Data Visualization

BANA – 6760 E01

University of Colorado, Denver

14th March, 2022
Project: US Accidents

Urvish Patel

Table of Contents

Problem Statement.

Upon searching for various datasets I found one that seems interesting to me and with which I can enhance my visualization, presentation, R coding skills. The dataset is from Kaggle.com (https://www.kaggle.com/sobhanmoosavi/us-accidents).

Now suppose this is year 2017 and I am an executive who works for Department of Transportation and with increase in Road accidents, the department head wants to know the following things which I will be working on.

1. Which state has the highest number of accidents?
2. Which Zip code/County that has the highest number of accidents?
3. At what time do accidents usually occur in the US?
4. At what day of week accidents occur most?
5. What are the Factors Affecting Accident Severity?
6. Predicting the location of the accident.
7. Predicting the accident prone zone in each state.
8. Generating the EDA for the data.
9. Predicting the number of accidents might occur in next few years.
10. Getting possible trends in Number of accidents. (Like every March of each year has highest number of accidents).

This project includes the cleaning of dataset and generating EDA for the data which consists of 1,516,064 rows and 47 columns which are listed below (It is a .csv file)

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).
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).
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 Boolean type which indicates presence of amenity in a nearby location.
32. Bump A Boolean type which indicates presence of speed bump or hump in a nearby location.
33. Crossing A Boolean type which indicates presence of crossing in a nearby location.
34. Give_Way A Boolean type which indicates presence of give_way in a nearby location.
35. Junction A Boolean type which indicates presence of junction in a nearby location.
36. No_Exit A Boolean type which indicates presence of no_exit in a nearby location.
37. Railway A Boolean type which indicates presence of railway in a nearby location.
38. Roundabout A Boolean type which indicates presence of roundabout in a nearby location.
39. Station A Boolean type which indicates presence of station in a nearby location.
40. Stop A Boolean type which indicates presence of stop in a nearby location.
41. Traffic_Calming A Boolean type which indicates presence of traffic_calming in a nearby location.
42. Traffic_Signal A Boolean type which indicates presence of traffic_signal in a nearby loction.
43. Turning_Loop A Boolean type 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.

This Analysis will be useful to the Road safety department, hospitals and response teams as they will be prepared for such accidents and will be knowing the measures to take beforehand the accident occur in an accident prone zone (after analysis). This will tell us what step must be taken to avoid such accidents. Will be recommending some actions to take when a similar condition arrives.

Methodologies I intend to use:

1. Regression models (linear/ multilinear/ logistic)
2. Time series forecasting.
3. Various libraries for less coding in RStudio.
4. Dashboard with great UI.
5. Exploratory Data Analysis.
6. Geospatial visualization.
7. Some Calculated Fields.
8. User Input parameter in tableau. (Will try to apply this feature).
9. Time series Visualization.

Tools I will be using:

1. RStudio
2. Tableau
3. PowerPoint

Tableau

In RStudio:

I will be using this for Exploratory Data Analysis and to build some regression models. Will also try to predict values (Number of Accidents in Future) in RStudio, while training and testing my dataset. I believe RStudio is a great tool for Statistical Analysis so there will be some statistics involved, like getting to know the normality of columns, mean, median, mode, etc.

In Tableau:

I will be using Tableau to make visualizations, Presenting to my boss and uploading to the department of transportation website for the citizens, so that they can evaluate and take precaution beforehand if something similar happens to them while on road.

In PowerPoint:

I will be presenting all my findings to my boss, so that the accident department could be alerted of this findings. And make several measures and plan to reduce these accidents in future. This will include my suggestions and views for the same.

References

https://public.tableau.com/app/profile/alexandre.papagiannidis4540/viz/AccidentsaMontreal/AccidentsinMontreal

https://public.tableau.com/en-us/search/all/%23accident

https://www.thedataschool.co.uk/conrad-wilson/creating-user-input-calculations-using-parameters-tableau-tips-and-tricks

https://smoosavi.org/datasets/us_accidents