Ariel Johnson

Sourcing Open Data

Exercise 6.1

Data Set: U.S Accidents (2016-2023)

https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

Data Source:

This data set is: U.S Accidents (2016-2023).

This is an external data source that I downloaded from Kaggle giving acknowledgements and citing:

Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. "A Countrywide Traffic Accident Dataset.", 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 Collection:

This dataset was collected in real-time using multiple Traffic APIs.

Data Contents:

This is a countrywide car accident dataset that 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. These APIs broadcast traffic data captured by various entities, including the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road networks. The dataset currently contains approximately 7.7 million accident records.

#	Attribute	Description
1	ID	This is a unique identifier of the accident record.
2	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).

#	Attribute	Description
3	Start_Time	Shows start time of the accident in local time zone.
4	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.
5	Start_Lat	Shows latitude in GPS coordinate of the start point.
6	Start_Lng	Shows longitude in GPS coordinate of the start point.
7	End_Lat	Shows latitude in GPS coordinate of the end point.
8	End_Lng	Shows longitude in GPS coordinate of the end point.
9	Distance(mi)	The length of the road extent affected by the accident.
10	Description	Shows natural language description of the accident.
11	Number	Shows the street number in address field.
12	Street	Shows the street name in address field.
13	Side	Shows the relative side of the street (Right/Left) in address field.
14	City	Shows the city in address field.
15	County	Shows the county in address field.
16	State	Shows the state in address field.
17	Zipcode	Shows the zipcode in address field.
18	Country	Shows the country in address field.
19	Timezone	Shows timezone based on the location of the accident (eastern, central, etc.).
20	Airport_Code	Denotes an airport-based weather station which is the closest one to location of the accident.
21	Weather_Timestamp	Shows the time-stamp of weather observation record (in local time).

#	Attribute	Description
22	Temperature(F)	Shows the temperature (in Fahrenheit).
23	Wind_Chill(F)	Shows the wind chill (in Fahrenheit).
24	Humidity(%)	Shows the humidity (in percentage).
25	Pressure(in)	Shows the air pressure (in inches).
26	Visibility(mi)	Shows visibility (in miles).
27	Wind_Direction	Shows wind direction.
28	Wind_Speed(mph)	Shows wind speed (in miles per hour).
29	Precipitation(in)	Shows precipitation amount in inches, if there is any.
30	Weather_Condition	Shows the weather condition (rain, snow, thunderstorm, fog, etc.)
31	Amenity	A POI annotation which indicates presence of amenity in a nearby location.
32	Bump	A POI annotation which indicates presence of speed bump or hump in a nearby location.
33	Crossing	A POI annotation which indicates presence of crossing in a nearby location.
34	Give_Way	A POI annotation which indicates presence of give_way in a nearby location.
35	Junction	A POI annotation which indicates presence of junction in a nearby location.
36	No_Exit	A POI annotation which indicates presence of no_exit in a nearby location.
37	Railway	A POI annotation which indicates presence of railway in a nearby location.
38	Roundabout	A POI annotation which indicates presence of roundabout in a nearby location.
39	Station	A POI annotation which indicates presence of station in a nearby location.
40	Stop	A POI annotation which indicates presence of stop in a nearby location.
41	Traffic_Calming	A POI annotation which indicates presence of traffic_calming in a nearby location.

#	Attribute	Description
42	Traffic_Signal	A POI annotation which indicates presence of traffic_signal in a nearby loction.
43	Turning_Loop	A POI annotation which indicates presence of turning_loop in a nearby location.
44	Sunrise_Sunset	Shows the period of day (i.e. day or night) based on sunrise/sunset.
45	Civil_Twilight	Shows the period of day (i.e. day or night) based on civil twilight.
46	Nautical_Twilight	Shows the period of day (i.e. day or night) based on nautical twilight.
47	Astronomical_Twilight	Shows the period of day (i.e. day or night) based on astronomical twilight.

Why I chose this dataset:

I chose this dataset because personally I find it interesting to explore where car accidents occur the most frequently and determine why those certain spots might be dangerous for drivers, and then come up with solutions to increase driver safety in those areas. I think that this project could provide insights to aid the National Highway Traffic Safety Administration (NHTSA). This data set specifically offers good information about locations and weather conditions during car accidents, which is why I chose this one opposed to other car accident data sets.

Data Profile

Cleaning :

- This data set initially had 7728394 rows and 46 columns
- I imputed 'Unknown' value into missing values for 'City', 'Zipcode', and 'Description' columns.
- Dropped 'End_Lat' & 'End_Lng' columns, as they were not crucial, and contained 3,402,762 missing values.
- Dropped unnecessary columns to analysis: Sunrise_Sunset, Civil_Twilight, Nautical_Twilight, Astronomical_Twilight, Timezone, Airport_Code, Weather_Timestamp, and Wind_Direction.
- Dropped missing street rows, resulted in dropping 10,869 rows. Knowing the street location is crucial for providing helpful insights about car accidents, so if that data is missing, I removed it.
- Imputed missing values with the mean for Temperature, Wind Chill, and Humidity columns.

- Imputed missing values with the median for Pressure, Visibility, Wind Speed, and Precipitation columns.
- Imputed missing values with the mode for the Weather Condition column.
- Changed data type of "Start Time" column from object to datetime format. This resulted in 740,771 missing rows. I dropped all missing rows from this column. This will be fine, because for a specific value I can revert to the 'End Time' column for reference. I needed to drop the missing rows from the 'Start Time' column, so that I could derive 'Year', 'Day', and 'Month' columns from the 'Start Time' column.
- There were 505577 rows containing outliers. After conducting descriptive statistics, I found that there were outliers in the Temperature and Wind Speed columns that exceeded historically accurate values. The max temperature was 207 F and the max wind speed was 1,087 mph. I understand that there are weather phenomenon's, but with extra research, I saw that these exceeded any recordings. Perhaps this is due to malfunctions in the recording technology, or input error. Rather than getting rid of all of the outliers, I set a cap, removing all records that exceed the capped maximums of 150 degrees Fahrenheit and a wind speed of 200 mph.
- After cleaning, dataset contains 6,976,702 rows and 39 columns.

Descriptive Statistics:

Severity

Count	6,976,702
mean	2.23
Min	1
Max	4
Std.	.49

Distance(mi)

Count	6,976,702
Mean	.51
Min	0
Max	441.7
Std.	1.74

Temperature (F)

Count	6,976,702
Mean	61.6
Min	-89
Max	143
Std.	18.7

Wind Chill (F)

Count	6,976,702

Mean	58
Min	-89
Max	140
Std.	18.9

Humidity(%)

Count	6,976,702
Mean	65
Min	1
Max	100
Std.	22.5

Pressure(in)

Count	6,976,702
Mean	29.5
Min	0
Max	58.6
Std.	.97

Visibility(mi)

Count	6,976,702
Mean	9.1
Min	0
Max	140
Std.	2.6

Wind Speed(mph)

Count	6,976,702
Mean	7.6
Min	0
Max	190
Std.	5

Precipitation(in)

Count	6,976,702
Mean	0.006
Min	0
Max	36.4
Std.	.09

Year

Count	6,976,702
Mean	2019
Min	2016

Max	2023
Std.	1.8

Month

Count	6,976,702
Mean	6
Min	1
Max	12
Std.	3.6

Day

Count	6,976,702
Mean	15
Min	1
Max	31
Std.	8.6

Limitations and Ethics:

Missing data: The dataset may be missing data for certain days, which could be due to network connectivity issues during data collection.

The dataset comes with a usage policy and legal disclaimer stating the following: “This dataset is being distributed solely for research purposes under the Creative Commons Attribution-Noncommercial-ShareAlike license (CC BY-NC-SA 4.0). By downloading the dataset, you agree to use it only for non-commercial, research, or academic applications. If you use this dataset, it is necessary to cite the papers mentioned above.” (Papers cited under “Data Source” section.

Define Questions to Explore

- What is the overall distribution of accidents by year, month, and day of the week?
- How do accident rates vary by time of day? Are there peak hours for accidents?
- Are accidents more likely to occur during specific weather conditions (rain, snow)?
- What geographical areas have the highest accident rates (cities, countries, specific intersections)?
- Are accidents more likely in urban or rural areas?

- Which roads or intersections have the highest number of accidents?

- How do road types (highways, urban streets) affect accident severity and frequency?

- What factors contribute most to the severity of accidents (weather conditions, location, etc according to descriptions)?
- How does severity vary by geographic area or road type?
- What is the correlation between weather conditions and accident severity?

- How do accidents and severity change between seasons?

- Do areas with more traffic calming measures (speed bumps, roundabouts) have fewer accidents?

- How does precipitation impact the likelihood of accidents occurring?
- Is there a correlation between wind speed and the severity of accidents?