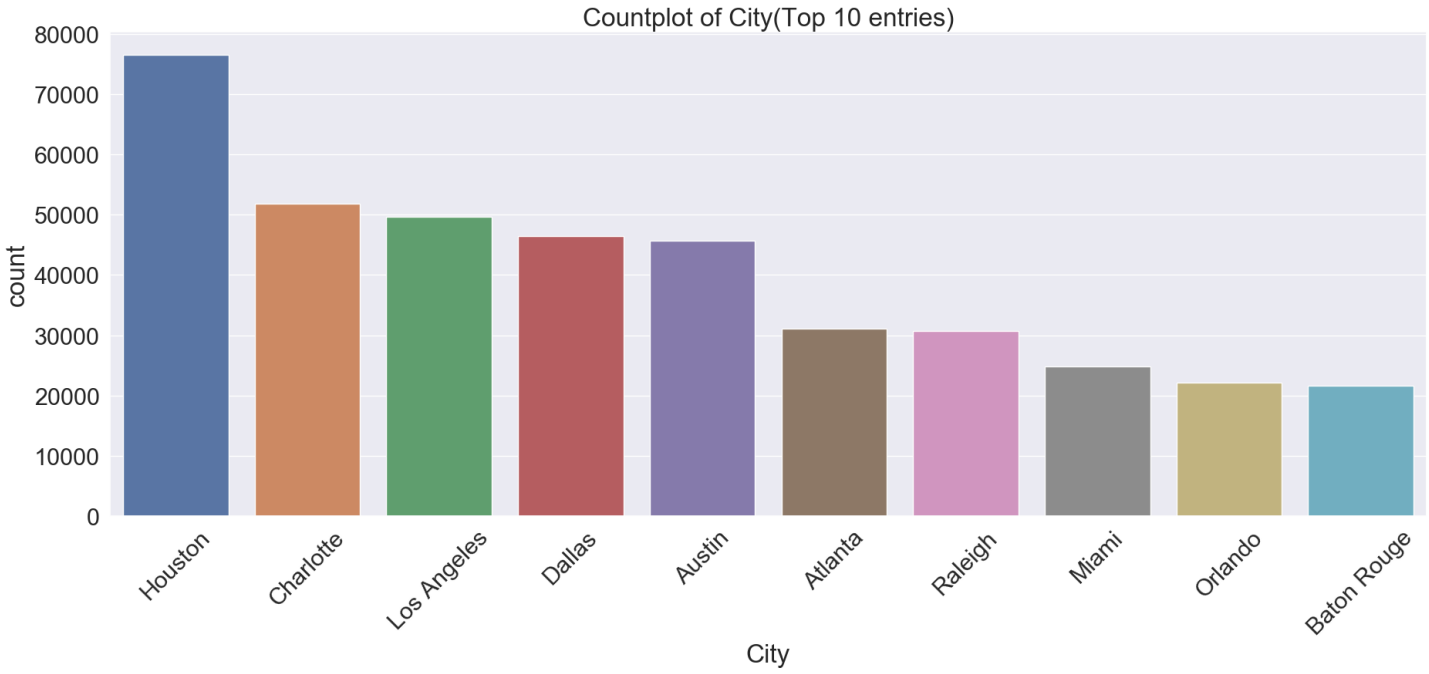
Now, let's look at the *Distance(mi)* feature. This feature tells the length(in miles) of the road extent affected by accident.





The above plot depicts that the impact of most of the accidents on the road is small.

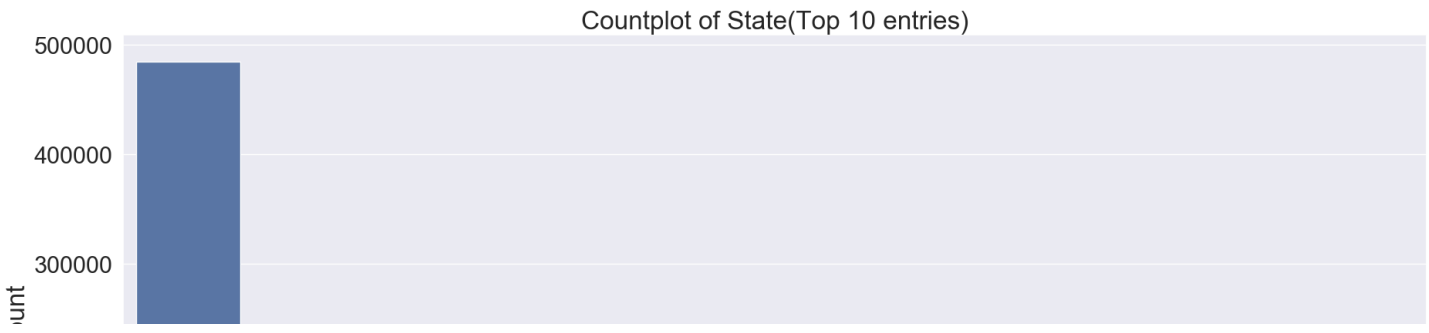
Now let's look at the most accident-prone cities in the USA.

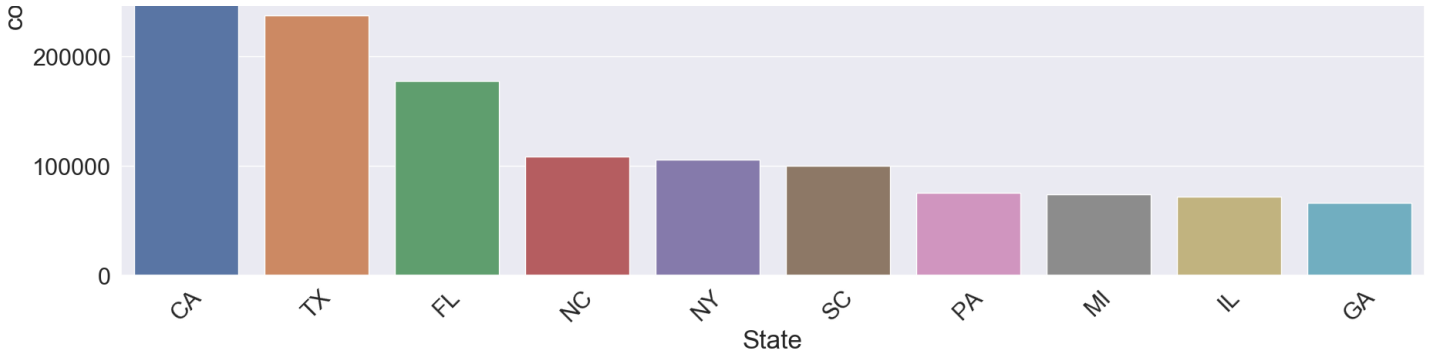


Top 10 cities in terms of number of accidents

We see that most of the accidents occur in Houston, followed by Charlotte and Los Angeles.

Let's look at most accident-prone states of USA





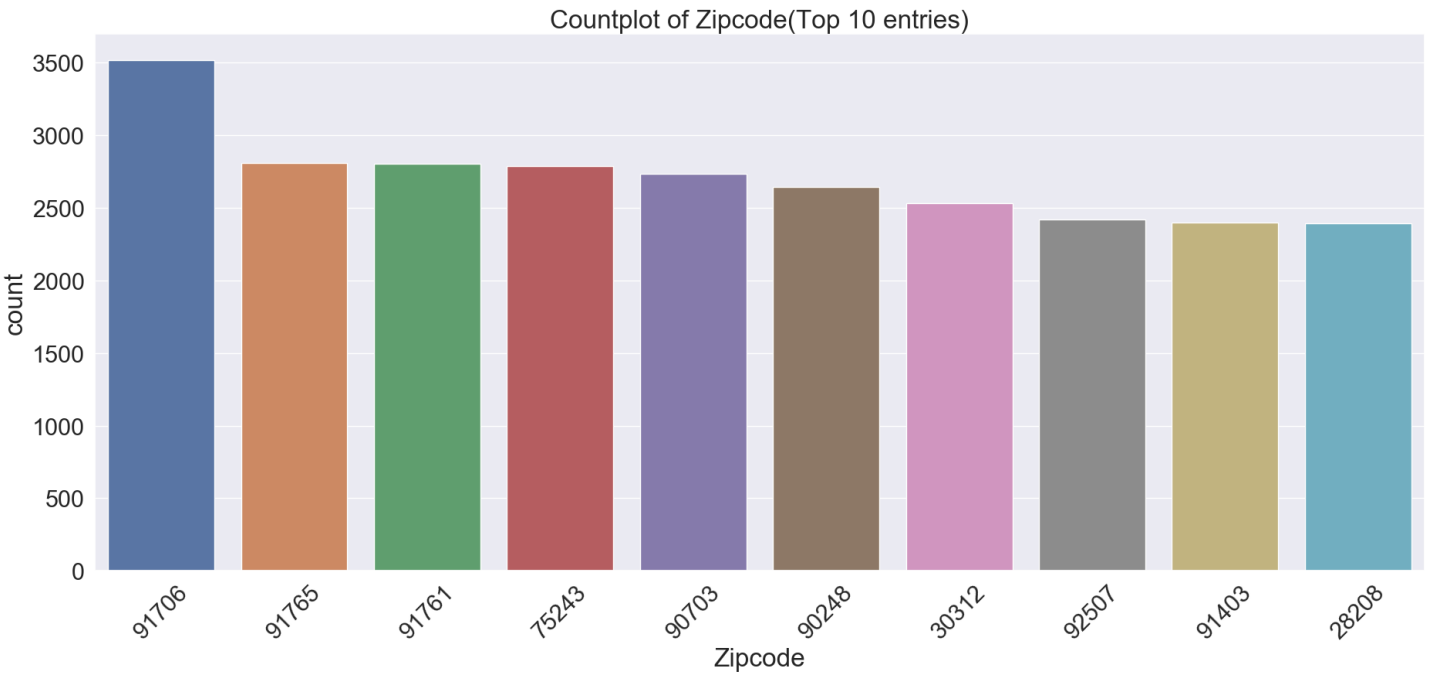
Top 10 states in terms of number of accidents

The list of states of the USA with their code is given here.

The plot depicts that California(CA) has the most number of accidents followed by Texas(TX) and Florida(FL). It is interesting to see that the number of accidents in California is almost twice the number of accidents in Texas.

We can see that the most accident-prone city in the USA is Houston which is in Texas followed by Charlotte(North Carolina — which is number 4) and Los Angeles(California).

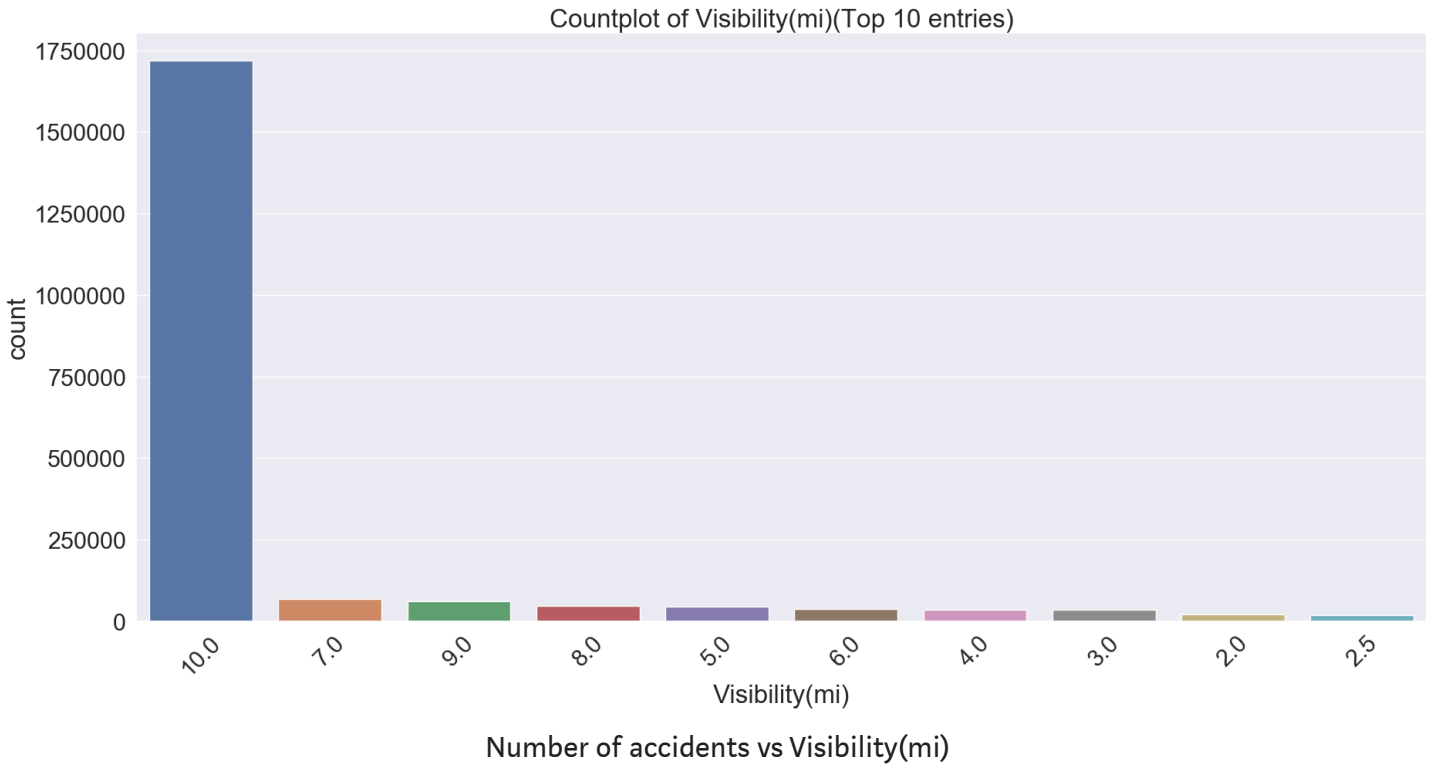
Let's go even deeper and plot the number of accidents with respect to zipcode.



Number of accidents vs zipcode

It can be observed that most numbers of accidents occurred in the region with zip code 91706 followed by 91765 and 91761. Refer to this link for more information about zip codes.

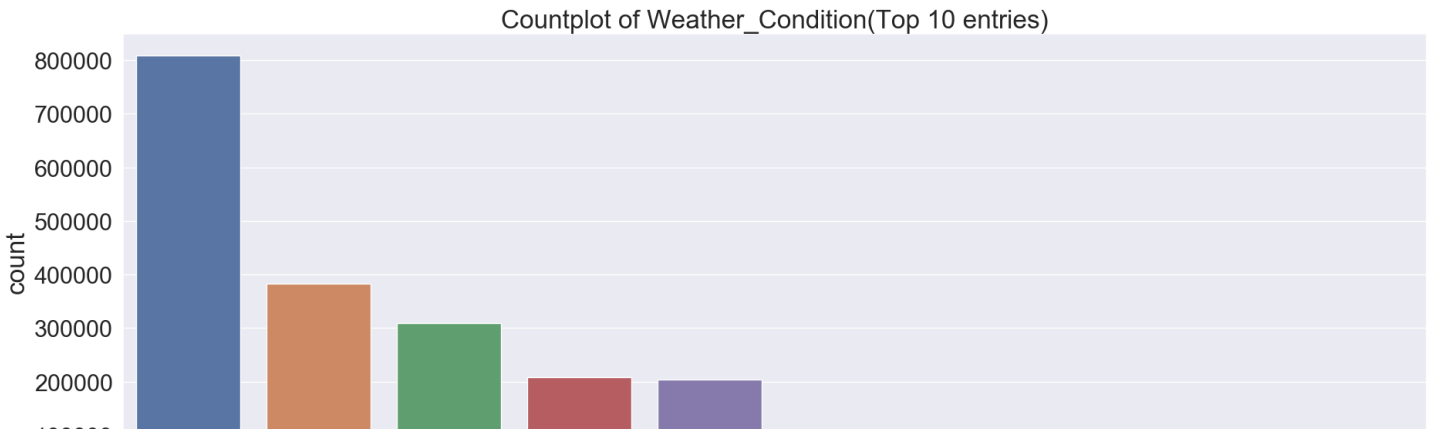
Now let's see the *visibility(mi)* feature. It denotes the visibility in miles.

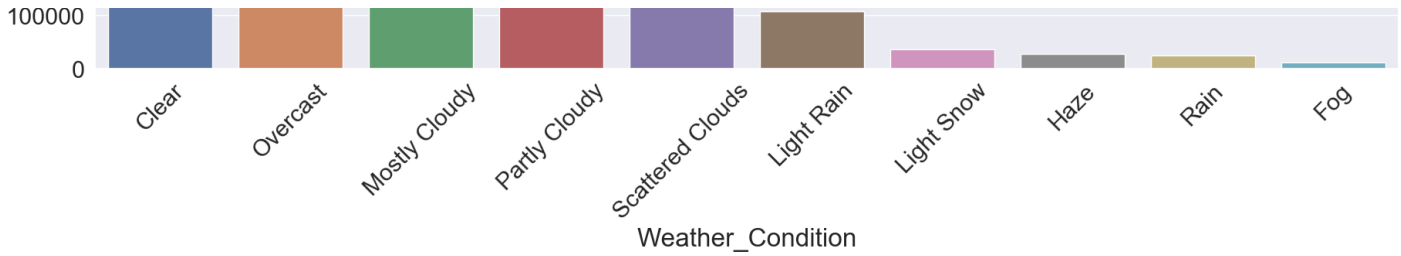


Number of accidents vs Visibility(mi)

The plot shows that most of the accidents occurred when visibility was high, which means that visibility is not a significant concern when it comes to accidents. This is obvious since low visibility is not the only factor.

Now let's see the weather conditions during the accidents.

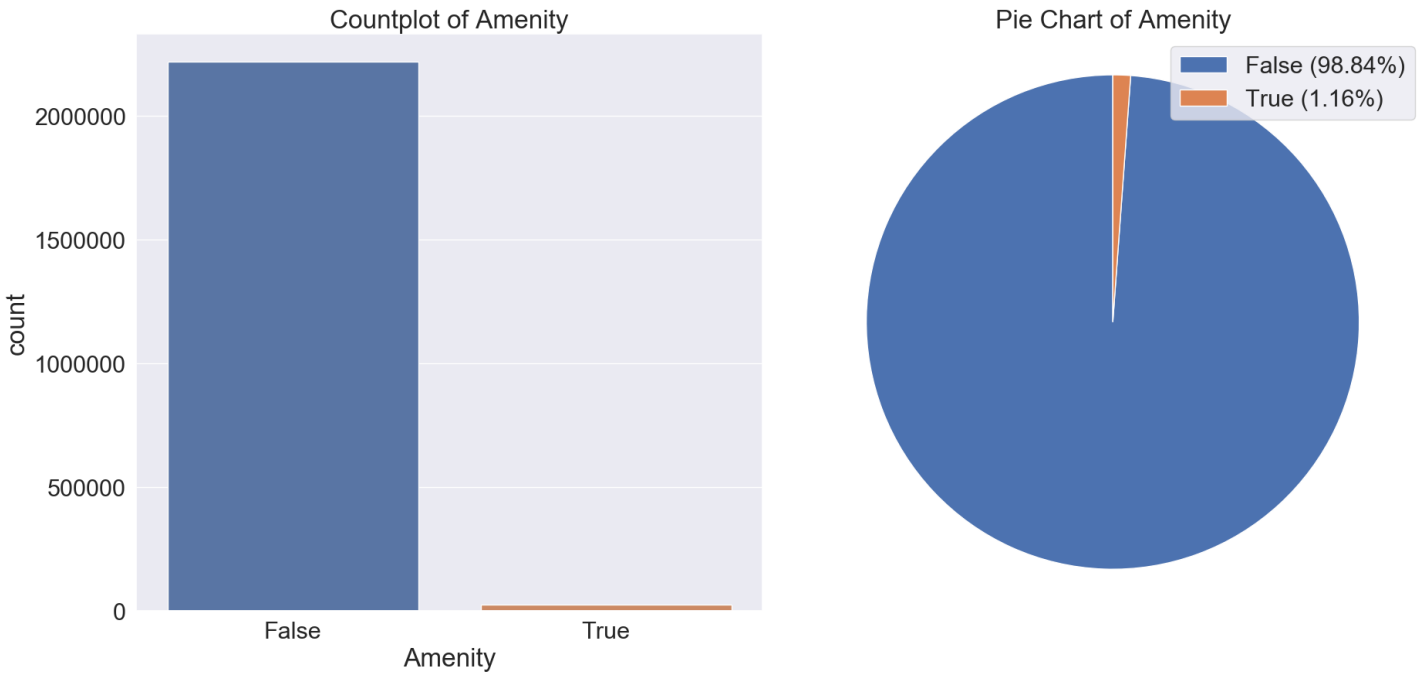




Number of accidents vs Weather_Condition

The plot depicts that the weather condition for most of the accidents was clear, followed by overcast and mostly cloudy. Overcast and mostly cloudy are reasonable factors for accidents unlike clear, which means that weather conditions also does not play an important role.

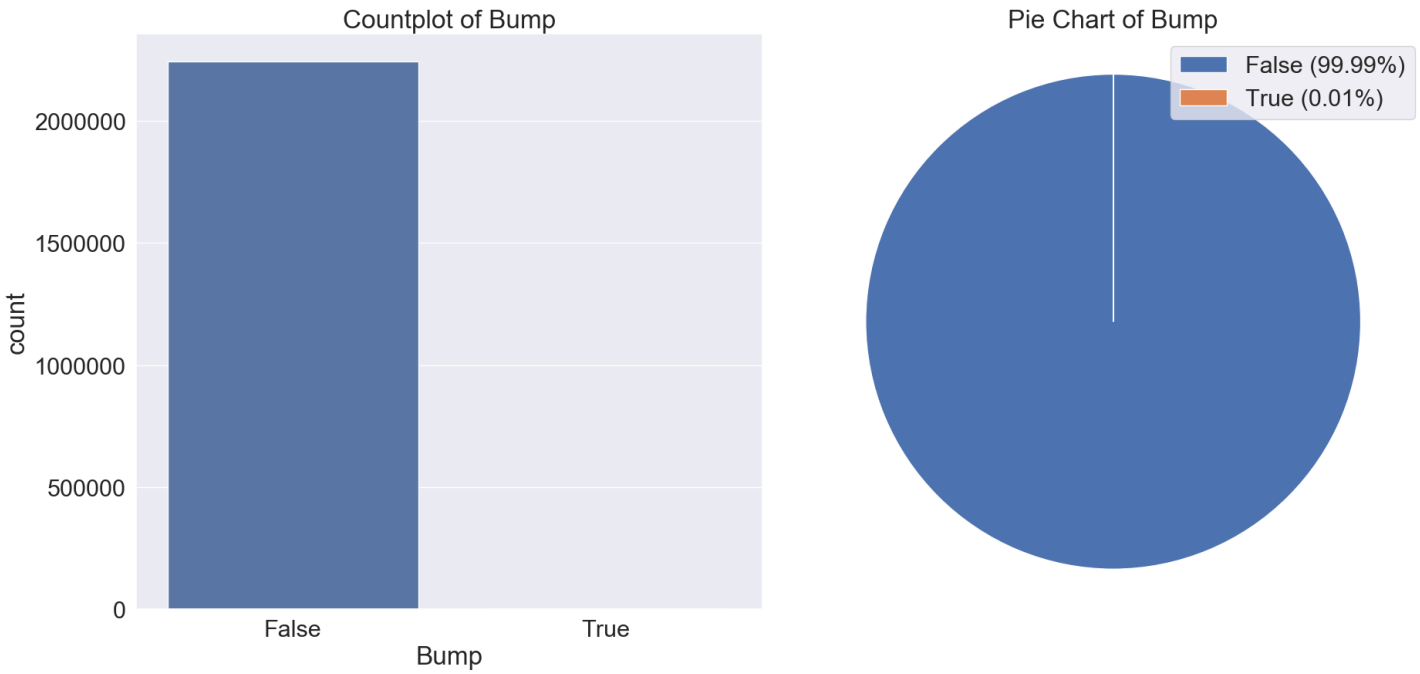
Let's look at the *Amenity* feature. This feature indicates the presence of amenity in a nearby location.



Number of accidents vs Amenity

We see that for almost all(98.84%) accidents, there was no amenity available, which is unfortunate.

Now let's look at the bump feature. This feature indicates the presence of a speed bump or hump in a nearby location.



Number of accidents vs Bump

We see that 99.99% of the accidents were not due to a speed bump.

With this, we come to an end of this analysis.

Conclusion

The USA accidents dataset, taken from Kaggle, was analyzed, and results were discussed above.

We came to a lot of exciting things like we came to know which city or state witnessed the most number of accidents in the USA, we even plotted the results on a map and also considered the severity of an accident.

Hope you got to know something and enjoyed the article!!

Machine Learning Data Analysis Data Visualization Artificial Intelligence Data Science

About Help Legal

Get the Medium app


Download on the
App Store


GET IT ON
Google Play