Analysis of US Accident Dataset

Author: Mehrdad Baradaran

Abstract

This analysis explores the car accident dataset, which spans 49 states of the USA, capturing traffic incidents recorded from February 2016 to March 2023. The dataset was collected through various APIs, providing real-time traffic incident data. While many have examined this dataset, this analysis takes a unique perspective by focusing on specific hypotheses related to COVID-19. The COVID-19 pandemic has had a profound impact on various aspects of life, and this report delves into how it has affected car accidents.

In this report, we pose two distinct hypotheses, exploring the relationship between the pandemic and car accidents. Our approach to these hypotheses highlights the exclusivity of this analysis. By examining this dataset through the lens of COVID-19, we aim to provide valuable insights into the lesser-explored intersections between the pandemic and road safety. Our findings are essential for understanding and adapting to the changing dynamics of traffic incidents during and beyond the COVID-19 era.

Introduction

This is a countrywide car accident dataset, which covers 49 states of the USA. The accident data were collected from February 2016 to March 2023, using multiple APIs that provide streaming traffic incident (or event) data.

I will use this US-Accidents dataset to extract cause that contribute to car accidents and how different factors affect the severity of the accidents differently.

Asks

1. What topic are we exploring?

   US car accidents; causality analysis; the environmeantla casue behind car accidents; the traffic behavior and accidents during COVID-19.

2. What is the problem we are trying to solve?

   How do different environmental factors affect the car accidents differently? Does COVID-19 have any impact on traffic behavior and accidents?

3. What metrics will we use to measure your data to achieve our objective?

   I'll use the environmental factors such as temperature, precipitation, windspeed and so on to analyze the impact of environmental factors on car accidents. Meanwhile, I'll use the Start_Time of the accidents, the location (Start_Lng and Start_Lat) of the accidents to help finding other possible insights.

4. My hypothesis

   Cold weathers, high precipitation, low visibility has a strong relation with the number of traffics, where low visibility has the most number of severe accidents.

   COVID-19 has an impact on traffic, and it makes the number of accidents fewer.

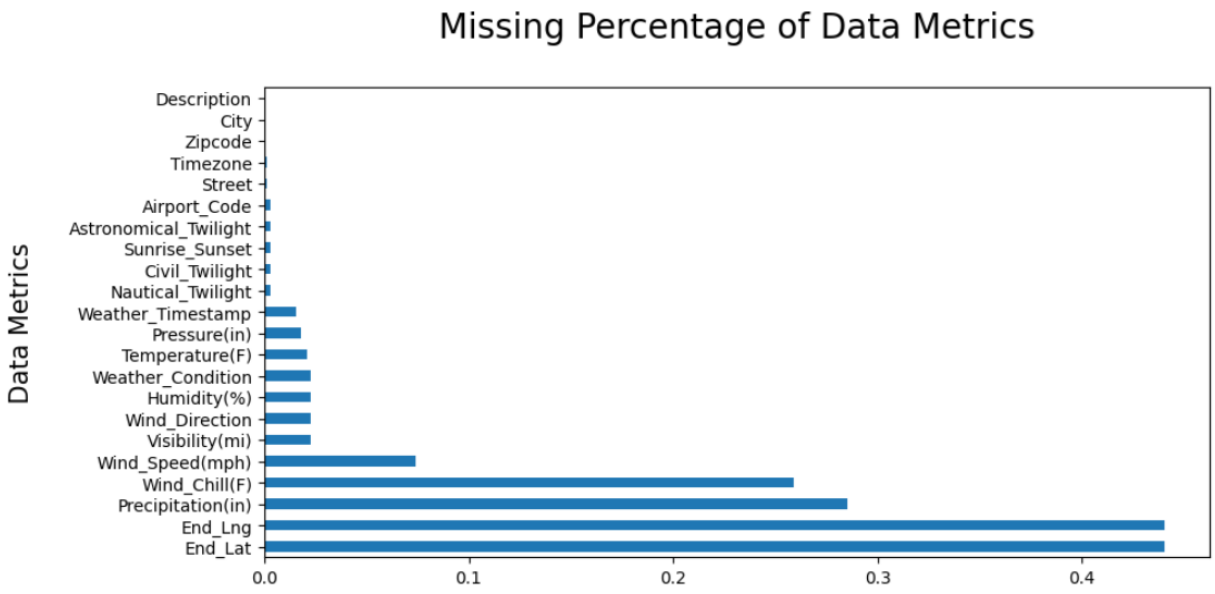
Prepare

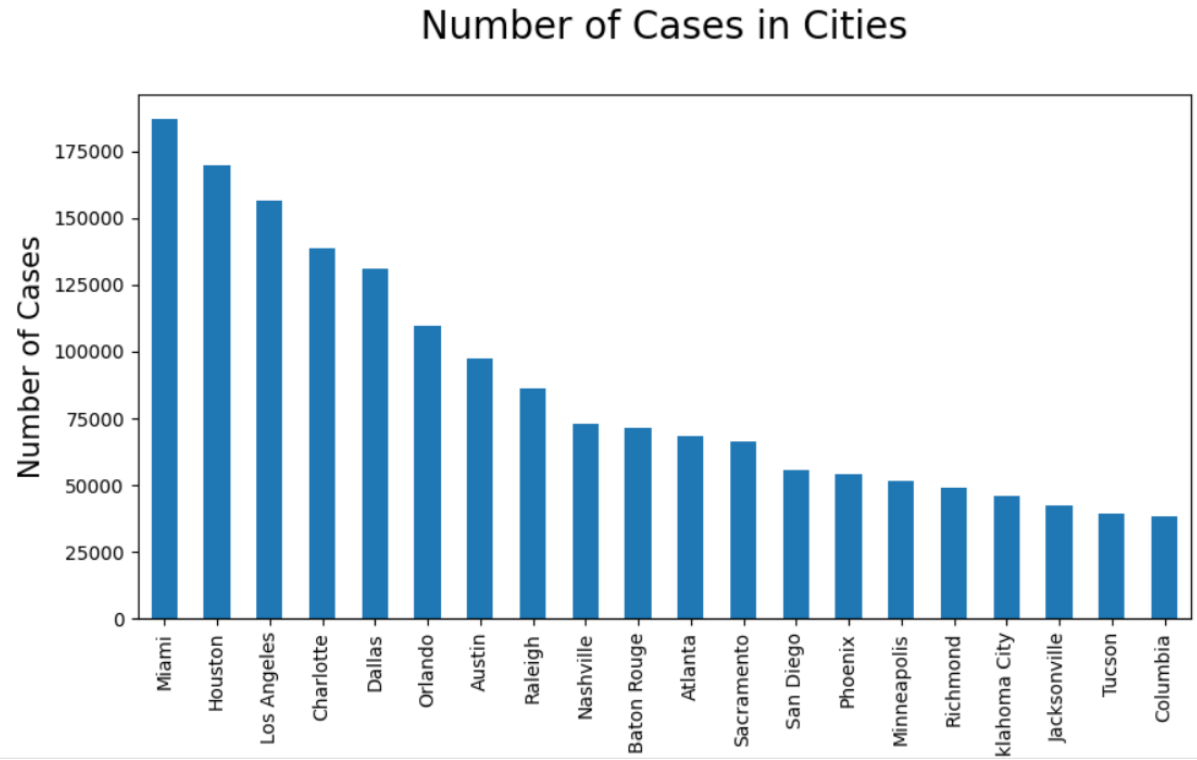
We will identify how the data is organized, understand the data by sorting and filtering the data, and determine the credibility of the data.

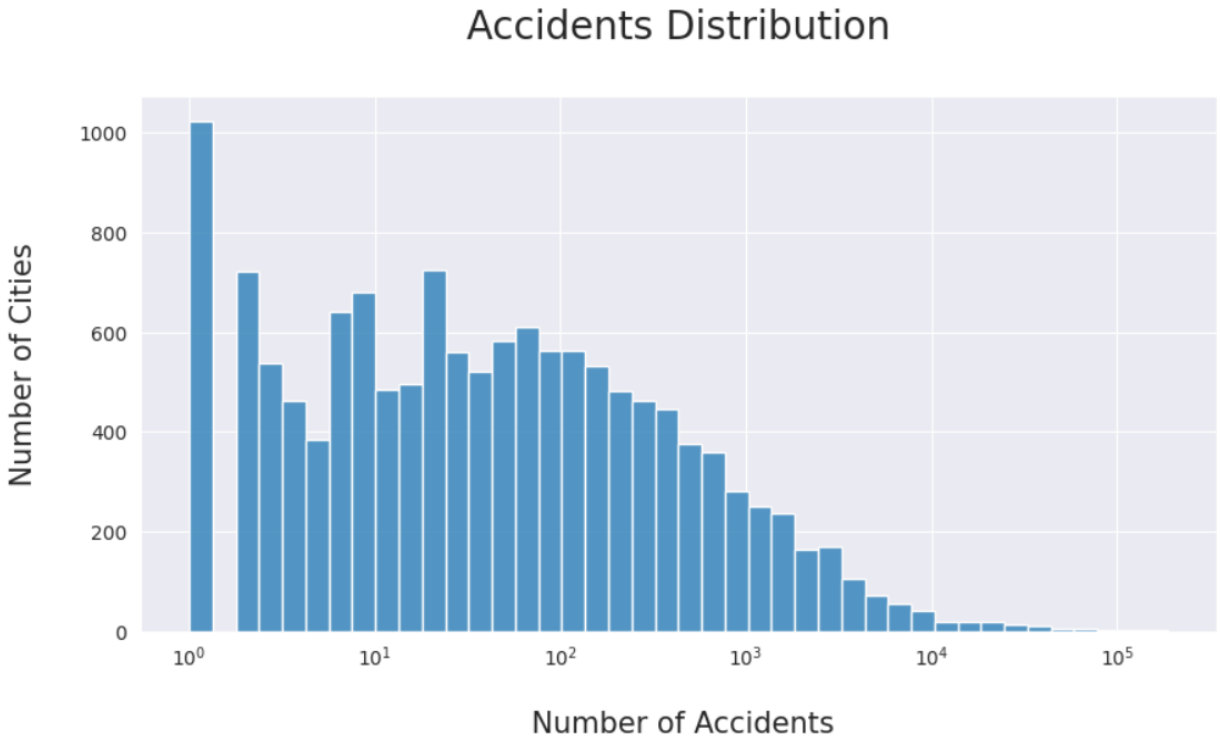
The dataset is a structured and organized dataset with 47 metrics. It's collected continuously, including APIs that provide streaming traffic event data. The parties that capture these data are US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Nearly around 7.7 million accidents were recorded from February 2016 to March 2023. The dataset is about describing environmental types of metrics and geospatial information of the reported US car accidents. Though the dataset may lack some crucial data that's mentioned below, it still has enough credibility.

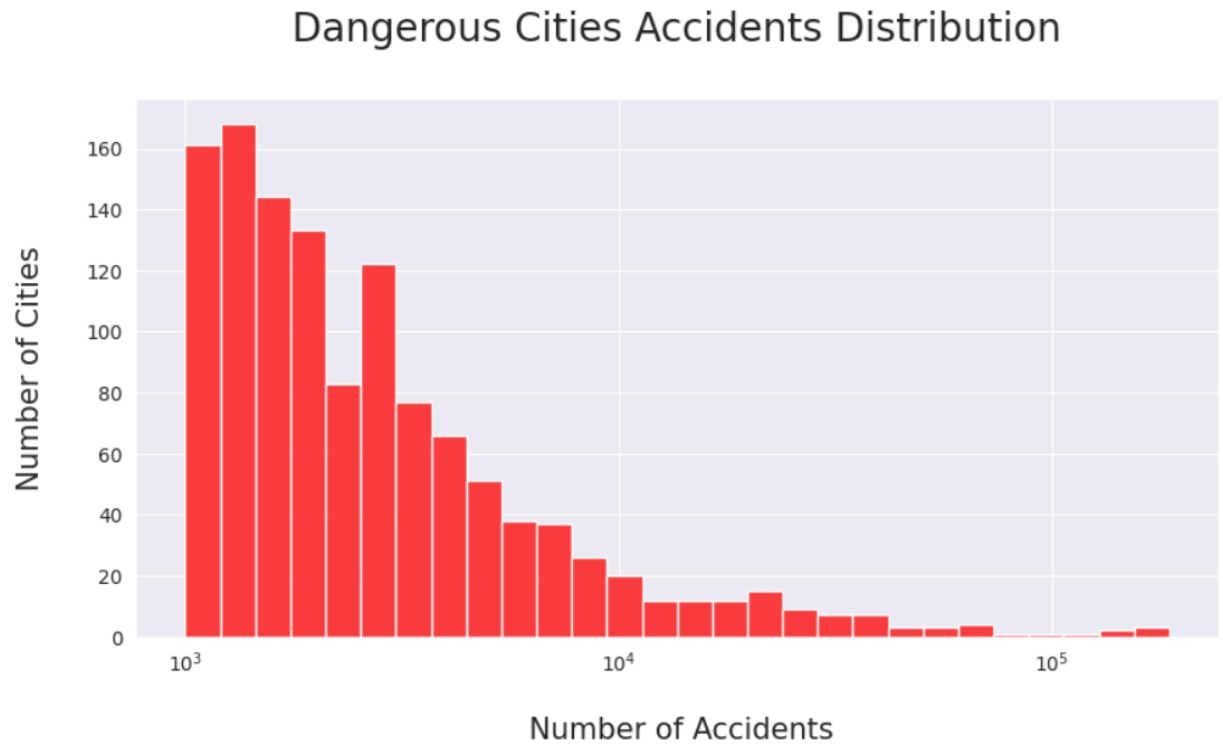
After briefly understanding the dataset, I find out that Miami has the most number of car accidents from Feb 2016- March 2023. However, the dataset does not contain any information about New York City, even being the most populous city in the states. The dataset does not cover the whole year of 2016, so there's a small lack of data. I define the dangerous city as a city that has more than 1000 accidents. Only 5% of all recorded cities in the US are dangerous in terms of the number of car accidents.

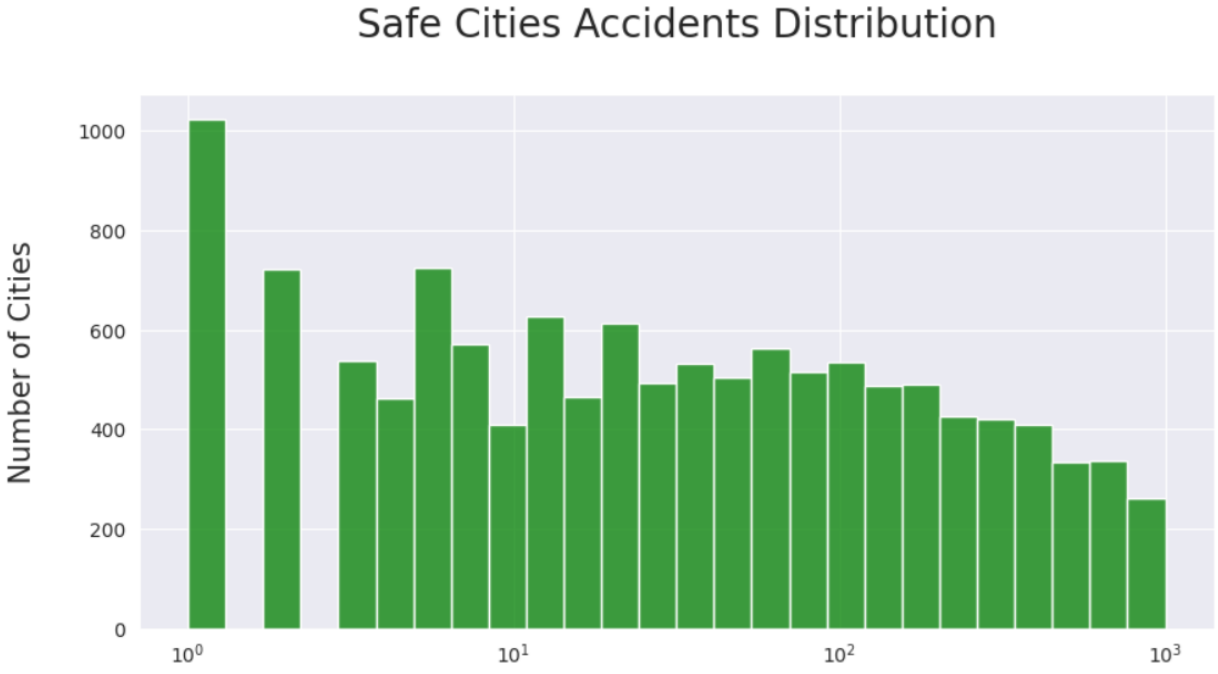
I then generated three histplot to see the number of cities that fall in each bin. The reason not to use distplot or kdeplot is to advoid unclear distribution. At first I see how all the accidents distribute, but then I divide it into two groups. Graph with red columns represents the distribution of cities with high number of accidents. Graph with green columns represents the distribution of cities with low number of accidents. However, it is strange to see more than 1000 cities only have 1 accident during those years.




Number of Cases in Cities
Number of Cases
175000
150000
125000
100000
75000
50000
25000
0
Miami
Houston
Los Angeles
Charlotte
Dallas
Orlando
Austin
Raleigh
Nashville
Baton Rouge
Atlanta
Sacramento
San Diego
Phoenix
Minneapolis
Richmond
Klahoma City
Jacksonville
Tucson
Columbia


Accidents Distribution
Number of Cities
1000
800
600
400
200
0
10^0
10^1
10^2
10^3
10^4
10^5
Number of Accidents


Dangerous Cities Accidents Distribution
Number of Cities
0
20
40
60
80
100
120
140
160
10^3
10^4
10^5
Number of Accidents


Safe Cities Accidents Distribution
Number of Cities
0
200
400
600
800
1000
10^0
10^1
10^2
10^3

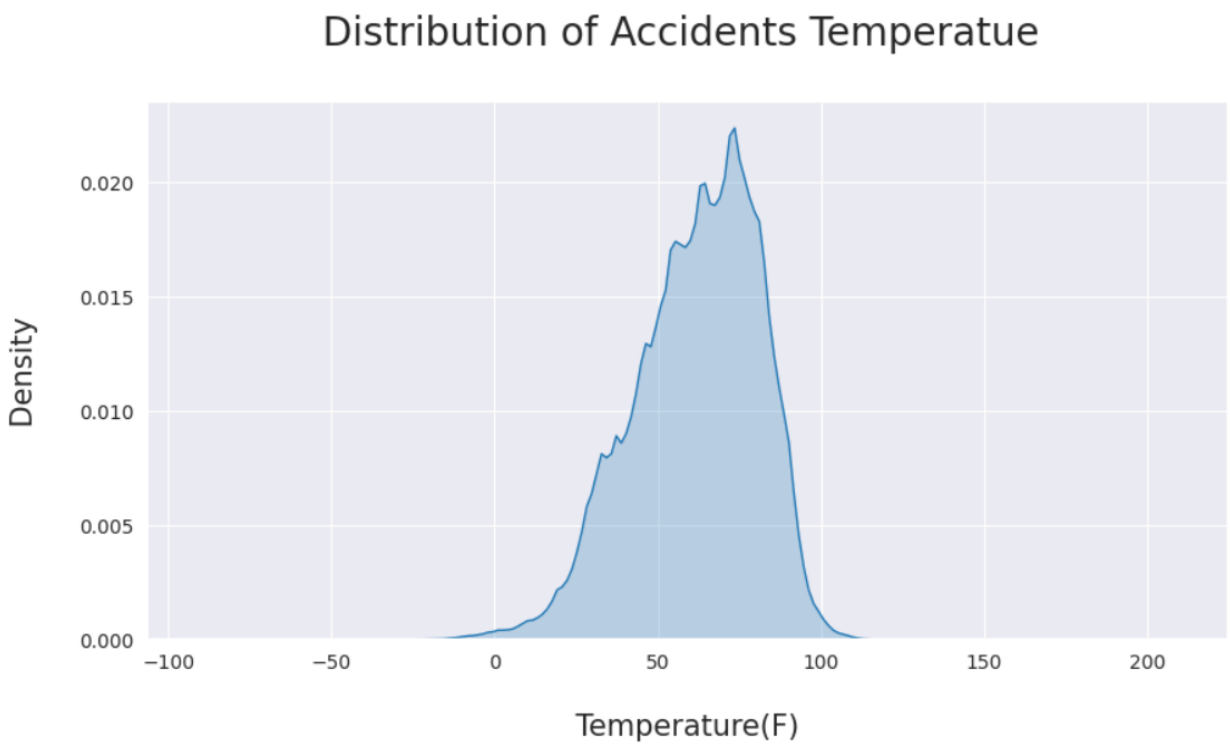
Process

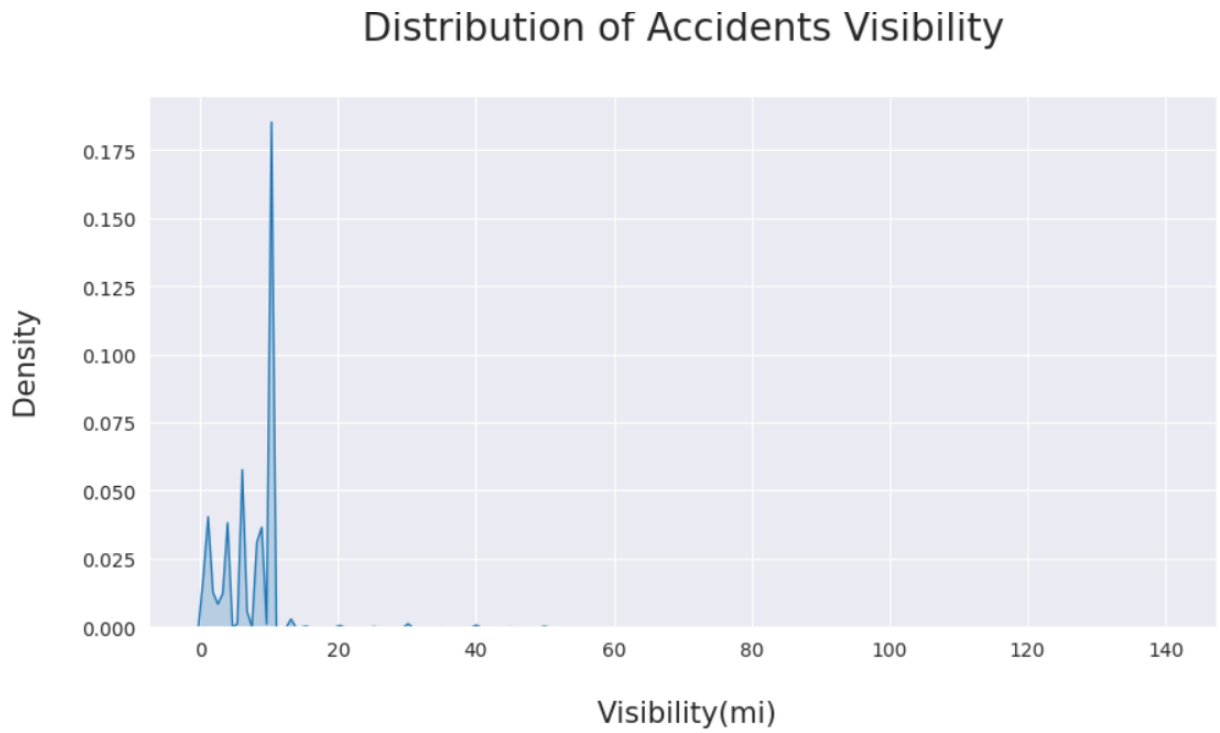
We will clean the data and extract what we need for the analysis.

Since we are only interested in how environmental factors and COVID-19 periods affect the traffic, We will only keep the following columns: ID, Start_Time, Start_Lat, Start_Lng, City, Temperature(F), Wind_Speed(mph), Humidity(%), Precipitation(in), Visibility(mi), Weather_Condition and drop the remaining columns for cleaner analysis. These other factors also have a large portion being null, so they are not very helpful for the analysis.

The first step is to fill the appropriate values in the columns with Null values. Depending on the distribution and the number of Nulls, I deploy different methods to clean the dataset.

Since temperature has less than 10% null values of the total number of values and they appear to be normally distributed. It might be a good idea to fill these empty data with the mean value. Whereas for Visibility(mi), it's right skewed. So replacing null values with a median value is more suitable.

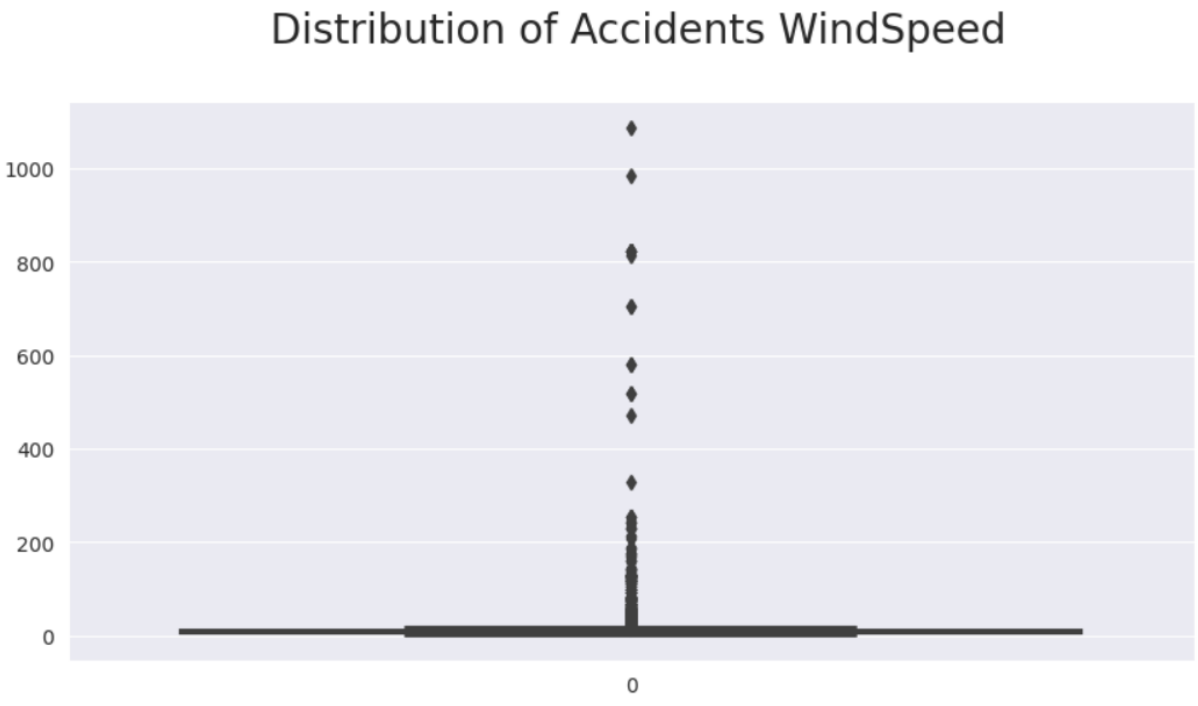


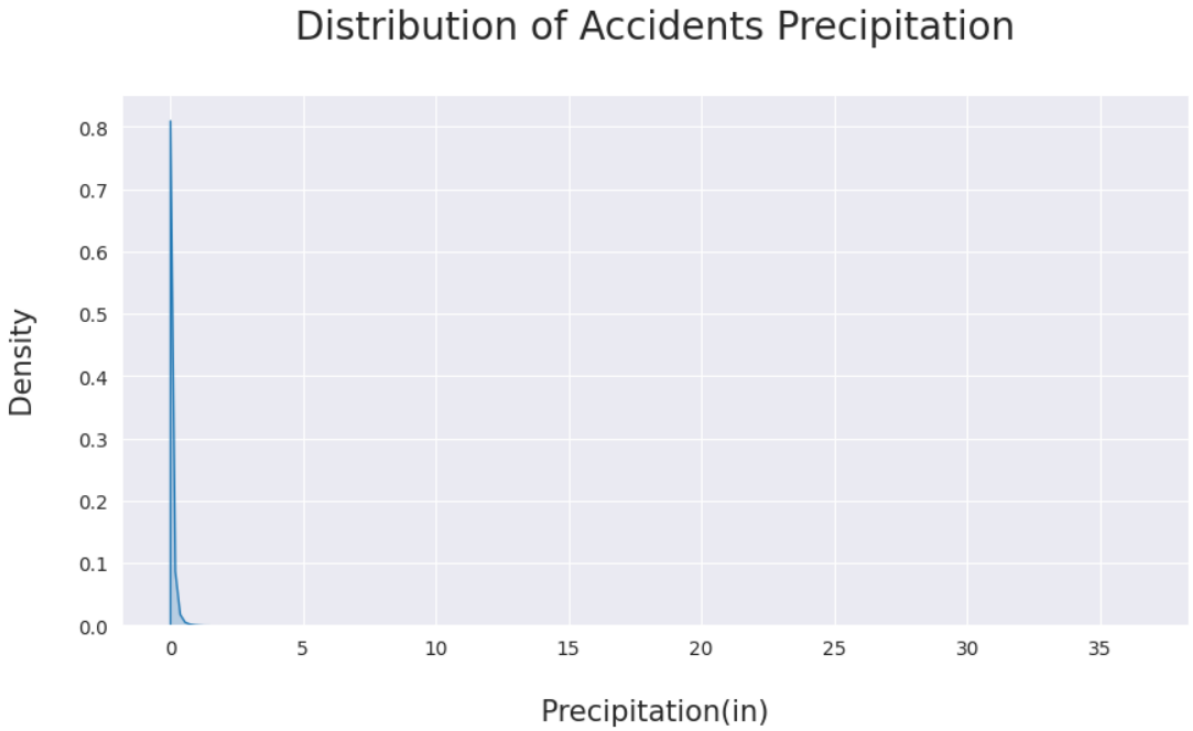


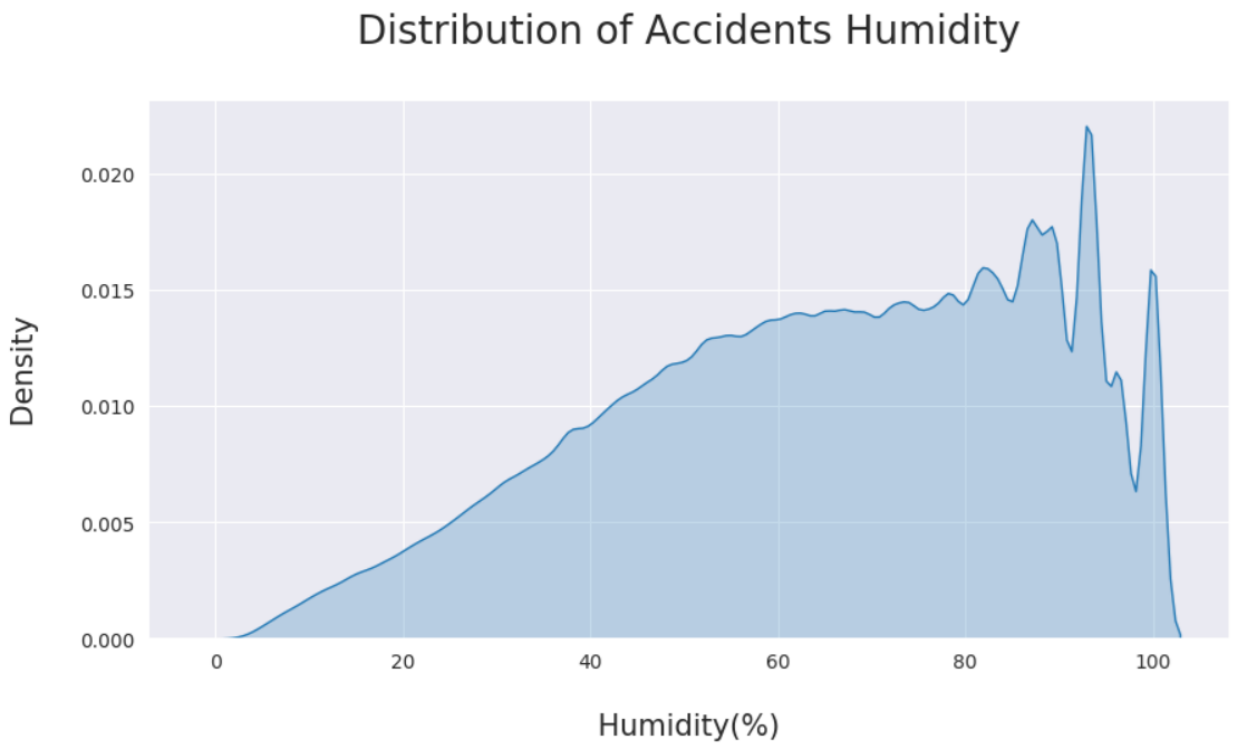
Since there are 144 types of weather condition and it's hard to differentiate and many of them can be consider as one big category. It's better to drop the column.

```
Weather_Condition
Fair                          2560802
Mostly Cloudy                 1016195
Cloudy                         817082
Clear                          808743
Partly Cloudy                  698972
                                ...
Heavy Sleet / Windy                 1
Sand / Windy                        1
Heavy Rain Shower / Windy           1
Blowing Snow Nearby                 1
Drifting Snow                       1
Name: count, Length: 144, dtype: int64
```

Since Precipitation(in), Wind_Speed(mph) have an right skewed distribution. It's better to use mode value to fill the Null value in these two columns. Humidity(%) though has a left skwed distribution. I still used the mode value to fill out the Null. It may not be accurate to fill out the Null value based on the previous or latter adjacent value, as every two accidents were hardly related.


Distribution of Accidents WindSpeed
1000
800
600
400
200
0
0
Wind Speed (mph)


Distribution of Accidents Precipitation
Density
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
5
10
15
20
25
30
35
Precipitation(in)



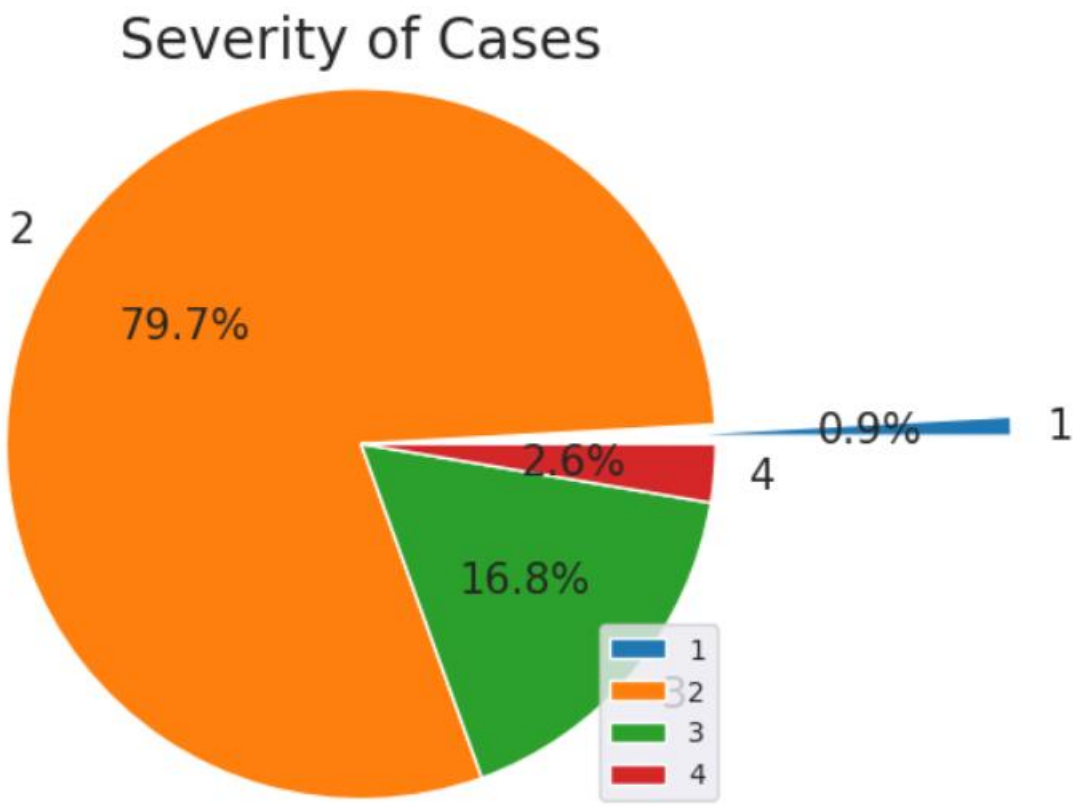
We find out that there are several factors such as Street, Zipcode have many missing values. In this case, these factors may not be able to contribute many information in our anlaysis. Since we want to find out how different environmental factors may contribute to the severity and potentially finding out which seems to be more responsible for car accidents. I extract only relevant columns. They are Precipitation(in), Wind_Speed(mph), Temperature(F), Visibility(mi), Humidity(%) and Weather_Condition.

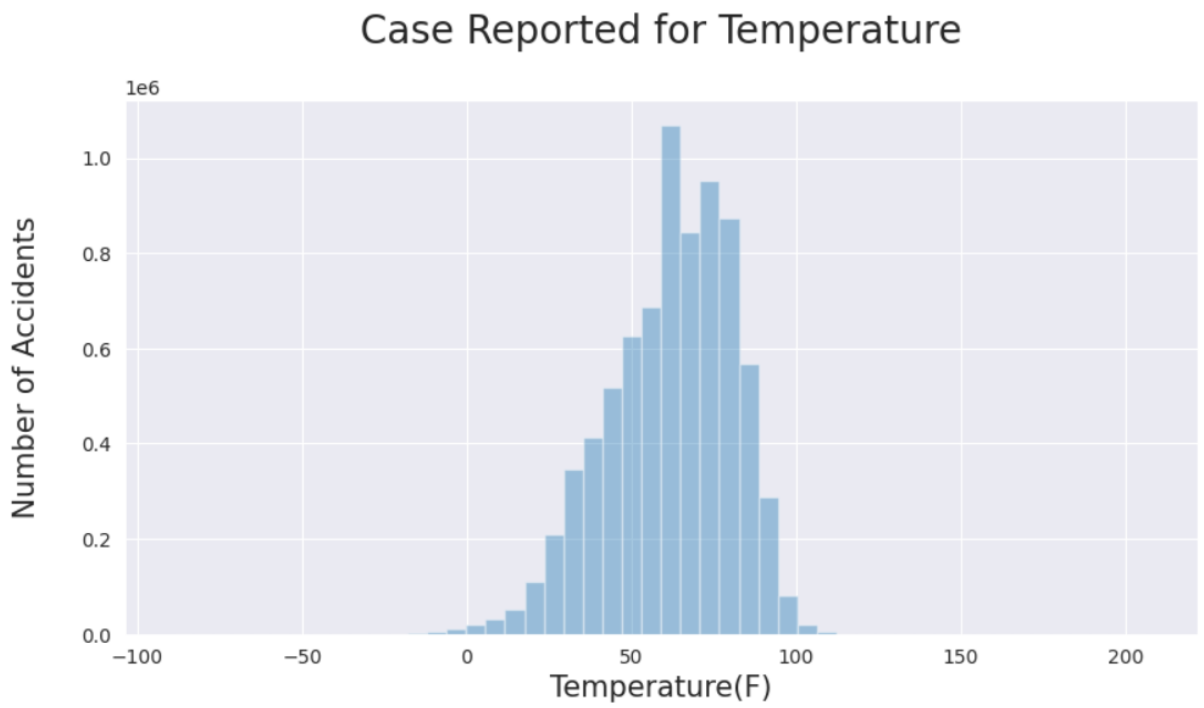
When undertanding these various environmental factors, I first check out the distribution of each. Temperature behaves normally while the other three skewed to the right. For each of these columns, the lack of value only counts less than 20% of total number of data. So it's not advisbale to delete the whole rows that may contain Null value. For Weather_Condition, there are 144 kinds of values. So it's not that useful for our data analysis. Finally, we drop the rows that have missing city because it's hard to infer which city this accident belonged to and unfair to .

Now we have cleaned and adjusted our data properly. I will discover the relationship between the factors and the severity of the car accidents and the number of accidents, the impact of covid on car number of accidents (value_counts) geolocation wise and time period wise.

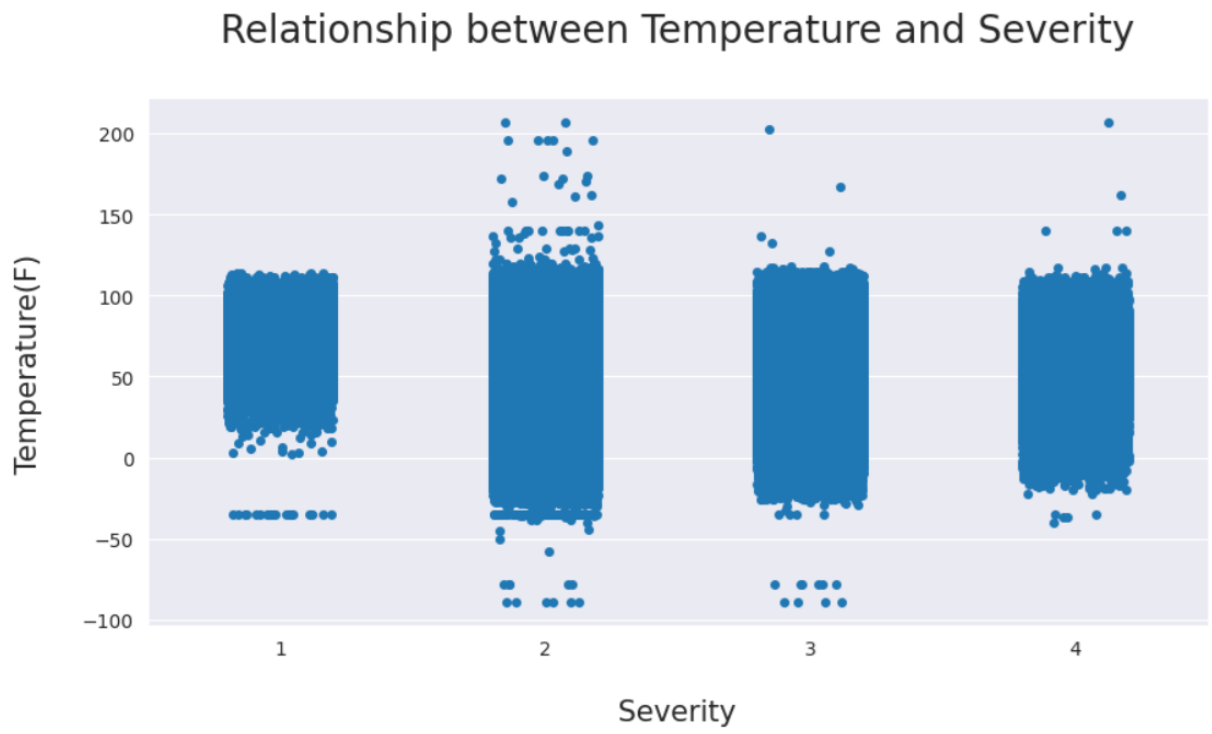
Analyze

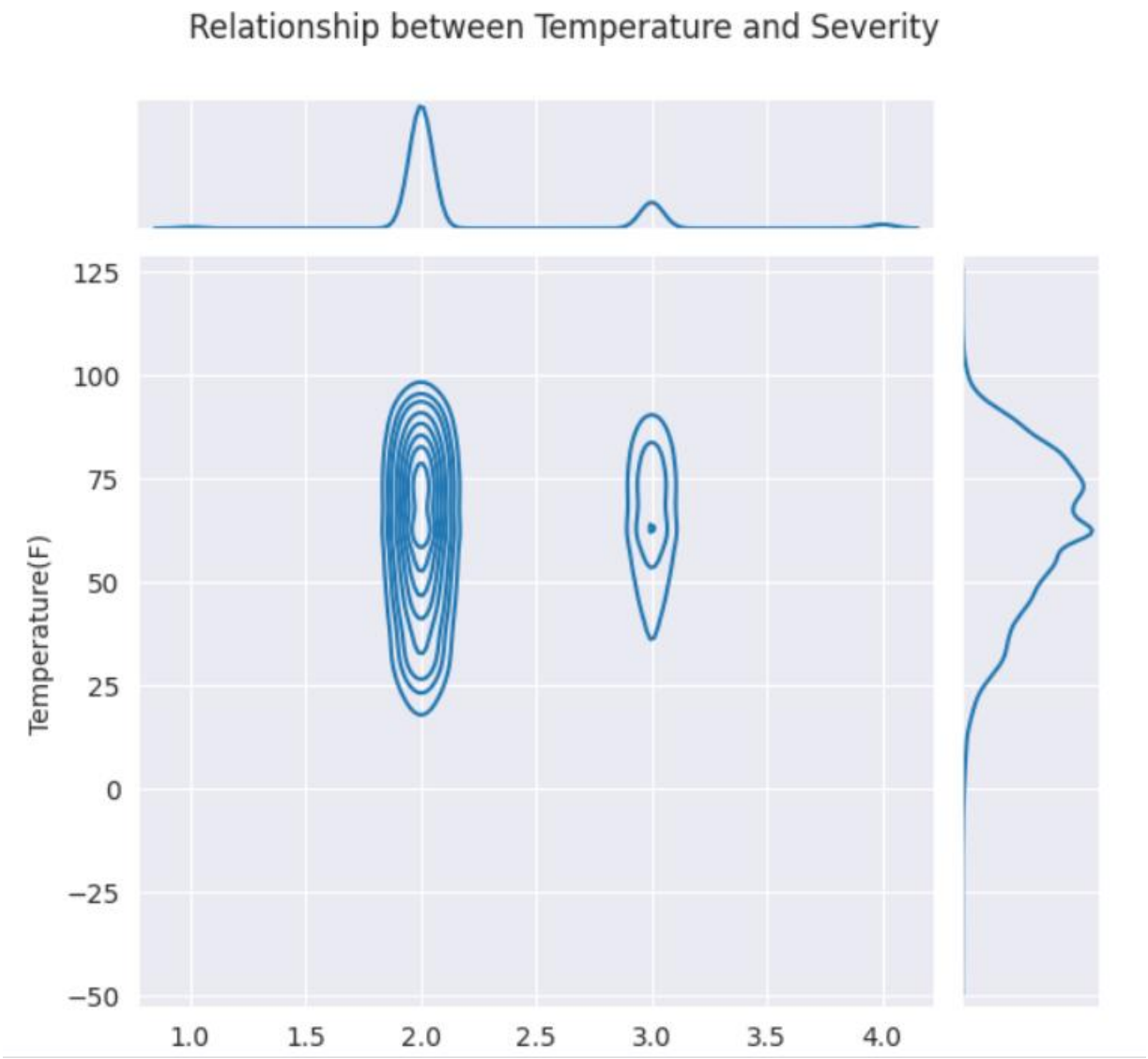
Environmental Factors Analysis



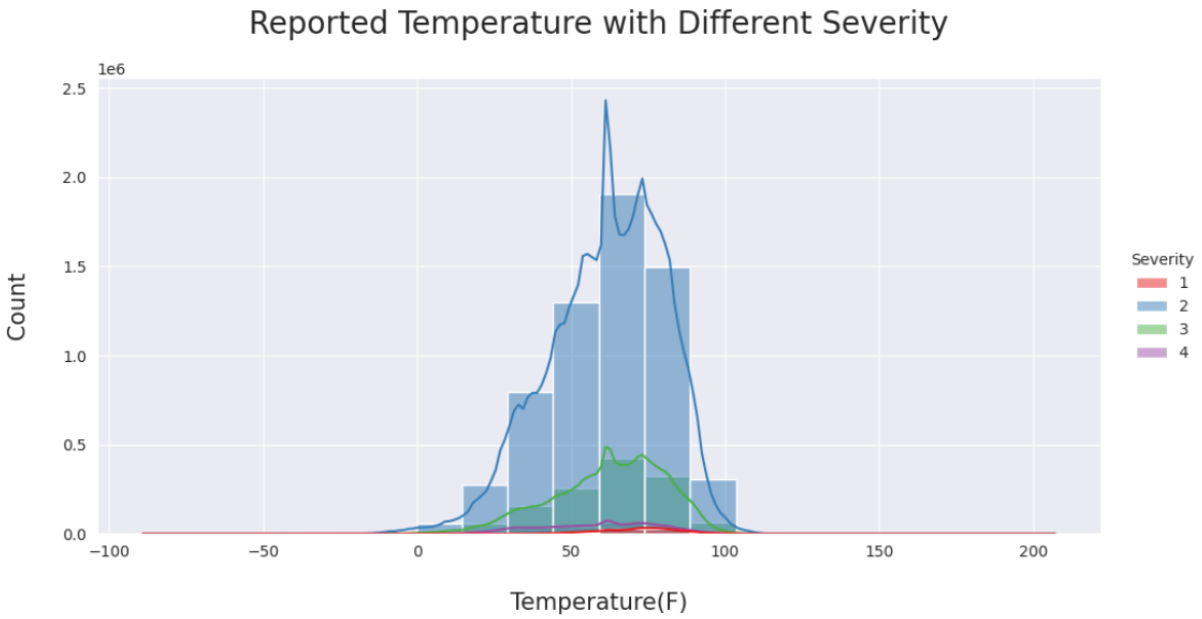


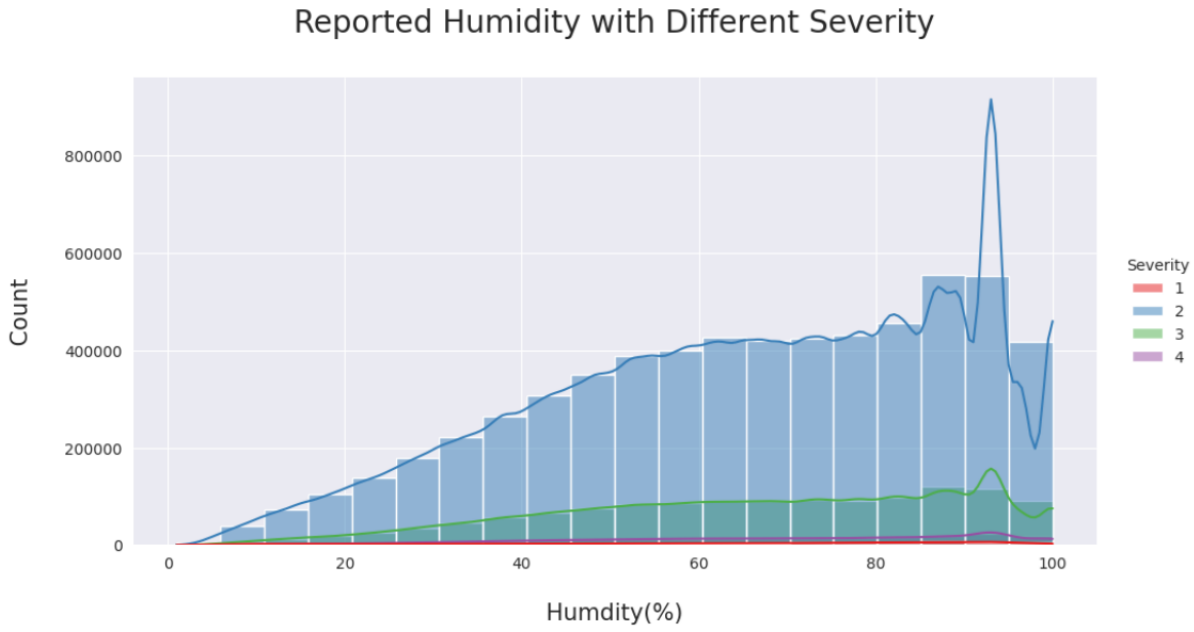
What is the relationship between temperature and the severity of accidents?



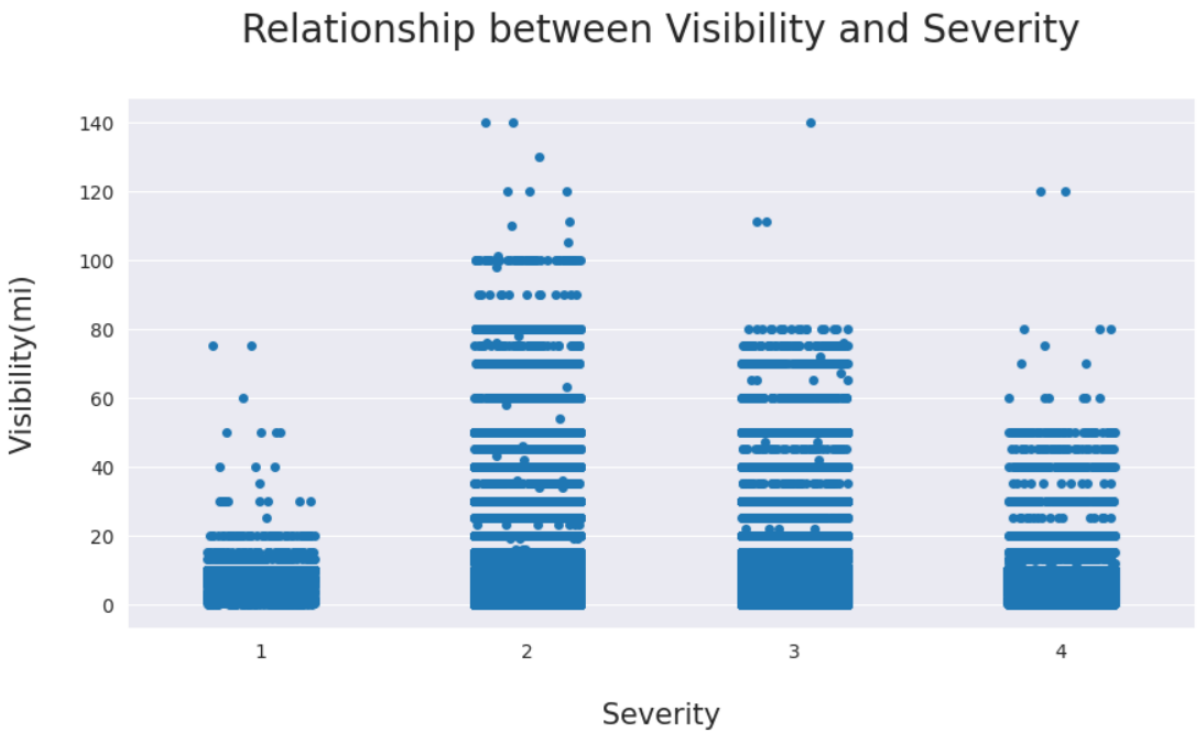


Do accidents with different severity scores behave the same, temperature wise?

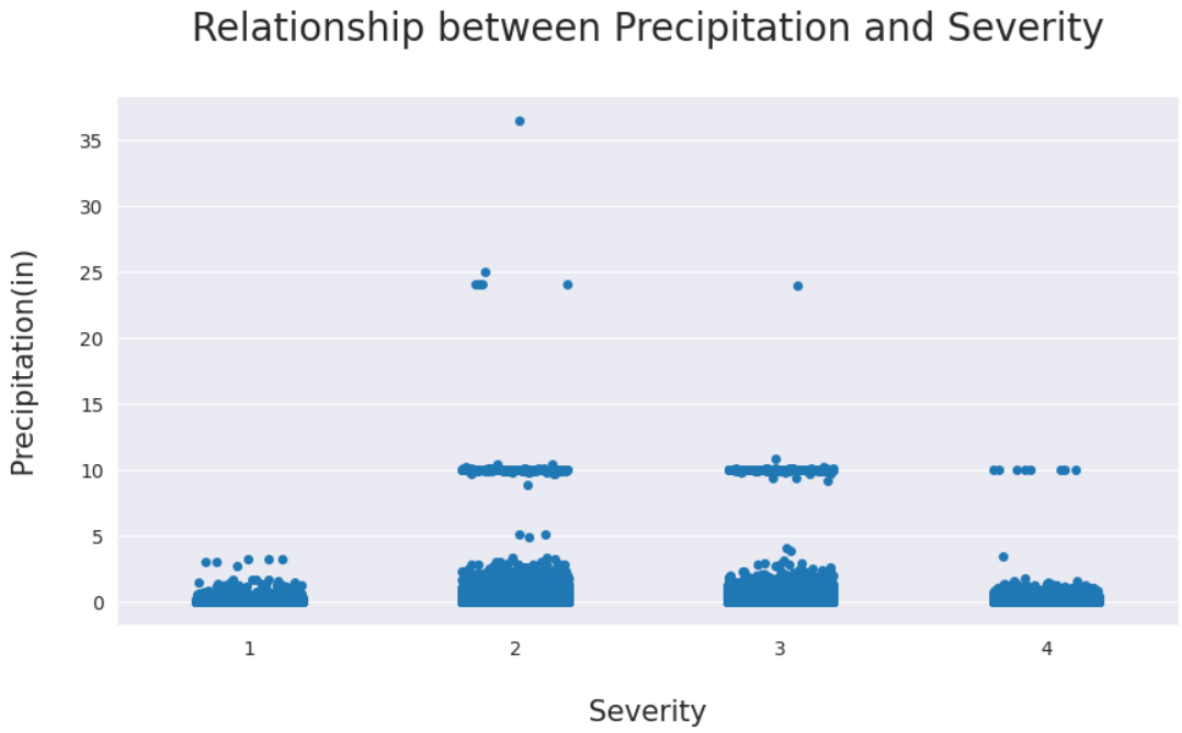




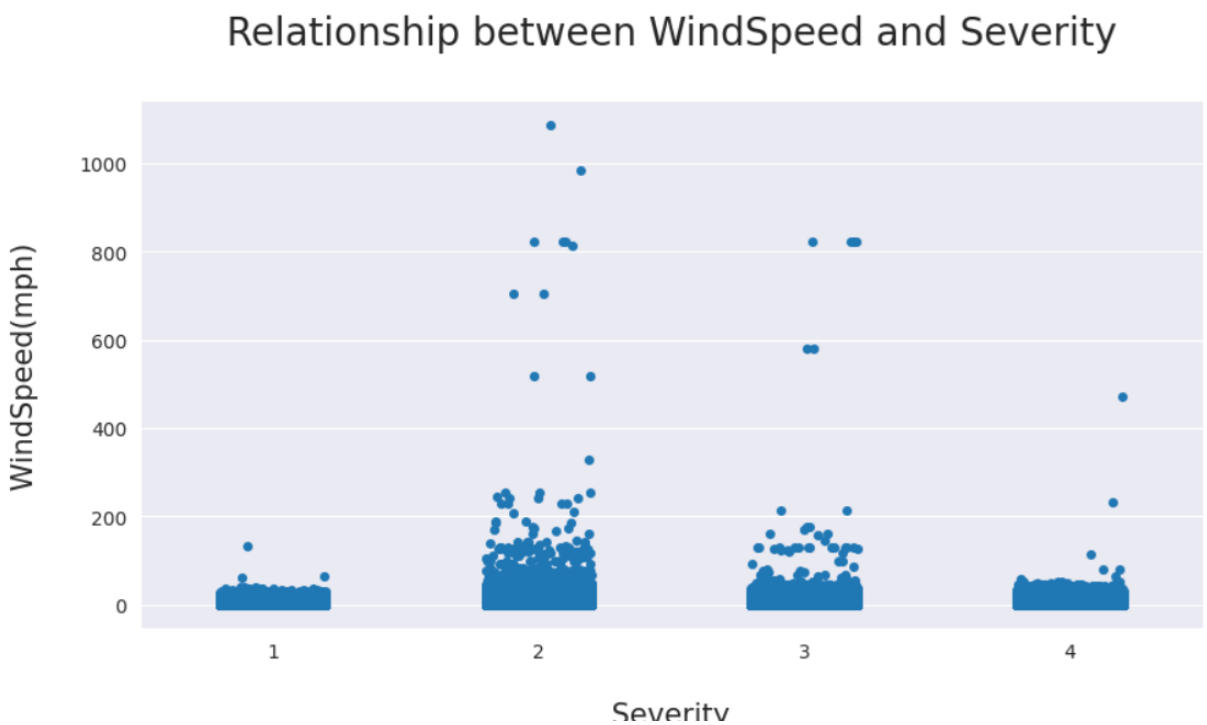
How many cases are there with different visibility? What's the relationship between visibility and severity scors?



How many cases are there with different precipitation? What's the relationship between precipitations and severity scors?



How many cases are there with different wind speeds? What's the relationship between the wind speed and severity scors?

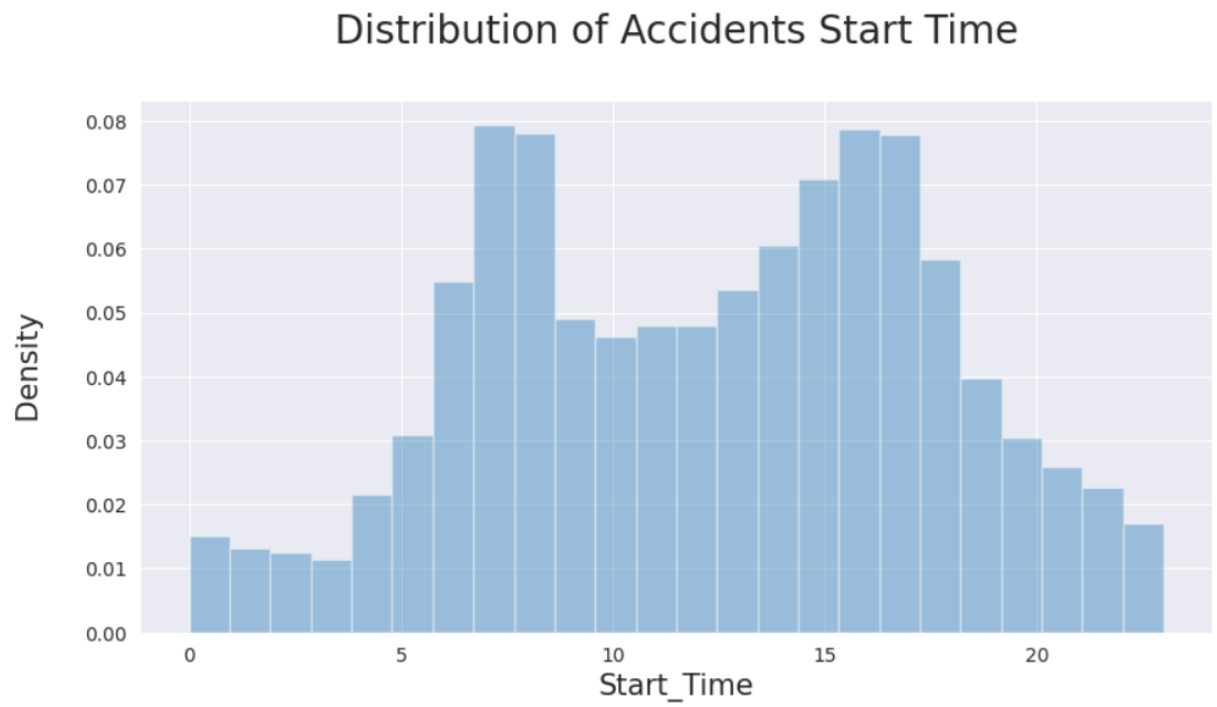


We've just analyzed how different environmental factors might contribute to the number of accidents, how different environmental factors may influence the severity of the accidents.

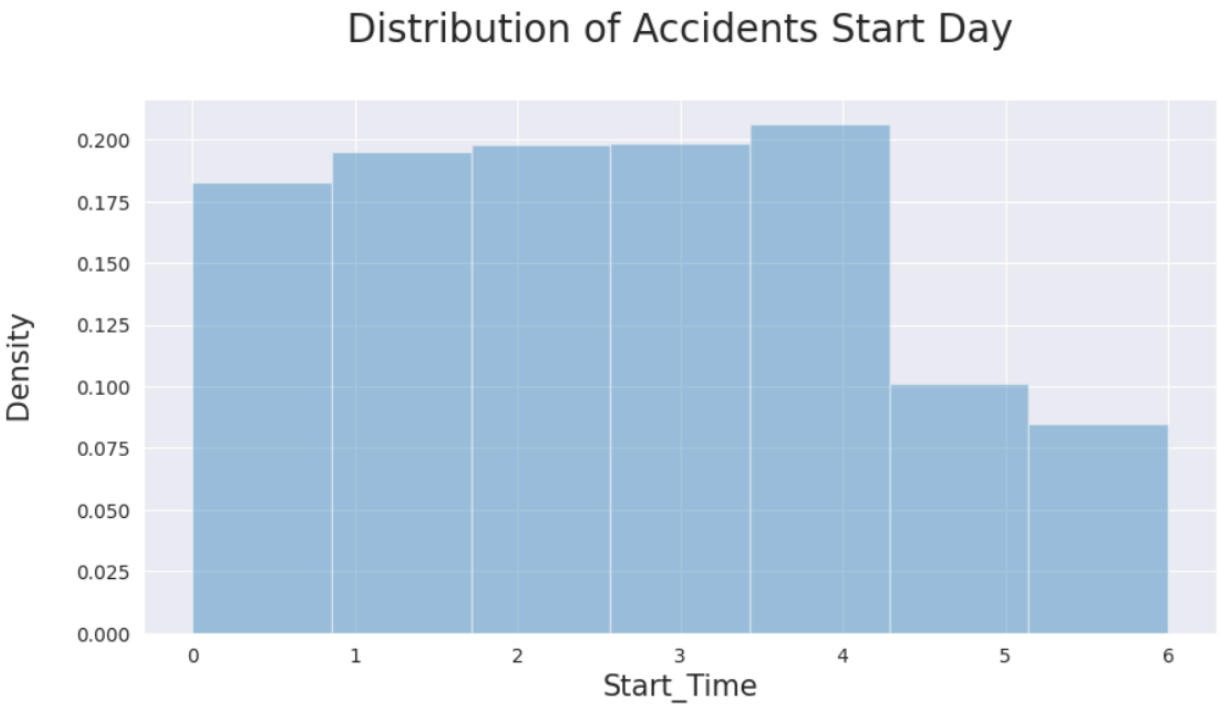
I have used pie chart, distplot, stripplot, jointplot, scatterplot, and etc to see the relationship between some environmental factors, severity and number of accidents. More conclusions and findings can be see in the last section.

Time Analysis: car accidents during the COVID-19 period

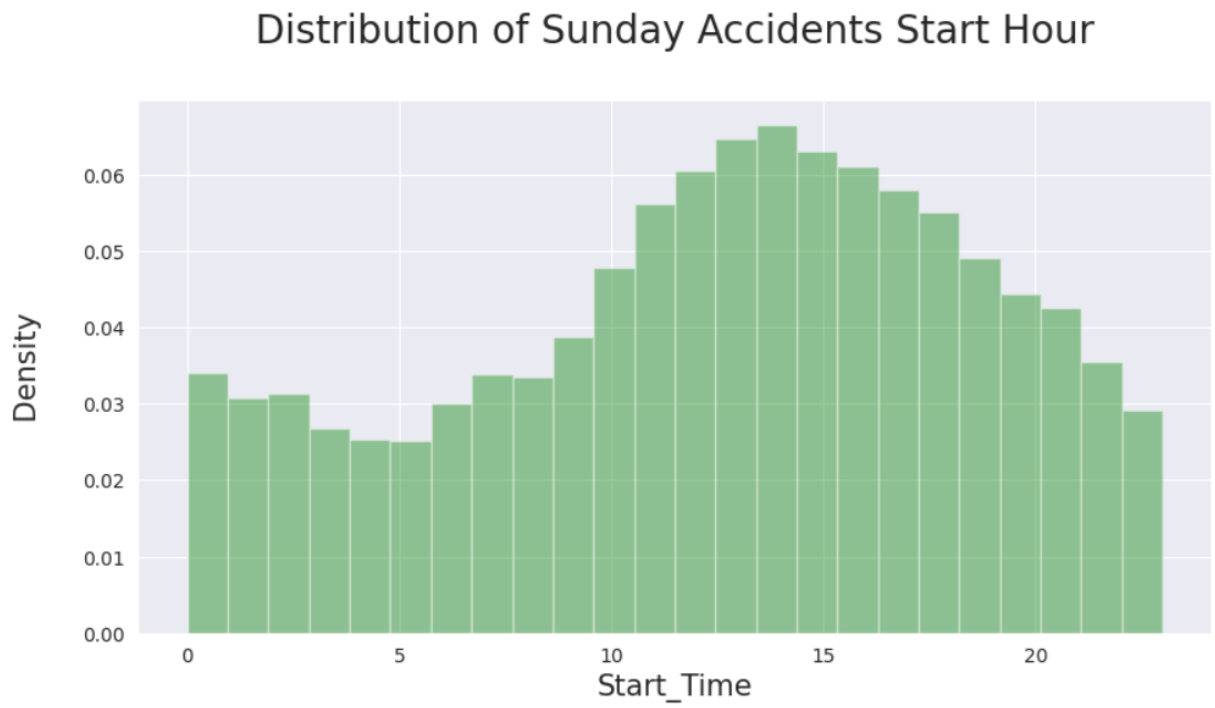
What time of the day have the most accidents?



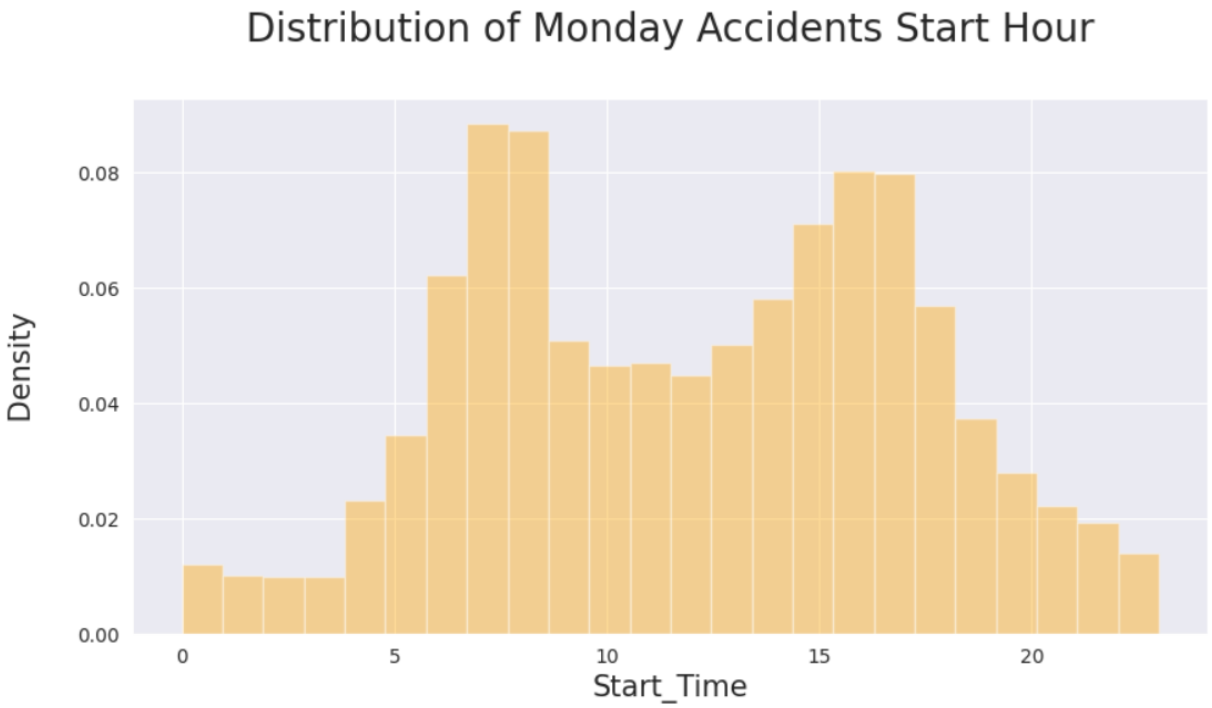
A high percentage of accidents occure between 15 to 18. Probably people are hurry to get home. The next highest one is around 6 to 8. This might be due to people going to work.



On weekends the number of accidents is lower. Is the distribution of accidents same on weekdays and on weekends?

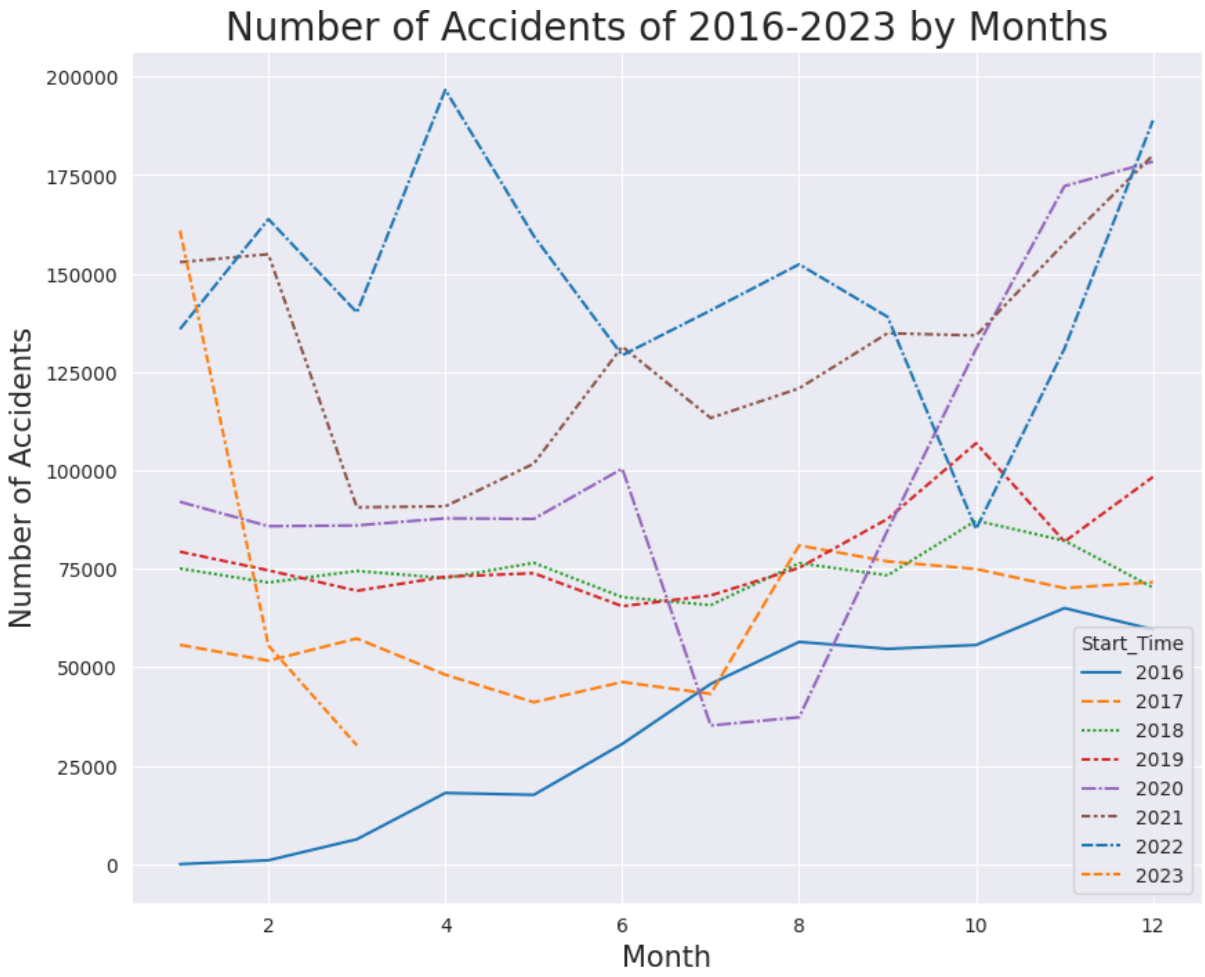


The distribution of Sunday accidents start time is different than the overall start time accidents distribution. Especiallt the density is higher at 0 o'clock for sunday accidents.

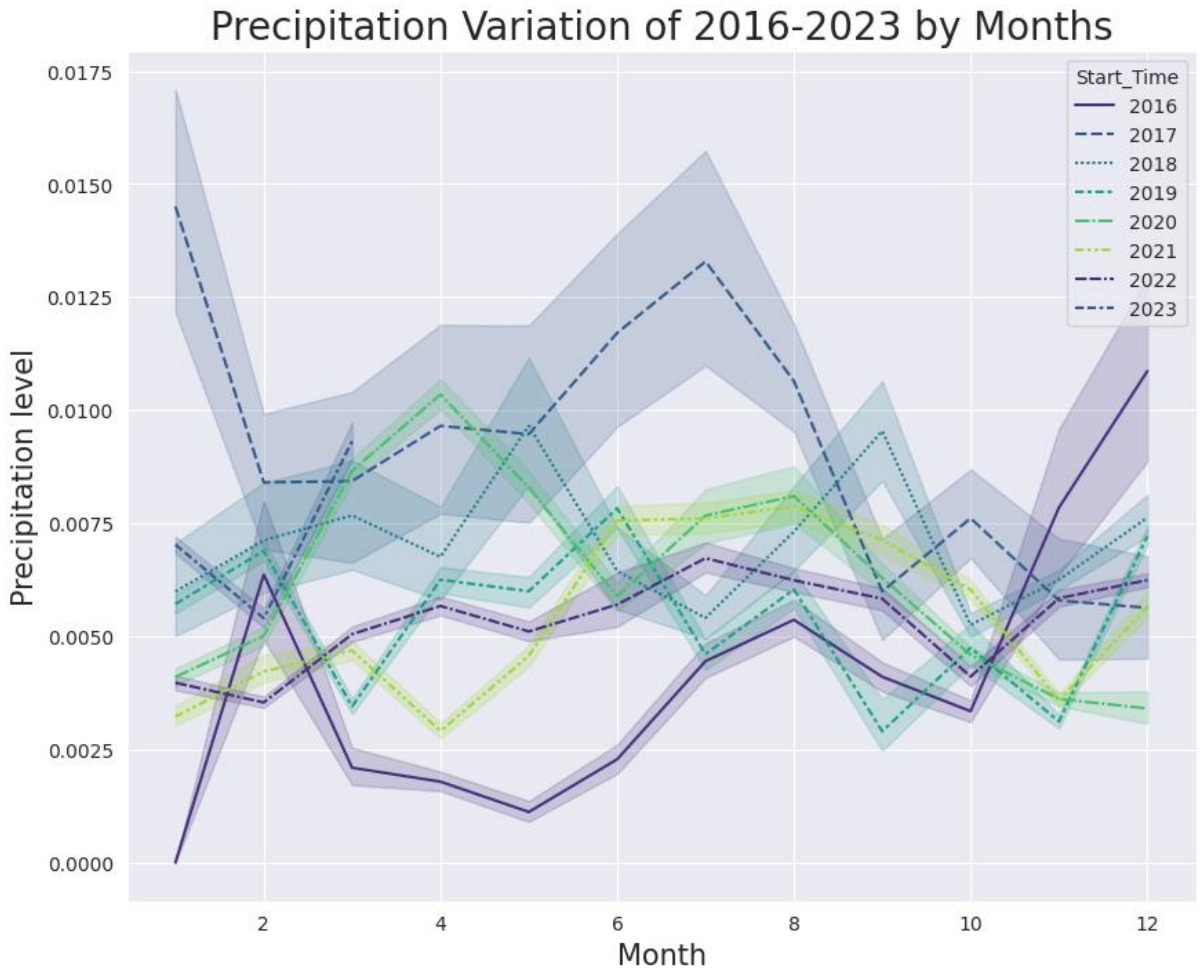


This monday accidents start time distribution is very similar to the overall one.

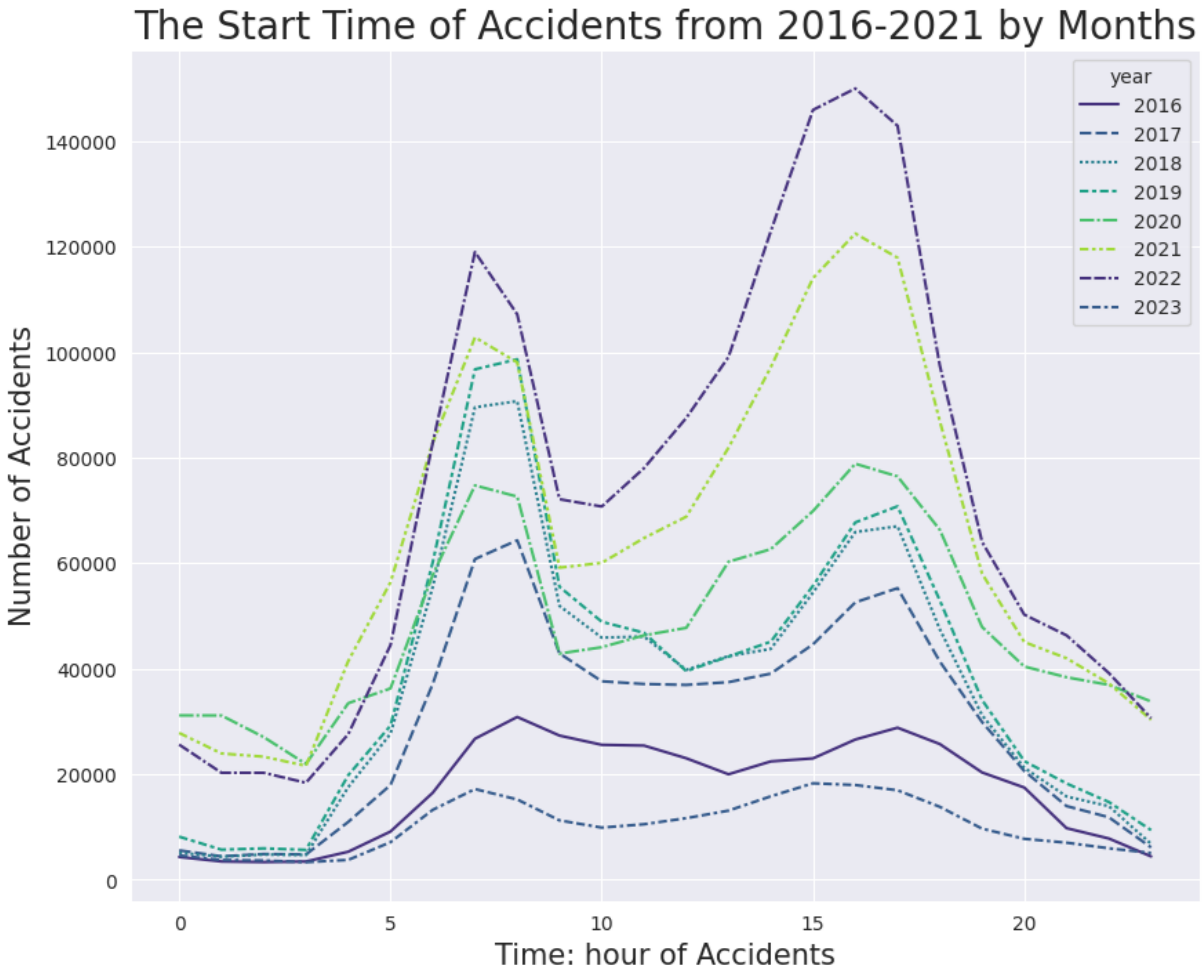
We see there's a rising number of car accidents in 2020 and 2021 (even 2022)-(Covid-19 period) But is it due to the environmental factors? We may need to have more data related to the number of positive cases, government policies etc. But we could check out the yearly environmental influence.



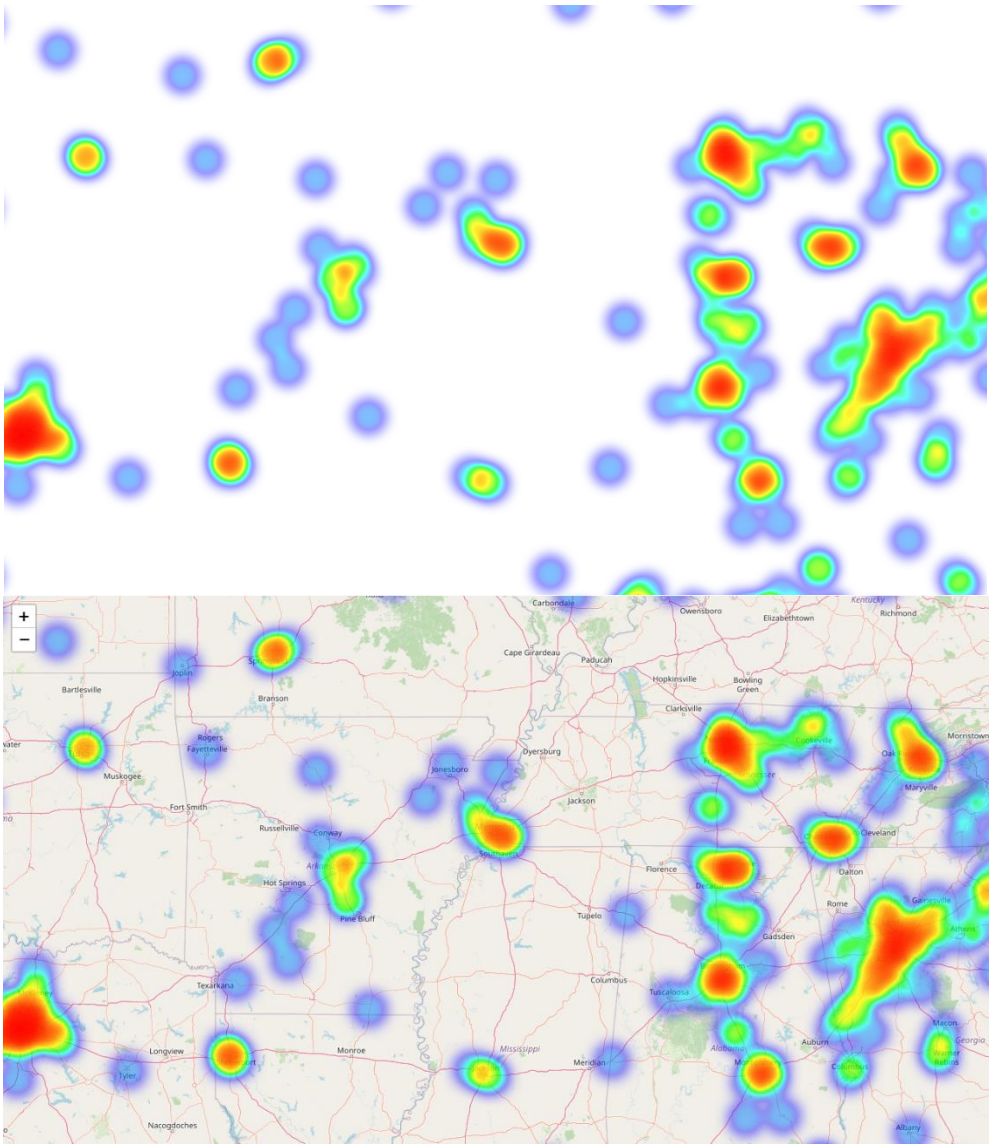
It seems like Precipitation level of every year varies drastically. 2017 has the overall highest precipitation level, whereas 2016 has the relatively low precipitation level. 2022 has the middle level of precipitation.



We also want to know if the start time of the accidents has change each of every year. During covid, the majority of jobs become work-from-home mode. Therefore, this new mode of working might have some impact on the number of accidents. Let's value count the number of accidents by start time per year in a new data frame.



I find out that 2022 still has the most number of accidents at every hour. Nevertheless, this may not reflect the reality because 2022 has siginificantly more reported accidents than in 2016 or previous year. Interestingly, the pattern of each year looks almost identical. Each year's line is just higher than the previous one (except the last year).



Above is another interesting map that shows where car accidents occur more frequently. Indeed, if we locate to New York City, we don't have any data to show. If we locate to Miami and Los angeles, the number of car accidents is high. East and west coast have the highest number of car accidents, whereas the middle of the states are relatively more 'safe.'

This map may not be able to tell much things about the COVID-19 though.

It can be much usable and interpretable in code this is the just a view of map

Act

Findings

- New York State car accidents data is not included in the dataset.
- Miami is the city with the most number of car accidents.
- The Temperature is almost normally distributed.
- Wind Speed, Visibility, Precipitation all follow a right skewed distribution.
- Humidity on the other hand follows a left skewed distribution.
- There are 144 types of weather condition in the dataset.
- Temperature does not have much of an impact on the severity of the accident. However, extreme weather temperature, like below -50 F degree, usually only causes accidents that have a severity score of 2.
- Car accidents are most likely to happen at the temperature around 50-80 F degree.
- Severe car accidents are mostly accompanied with a low visibility (below 50 mi). Meanwhile, The least severe car accidents tend to have the poorest visibility (below 20 mi).
- Precipitation and the number of accidents have a simple inverse relationship. The heavier the rain, the fewer the number of car accidents.
- Precipitation does not affect the severity of the accidents as much as the previous factors. However, during heavy rain(25inch), accidents mostly have a severity score of 2.
- Wind Speed affects the severity of the accidents almost equally.
- High wind speed, above 20mph, accompanies with 0 reported car accidents. Low wind speed, below 20 mph, accompanies with more car accidents as it decreases.
- A high percentage of accidents occur between 15 to 18. Probably people'r hurry to get home. The next highest one is around 6 to 8.
- On weekends the number of accidents happened at 0 o'clock is higher than weekdays'.
- Weekday accidents distribution is almost the same as the overall accidents distribution.
- For every year from 2016-2023, the number of reported car accidents increases. 2022 has the most number of car accidents.
- Temperature of accident trends of every year are stable.
- 2017 being the year has the most precipitation.
- 2022 still has the most number of accidents at every hour.reported accidents than in 2016 or previous year. Interestingly, the pattern of each year looks almost identical. Each year's line is just higher than the previous one.
- During covid 19 months, the number of car accidents has increased significantly. Feb 2020 is the start of COVID.
- During covid 19 months, though most jobs become work from home. The number of car accidents happened at every hour has the same pattern as previous years though much higher.
- Cities with the most Covid-19 cases have more car accidents.

Answers to the hypothesis

- Cold weathers does not have a strong relation with the number of traffics. Moderate temperatures, 50-80F, correspond with high number of accidents. Extreme weather temperature, higher than 90 or lower than -50, do result in more severe accidents.
- Low visibility (<5mi) has a strong relation with the number of accidents. Low visibility affects the severity of accidents almost equally, no matter what the severity score is.
- A high precipitation level does not cause more accidents. People may stay at indoors more often when rains are heavy. A high precipitation does cause more severe accidents more often.
- A high wind speed does not cause more accident. People may stay at indoors more often when winds are strong. A high wind speed does cause more sever accidents more often.
- No matter what the humility level is, it does not hava a strong relation with the severity of accidents. Though most of the reported accidents due to humility have a severity score of 2.
- COVID-19 could have some impacts on the increased number of car accidents. Nevertheless, the increased number of car accidents might be due to other factors that are not shown in the dataset. During this COVID-19 period, many factors could account for the increased number of car accidents. COVID-19 is just one of the explanations. During the COVID-19 period, starting from Feb 2020, the number of car accidents increase so much faster and higher than previous years. Since the temperature and humidity level do follow very similar trends across the years, the number of increased car accidents should not be due to these two environmental factors. The precipitation level varies yearly drastically, so it may have some minimal impact on number of car accidents during the COVID-19 period. During the COVID-19 period, people work from home, but the number of accidents at every hour in 2021 is still the highest, so the working from home mode does not decrease the number of car accidents during this COVID-19 period.

At the conclusion of this work, normal hypothesis tests will be conducted to gain further insights into the factors influencing traffic accidents in the USA.

Hypothesis : The Average Severity of Car Accidents in the USA is above 2.5

```
Results for Hypothesis Test 1:
Null Hypothesis: The average severity of car accidents in the USA is 2.5 or lower.
Alternative Hypothesis: The average severity of car accidents in the USA is above 2.5.
Sample Data Mean: 2.212384970719349
T-Statistic: -1640.021871231594
P-Value: 0.0
Hypothesis Accepted: alternative
```

Reject hypothesis : doesn't really work because the variance of the data is very low.

Hypothesis, the average distance that an accident affects the road would be one mile.

```
Results for Hypothesis Test 2:
Null Hypothesis: The average distance that an accident affects the road is one mile or less.
Alternative Hypothesis: The average distance that an accident affects the road is more than one mile.
Sample Data Mean: 0.5618422831523752
T-Statistic: -685.5418229790475
P-Value: 0.0
Hypothesis Accepted: alternative
```

Reject hypothesis, average accident distance is less than one mile