Data Preparation- Project Milestone 1

Guruprasad Velikadu Krishnamoorthy

College of Science and Technology, Bellevue University

DSC540-T301: Data Preparation

Professor. Catherine Williams

June 18, 2023

Data Preparation- Project Milestone 1

Project Description and the Subject Area

Subject Area: Fatality Analysis Reporting System

Description: The project analyzes the data about Fatalities in US Road accidents and studies patterns from the data by plotting visualizations. This can be vital input for building safer roads and infrastructure as thousands of lives are lost and billions of dollars of damage are caused every year. In this project, data is sourced from various sources and different types of data such as Accident data, Information about the victims, weather, etc. from 2021, and then merge to plot visualizations.

Data sources

Three flat files will be used in this project. They are Person data, Data about Races, and Weather data for the year 2021.

1. **Person File:**

Description:

The Person file includes data about the Motorists and Non-motorists involved in the crash. It has details about the fatalities such as their age, sex, time, and day of the accident if they were intoxicated or under the influence of drugs, the status of the seat belt, type of vehicle they were driving, to name a few. Every crash is identified by a unique number ST_CASE. This along with the Vehicle number (VEH_NO) and the person id (PER_NO) can be used as a unique identifier in the Person file. The file also contains other details such as State, county, and city where the crash occurred. This file has 126 columns and about 96500 rows.

Link:

https://static.nhtsa.gov/nhtsa/downloads/FARS/2021/National/FARS2021NationalCSV.zip. Within the zip file, the file is named as "person.csv"

2. **Races file:**

Description:

The race file contains the race of each deceased person from the crash as indicated in the death certificate. The fields ST_CASE (unique identifier of the crash), Vehicle number (VEH_NO), PER_NO (Person number), and the Order listed (ORDER) can be used as the Unique identifiers of each record. The file has 11 columns and about 97000 rows.

Link:

https://static.nhtsa.gov/nhtsa/downloads/FARS/2021/National/FARS2021NationalCSV.zip. Within the zip file, the file is named as "race.csv"

3. **Weather file:**

Description:

The weather file contains data about the atmospheric condition at the time of the crash. The fields ST_CASE and WEATHER can be used as unique identifiers of the file. The field WEATHER contains unique codes for each type of weather event such as Cloudy, Rain, etc. The file has 5 columns and about 40000 rows.

Link:

https://static.nhtsa.gov/nhtsa/downloads/FARS/2021/National/FARS2021NationalCSV.zip. Within the zip file, the file is named as "weather.csv"

4. API:

Description:

Accident data is pulled using the API. This data includes the details of the crash from 2021. The ST_CASE is the unique identifier of each record. This data contains other information such as Crash date and time, state, county, Number of vehicles involved, number of persons involved, Number of fatalities, etc. The API returns 11 columns and about 40000 rows.

Link:
https://crashviewer.nhtsa.dot.gov/CrashAPI/crashes/GetCaseList?states=1,51&fromYear=2021&toYear=2021&minNumOfVehicles=1&maxNumOfVehicles=6&format=json

5. **Website**

Description:

The website data provides various crash statistics of the 50 states along with the population. Some stats are crash totals in 2021 from each state, Motor vehicle crashes by road user type in 2021 by state, and Deaths by crash type such as single-vehicle, and multiple-vehicle crashes. Some tables have about 8 to 10 columns and 51 rows in each.

Link: https://www.iihs.org/topics/fatality-statistics/detail/state-by-state#crash-types

Relationships

As the project is about analyzing the crash data, most of the connections happens through the field ST_CASE (State Case) which is a unique identifier for each crash per state. The ST_CASE is a unique number for each crash that can uniquely identify the crash and the state where it occurred. For instance, the ST_CASE 360115 is the 115th crash in the state of New York and 36 is the State ID for New York state. Also, there are additional columns by which the datasets are being connected which are discussed below.

Relation between the CSV files:

The CSV files person.csv and race.csv are connected by the fields ST_CASE, VEH_NO, and PER_NO. These fields identify the person by the vehicle for each crash. The CSV files person.csv and weather.csv are connected via the ST_CASE field.

Relation between API and CSV files:

The API returns the details about the accident and hence the API and the person.csv and other csv files can be connected by the ST_CASE field which uniquely identifies each accident. The API also returns county information that includes the County name and the County number in the state (Example: ST. CLAIR (115). The County number in the state can be extracted and it can be used to join with person.csv data to create a summary at the County level in each state.

Relation between Website data, API, and CSV files:

The website has the summary data per state such as the summary of crashes by road user type, type of vehicle involved, etc. Hence the data from website tables can be combined with the API and CSV files using the State_ID field and by the State name. There are about 96,500 records in the Person file while each table on the website has 51 rows (50 US states and Washington DC). Hence the website data can be used in creating Visualizations that show summary data per state when combined with the data from the CSV files or API.

Project plan:

The project aims to create Visualizations from the crash data and study the pattern across 50 states. The website has data about the state population, and it can be useful to compare the trends with the population for each state. Some of the visualizations that will be created in this project are:

1. Summary of crash victims by state

2. Summary of Crash victims by state and Person type such as Driver, Passenger, Motorcyclist, Pedestrian, etc.

3. Summary of People killed by state and by Age group. This can help us find which state has the most teenage victims in the Road crash, for example.

4. Summary of victims by state and Vehicle type

5. Summary of the data from different counties of a state.

6. Summary across different times of the day to identify the deadliest hour when most crashes occur. There is a belief that most crashes occur between midnight to 6 AM when the drivers are most tired. The study can help us find out if it is true.

7. Summary of crashes across 12 months to see which month had the most crashes. Also, we can find the deadliest day of the year that had the greatest number of people killed in road crashes.

Ethical implications:

1. One of the major Ethical implications of this study is the ability to differentiate between the acts that are considered criminal charges versus not criminal offenses. For instance, speeding is not considered to be a major offense, however, most crashes and fatalities from crashes occur due to vehicles going at high speeds thus killing people. Hence the data collection and research must be done keeping this in mind.
2. The other ethical implication is the Alcohol limit. For instance, in Texas, the legal alcohol limit is 0.08% for regular drivers and 0.04% for commercial drivers. Hence reporting the alcohol limits and age limits must be done accurately.
3. Wearing seat belts and helmets has saved so many lives during crashes. However, there is an ethical implication that it may invade someone's freedom by not letting them decide for their safety but imposing laws to wear seat belts and helmets.
4. Finally, the project deals with people that lost lives during crashes. Hence the data must be researched and reported with empathy and not just as mere numbers, especially while reporting children's deaths.

Potential Challenges

1. The large volume of data may be a challenge. There are about 96500 records in Person dataset and about 40000 records from the Accident API. Hence joining these records may be a challenge. Also, the API has a limit of 5000 records for each request. Hence the API calls must be done in batches.
2. Data Quality may be a challenge. If the data has many missing or invalid values, it may lead to erroneous results.

What is required to achieve all milestones and the Interpretation of the data?

1. Performing Exploratory analysis to understand the data is one of the essential steps to achieving the milestones. Identifying the key fields that are required in each dataset and identifying how to join the datasets to derive meaningful relationships between them is a key step in the process.
2. Data cleansing is another vital step in the process by identifying outliers and removing them as needed. Fields with missing or erroneous values must be handled.
3. Some of the date and timestamp fields may not be in the desired format and hence they must be converted to the expected format before processing them.
4. While reading the data from API in batches, caution must be taken as the data will be truncated due to the limit set as 5000 for each request. Hence it will require multiple trials and errors to identify filters during each API call to avoid data loss.