R Group 5 Project Report

Jakob Anderson, Nicole Coates, Carrie Cox

U.S. Car Accidents (2019-2021)

Situation

For this project, we analyzed three years of traffic accident data in the United States from 2019-2021. The data was collected using Traffic APIs in real-time from February 2016 to December 2021 and was made publicly available on Kaggle. The dataset originally contained roughly 3 million observations, so we focused on only the final three years for analysis. The data set has variables related to location, time of day, road features, weather conditions, and severity of traffic accidents. We are analyzing traffic data to determine which factors tend to be correlated with increased frequency or severity of accidents.

The Questions to be answered

1. What state has the most accidents?
2. How does the length of road affected by an accident relate to road closures?
3. What patterns are there between different weather conditions and frequency of accidents?
4. How many accidents tend to be severe vs not severe? For each year?
5. What time of day has the most accidents?
6. What road feature is associated with the most accidents?

Analysis & Answering Project Questions

Accident Overview

To examine traffic accidents by state, we wanted to use choropleth maps to visualize the number of accidents. First, our data set did not have the FIPS code, so it had to be merged with state.region data frame in R (based on state abbreviation) to provide choropleth maps with the

correct region code. **Figure 1** below visualizes the number of reported traffic accidents over 3 years by state.

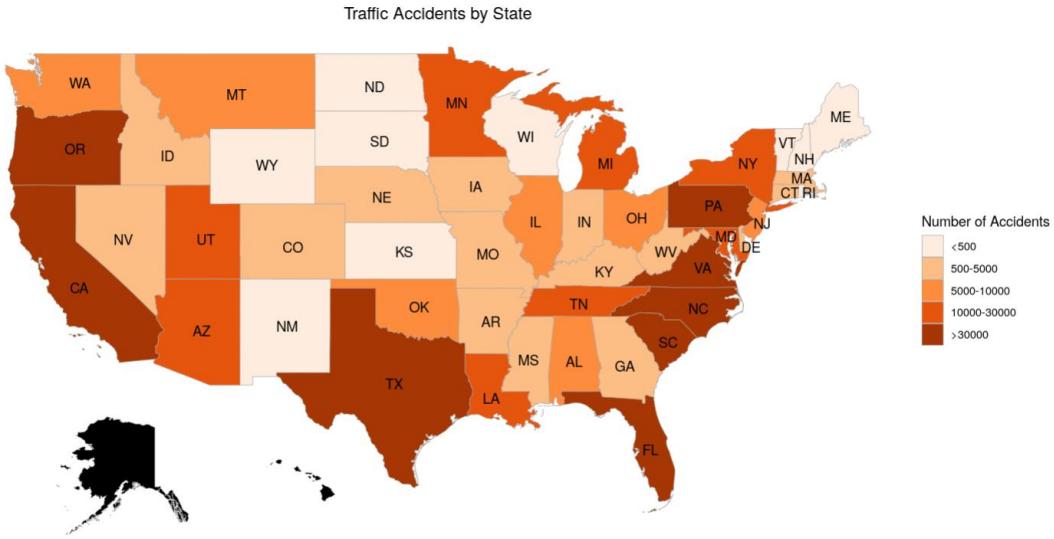

Figure 1

From 2019 – 2021, Oregon, California, Texas, Florida, Pennsylvania, Virginia, North Carolina and South Carolina were the states with the highest number of traffic accidents. Each of these eight states had over 30,000 accidents over the course of 3 years.

Florida has the highest number at 217,520 traffic accidents. Of 9 counties in Florida with more than 5,000 accidents, 5 of them are bordering each other and on the inner coast of the Florida peninsula, suggesting this may be an area to target for increase safety measures.

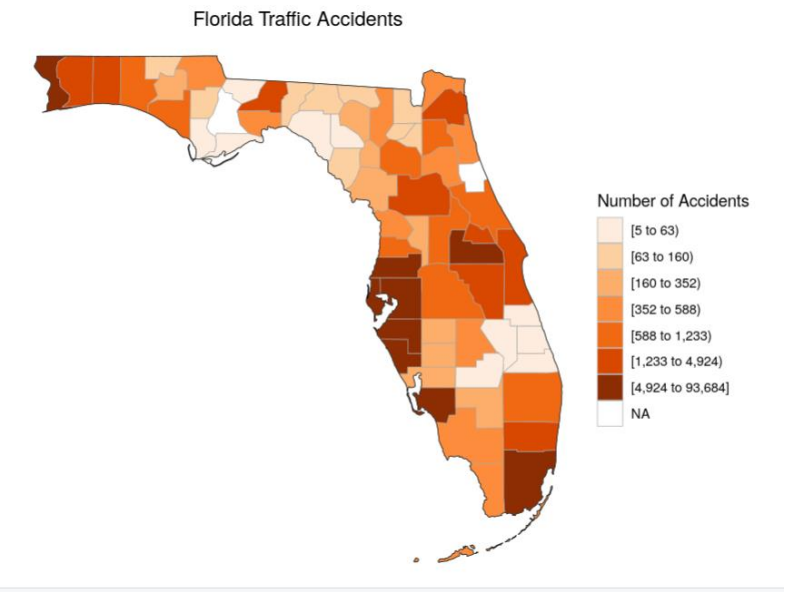

Figure 2

Comparing Florida's accident count by month is also revealing. While accidents are similar in number from January – August, there is a noticeable increase from October – December. Florida does not experience snowy winter weather like the northern states, so an increase may seem unexpected. However, this data was collected from Lyft, a rideshare service. Rideshare services see their busiest season from October—December around the holidays. It's likely that this increase in accidents is simply due to an increase in rideshares being used.

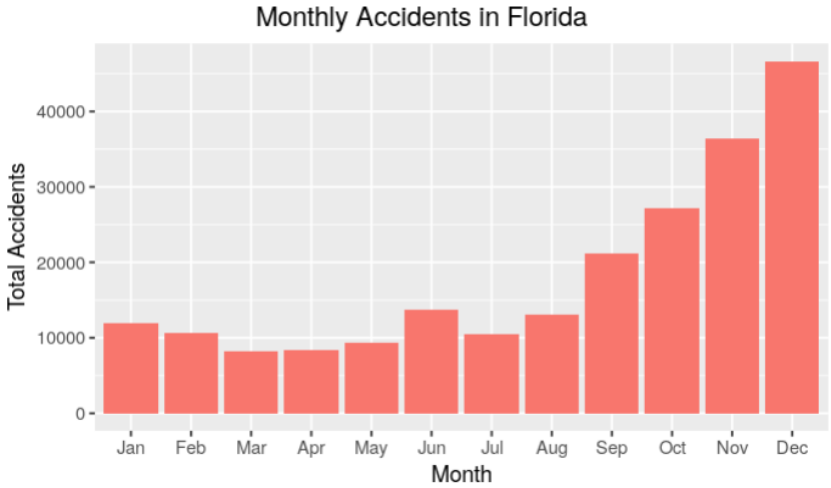

Figure 3

Severity of Accidents

There is a discrepancy between different years in this data set and the total number of accidents as shown in **Figure 4**. This could be because of a higher frequency of accidents in certain years; however, it is likely due to the amount of data collected.

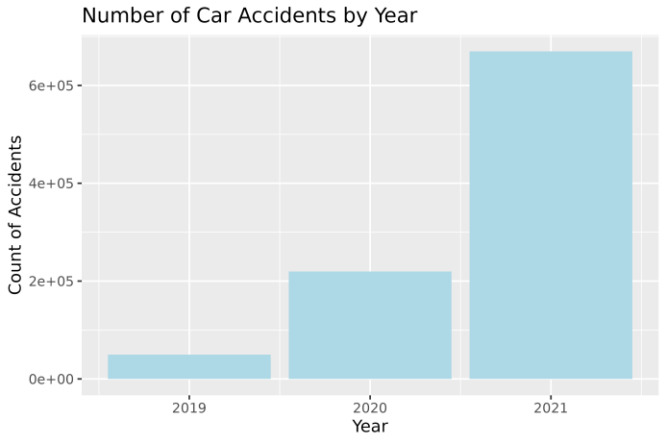

Figure 4

For that reason, when comparing the three years in our project, we chose to look at the severity of accidents rather than the frequency. The severity score is a score of 1-4 on how intense the accident was. We can see in **Figure 5,** that most accidents are labeled as a severity level of 2.

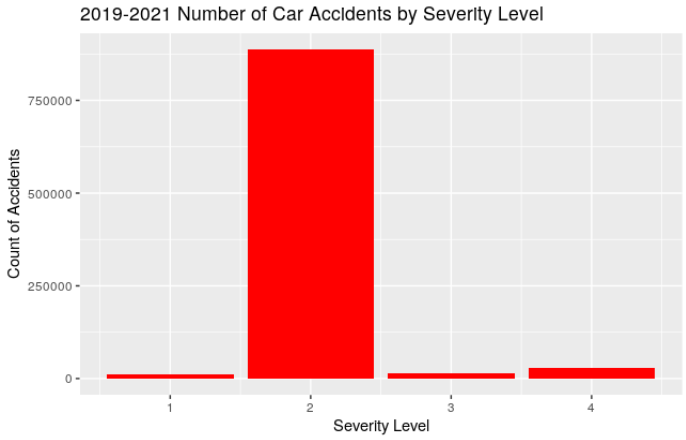

Figure 5

Next, we decided to look at accidents with a score of 1 and 2 as “not severe” and scores of 3 or 4 “severe”. In **Tables 6-7** below, you can see the number of accidents, and percentage of severe and non-severe accidents for each year. Overall, we see that 2021 had the most accidents, but 2019 has the highest percentage of severe accidents.

Year	percent_of_severe_accidents	percent_of_nonsevere_accidents	num_accidents
2019	0.2130577	0.7869423	49672

Year	percent_of_severe_accidents	percent_of_nonsevere_accidents	num_accidents
2020	0.08559328	0.9144067	219585

Year	percent_of_severe_accidents	percent_of_nonsevere_accidents	num_accidents
2021	0.01672215	0.9832778	669531

Figure 6-7

Weather Conditions

In this data set, we chose to analyze the following variables relating to weather: Weather Condition, Temperature, Precipitation, Wind Direction, Visibility, Wind Speed, Wind Chill, Humidity, and Pressure. However, the majority of accidents occur in more populated areas of the country and these areas tend to have milder, temperate weather. Because of this, many of these variables did not show higher frequencies or severity scores of accidents with worse weather conditions. Of the nine weather variables, Wind Direction and Wind Speed showed no significant or relevant trends. This conceptually makes sense because these two factors do not typically make driving conditions worse. Below in **Figures 8-9** you can see two tables for Weather Condition and the average severity score (left) and the number of accidents (right). The table on the right is ordered by number of accidents and only shows the top 10 weather conditions. There did not seem to be any trends in more accidents when weather conditions are worse. For that reason, I created a data frame with a few different conditions and their average

severity score. Heavy Snow followed by Light Snow had the highest severity score on average. From this, we can conclude that snow does cause accidents to be more severe in the U.S.

condition	severity
Fair	2.052782
Cloudy	2.080695
Mostly Cloudy	2.054762
Partly Cloudy	2.058686
Light Rain	2.077479
Light Snow	2.122966
Heavy Snow	2.135048
T-Storm/Windy	2.010309

Weather_Condition	num_accidents
Fair	465574
Cloudy	139599
Mostly Cloudy	123370
Partly Cloudy	84892
Light Rain	38436
Fog	17225
Haze	11161
Light Snow	9710
Rain	7191
Fair / Windy	6107

Figure 8-9

Throughout this analysis, all of the weather-related variables were analyzed. However, most of these only showed trends of more accidents, and higher severity scores for milder conditions. Most accidents happen in temperatures between 50-90 degrees, with no precipitation, and a higher visibility of between 6-10 miles. There are similar patterns within the variables Pressure, Wind Chill, and Humidity. All of these variables show more accidents and more severe accidents when the weather is nice out. From this analysis, we concluded that there are other factors that need to be considered when comparing car accidents and weather conditions. First, more accidents occur in populated areas, and these areas tend to have warmer weather without extreme weather conditions. Second, it is important to consider how many days of the year have clear weather compared to poor weather. And thirdly, each state has different weather conditions and that could affect the severity score and frequency of accidents (for example, some states have snow and that could lead to more or worse accidents).

Accident Duration and Time of Day

A new column was created that indicates what hour the accident started in. Utilizing this new column “hour”, we created a summary table and visualization that shows accident count in each hour.

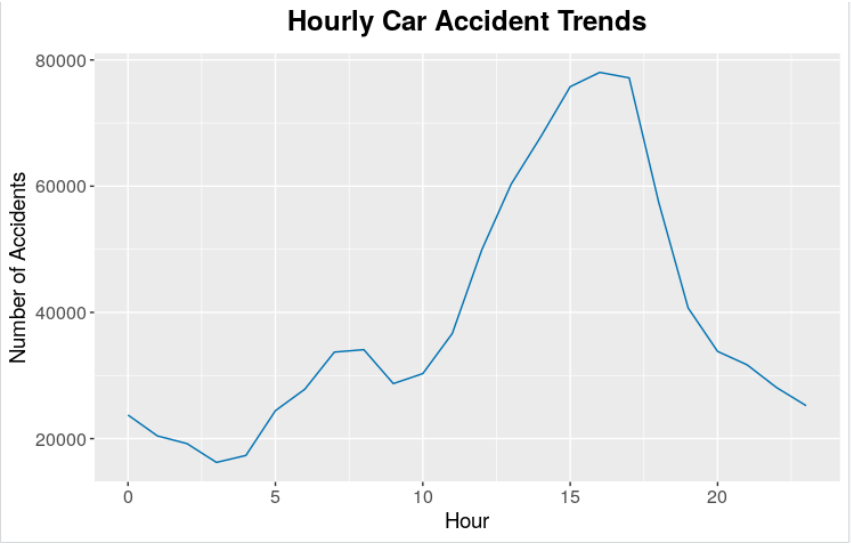

Figure 10

Accidents spike from 3pm-6pm with the peak at 4pm. This is in line with rush hour traffic as people are leaving the workday. We probably aren’t seeing the same spike in the morning rush hour because the dataset isn’t on total US accidents, and Lyft drivers might not be out yet.

Road Features

There were 13 road features in the dataset. We analyzed each road feature to see where Lyft drivers were getting into accidents. From Figure 11, we can see that the majority of accidents are from Traffic Signals and Crossing. We wanted to take a deep dive into these variables to see if these accident trends are consistent with our overall accident reports.

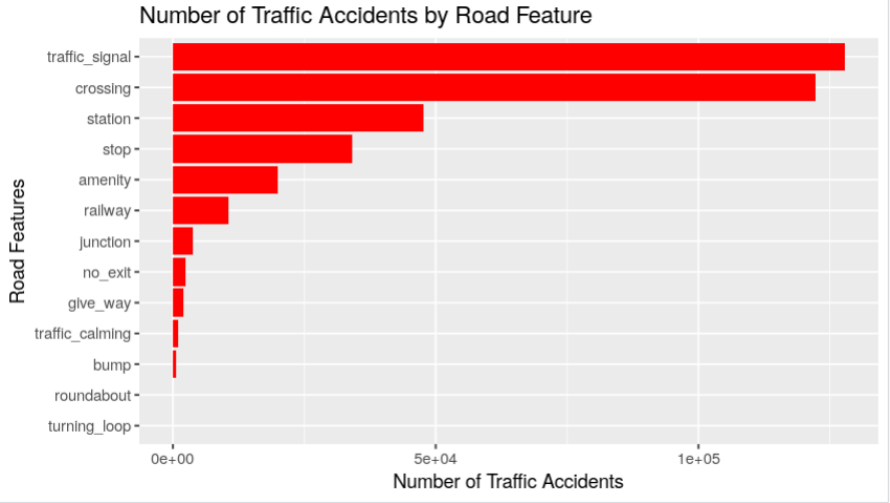

Figure 11

These figures showcase that there is no discernible difference between Traffic Signals, Crossing, and our national accident report. Therefore, we can conclude that there isn't a regional difference. People everywhere naturally get into accidents at these locations, and cities with higher levels of accidence have a proportion increase of accidents at these locations.

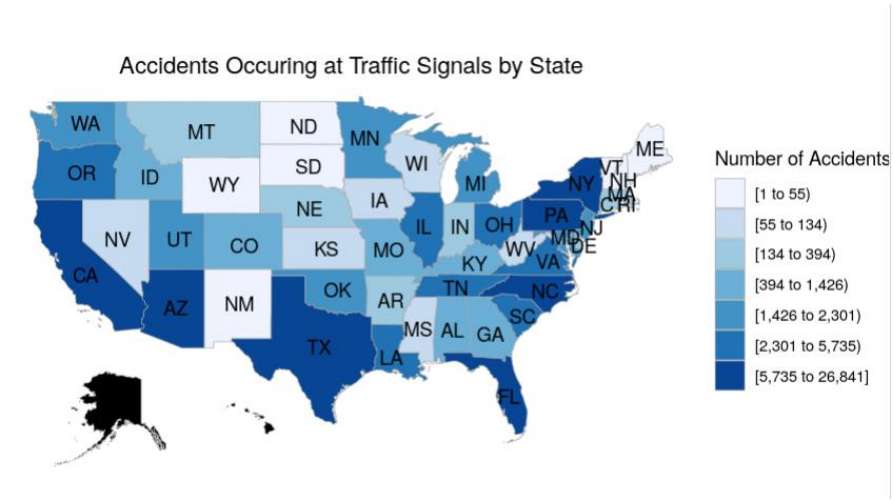

Figure 12

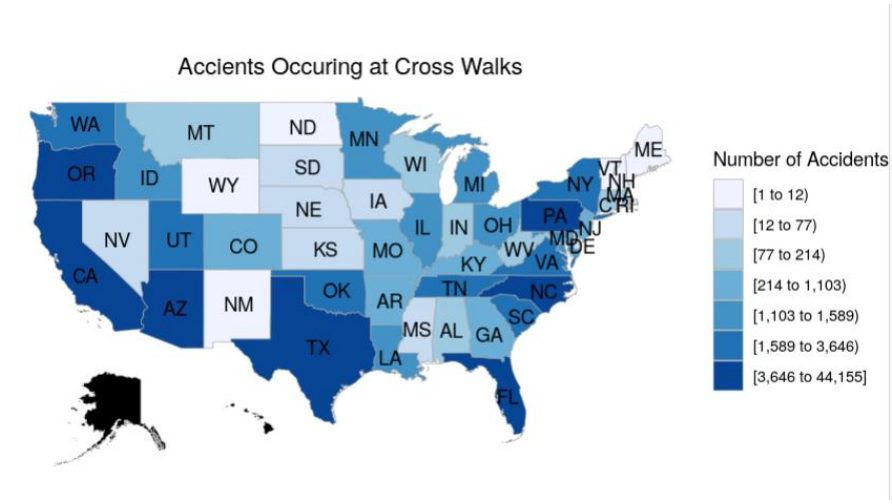

Figure 13

Finally, we looked at accident duration, the time in minutes it takes for traffic to return to normal. It's difficult to draw a conclusion without a greater description of the accident, but we can look at aggregate data. On average, railways cause the greatest traffic disturbance.

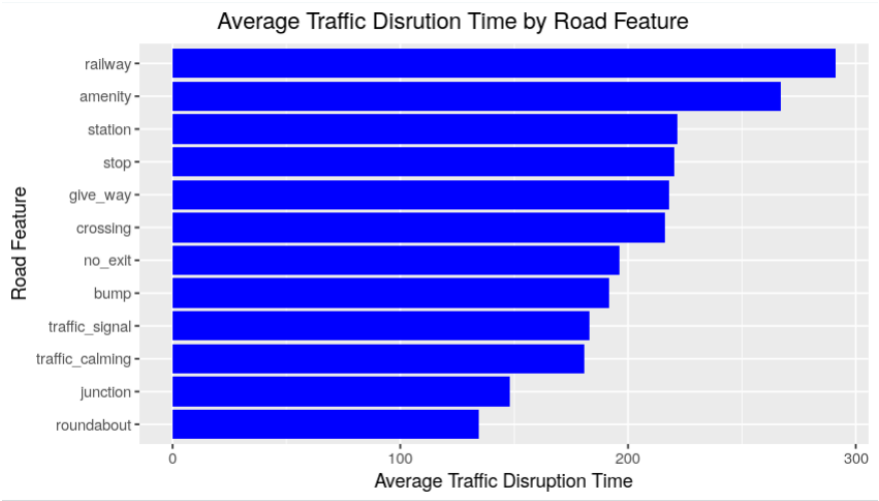

Figure 14

Analysis Conclusions

After analysis, we see some clear trends in the data that could be used to guide new safety initiatives. As more traffic accidents occur on coastal states, Lyft may want to get more insurance coverage for those specific areas. We also see a significant increase in accidents around 4pm in the evening. During these times perhaps a notification could go out to drivers to be extra cautious when driving, a timely reminder to try and decrease accidents. From October – December, a busier time for rideshare, there could be reminders to drivers and passengers about travelling safely in cars. Finally, although there has been an increase in reported accidents from 2019 to 2021, the percentage of severe accidents has decreased, it could be beneficial to see what changes have occurred to limit more serious accidents. Analysis of traffic accidents has potential to help rideshare companies, insurance companies, emergency services and drivers alike to improve safety on the roads.

Data Processing and Preparation

This data set included over three million rows. For that reason, we downloaded the data set from Kaggle (US Accidents (2016 - 2021) | Kaggle) and uploaded it directly into R without opening it. We found that opening the data set in Excel deleted half of the data because it exceeded the row maximum. After reading the data into an R script, we found that there were 3,414,239 total NAs, and the following row and column counts shown below in **Table 15.**

NA_total	Row_count	Column_count
3414239	2845237	47

Table 15

All NAs were dropped, and the following columns were removed because we did not want to incorporate them in our analysis: Civil Twilight, Nautical Twilight, Astronomical Twilight, End Latitude, Eng Longitude, Airport Code, and Weather Timestamp. Next, we created a subset of

the data to only look at the years 2019-2021. Since this was such a large data set, we wanted to focus on only a couple years to be able to do yearly analysis. The last steps of data cleaning included renaming columns for our own convenience and creating the new columns Date and Year. **Table 16** shows the cleaned data NA**,** row and column count. **Table 17** shows the mean value for all continuous variables in this data set (after being cleaned).

NA_count	Row_count	Column_count
0	938788	42

latitude_mean	longitude_mean	distance_mean	temperature_mean	windchill_mean	humidity_an	pressure_mean	visibility_mean	windspeed_mean	precipitation_mean
35.04027	-95.0968	0.2724135	63.9391	63.07302	64.46602	29.41997	9.213883	7.119864	0.004278474

Figure 16 & 17

Project Obstacle Summary

While we overall had a successful project, there were still a few obstacles that we were faced with. First of all, our data set had almost 3 million rows. Our first mistake was downloading and opening this CSV file in Excel, because it deleted half of our data to not exceed Excel's row maximum. This was confusing when we were trying to look only at the years 2019-2021 and the only data that was showing was 2019. However, we realized that we should download the data and then directly upload it into R to avoid this situation. Another challenge was using choropleth maps, since our dataset only had state and county names, not FIPS codes which are necessary for choropleth maps. It required us to merge two datasets together to get a count of accidents in various counties and states, after counties in our dataset to lowercase, since both datasets had to have a direct match to merge based on county names. In addition, there were numerous counties with the same name. To focus in on one state, first the dataset needed to be subset to only see counties for Florida and zoom_state had to be set equal to "florida". Both of these conditions had to be added or it would not show the correct data in the choropleth map. A final challenge in this project was that code was reused multiple times to do the same process in different areas (ex:

state and county choropleth maps or bar chart of accidents for different weather conditions). It was easy to miss updating the code to the new section and was something we had to be extremely careful about.

Appendix

	column_name	description
1	ID	Unique Identifier
2	Severity	Severity of the accident - range (1-4), 1 indicates the least impact on traffic
3	Start_Time	Start time of accident
4	End_Time	End time of accident; when traffic returned to normal
5	Latitude	Latitude location of the accident site
6	Longitude	Longitude location of the accident site
7	Accident_Distance	The legnth of the road affected by the accident
8	Description	Human provided description of the accident
9	Street_Number	Street number in address field
10	Street	Street name
11	Road_Side	Lists what side of the street the accident occured (R or L)
12	City	City name
13	County	County name
14	State	State name (2 letter abbreviation)
15	Zipcode	5 diget Zipcode
16	Country	Country name
17	Timezone	Timezone in which the accident occured
18	Temperature	Temperature in Fahrenheit
19	Wind_Chill	Wind chill in Fahrenheit
20	Humidity	Humidity %
21	Pressure	Air pressure in inches
22	Visibility	Visibility in miles
23	Wind_Direction	Direction the wind was blowing
24	Wind_Speed	Wind speed in miles per second
25	Precipitation	Precipitation levels in inches
26	Weather_Condition	Weather condition during accident (Rain, Snow, etc.)
27	Amenity	States if the accident happend in an Amenity Zone (True/False)
28	Bump	States if there was a presence of a speed bump nearby (True/False)
29	Crossing	States if the accident happened near a crossing (True/False)
30	Give_Way	States if the accident happened near a yeild sign (True/False)
31	Junction	States if the accident happened at a junction (True/False)
32	No_Exit	States if the accident happened on a no exit street (True/False)
33	Railway	States if the accident happened near a railway (True/False)
34	Roundabout	States if the accident happened near a roundabout (True/False)
35	Station	States if the accident happened near a station (True/False)
36	Stop	States if the accident happened near a stop sign (True/False)
37	Traffic_Calming	If it happened near a traffic calming method (True/False)
38	Traffic_Signal	States if the accident happened near a traffic signal (True/False)
39	Turning_Loop	States if the accident happened near a turning loop (True/False)
40	Day_or_Night	States if the accidend happened during the day or at night
41	Date	Date of the accident (YYYY-mm-DD)
42	Year	Year of the accident

Cleaned Data Dictionary