Proposal: Exploring Road Safety Factors and Accident Patterns in New Jersey

DATA 450 Capstone

Sai Harshitha Dalli

February 8, 2024

1 Introduction

Every day, road accidents are alarmingly common. Often, we overlook the frequency of crashes on the routes we take regularly and the tragic toll they take on human lives. I believe it's crucial to address this issue as it concerns both public safety and public health. By looking into the data, I want to analyze specific trends and contribute to preventing deaths resulting from vehicle crashes. This research has the potential to inform future policies, to reduce mortality rates on roads.

2 Dataset

The dataset utilized for this analysis is sourced from The Fatality Analysis Reporting System (FARS), a comprehensive nationwide database maintained by the National Highway Traffic Safety Administration (NHTSA) in the United States. FARS provides detailed information on fatal injuries sustained in motor vehicle traffic crashes, offering valuable insights for research and policymaking.

The dataset can be downloaded from here.

The dataset, spanning from 1975 to 2021, encompasses various files related to accidents, persons involved, vehicles, weather conditions, distractions, damages, and more. While data is available for multiple years, I have chosen to focus specifically on the year 2021 due to its relevance and recency. Leveraging pandas, I will merge key datasets using the unique identifier "ST_Case," which is consistent across all relevant files.

The selected datasets include accidents.csv, drimpair.csv, nmcrash.csv, person.csv, and weather.csv, each containing essential variables pertinent to addressing research questions

Some variables I will be using are:

- ST_Case: The unique identifier of each case.
- STATE: The states unique number(1-50)
- PERSONS: This is how many people were in the vehicle at the time of accident.
- COUNTYNAME: County where the accident occurred.
- ROUTE: Type of roadway (e.g., Interstate, State Highway, County Road).
- CITYNAME: City or locality where the accident occurred.
- MONTH: The month the crash took place(1-12).
- DAY: The day the crash took place(1-31).
- DAY_WEEK: The day of the week the crash took place(1-7).
- HOUR: The hour of the day that the crash happened(1-24).
- TWAY_ID: The highway ID where the crash took place.
- FATALS: The number of people who died in the crash.
- PEDS: Number of pedestrians involved.
- PERNOTMVIT: Number of non-motorist fatalities.
- DRIMPAIRNAME: The driver's impairment during the time of the crash(under the influence, depression, etc).
- MOD_YEAR: The year that the model of the vehicle was made.
- VPICMAKENAME: The vehicles make name.
- AGE: The age of the person in the vehicle.
- SEXNAME: The sex of the person in the vehicle.
- INJ_SEVNAME: The severity of the individual's injury.
- RUR_URBNAME: Rural or Urban classification of the accident location.
- FUNC_SYS: Functional system classification of the roadway.
- LGT_COND: Light condition at the time of the accident.
- TYP_INT: Type of intersection or non-intersection.
- WEATHERNAME: Weather conditions at the time of the accident.

3 Data Acquisition and Processing

I will download the necessary CSV files from the FARS website. These files contain detailed information on accidents, persons involved, vehicles, weather conditions, and other relevant variables for the year 2021.

After downloading the CSV files, I will then merge the relevant datasets using a common key variable, "ST_Case," to create a unified dataset for analysis. I will then encode the categorical variables by either using

Data Processing and Cleaning: Beacause there is a lot of data and everything might run slow if I use all the observations, I will filter to only have the accidents from NJ. This is the most relevent to us since I live in NJ and this will be useful since I know the driving laws for NJ.

I will drop the all the columns that are not relevant to my research questions. The variables listed in the previous section will stay. I will then check for missing values and impute them or remove them if necessary. We have abundance of data, so I will remove missing values if not much of the data is being lost. I would encode the categorical variables either using dummy variables or if binary encoding them to 0 and 1. I will also standardize some categorical variables by converting the text to all lowercase. I will check for any duplicate rows and remove them.

4 Research Questions and Methodology

1. Are there specific road conditions or locations where pedestrian accidents are more prevalent in New Jersey?: To answer this question, I will use the columns, PEDS, PERNOTMVIT, ROUTE, RUR_URB, FUNC_SYS, LGT_COND, WEATHER, COUNTYNAME, CITYNAME, TYP_INT, LATITUDE, and LONGITUDE. I will first filter the data only to include pedestrian accidents. Then I will group the filtered dataset by the variables related to road conditions and locations that you have identified. These might include ROUTE, RUR_URB, FUNC_SYS, LGT_COND, WEATHER, COUNTYNAME, CITYNAME, and TYP_INT. I will count the number of pedestrian accidents in each category of the grouped variables using group along with the count function in pandas. I will create a choropleth map by first obtaining geospatial data of the boundaries of the counties for New Jersey online. I will merge the pedestrian accident data with the geospatial data based on common geographic variables like COUNTYNAME.

2. Is there a correlation between reported driver impairment and the severity of accidents?: To answer this research question, I will use the variables FATALS and DRIMPAIRNAME. I have to make a function to define severity levels, an accident's severity is considered low if there are 0 fatalities, moderate if there is 1 fatality, and high if it is anything more than 1. Then I will calculate summary statistics by grouping the accident data by reported driver impairment and severity of accidents, computing the frequency of each severity level for each reported driver impairment category. Then, I perform a chi-square test of independence to determine whether there is an association between reported driver impairment and accident severity. Then, I will create a contingency table to organize the frequencies of reported driver impairment and severity levels. Finally, I compute the correlation coefficient between reported driver impairment and the number of fatalities to understand the relationship between these two variables. To visualize the correlation between reported driver impairment and fatalities, I will make a scatter plot where each point on the plot represents an accident, with the reported driver impairment on the x-axis and the number of fatalities on the y-axis.

3. Do certain vehicle types have higher rates of involvement in fatal accidents? For this question, I will use the VEH_MAKE and FATALS to see which make of the vehicle has higher or lower fatalities and see if that is somehow helpful when making decisions to

purchase new cars. I will first filter data for fatal accidents which is any observation with fatals greater than 0. I will then use groupby to group vehicle make and fatals and then calculate involvement rates. To visualize, I will create a stacked bar graph where x will be the vehicle make and y will be the number of fatal accidents. Each bar will represent the total number of fatal accidents, and the length of the bar is divided into segments representing different vehicle types. Each segment within the bar corresponds to a specific vehicle type, and the height of the segment corresponds to the number of fatal accidents involving that vehicle type.

4. Are accidents more common in urban or rural areas, and how does road type (e.g., interstate, state highway) affect accident rates? To answer this question, I will group the variables RUR_URBNAME and FUNC_SYSNAME. I will then make a grouped bar plot which will allow me to compare the accident rates across different road types for both urban and rural areas side by side. Each road type will have a pair of bars representing urban and rural accident rates, making it easy to compare. Urban/Rural areas will be on the x-axis and the number of accidents will be on the y-axis.

5. What is the relationship between different weather conditions and the severity of fatal accidents? To answer this question, I wanted to create an interactive dashboard using Dash to visualize the connection between weather conditions and fatal accidents. I would import the required libraries like Dash, Plotly Express, and pandas. I will use the variables WEATHER, LONGITUDE, LATITUDE, and FATALS.

5 Work plan

Week 4 (2/12 - 2/18):

- Data Processing and Cleaning(5 hours)
- Question 3 (2 hours)

Week 5 (2/19 - 2/25):

- Question 1 (4 hours)
- Question 4 (3 hours)

Week 6 (2/26 - 3/3): * Question 2 (2 hours) * Question 5 (5 hours)

Week 7 (3/4 - 3/10): * Fixing any errors and working on any questions if they don't work out (3 hours) * Presentation prep and practice (4 hours)

Week 8 (3/11 - 3/17): *Presentations given on Wed-Thu 3/13-3/14. Poster Draft due Friday 3/15 (optional extension till 3/17).*

- Work on poster draft (2 hours)
- Poster prep (4 hours)

- Presentation peer review (1.5 hours)

Week 9 (3/25 - 3/31): *Final Poster due Sunday 3/31.*

- Peer feedback (3.5 hours)
- Poster revisions (3.5 hours)
- [Do not schedule any other tasks for this week.]

Week 10 (4/1 - 4/7):

- Prepare for the DMC fair presentation (4 hours)
- Start working on the blog post (3 hours)

Week 11 (4/8 - 4/14):

- Research how to write the blog post (3.5 hours)
- Put all the research questions into the blog post (3.5 hours)

Week 12 (4/15 - 4/21):

- Revise the blog post (4 hours)
- Fix any errors that occur during the time (3 hours)

Week 13 (4/22 - 4/28): *Blog post draft 1 due Sunday night 4/28.* [All project work should be done by the end of this week. The remaining time will be used for writing up and presenting your results.]

- Draft blog post (4 hours).

Week 14 (4/29 - 5/5):

- Peer feedback (3 hours)
- Blog post revisions (4 hours)
- [Do not schedule any other tasks for this week.]

Week 15 (5/6 - 5/12): *Final blog post due Weds 5/8. Blog post read-throughs during final exam slot, Thursday May 9th, 8:00-11:20am.*

- Blog post revisions (2 hours)
- Peer feedback (2 hours)
- [Do not schedule any other tasks for this week.]

6 References

"NHTSA." NHTSA, 14 Nov. 2016, www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars. Accessed 8 Feb. 2024.