ANALYSIS OF ACCIDENTS IN UNITED STATES

CSE-564 Project Report

Submitted By:
Anuroop Katiyar (112609023)
Gourav Mangla (112958716)
[Group No. – 7]

INTRODUCTION

Approximately 1.35 million people die in road crashes each year, and around 3,700 people lose their lives every day on the roads. An additional 20-50 million suffer non-fatal injuries, often resulting in long-term disabilities. Road traffic injuries are the leading cause of death among young adults aged 15-44 account. Road crashes are the single greatest annual cause of death of healthy U.S. citizens traveling abroad. These are some really staggering numbers which rise great concerns. The government wants to reduce these numbers with the help of some tool that would help state and federal police to take preventive measures.

OBJECTIVE

The main goal of our project was to create an interactive web visualization which could be helpful for the government to gain insights about the accidents across the United States. We planned to visually analyze the accidents state wise so that the different state police can have access to information about particular state through interactive user interface. We will also try to find some correlation between the accidents and the population of different states to get any interesting relation if it exists. As per the feedback on our initial project proposal, we have also incorporated some county-wise data for each state.

Overall, we tried to answer the following:

- Analyze which states have high number of accidents reported.
- Group accidents as per severity of the accident i.e. fatal, incapacitated, non-incapacitated and minor.
- Visualize different plots depicting state-wise statistics related to accidents like the number of accidents, population of state, impact of weather on accident, etc.
- Visualize plots for recent years to see the trend increase/decrease in number of accidents over several years.
- Find some correlation between the population density of a particular state with the number of accidents in that state.

DATASET

Dataset	Timeline	No. of Entries	Major Columns/Features	Data Source
US Accidents	2011-2019	300000 x 49	Severity, Time, State, Temperature, Visibility, etc.	https://www.kaggle.com/sobhanmoosavi/us- accidents#
US Population Breakdown	2019	51 x 4	State, Population, Year, etc.	https://www.kaggle.com/fireballbyedimyrnmom/us-population-breakdown

For our project, we are using above datasets available on Kaggle. The US accidents dataset is around 1 GB in size containing approximately 3 million rows. Each row represents information like the year, exact spot of accident, weather condition, start/end time, state name, county name, etc. The other dataset consists of population count of every state in 2019. We have used some of the important attributes from these datasets for our dashboard.

METHODOLOGY

The following tools were used in the implementation of the project:

- **Python:** Python has been used for data cleaning and pre-processing.
- **Server:** Python Flask Framework has been used to run server on localhost.
- **Client:** For the client-side interface, we have used javascript, html, CSS and d3.js libraries for visualization.
- **External Libraries:** We also used some python libraries like sklearn, numpy, pandas, etc. for our data analysis and JavaScript libraries like Crossfilter.js and queue.js to filter data and design dashboard.

DATA CLEANING

As the original datasets were approximately 1 GB in size, we cleaned the dataset by extracting only features for our use. After cleaning the data, below 3 files were generated having following features:

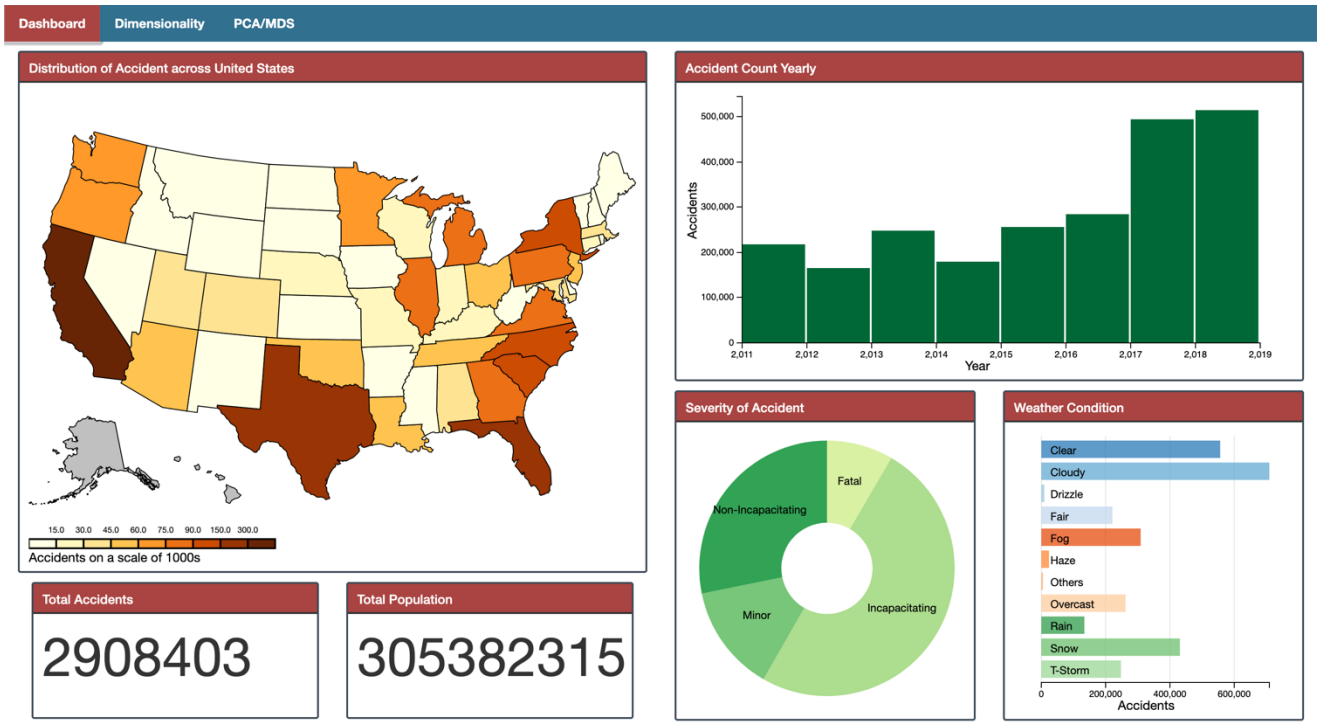
1. **Accident.csv –** This file contains complete dataset with each row representing one accident. Features included in this file are:
    - State Name
    - County Name
    - Year of Accident
    - Severity of Accident
    - Weather Condition
2. **State_accident.csv –** This file contains state wise grouped data where each row represents following features of each state:
    - State Name
    - Total Population
    - Total Accidents in State
    - Spot of accident – crossing, junction, bump, etc
3. **County_accident.csv -** This file contains county wise grouped data where each row represents following features of each county:
    - County Name
    - State Name
    - Total Accidents in State
    - Spot of accident – crossing, junction, bump, etc

ANALYSIS OF ACCIDENTS IN US

After thorough data cleaning, our next step was to create a meaningful visual representation of our dataset using different visualization techniques. We have used a combination of simple and advanced visualizations as below:

- **Dashboard:** A dashboard provides at-a-glance view of key performance indicators relevant to our analysis. It represents a central location for users to access, interact and analyze updated information to help them make smarter data-driven decisions. During our analysis, we mostly focused on visualizing the data on dashboard which would get user engagement within seconds and provide clear understanding of the visualization.

We designed our dashboard using the accidents dataset having 3 million records. The dashboard consists of 6 different charts which helps the user to visualize and analyze accidents across different states in United States.



a) Distribution of Accidents (Data Map): The US datamap show distribution of accidents across the different states. We have used colors scale to represent the number of accidents in a state, where lighter color means lesser accidents and vice versa. The user can select a particular state or a group of states to filter out information as required. Additionally, when we hover over any particular state additional information about that state like total population, total accidents and counties with most accidents is displayed.

b) Total Accidents: This area represents the number of accidents for the selected configuration. The value displayed gets updated as the user makes different selection.

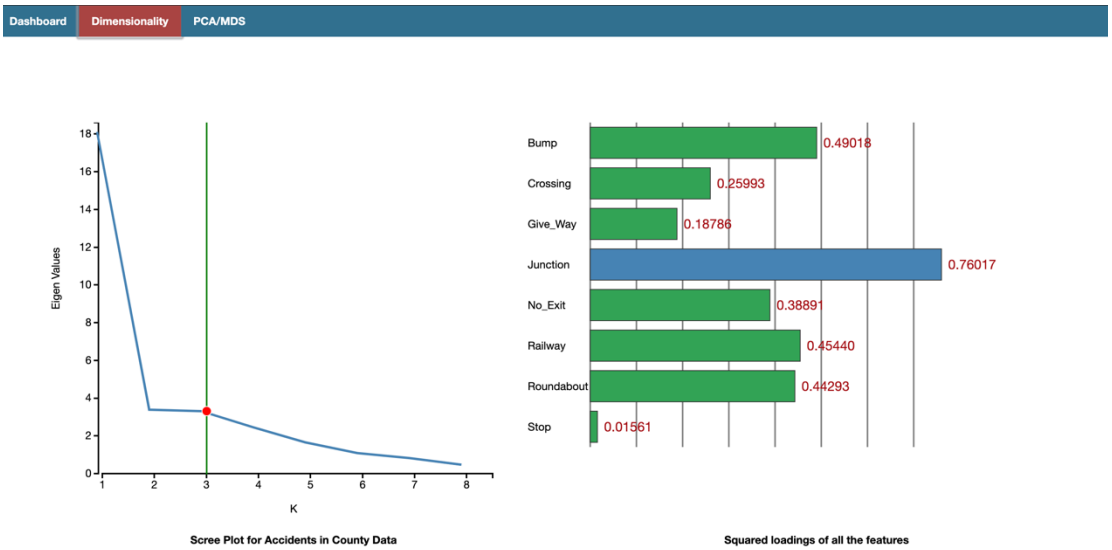
c) Total Population: This area displays the total population of the selected configuration on the dashboard. Similar to total accidents, this count is also updated as the user makes different selections.

d) **Accident Count Yearly (Histogram):** This chart shows the number of accidents reported between 2011-2019. Users can select a range of bars using a slider which will update the other charts accordingly on the dashboard.

e) **Severity of Accident (Pie Chart):** The pie chart shows the percentage of severity of accidents i.e. minor, fatal, non-incapacitating and incapacitating. The user can select a particular slice to display information about that accidents of the selected severity.

f) **Weather Condition (Bar Chart):** This chart displays the weather condition on the day of accident. The x-axis represents the count of accidents occurred during the given weather conditions. The user can select a particular bar to see in which states accident occurred during the selected weather condition.

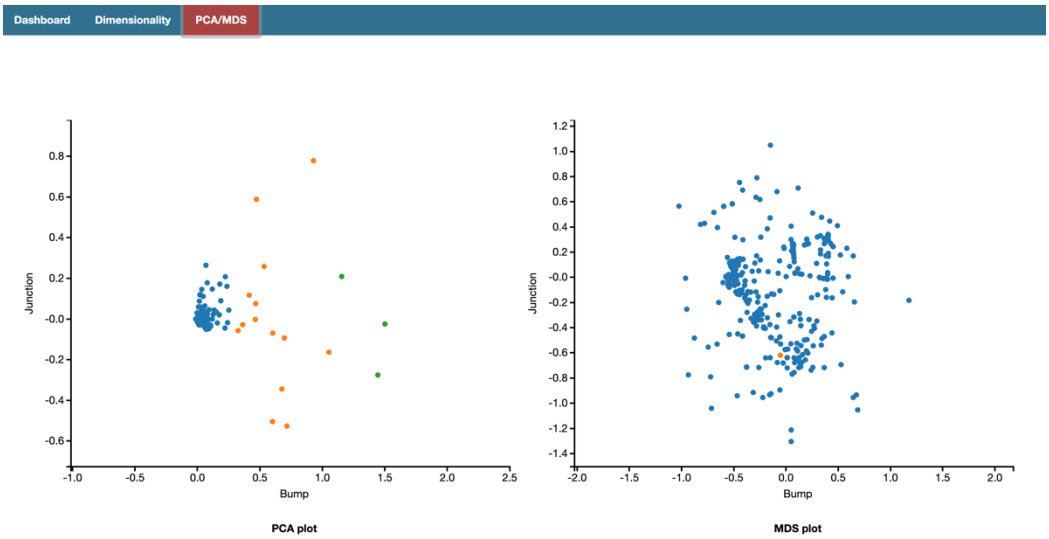
❖ **Dimensionality:**



a) **Scree Plot:** We calculated the intrinsic dimensionality of the dataset using a scree plot of the eigen values for each dimension. As can be seen from the above plot, 3 features are considered as the primary components of data as they reflect more than 98% of the total variation in the data.

b) **Squared Loadings:** We now find out which features have the maximum intrinsic dimensionality. The squared loadings of all the features are plotted and it is found that junction, bump and railway are the top 3 components.

❖ **Multidimension Visualizations:**



a) Principal Component Analysis (PCA): We have plotted the top 2 principal components with highest variation on the x-y plane. Using k-means clustering and elbow technique k value was found as 3. The data was clustered into 3 components and was found to be highly correlated.

b) MDS Correlation: MDS technique was used for dimension reduction using correlation distance on random samples. The MDS plot has its center at origin and shows the variation along top 2 components.

CONCLUSION

Using our analysis, we identified the states with the greatest number of accidents across the years. We were also able to list out the top 5 counties in a state based on their accident counts. Due to the interactive nature of our dashboard, we could select single or multiple states at a time on the data map using which we found out trends in the accidents reported for the selected states over a particular time frame. Using the weather data, we tried to find a correlation between the weather condition on the accident day and found that weather could infact be the reason for the accident in some cases. The below are some of the observations drawn from the dashboard:

- Populated states like California, Texas, New York, etc. have high number of accidents.
- There is a huge increase in accidents reported in 2018-2019.

- Almost half of all the accidents reported left the victim in an incapacitated state requiring medical attention.
- Most accidents took place on days when weather was snowy, clear or cloudy.

The dataset we started off with had around 3 million records, which made it very difficult to analyze. Visual analytics not only helped us to represent the large dataset in a more meaningful manner, moreover it also helped to extract useful information and find interesting correlations.

REFERENCES

- https://www.cdc.gov/injury/features/global-road-safety/index.html
- https://en.wikipedia.org/wiki/Motor_vehicle_fatality_rate_in_U.S._by_year
- https://www.asirt.org/safe-travel/road-safety-facts/
- https://github.com/d3/d3/wiki
- https://getbootstrap.com/docs/4.4/getting-started/introduction/