DS 150 Milestone 3

Abstract

This paper will investigate and analyze fatal car accidents in the United States during 2021 by utilizing the National Highway Traffic Safety Administration's (NHTSA) Fatality Analysis Reporting System (FARS). The large dataset, containing 39,508 fatal accidents and 80 columns of various data recorded for each accident, was cleaned and organized in order for a comprehensive analysis and exploration to be achieved. Initial exploratory analysis also involved statistical summaries and various graphical representations of the data. Preliminary analysis revealed that a majority of fatal accidents in 2021 resulted in the death of a single individual, while there were a few outliers that skewed the data. Further analysis revealed relationships between fatalities and factors such as date, light conditions, weather, and county. A scatter plot unveiled an exponential correlation between the date of accidents and fatalities. Both light conditions and weather patterns at the time of accidents in 2021 revealed distinct patterns, with a majority of fatalities occurring in broad daylight and under clear skies. Additionally, the examination of fatalities counties in the United States revealed that highly populous areas like Los Angeles and Maricopa had higher fatality rates, potentially influenced by population size.

The report emphasizes that while these insights provide valuable information regarding fatal accidents in 2021, there are certain limitations that exist within the exploratory analysis. The absence of an adjustment for population size when analyzing fatalities by county could lead to skewed conclusions and inaccurate representations of the data, especially for densely populated areas.

This report aims to contribute to the understanding of dynamics of fatal car accidents in 2021, potentially laying the groundwork for more comprehensive analysis as well as potential modeling to assist in enhancing traffic safety measures within the United States.

Background

As a brief background, the National Highway Traffic Safety Administration reported a total number of 39,508 fatal car accidents that had occurred within the United States during 2021, which resulted in the deaths of 42,939 people. The NHTSA is an agency within the US that is responsible for collecting and analyzing traffic safety data in order to promote and improve traffic safety laws. NHTSA is authorized by federal law to collect data on motor vehicle accidents in order to aid in the identification of issues, as well as aid in the development, implementation, and evaluation of motor vehicle and highway safety measures to reduce fatalities and property damage that are associated with these accidents (Economics Forum, 2023). The data that will be looked at for this project comes directly off of the NHTSA's website. NHTSA’s Fatality Analysis Reporting System (FARS). More specifically, the 2021 National FARS report will be used for this project, which was pulled directly from the NHTSA database. The dataset is 39,508 rows long by 80 columns wide, and contains various data collected by the NHTSA in 2021 in regards to fatal car accidents. For reference, there was no dictionary provided for the variables within the dataset, so there had to be some user interpretation for their meaning.

Methodology

For this section of the report, the methodology of analysis will be outlined in detail. This section will also explain what libraries were used to perform the analysis. Within this report, there are several key steps that were involved in the methodology: data collection, data cleanup, preliminary analysis, exploratory analysis, modeling of data and analysis, conclusion, and potential limitations. The dataset used for this project was sourced from the NHTSA's FARS database, and it specifically focuses on fatal car accidents that occurred during 2021. The dataset was first downloaded as a '.csv' file, then uploaded into a public-facing directory on github, and then was uploaded from github into the project's jupyter notebook with the use of the 'pandas' command 'pd.read_csv()'. The next step in the methodology was the data cleanup, which involved dropping unnecessary columns as well as the addition of new columns. In this phase, nearly 50 columns from the dataset were dropped as they were considered unnecessary for the project. The dropping of unnecessary columns aided in streamlining the analysis of the data. A new column, 'DATE' was created with the assistance of the 'datetime' library, and it was a combination of three columns. The preliminary analysis phase of this project had a goal of understanding basics about the data as well as uncovering different relationships between fatalities and various factors.

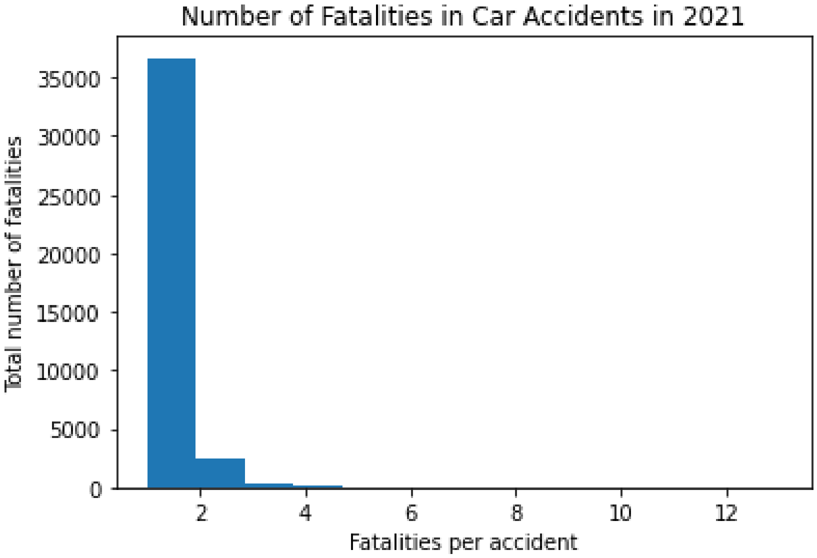
Exploratory Analysis

Data Cleanup

Prior to beginning any of the exploratory analysis for this project, the dataset needed to be cleaned up. To start the process of data cleanup, close to 50 of the columns were dropped from the dataframe because they were not going to be used for the project. After dropping the unnecessary columns from the data frame, it became much easier to parse through to get the data that was needed. A new column, titled 'DATE' was added to the data frame and is a combination of three columns within the data frame, 'YEAR', 'MONTH', and 'DAY'. The 'DATE' column records the full date for each fatal accident within the dataset. Data cleanup was crucial for this project so that the data frame contained the necessary information and wasn't too overwhelming for an individual to glance over.

Preliminary Analysis

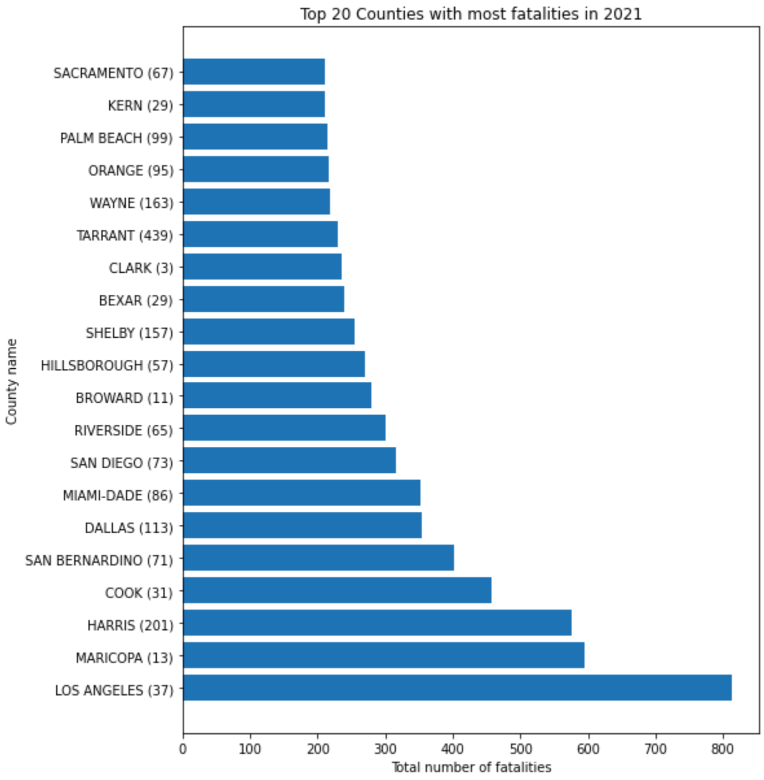
Prior to doing more in-depth exploratory analysis on the data, or to creating a model for the data, some preliminary analysis was conducted to gain a better understanding of the data. To start, the command 'df.describe()' was used to gain some basic statistical information on the numerical columns within the data frame. For the 'FATALS' column, which records the number of fatalities per fatal accident in 2021 (basically how many people died each accident), the max number of fatalities in an accident was 13, and there was a standard deviation 0.354. As part of the preliminary analysis, the sum of the 'FATALS' column was calculated, and it was discovered that 42,939 deaths were due to fatal car accidents in 2021. Furthermore, the histogram shown below, titled Figure 1.0: 'Number of Fatalities in Car Accidents in 2021', depicts that the majority of fatal accidents only caused one death:



(Figure 1.0: 'Number of Fatalities in Car Accidents in 2021')

As seen above, over 35,000 of the deaths from car accidents occurred when there was only one death per accident. The histogram above is not a good representation of the data because the majority of the data lies within one bin, and outliers within the data (i.e the accident that caused 13 fatalities, the one that caused 10 fatalities, and the accident that caused 9 fatalities) make the plot severely skewed.

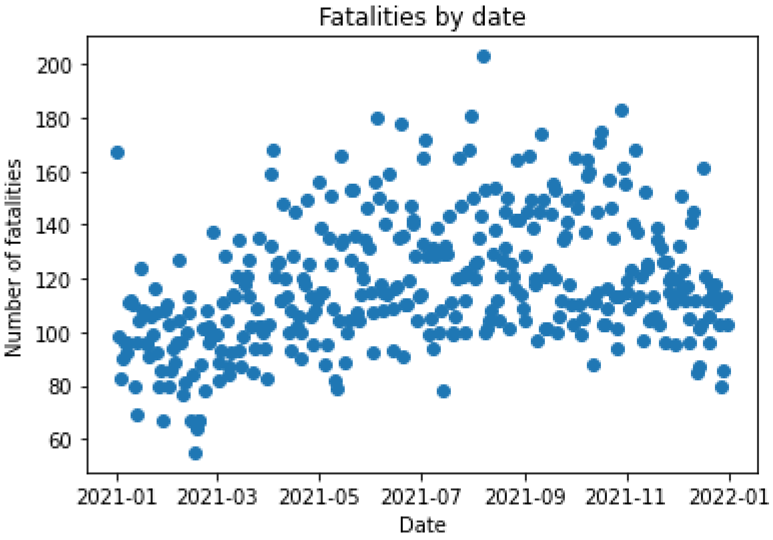
The next step in the preliminary analysis was to investigate the relationship between different counties in the United States and the fatalities per accident. To investigate this relationship, a new pandas data frame, known as ‘countynames’, was created by sorting the ‘FATALS’ column using the command ‘groupby()’ and grouping it by the ‘COUNTYNAME’ column from the dataframe and by totaling the number of times each county name appeared within the data. Instead of investigating all of the counties, only the top 20 counties were used for this analysis. In the bar plot below, titled Figure 2.0: ‘Top 20 Counties with most fatalities in 2021’, the total number of fatalities in the 20 top counties is depicted:



(Figure 2.0: ‘Top 20 Counties with most fatalities in 2021’)

As seen above in the bar plot, Los Angeles county had the most number of fatalities from car accidents with 813 fatalities, followed by Maricopa county with 596 fatalities. Each one of the 20 counties depicted above in the bar chart account for over 200 fatalities that resulted from car accidents.

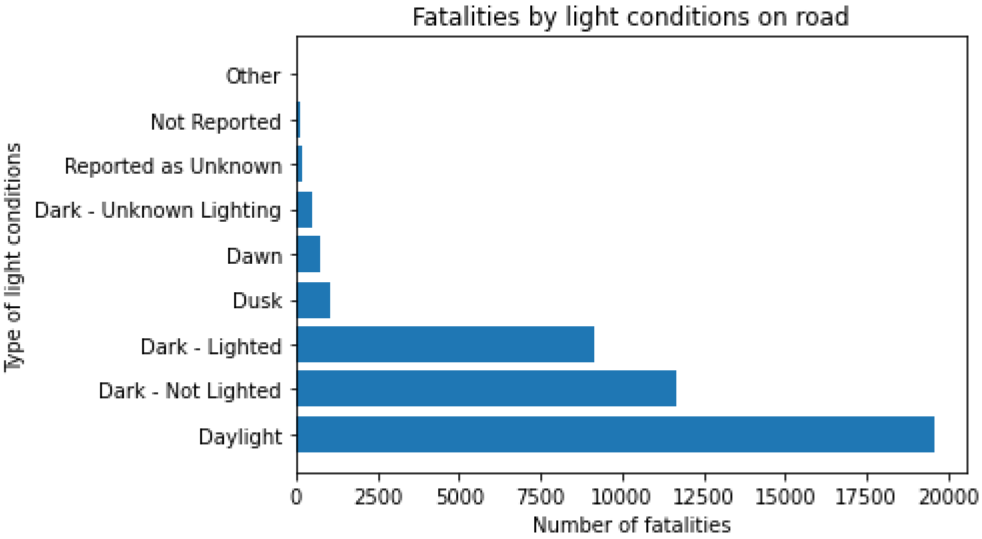
Next, for the next part of the preliminary analysis of the data, the relationship between the date of the recorded accidents and the fatalities was investigated. In order for this relationship to be explored, another data frame was created, this one named 'fataldate', by sorting the 'FATALS' column by the 'DATE' column and then by totaling the number of times each date was recorded in the data frame. As seen below, the scatter plot, titled Figure 3.0: 'Fatalities by date', depicts the total number of fatalities by date:



(Figure 3.0: 'Fatalities by date')

As seen above in the scatter plot, there seems to be a quadratic or exponential relationship between the date in which an accident was recorded and the number of fatalities that occurred. The code for the predictions (quad or exponential) is still being debugged, but it is interesting to notice this relationship between the day of year and the number of fatalities.

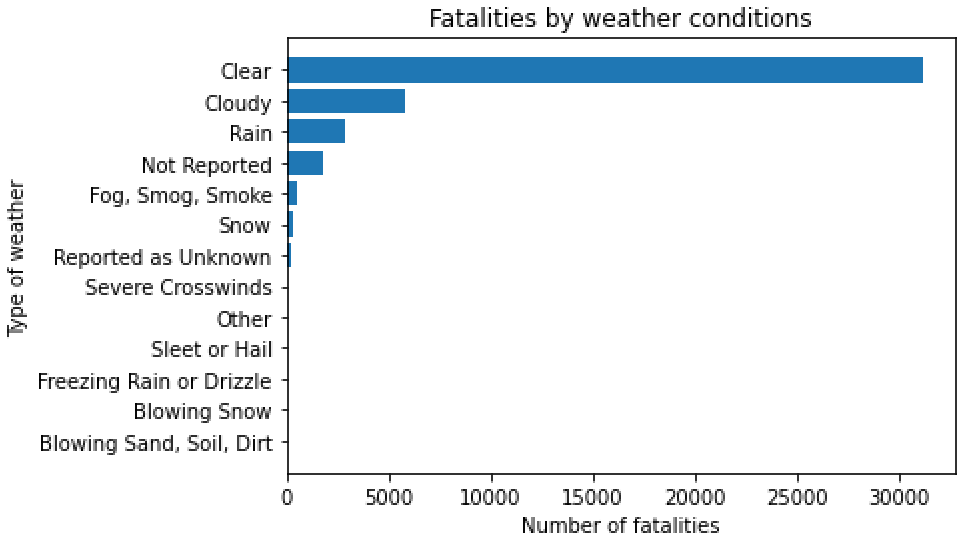
As the last part of the preliminary analysis, the relationship between fatalities and light conditions on the road during the time of the recorded accident was investigated, and the relationship between fatalities and weather conditions was also investigated. Both investigations were done in a similar fashion as listed previously, and two bar charts were created subsequently during the investigations. As seen below in the bar plot, titled Figure 4.0: 'Fatalities by light conditions on road', the relationship between light conditions and fatalities is depicted:



(Figure 4.0: 'Fatalities by light conditions on road')

As seen above in the bar plot, a large number of the total fatalities occurred during the broad daylight. The second most number of fatalities occurred during night time on roads that were not well lit. This information was surprising due to the fact that it was presumed that more fatal accidents would occur in the dark and not in broad daylight.

The bar plot below, titled Figure 5.0: 'Fatalities by weather conditions', depicts the relationship between the total number of fatalities and various types of weather conditions in 2021:

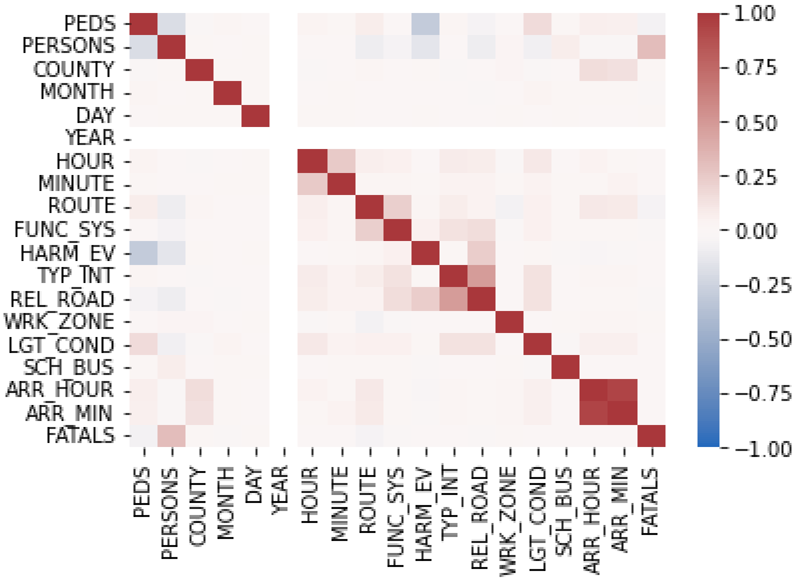


(Figure 5.0: Fatalities by weather conditions)

As seen above in the bar chart, the majority of fatalities occurred when there were clear skies at the time of the car accident. In comparison to fatalities during clear skies, there are significantly less fatalities that occurred during weather conditions such as cloudy skies, rain, fog or smoke, or snow. This was interesting because it was initially presumed that harsh weather conditions would

Exploratory Analysis

In order to begin a more in-depth analysis of the data frame, a heat map was created to determine whether there was any sort of correlation (positive or negative) between 'FATALS' and any of the other columns. To investigate the correlation between columns, a heatmap created with the 'seaborn' library was created. The heatmap below, titled Figure 6.0: Heatmap of Correlation shows whether different variables have any correlation, positive or negative to each other:



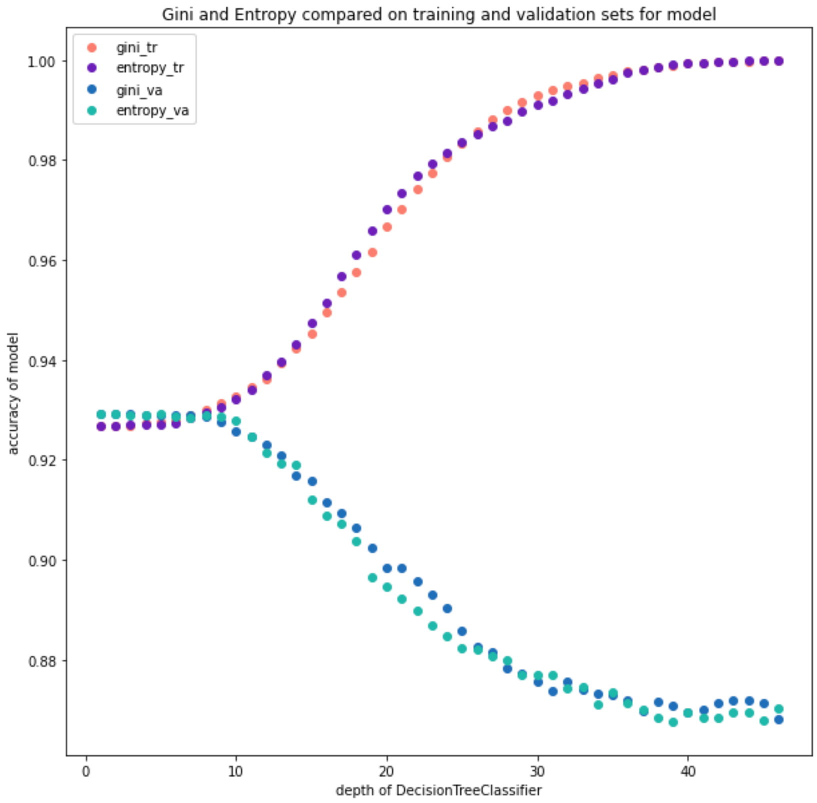
(Figure 6.0: Heatmap of Correlation)

As shown above in the heatmap, 'FATALS' is located on the bottom of the left list of columns from the dataset. In a heatmap, the color corresponding to each rectangle indicates the correlation between one variable and another, where a positive correlation of 1.0 (which means there is a strong correlation) is a dark red, and a negative correlation of -1.0 is a deep blue. A correlation of 0, which means the variables have little to no correlation is indicated between variables, is a neutral cream color. Interestingly enough, this heatmap above shows that most of the numerical variables in the data have little to no correlation with the fatalities. This heatmap was not particularly useful in determining any correlation between 'FATALS' and other factors in the dataset, but there seems to be some positive correlation between 'PERSONS' and 'FATALS', as well as some negative correlations between 'FATALS' and 'PEDS' and 'ROUTE'.

To conduct further analysis on the relationship between the fatalities and other factors, a decision tree classifier was used to predict fatalities based on several features. Although the heatmap wasn't particularly useful in identifying viable numerical variables from the data to use to predict, features were still chosen and used to build a tree. The features that were selected for prediction purposes were the columns ‘DATE’, ‘COUTNYNAME’, ‘WEATHERNAME’, ‘ROUTE’, ‘PERSONS’, ‘LGT_CONDNAME’, ‘WRK_ZONE’, and ‘PEDS’. The target to predict was of course the fatalities in ‘FATALS’. When creating the decision tree, ‘random_state = 0’ was used to ensure that the results of the tree could be repeated by another individual. The data was split into training, validation, and testing sets and then the tree was scored on the validation set after it was fit to the training set. There is no depiction of the decision tree due to it having a total of 46 layers, which is too many to fit into a readable graphic.

Analysis/Modeling Results

After conducting the exploratory analysis through the creation of a decision tree classifier, the model was scored on its validation set to determine the accuracy of its prediction for fatalities. The initial tree, named ‘clf’, had an accuracy of 87.1 percent, which means it was reasonably accurate in predicting fatalities based on the features selected. The tree had 46 layers in which it asked various questions on the data to make predictions on the target, fatalities from car accidents in 2021. After the tree was scored on the validation set, a for loop was created to find the optimal number of branches for the best accuracy of the tree. The accuracies for the gini impurity and the entropy for each layer of the decision tree were compared and then stored in a dictionary. After the accuracies were stored in the dictionary, a scatter plot was created to visually compare the accuracies of the tree on the training set versus the validation set. In the scatter plot below, titled ‘Figure 7.0: Gini and Entropy compared on training and validation sets for model’ compares each impurity and accuracy of the decision tree at each layer:



(Figure 7.0: Gini and Entropy compared on training and validation sets for model)

As seen above in the scatter plot, the accuracy recorded for each depth of the decision tree is recorded, labeled, and color-coordinated based on whether the accuracy is for the gini model on the training set, ‘gini_tr’, the gini model on the validation set, ‘gini_va’, the entropy model on training set, ‘entropy_tr’, and the entropy model on the validation set, ‘entropy_va’. Notably, the accuracy for both the entropy and the gini models increased to 99 percent on the training set as more layers were added to the tree. This seems reasonable due to the fact that the tree is learning on this data set and should be able to predict very accurately on it. On the other hand, the accuracy on both models decreased on the validation set as more layers were added to the tree. After analyzing the scatter plot, it seems that the most accurate decision tree for the validation set is created when there are no more than 10 or 11 layers.

With this information, a new decision tree classifier was created with a max depth of 10 layers. Given that the accuracy for both gini and entropy was nearly identical at this layer, the gini impurity was used for the creation of this tree. After this new tree was created, it was scored on the validation set, and it had an accuracy of 92.6 percent. This accuracy was much higher than the original tree, which meant that it was able to better predict fatalities. Finally, to test the model's performance on new, unseen data, it was scored on the training data set. The model had an accuracy of 92.8 percent on the test data set, which means that it was very accurately able to predict fatalities in car accidents in 2021 based on the features selected for the model.

Limitations

Limitations in Predictions

As with every data set, there were limitations in predictions due to certain factors. When examining the number of fatalities per county, the data was not adjusted by the population size of each county, so it is unsurprising that Los Angeles county has the highest number of fatalities (813). The number of fatalities per county may be skewed due to this, especially when the population is taken into consideration. Los Angeles and Maricopa are both populous counties, so it is not at all surprising that they have more fatalities than Cochran county, which is the last county recorded in the 'countynames' data frame.

Other Limitations

There are also potential limitations in this project due to a lack of a dictionary for the dataset. In typical cases, a dataset will have a dictionary to define what each of the variable names mean to help individuals gain an understanding of the data that is being examined. In the case of the 2021 FARS dataset, there was no dictionary provided, so there was a substantial lack of understanding in regards to what a large portion of the column names meant. This could be a limitation in the prediction of fatalities due to the fact that some of the columns that were dropped in the initial cleanup of the data could have been good features to use for prediction.

References

(n.d.). *Fatality Analysis Reporting System (FARS) and Non-Traffic Surveillance -- NHTSA invites comments to OMB (by 8/25)*. Economics Forum. https://www.aeaweb.org/forum/2845/fatality-analysis-reporting-traffic-surveillance-comments

(n.d.). *Fatality Analysis Reporting System (FARS)*. NHTSA. https://www.nhtsa.gov/research-data/fatality-analysis-reporting-system-fars

(2017, July 11). *Scikit-learn DecisionTreeClassifier with datetime type values*. Qiita. Retrieved December 13, 2023, from https://qiita.com/bmj0114/items/90fae0e30cd6ee8de6db

Newman, R. (2023, December 3). *Rachek253/fatal-car-crashes-2021*. GitHub. https://github.com/rachek253/fatal-car-crashes-2021