Lydia Teinfalt

Individual Report

DATS 6103 Spring 2021

## Introduction

In 2019, the Washington, DC area was ranked 5th for worst traffic in the United States (INRIX, 2020). During 2020 with the Covid-19 pandemic, the DC area saw 77% less congestion and improved its status by no longer being in the top 10 cities with worst traffic (INRIX, 2021). Our group analyzed DC Crash data available at the Open Data DC website from DC Government. With the introduction of the Covid-19 vaccine and more cars coming back to the roads, we wanted to understand the factors involved in traffic crash where major injuries or fatalities were incurred.

## Description

Our group was going to use two datasets: Crashes in D.C and Crashes Details Table from the D.C. government—The District Department of Transportation (DDOT) and The Metropolitan Police Department (MPD). *Crashes in D.C.*  dataset provides overall information about the crash event: the location of the crash, the number of injuries (major or minor), number of fatalities, types of road user involved (car, bike, pedestrian), if any of the involved parties were impaired, and if speeding was involved. *Crashes Details* dataset is a companion and provides information about the person(s) involved in the crash.

District Department of Transportation, Metropolitan Police Department, *Crashes Details Table*, Open Data DC, (District of Columbia): Vision Zero Data Planning Work Group, 2020. Accessed on: Mar. 14, 2020. [online]. Available:[**https://opendata.dc.gov/datasets/crash-details-table**](https://opendata.dc.gov/datasets/crash-details-table)

When we had two datasets, my task was to create readdata.py. The class created read csv files into pandas dataframes and use the merge method to combine the two datasets using CRIMEID. The resulting table was a cartesian product of the two datasets -- Crashes in D.C. has a one-to-many relationship with Crashes Details table. The merge method used kept data from both dataset therefore, creating multiple Longitude/Latitude coordinates with a single crash event. I spent time beginning of the project focused on mapping where crashes occurred and creating a heatmap of month and year of when DC crashes as part of the exploratory data analysis.

The following tables are incorrect because they contain duplicate longitude and latitude data rows from merging two datasets.

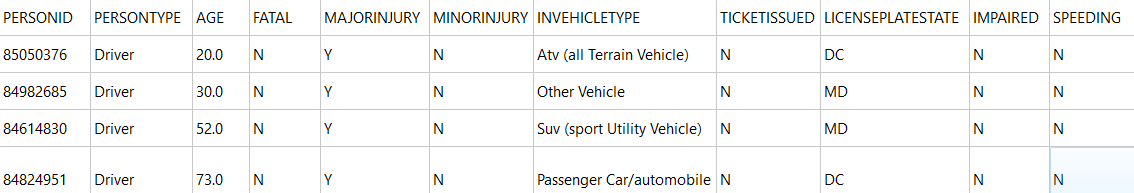
Chart

Description automatically generatedShape

Description automatically generated with medium confidence

When we found flaw with the way we were merging the two datasets, our group focused on analyzing Crashes in Details dataset. As I explained in our group final report, dataset has 10 categorical data except for PERSONID and AGE.

Table 1 Data Sample of Crashes Details DC from March 2020

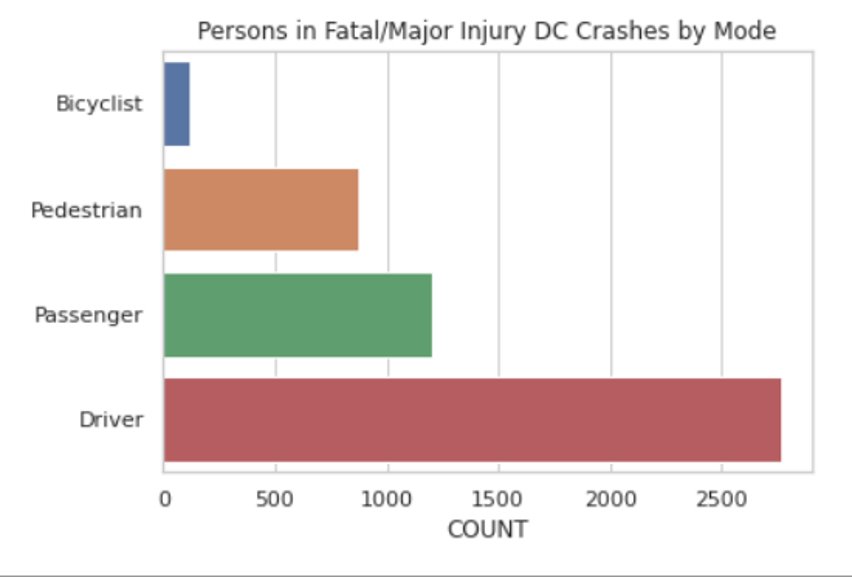


The PERSONID is a unique identifier of the person(s) in the crash.  If the person was involved in more than one crash, the PERSONID would be in another data row. The age is associated with each person identified in the traffic accident. There can be multiple individuals involved in a crash with their four modes of transport to be driver, passenger, pedestrian, and bicyclist. FATAL, MAJORINJURY, MINORINJURY, TICKETISSUED, IMPAIRED, and SPEEDING are binary columns with possible values being Y for YES and N for NO. INVEHICLETYPE represents the type of vehicle involved in crash associated with PERSONID. If a vehicle, the LICENSEPLATESTATE contains state abbreviation where the license plate was issued.

I created the readdata.py class to read in data into a dataframe and created a new column that is also a target called "FATALMAJORINJURIES". This column contains 1 of any the two columns original from the dataset “FATAL" or "MAJORINJURY" is equal to Y (YES).

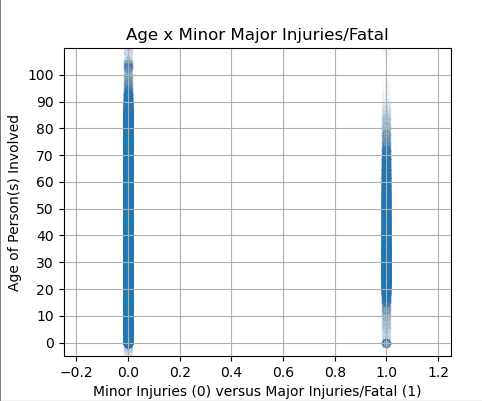
## EDA

My EDA work was eliminated when we shifted from using two datasets into focusing on the crashes details table. We kept the code in the eda.py class but commented it out. I created one chart describing Fatal/Major Injury crashes by the mode of the persons involved. A clear majority of the cases involved was the driver and second, the passenger. The number cases involving pedestrians versus bicyclists were quadruple.

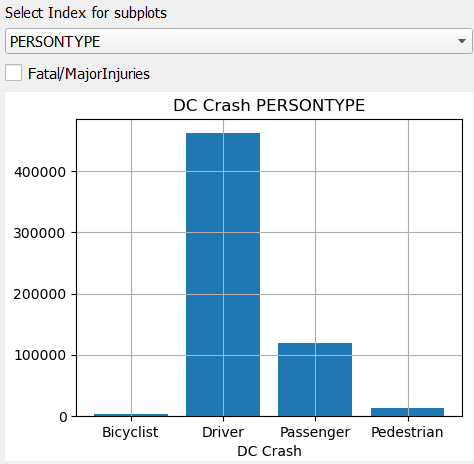


For the project, I did not code any models. Being the developer in group, my task was to develop a GUI application to showcase our project.

The project’s GUI desktop application is accessed by executing Main.py. The menu items are File – Data - EDA – ML Models. From the File menu – Quit (CTRL + Q) will close the GUI application. Data menu – Sample displays the a few rows data from DC Crash Details before preprocessing. Under EDA menu are three exploratory data plots.  The first is Age Histogram which displays “Age of People Involved in Traffic Accidents with Fatalities or Major Injuries” where Age is on the x-axis and number of crashes on the y-axis. The second EDA is “Age Scatter” plot if ages if persons involved in crashes that resulted in Minor Injuries versus crashes that resulted in Major Injuries or Fatal. The range of ages in the data show that we had negative ages and ages above 100 years old.



The third EDA menu option Crash Graphs allows user interaction where an index can be selected by one of the following options to see a resulting graph: IMPAIRED, SPEEDING, TICKETISSUED or PERSONTYPE.  Mark the checkbox to see FATAL/MAJORINJURIES crashes.



The ML Models menu has three models developed by the group: Logit, Random Forest Classifier and XGBoost Classifier. For the models the feature columns are: PERSTONTYPE, AGE, INVEHICLETYPE, TICKETISSUED, LICENSEPLATESTATE, IMPAIRED and SPEEDING. For the Logit model displays findings implemented in the Logit.py class and discussed in previous sections. For both the Random Forest Classifier and XGBoost Classifier displays findings implemented in the randomforest.py and xgboost.py classes but without cross validation.

## References

INRIX: Congestion Costs Each American Nearly 100 hours, $1,400 A Year (2020, March 9). INRIX Press Release, <https://inrix.com/press-releases/2019-traffic-scorecard-us/>, (Accessed April 28, 2021)

INRIX: Downtown Travel Plummets 44% in 2020 amid COVID-19 Pandemic (2021, March 9). INRIX Press Release, <https://inrix.com/press-releases/2020-traffic-scorecard-us/> , (Accessed April 28, 2021)

District Department of Transportation, Metropolitan Police Department, *Crashes Details Table*, Open Data DC, (District of Columbia): Vision Zero Data Planning Work Group, 2020. Accessed on: Mar. 14, 2020. [online]. Available:[**https://opendata.dc.gov/datasets/crash-details-table**](https://opendata.dc.gov/datasets/crash-details-table)