Traffic Incident Project

Data Overview:

Data Location : https://www.kaggle.com/sobhanmoosavi/us-accidents
Website for information on data: https://smoosavi.org/datasets/us_accidents

Site for GeoJSON files: https://eric.clst.org/tech/usgeojson/

Request from collector to cite the following papers if dataset used:

- Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", arXiv preprint arXiv:1906.05409 (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.

Data Source : The data providers 'MapQuest Traffic' and 'Microsoft Bing Map Traffic' APIs broadcast traffic events. These events are captured by entities such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks.

Data Collection : Collector set up a process that monitored the output from the MapQuest and Bing Map traffic APIs. The process "pulled data every 90 seconds from 6am to 11pm, and every 150 seconds from 11pm to 6am. In total, we collected 2.27 million cases of traffic accidents between February 2016 and March 2019". (Per A Countrywide Traffic Accident Dataset paper noted below)

Data Contents : The data contains the time, location, and severity of a traffic accident. It also includes the nearest weather report and if various points of interest are nearby.

Why choose this data?
How people drive and what they do while driving is a common conversation topic in society. "People in Washington drive slow", "Californians can't drive in the rain", " Kids today are all driving while on their cell phones".
People often consider themselves good drivers in a manner similar to the cognitive bias of the above average effect where a person overestimates their own qualities and abilities. However, many dangerous behaviors such as cell phone use, driving under the influence, or speeding are commonplace. The continuation of these dangerous behaviors would seem to be an indication that while these behaviors raise the likelihood of accidents they do not guarantee them. This

leads me to think that per operant conditioning, people are often positively reinforced or at least not negatively reinforced for the behaviors and make them more likely to occur in the future.
In my personal experience in these conversations, a fairly universal measure of a safe driving record would be the number of traffic accidents one has had. I am curious to take the individual factor out of this to find out how common traffic accidents are. Having data from 49 states across the USA reported by local governments, law enforcement, and traffic sensors looks like a good place to start.

Data Profile

Quantitative

Variable	Description	Min	Max	Mean
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).	1	4	2.3
Start_Time	Shows start time of the accident in local time zone.	2/8/2016 0:37	12/31/2020 23:28	NA
End_Time	Shows end time of the accident in local time zone. End time here refers to when the impact of accident on traffic flow was dismissed.	2/8/2016 6:37	1/1/2021 0:00	NA
Start_Lat	Shows latitude in GPS coordinate of the start point.	24.6	49	36.5
Start_Lng	Shows longitude in GPS coordinate of the start point.	-124.6	-67.1	-96.4
End_Lat	Shows latitude in GPS coordinate of the end point.	24.6	49	36.5
End_Lng	Shows longitude in GPS coordinate of the end point.	-124.6	-67.1	-96.2
Distance(mi)	The length of the road extent affected by the accident.	0	333.6	0.4
Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).	2/8/2016 0:53	12/31/2021 23:35	NA
Temperature(F)	Shows the temperature (in Fahrenheit).	-77.8	132.6	61

Humidity(%)	Shows the humidity (in percentage).	1	100	65.4
Pressure(in)	Shows the air pressure (in inches).	25.7	31.2	29.7
Visibility(mi)	Shows visibility (in miles).	0	140	9.1
Wind_Speed(mph)	Shows wind speed (in miles per hour).	0	110	7.8
Precipitation(in)	Shows precipitation amount in inches, if there is any.	0	24	0.0064

Qualitative

Variable	Description	Value Count	Mode
ID	This is a unique identifier of the accident record.	2906610	NA
Description	Shows natural language description of the accident.	1484192	NA
Number	Shows the street number in address field.	41617	1
Street	Shows the street name in address field.	175527	I-5 N
Side	Shows the relative side of the street (Right/Left) in address field.	2	R
City	Shows the city in address field.	11790	Los Angeles
County	Shows the county in address field.	1729	Los Angeles
State	Shows the state in address field.	49	CA
Zipcode	Shows the zipcode in address field.	372019	91761
Country	Shows the country in address field.	1	US
Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).	5	US/Eastern
Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.	2014	KCQT
Wind_Direction	Shows wind direction.	19	CALM
Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)	129	Fair

Amenity	A POI annotation which indicates presence of amenity in a nearby location.	2	FALSE
Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.	2	FALSE
Crossing	A POI annotation which indicates presence of crossing in a nearby location.	2	FALSE
Give_Way	A POI annotation which indicates presence of give_way in a nearby location.	2	FALSE
Junction	A POI annotation which indicates presence of junction in a nearby location.	2	FALSE
No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.	2	FALSE
Railway	A POI annotation which indicates presence of railway in a nearby location.	2	FALSE
Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.	2	FALSE
Station	A POI annotation which indicates presence of station in a nearby location.	2	FALSE
Stop	A POI annotation which indicates presence of stop in a nearby location.	2	FALSE
Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.	2	FALSE
Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby location.	2	FALSE
Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.	1	FALSE
Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.	3	Day
Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.	3	Day
Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.	3	Day
Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.	3	Day

Derived Columns

Column Name	Column Derived From	Logic Used
start_date	Start_Time	Split Start_Time to return date portion of timestamp
start_time	Start_Time	Split Start_Time to return time portion of timestamp
time_of_day	start_time	After 21:00 or before 03:00 = Night After 03:00 and before 09:00 = Morning After 09:00 and before 15:00 = Day After 15:00 and before 21:00 = Evening
avg_state_sev	state & Severity	Average severity value per state
avg_state_temp	state & Temperatur(F)	Average temperature value per state
avg_state_rain	state & Precipitation(in)	Average precipitation value per state
start_time_float	Start_Time	Time of day from timestamp as floating point data type

Questions to Explore:

- In which states are accidents most common?
- What collection of POI or geographic variables have the highest rate of accidents?
- What is the frequency of accidents? Is there a relationship between time of day and accidents?
- Can we find the likelihood of an accident per region(city, state, region)?
- Are accidents more common in inclement weather?
- Do accidents occur more often during low light conditions?
- Where is the safest state to drive?
- What is the safest route from point A to point B?
- Not a single recorded value for accidents near a turning loop. Are any POI safer than others?
- What is the time gap between the accident and the weather timestamp?
- What is the most common severity of accidents?

Additional Questions after initial exploration

- Is there any difference in accidents per month or year? Overall or split into the time of day categories.
- Are accidents going up or down by state by year?

- What are the average weather values for accidents when grouped by State? Or perhaps what are the most common weather values for accidents per state?
- Is it possible to combine weather values into a single column that defines driving danger? This could be added to the above question of grouping by state.
- When POI are near, accidents most commonly occur where traffic crosses (Signal, Junction, Crosses). Do accidents more commonly occur on or off highways/freeways?

Hypothesis Ideas

- Developing hypothesis ideas
- Accidents will occur regardless of weather. The most commonly experienced weather conditions for a state will result in the most number of accidents.
- The worst weather conditions will result in a higher proportion of severe accidents than good weather conditions.
- If you are in state x, then you will be most likely be in an accident when weather conditions are y, the time of day is q, and your location is near z(likely either POI or highway vs roadway). Could also do least likely.
    - “Your last hypothesis idea is really good, I'm interested to see where this goes. My mind automatically goes to multivariate regression analysis, but that may be a bit too complicated if you are not used to the statistics.”

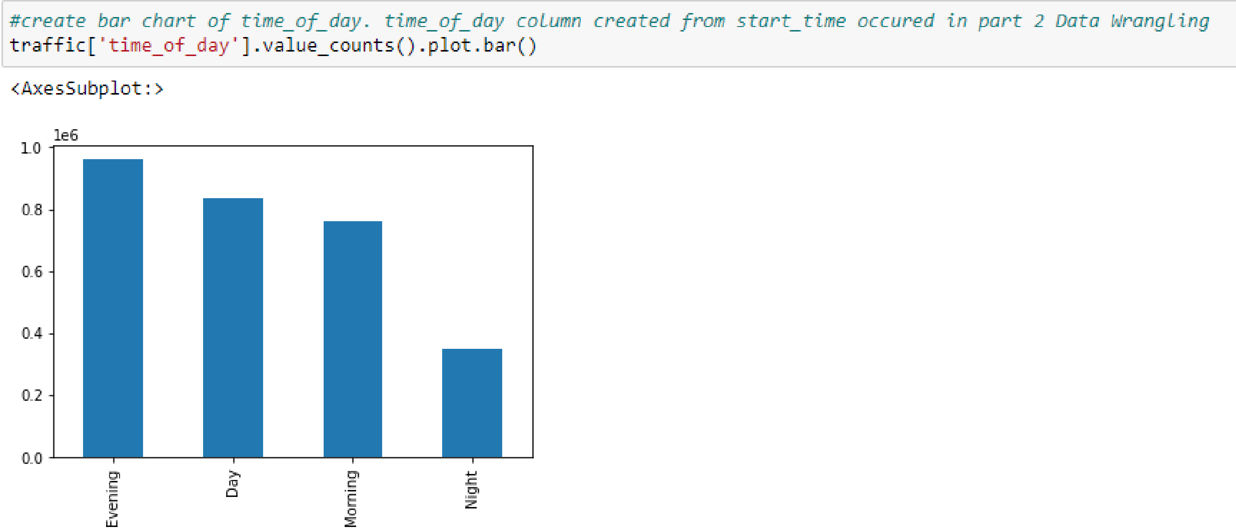
Dashboard Idea

- Condensed guideline questions
    - Who is it for?
        - National Auto insurance company
    - Why is it being built?
        - Used to assess differences between states to apply rates.
    - What will it consist of?
        - Where are most accidents occuring by state map
        - Accident count/trend over time. National and by state
        - Accidents by time of year / week / holidays
        - Does weather impact accidents?
            - Idea for hypothesis. By state, weather will be tied to the count of accidents
    - When will it be used?
    - Where will it be hosted?

Interesting Notes

- Most common severity is 2
    - 2: 2129263
    - 3 :629452
    - 4: 119144

    - 1: 28751
    - Note: wyoming has far fewer accidents than most states. But the majority of accidents are severity 4. Highest average severity for a state by far
- Accidents most common in CA. But eastern timezone has most accidents overall for timezone
- Wind direction most common value is calm
- Most common weather condition is Fair
- Most common ‘Street’ for accidents is I-5
- Most common city and county for accidents is LA
- Evening and day are most common accident times. Night is lowest



```
#create bar chart of time_of_day. time_of_day column created from start_time occured in part 2 Data Wrangling
traffic['time_of_day'].value_counts().plot.bar()
```

```
<AxesSubplot:>
```

    -
- Traffic Accidents near POI
    - The only times that accidents are located at points of interest more than 2 % of the time are:
        - Crossings at % 7.5
        - Junctions at % 9.4
        - Traffic Signals at %15.6
        - Near a roundabout accidents were reported 142 times or % 0.004
        -

Limitations and Ethics:

Data was collected by automated systems. The only limitations expected would be errant or extreme recordings from sensors or other automated systems that dont make sense but were not caught due to collection methods. I have addressed any values that seemed beyond the ordinary in the Accuracy section below.
Data is anonymous and is pulled from trustworthy and largely automated services. I foresee no ethical issues with data.

Cleaning Notes:

Missing Values

- Visibility(mi) removed 13464 missing values to allow regression test.
- End_Lat and End_Lng each have 282821 missing values on the same rows. This is almost 10% of the data. Descriptive stats like mean are much lower on the Distance(mi) variable for when End_Lat and End_Lng are null. All other numerical variables stay pretty similar. This coincides with End_Lat and End_Lng not being reported because the accidents started and ended close enough to not warrant reporting the end point. These values will be left alone.
- Number has 1891672 missing values (65% of data). This makes sense as many accidents can occur on highways where a number would not be recorded. These values will be left alone.
- Wind_Chill(F) has 1183859 missing values (41% of data). This column will be removed. Amount of missing values is too high for analysis not to be affected.
- Wind_Speed(mph) has 307163 missing values (11% of data). Unable to find a clear assumption to input a value based on Weather_Condition. Majority of values are conditions that would match having low wind levels. This matches the overall Wind_Speed(mph) average of 7.8 which falls between light and gentle breeze.These values will be left alone.
- Precipitation(in) has 1301326 missing values (45% of data). Top 7 Weather_Condition values where Precipitation(in) is null are Clear, Overcast, Mostly Cloudy, Partly Cloudy, Scattered Clouds, Fair, and Haze. These would make sense to not be rainy conditions and make up over 90% of missing values. Will mark these as 0 rain.
- All other variables with missing values (City, Zipcode, Timezone, Airport_Code, Weather_Timestamp, Temperature(F), Humidity(%), Pressure(in), Wind_Direction, Weather_Condition, Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight) are under 2% missing. Since they all have Start_Lat and Start_Lng values these variables with missing values will not be expected to affect calculations and will be kept.

Duplicates

- No full rows duplicates found.
- ID column had no duplicates and is only column that would be required to have unique values.

Mixed Data Types

- The following columns had multiple data types and were set to string data types.
    - City
    - Zipcode
    - Timezone
    - Airport_Code
    - Weather_Timestamp
    - Wind_Direction
    - Weather_Condition
    - Sunrise_Sunset

    - Civil_Twilight
    - Nautical_Twilight
    - Astronomical_Twilight

Accuracy

- Temperature(F) has values outside min(-80 F) and max(134 F) recorded in USA. Set these to Nan values.
- Pressure(in) has values outside min(25.69 in) and max(32.01 in) recorded in World. Unable to set these to NULL values.
- Set values for Wind_Direction below to match formats:
    - Calm to CALM
    - South to S
    - West to W
    - North to N
    - East to E
    - Variable to VAR
- Set the Side value not equal to R or L to NULL
- Removed 63 Wind_Speed(mph) values over 111mph as values indicated largely clear weather and that is the lower level for the designation of major hurricaine.

Other Notes

- Country only has 1 unique value ‘US’ for all rows. Removing column.

Changes to Data:

Changes to Row Numbers:

- Original Rows = 2906610
- Removed rows where Precipitation(in) is greater than 10
    - Rows = 2798952
- Removed ros where Visibility(mi) is null.
    - Rows = 2785488

Changes to Column Numbers

- Original Columns = 47 columns
- Post cleaning = 45 columns
    - Wind_Chill(F) and Country removed
- Post wrangling and exploring =
    - Added time_of_day, start_time, and start_date
    - Removed 'Sunrise_Sunset', 'Civil_Twilight', 'Nautical_Twilight', and 'Astronomical_Twilight'

Data Value Changes:

- Updated NULL values in Precipitation(in) column to 0 where Weather_Condition values are Clear, Overcast, Mostly Cloudy, Partly Cloudy, Scattered Clouds, Fair, or Haze.

Data Exploration:

- The column pairs of starting and ending latitude and longitude show a correlation of 1 but the columns themselves are not equal.