CMPT 732 – BIG DATA PROGRAMMING I

Project Tile: US Accidents - A Countrywide Traffic Accident Dataset Analysis

Web Based Report Link: https://usa-accidents.herokuapp.com/

Contributors: Akash Sindhu, Bilal Hussain, Sakina Patanwala

1. Introduction

US-Accidents can be used for numerous applications such as real-time car accident prediction, studying car accident hotspot locations, casualty analysis and extracting cause and effect rules to predict car accidents, and studying the impact of precipitation or other environmental stimuli on accident occurrence. The most recent release of the dataset can also be useful to study the effect of COVID-19 on traffic behavior and accidents. The dataset has been extracted from Kaggle and can be found **here**. It is a countrywide traffic accident dataset, which covers 49 states of the United States. The data is continuously being collected from February 2016 to June 2020, using several data providers, including two APIs which provide streaming traffic event data. These APIs broadcast traffic events captured by a variety of entities, such as the US and state departments of transportation, law enforcement agencies, traffic cameras, and traffic sensors within the road-networks. Currently, there are about 3.5 million accident records in this dataset. The data is provided in the form of a CSV file.

2. Methodology

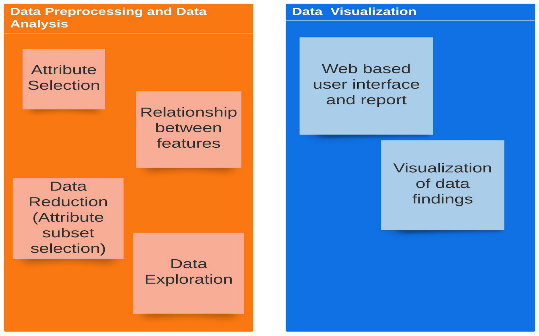

Data Preprocessing and Data Analysis
Attribute Selection
Relationship between features
Data Reduction (Attribute subset selection)
Data Exploration
Data Visualization
Web based user interface and report
Visualization of data findings

Fig 1: Workflow Chat

2.1.Data Processing and Data Analysis

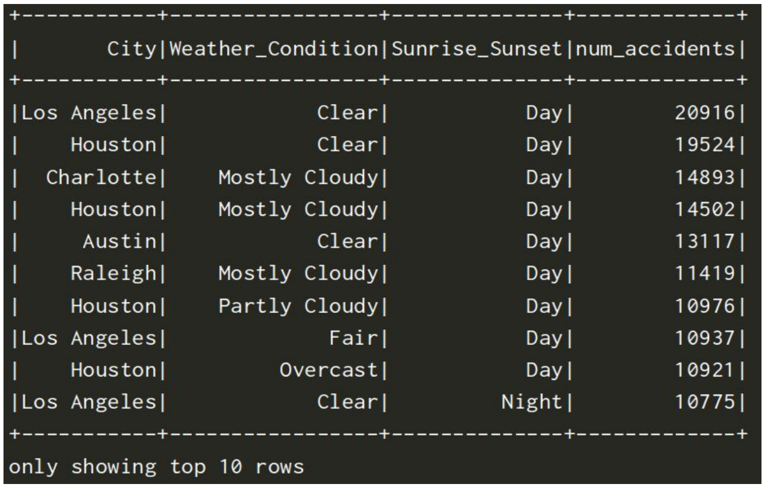
The dataset is 1.23 GBs in CSV format with 3.5 million rows. Since the data was already clean, we spent plenty of time understanding the 49 features and eliminating columns with discrepancies in values or missing values. We found relationships amongst the features based on the problem we are solving. Following are the accident trends analyzed and plotted using the dataset:

1. Most and least accident-prone states in the US: The dataset contains 3.5M rows hence 3.5M accidents occured in the given timeframe in the US. To get the total number of accidents, in each state, we grouped them by ‘state’ and aggregated them to get a count for the ‘number of accidents’ in each state. Results showed that California had the highest and North Dakota has the least number of accidents amongst all the states in the country.

2. Accident count per state using severities from low to high in the US: We grouped by ‘state’ and ‘severity’ of the accident to get the total count for accidents of very low, low, high, and very high types of severities for each state. California had the highest number of accidents for all types of severities, while the Northwestern US had a fewer number of accidents.
3. Accident count per year in the US: We used PySpark’s ‘functions.to_timestamp’ to convert the original string ‘Start_Time’ column to timestamp. Next, ‘functions.year’ was used to extract the year and aggregated to count the ‘number of accidents’ per year for the US. Since data for some months of 2016 and 2020 was missing, their total count was comparatively low. 2019 is the year with the highest number of accidents.
4. Accident count per month of the year in the US: Using a similar approach as (3), we used ‘function.month’ to get the total number of accidents happening in each month for the year 2016-2020. All the months had about the same number of accidents while July had the lowest and October had the highest.
5. Accident count per day of the week in the US: We used ‘functions.dayofweek’ to extract the ‘day’ on which each accident happened, aggregated to get the count for the number of accidents for each day. Most numbers of accidents occurred between Monday to Friday, which are working days in the US.
6. Accident count by the hour of the day in the US: Hour of the day was extracted using ‘functions.hour’ and aggregated to count the ‘number of accidents’ which happened during each ‘hour’. Most accidents occur between 7 am to 8 am and 4 pm to 5 pm, which is considered rush hours in the country.
7. Accidents caused by different weather conditions in different months of the year in the US: The column ‘Weather_Condition’ has 128 different values in the dataset along with null values. Hence, we extracted the top 10 weather conditions after filtering out the null values and aggregated them to get the total number of accidents for each weather condition. Most accidents occurred when the weather was clear and fair, while the least occurred during rain and light snow.
8. Aggregated monthly accident count in the top twenty cities in the US: We found the top twenty cities with the highest number of accidents. Then we extracted the month from the timestamp and grouped it by month and city to get the aggregated count of all the accidents occurring in each city for all twelve months. Houston has the highest number of accidents for all twelve months, while Los Angeles had the second-highest number of accidents.
9. Accidents caused by different severity levels due to low visibility in the US: The ‘visibility’ column has values from 1 to 10 and also null values. A limit of 5 was set on ‘visibility’ to be ‘low visibility’. Next, we grouped by ‘state’ to find all the low and high severity accidents separately which occurred during low visibility. California has the highest number of accidents for high and low severity due to low visibility.
10. Accidents caused by different weather conditions at various times of the day in Houston: Houston had the highest number of accidents amongst all the cities in the country so we explored it further. First, we

extracted the ‘hour’ from the timestamp and divided it into categories like ‘Early Morning’, ‘Late Morning’, ‘Early Afternoon’, ‘Late Afternoon’, ‘Early evening’, and ‘Night’. Second, we extracted the top 20 weather conditions when most accidents occurred. Second, we grouped by ‘City’, ‘Weather_Condition’, ‘hour’ and aggregated to get the sum of the number of accidents that occurred for the same weather condition at different times of the day. Most accidents occurred during early and late mornings when the weather was clear while the least number of accidents occurred during the night. Surprisingly, fewer accidents occurred during heavy rainfall.

11. Accidents caused by different severity levels at various times of the day in Houston: For this query, we aggregated by ‘City’, ‘Severity’, ‘hour’ to get the number of accidents for different types of severity occurring at various times of the day. Low severity accidents were common during all times of the day, while there were only a few very high severity accidents that occurred during the night or early evening.
12. Twenty most accident-prone streets in Houston: For this query, we grouped by ‘Street’ to get the twenty accident-prone streets in Houston. Among these top 20 streets, I-45 N was the most dangerous with the highest number of accidents while N Sam Houston Tollway W had the least number of accidents.
13. Accident count due to nearby road features in Houston: There were common road features in the dataset with boolean values, showing the presence of a certain feature at the location of the accident. We aggregated all those features to get the common feature that was present at the location of the accident. Most accidents occurred near the traffic signal and crossings, while a few occurred near no exits and railway stations.
14. Which city has the most number of accidents in the given weather condition(rain, snow, thunderstorm, fog, etc.) and period of day (i.e. day or night): We found out that Houston and Los Angeles are the most dangerous cities during the day. All top 9 accidents occurred during day time.

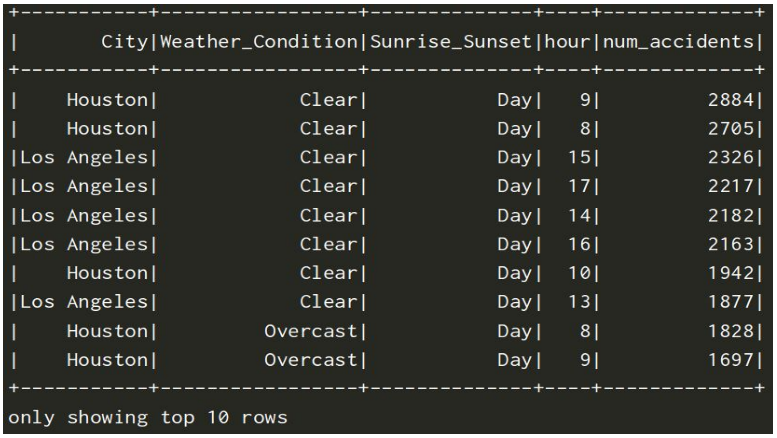


```
+-----------+-----------------+--------------+-------------+
|       City|Weather_Condition|Sunrise_Sunset|num_accidents|
+-----------+-----------------+--------------+-------------+
|Los Angeles|            Clear|           Day|        20916|
|    Houston|            Clear|           Day|        19524|
|  Charlotte|    Mostly Cloudy|           Day|        14893|
|    Houston|    Mostly Cloudy|           Day|        14502|
|     Austin|            Clear|           Day|        13117|
|    Raleigh|    Mostly Cloudy|           Day|        11419|
|    Houston|    Partly Cloudy|           Day|        10976|
|Los Angeles|             Fair|           Day|        10937|
|    Houston|         Overcast|           Day|        10921|
|Los Angeles|            Clear|         Night|        10775|
+-----------+-----------------+--------------+-------------+
only showing top 10 rows
```

Fig 2: Groupby city, sunrise_sunset and number of accidents and sorted in desc order by num_accidents.

15. Further which cities are most dangerous during the day and what time: We did data preprocessing to extract hours in 24-hour format and found that morning time between 8 am to 9 am and evening

time between 4 pm to 6 pm is the busiest.



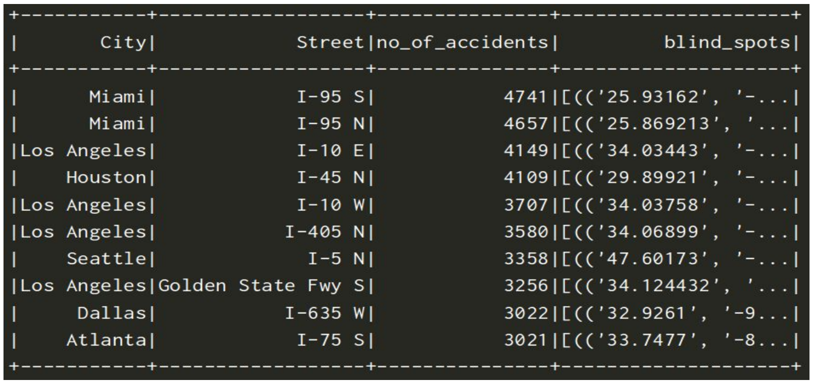
```
+-----------+-----------------+--------------+----+-------------+
|       City|Weather_Condition|Sunrise_Sunset|hour|num_accidents|
+-----------+-----------------+--------------+----+-------------+
|    Houston|            Clear|           Day|   9|         2884|
|    Houston|            Clear|           Day|   8|         2705|
|Los Angeles|            Clear|           Day|  15|         2326|
|Los Angeles|            Clear|           Day|  17|         2217|
|Los Angeles|            Clear|           Day|  14|         2182|
|Los Angeles|            Clear|           Day|  16|         2163|
|    Houston|            Clear|           Day|  10|         1942|
|Los Angeles|            Clear|           Day|  13|         1877|
|    Houston|         Overcast|           Day|   8|         1828|
|    Houston|         Overcast|           Day|   9|         1697|
+-----------+-----------------+--------------+----+-------------+
only showing top 10 rows
```

Fig 3: Groupby city, weather condition, Sunrise sunset, hour of day and sort on number of accidents.

16. Accident-prone areas in hundred most busy streets in the US:

We found the top hundred cities and their blind spots/accident prone areas using latitudes and longitudes of the accidents which occurred within 10 meters of each other. We concluded that if within 10 meters, three accidents occur then that area is accident-prone and should be avoided for the school zones and should have crosswalks,etc so people can avoid jay-walking.

Sample blind spot/accident prone area:
[(('33.744976', '-84.390343'), ('33.754379', '-84.3787'), ('33.766743', '-84.388107'))]
Here, ('33.744976', '-84.390343'), ('33.754379', '-84.3787'), ('33.766743', '-84.388107') are the three latitudes which are within 10 units to each other. We only looked for 3 accidents as finding distance is a costly function. Euclidean distance is used to calculate the distance between two points. We used dash to plot the latitudes and longitudes on the graph. The streets of I-95 S and I-95 N in Miami city have the most number of accidents followed by streets in Los Angeles and Dallas. This accident lookup threshold is the number of accidents we looked to consider an accident-prone area and can be scalable to find exact locations of blind spots/accident prone areas.



```
+-----------+------------------+---------------+--------------------+
|       City|            Street|no_of_accidents|         blind_spots|
+-----------+------------------+---------------+--------------------+
|      Miami|            I-95 S|           4741|[(('25.93162', '-...|
|      Miami|            I-95 N|           4657|[(('25.869213', '...|
|Los Angeles|            I-10 E|           4149|[(('34.03443', '-...|
|    Houston|            I-45 N|           4109|[(('29.89921', '-...|
|Los Angeles|            I-10 W|           3707|[(('34.03758', '-...|
|Los Angeles|           I-405 N|           3580|[(('34.06899', '-...|
|    Seattle|             I-5 N|           3358|[(('47.60173', '-...|
|Los Angeles|Golden State Fwy S|           3256|[(('34.124432', '...|
|     Dallas|           I-635 W|           3022|[(('32.9261', '-9...|
|    Atlanta|            I-75 S|           3021|[(('33.7477', '-8...|
+-----------+------------------+---------------+--------------------+
```

Fig 4: Group by city, Street, sort on number of accidents.

2.2 User Interface and Visualization

We developed a web based report in this project as this skill is somewhat necessary for data scientists to know how to make engaging UIs. We used PySpark to extract interesting trends and build queries and saved the achieved dataframes into CSV files. Later, by using Dash plotly, plots and graphs were generated from these CSV files. Dash is also used to generate web based reports. The link is then published and made public using heroku. The web based report can accessed from this **https://usa-accidents.herokuapp.com/**

3. Problems and Challenges

Some of the problems we faced during the project implementation are:

3.1 Code quality over code quantity

We spent time in furnishing the code quality and the PySpark efficiency. Since the data has 49 columns, we tried not loading all the columns into the dataframe when not needed. We understood each query and loaded the minimum amount of required data. This is one of the reasons we decided to not go for Cassandra because we had many different types of queries that required different columns, it was not worth using Cassandra as a database for this project.

3.2 Understanding the queries thoroughly

When we were solving the problem of accident prone areas/blind spots we tried understanding the differences and significance of various distance functions like euclidean distance, spatial distance, haversine distance and how to use it in our project. We spent some amount of time on solving spatial distance.

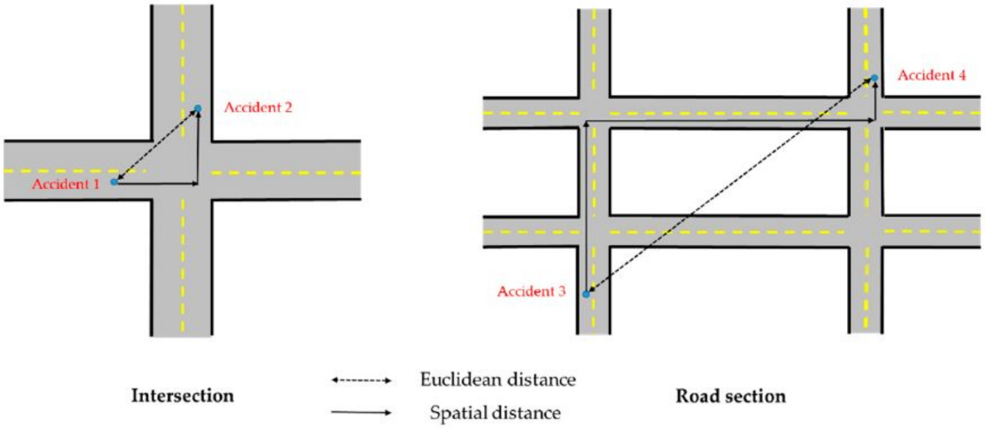


Fig 5: Comparison of different distances.

1. Learning new technologies: Understanding Dash by plotly was time consuming, we utilized its capabilities to generate interactive graphs and plots in the web browser.

4. Results

From analysis we tried to understand the relationship between different columns which helped us to come up with some problems to solve efficiently. We can now comprehend what factors are responsible for most of the accidents in the US and this can be useful in future when we are working on computer vision based problems like self-driving cars. Besides this, with a detail oriented mindset we understood how to make conclusions using web based visualization with PySpark and Dash. By combining new technologies that we researched for this project along with PySpark that we have been using through the coursework we learned how to amalgamate different technologies with each other. Also, we learned how to effectively utilize the capabilities of parallel computing.

4.1 Project Summary

Category	Points
Getting the data: Exploring and downloading a dataset from Kaggle.	1
ETL: Understanding the data and making required queries to solve, PySpark job for cleaning the dataset and performing Extract-Transform-Load.	2
Problem: Working on defining the various accident trends to be executed and calculated and conducting analysis for these trends.	1
Bigness/parallelization: 1.24 GBs of data with 3.5 M rows and 49 columns is the bigness of the data. Once, Dash libraries and dataset is on the cluster the project is efficient on the cluster and is scalable if the dataset increases in size.	3
Logic/Algorithms: Developed an algorithm to extract distance between latitude and longitude and logic to use it in our project to find nearest points efficiently.	2
UI: Dash Plotly retrieves .csv files and plots it to UI and hosts a website using Heroku.	4
Visualization: Horizontal and vertical bar graphs, scatter plot graphs, heat maps, mapbox plots, choropleth maps plotting of accident trends and external factors causing them.	4
New Technologies: Dash, Heroku, Pandas	3
TOTAL	20

5. References

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.