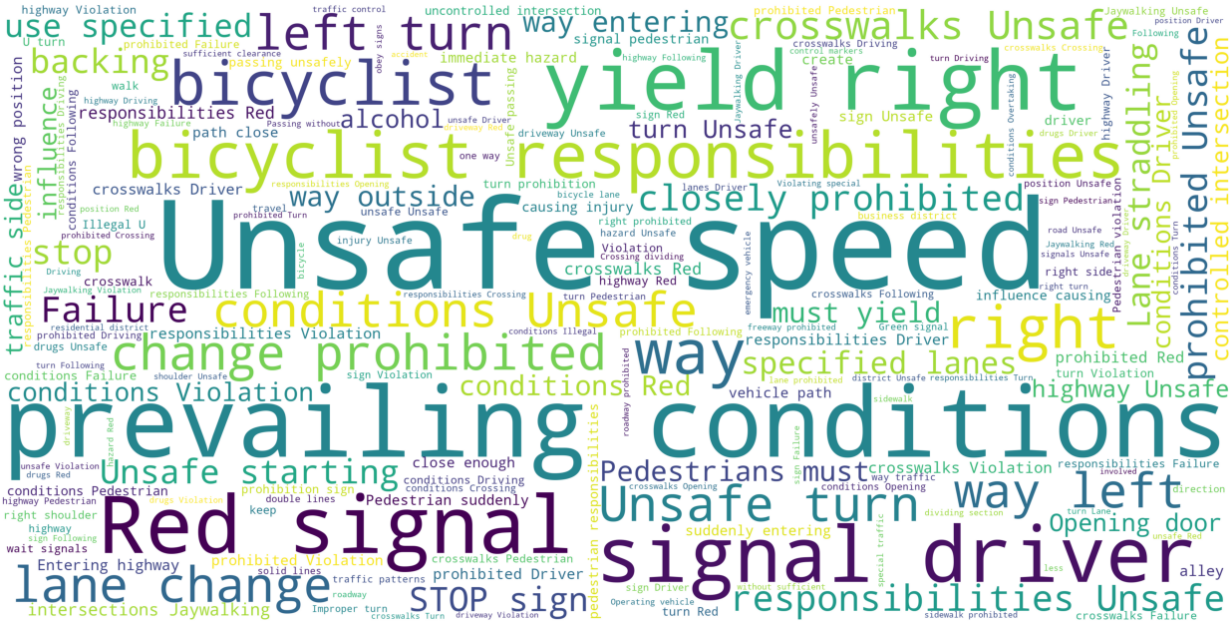
BUDT704 Project Proposal

10.13.2023
Section 0503
Team 5

Project Title: Accident Metadata for Enhanced Analysis

Team Accident Analysis Avengers (AAA): Ansh Dugar, Balaji Udayakumar, Ei Aung, Harshit Kavathia, Joshua Mohammed, Lirong Huang, Sanskriti Yadav

Introduction



In this project, we are set to analyze 18 years of data (2005-2023) containing data from all crashes resulting in injury in San Francisco, California. The issue of collisions is a major threat to the lives of humans and can create lifelong consequences. After conducting our analysis, we hope to provide valuable information related to public safety, create predictive models to reduce the risk of collisions and help the U.S. government make more informed decisions. With this analysis and information, we can increase the awareness of the public on how to avoid collisions and be more cautious. We will do this through analyzing specific traffic violations that have been breached that impacted the risk of collision. After a successful analysis, this can be implemented for other regions throughout the United States.

Questions of Interest

In this project, we will look to answer the following three main questions:

- Are there any geographic hotspots or any other patterns in collisions?
- How do vehicle brands, their safety equipment, and their year of manufacturing affect collisions?
- How can pedestrians and vehicle drivers be more cautious to avoid collisions?

Through analyzing these three questions, we aim to build a model that will return an index (out of 10) suggesting how likely it is for a collision to occur based on different factors.

Data Processing and Analysis

Dataset Description:

Inside the dataset, "Traffic Crashes Resulting In Injury", there is data that has been obtained from the Office of the Chief Medical Examiner (OME) death records, San Francisco Police Department (SFPD) Interim Collision System for 2018 through 2023, and the Crossroads Software Traffic Collision Database. This dataset is located in the data catalog for the U.S. Government (catalog.data.gov). In the "Traffic Crashes Resulting in Injury" dataset there are 56,010 observations, 60 variables and 3,360,540 data points that contain information related to all crashes resulting in an injury in the City of San Francisco. This dataset is in the form of a CSV file.

Another dataset is "Traffic Crashes Resulting In Injury: Parties Involved" that is similar to the first dataset but it has victims involvement and has more variables. The location of this dataset is the same as the corresponding dataset: catalog.data.gov. There are 80 variables, 118, 087 observations, and 9,446,880 data points. This data is collected from the California Highway Patrol 555 Crash Report as submitted by the police officer within 30 days after the crash. This dataset is in the form of a CSV file.

In both datasets, there are variables such as a unique ID for each collision, the longitude and latitude of the collision, collision severity, type of collision, the vehicle type, and the number of victims in the party, and more. Here are both link of all the column descriptions that are included in both datasets:

1. https://data.sfgov.org/Public-Safety/Traffic-Crashes-Resulting-in-Injury/ubvf-ztfx
2. https://data.sfgov.org/Public-Safety/Traffic-Crashes-Resulting-in-Injury-Parties-Involv/8gtc-pjc6

Data Processing Tasks:

- Indexing, selection, and filtering: We will be using indexing in order to optimize the data access, selection to gather specific columns and rows of data, and filtering to extract the data.
- Remove Duplicate Data in order to ensure that our data is accurate and consistent.
- Manage Missing Data: First we will check how much missing data is involved in the dataset by using .isnull(). If it is a small proportion, we'll use a filtering method such as .dropna(). If it is a big proportion, an imputing method will be suitable to manage missing data.
- Data transformation: Before we analyze data with visualizations, we will transform the dataset including random sampling, being discretized to get ranges and binning to identify the size/group of the value to put in the range. We are going to use 'dummy variables' when we want to convert some category variables into numerical variables which we want to include in our analysis. Furthermore, we will be renaming existing variables, creating new variables, and deleting unnecessary variables using Python libraries. For example, in the **vehicle _make** variable in our *parties involved* dataset we are only interested in finding the brand of the vehicle. But in our data the values are all in different cases and also they are not very consistent. So we will have to clean the particular column, impute the column with reasonable values, remove unnecessary parts of the values like its model, transform each value to get only the brand name of the vehicle out of that column and we will add it in a different column.
- Text processing: We will use vectorized string methods when data needs to be grouped and extracted, finding matches, splitting, and replacing. Also, we will create dummy variables from the text data in which we can make inferences or comments of the output results. We will process the texts that are given in the column description in order to utilize that data for the purpose of modeling.
- Data Visualization: We will be utilizing data visualization for the purpose of telling a story about the data. When determining geographic hotspots where collisions occur, we aim to create a map of San Francisco to present the frequency of collisions in every district.

We will be using these processing tasks to clean the dataset, make the dataset appear more organized, and make it easier to analyze.

Data Analysis:

- Are there any geographic hotspots or any other patterns in collisions?

  We will display the geographic hotspots for collisions in San Francisco through the map function. The map will show the frequency of collisions in each unique district from the collision_distrct column. We will also apply color to indicate the magnitude of the frequency, which can be the larger the frequency, the darker the color. The hotspot map can visually display any geographic hotspots for collisions.

  We will use a bar chart to display the number of each collision severity that happened under each kind of weather condition, with the comparison of the number of different collision severity under different weather conditions, we can determine if there is a specific pattern in collision severity based on weather conditions.

- How do vehicle brands, their safety equipment, and their year of manufacturing affect collisions?

  We will create a pie chart to display the different portions of the vehicle brand in the collisions dataset, which can help indicate if the vehicle brand would affect collisions through the area distribution of each brand in the pie chart. For the year of manufacturing, we will also create a pie chart to display the frequency or percentage of each year of manufacturing in the collision dataset, which can help indicate if the year of manufacturing affects the collision by looking at the area distribution of each year of manufacturing of the crash cars.

  We will use a bar chart to display the efficiency of safety equipment in collisions in San Francisco. The x-axis will display different safety equipment from the party_safety_equip column and the y-axis will display the frequency or percentage of different collision severity when certain safety equipment is used, which can help determine the efficiency of safety equipment in collisions.

  We will create a pie chart to display the different portions of the vehicle brand in the collisions dataset, which can help indicate if the vehicle brand would affect collisions through the area distribution of each brand in the pie chart. For the year of manufacturing, we will also create a pie chart to display the frequency or percentage of each year of manufacturing in the collision dataset, which can help indicate if the year of manufacturing affects the collision by looking at the area distribution of each year of manufacturing of the crash cars.

- How can pedestrians and vehicle drivers be more cautious to avoid collisions?

We will use a bar chart to present the different types of collisions and their frequencies. The x-axis represents the types of collisions, whereas, the y-axis represents the number of collisions. The frequency or the percentage of each type of collision can help indicate what type of road conditions the pedestrians and vehicle drivers should be more cautious about.

We will also use a scatter plot to present if the collision has anything to do with the combination of weather and lighting at a certain time, which can be extracted from the weather column and the collision datetime column. Each collision from the dataset will be presented as a dot in the graph. The density of the dots will tell in which combination of the weather and the daytime the pedestrians and the vehicle drivers should be more cautious about.

We will also create a line chart that will be utilized to discover trends in collision frequency over time.

Expected Findings

After conducting our analysis, we expect to find various answers to our questions of interest. One issue that we expect to find is what are the major mistakes that people make before a collision like not using safety equipment, being drunk on the road, not stopping at the red signals, and crossing the road in the wrong direction. Another issue that we expect to find out is that there are some specific geographic locations in San Francisco where collisions are most likely to occur so that people are more cautious of those locations and the government can implement better safety measures. A third conclusion that we are hoping to make is that collisions are more prevalent in one particular vehicle brand and its year of making.

Project Timeline

To work on all three analysis questions, we will divide our team into 3 teams. Each team will work on each question for around 2 weeks. We think that way we will be able to do all the coding quicker and it will give us more time to analyze and brainstorm our findings. And once we have worked on all three questions we will all gather to work on the data modeling part because it is the new topic for all of us and that is quite important for all of us to learn.
Team 1 (Question 1): Lirong, Balaji
Team 2 (Question 2): Harshit, Sanskriti

Team 3 (Question 3): Ei, Josh, Ansh

Task	Task Lead	Due Date
Managing Missing Data and Inconsistencies	Lirong	10/20/2023
Renaming and Deleting Certain Columns	Sanskriti	10/20/2023
Eliminate Duplicates and Outliers	Ansh	10/20/2023
Transforming certain columns to fit our needs	Josh, Harshit	10/20/2023
Extracting New Features from Dataset	Balaji, Ei	10/20/2023
Discover patterns and factors contributing to collisions	Team 1	11/10/2023
Mapping hotspots for collisions	Team 1	11/10/2023
Determine collision severity under each type of weather condition	Team 1	11/10/2023
Create a hotspot map and bar chart for Question 1	Team 1	11/10/2023
Discover relationship between vehicle brand and collisions	Team 2	11/10/2023
Discover relationship between vehicle equipment and collisions	Team 2	11/10/2023
Discover relationship between the year of manufacturing of vehicle and year of collisions	Team 2	11/10/2023
Create a pie chart and bar chart for Question 2	Team 2	11/10/2023
Determine the impact of weather and lighting involved in collisions	Team 3	11/10/2023
Create bar charts, scatter plot, line chart for Question 3	Team 3	11/10/2023
Trend Analysis for collisions over time using the line chart	Team 3	11/10/2023
Preparation for Data Modeling: Splitting data and handle categorical features	Balaji, Sanskriti	11/17/2023
Creating features matrix and labels	Ei	11/17/2023
Choosing and comparing the ML Models	Lirong, Ansh	11/24/2023
Implement the right ML Model	Josh	11/24/2023
Hyper parameter tuning	Harshit	11/24/2023