SPARK INDIVIDUAL ASSIGNMENT

Final Report

ABSTRACT

In this final report, I discuss the background, goal of analysis, the analysis, and conclusion of a dataset regarding roughly 3.5 million car accidents that took place in the United States between February 2016 and June 2020.

Nelson, Ckalib

Spark – Professor Raul Marin Perez

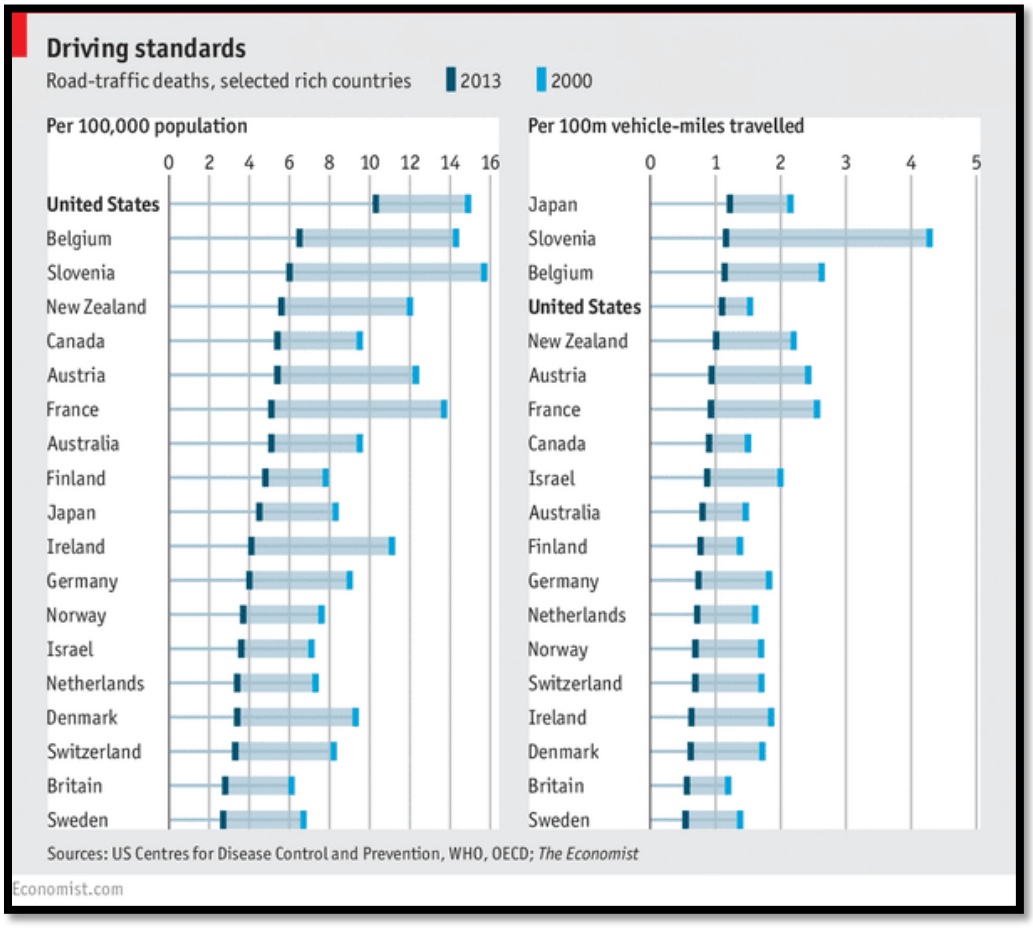
Table of Contents

Background/Scenario Description

Background Information

During the early 2010s, many Americans benefited from a better economy and significantly cheaper oil prices. Both of these trends greatly incentivize individuals to travel via automobile. Additionally, Americans tend to drive much more than others around the world, especially on rural roads with insufficient lightening and barriers.

Unfortunately, the United States had one of the worst road-safety records in the world during this time period. Per The Economist, America's road deaths rose by the highest annual percentage increase in 50 years in 2013.[1] Furthermore, the rate of death per 100,000 people per year was roughly 10.9, which is twice as high as Belgium, the country with the second worst record. Surely, these horrifying incidents prompted many public institutions to collect data on accidents in an effort to prevent deaths on the road.



[1] The Economist: America's road-safety record is the worst in the rich world. https://www.economist.com/graphic-detail/2016/09/05/americas-road-safety-record-is-the-worst-in-the-rich-world

Data Description

To complete the individual assignment, I decided to use the 3.5 million records of US Accidents per Kaggle. The data was collected in real-time via APIs from various local and government institutions and represents traffic accidents in 49 states from February 2016 to June 2020. Per the notebook, here is an illustration of the data on accidents in the United States:

Exploratory Data Analysis

In [2]:
```
#Count number of accidents
us_accidents.count()
```

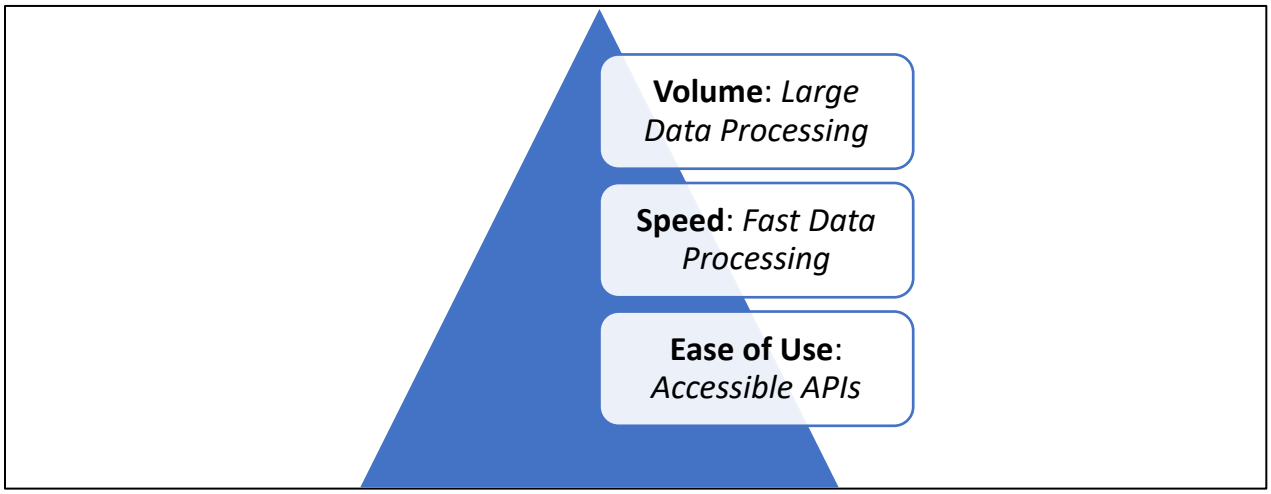
Out[2]: 3513617

Here, we can confirm that there were roughly 3.5 million accidents that occured in the US.

In [3]:
```
#Preliminary View of Dataset
us_accidents.show(1)
```

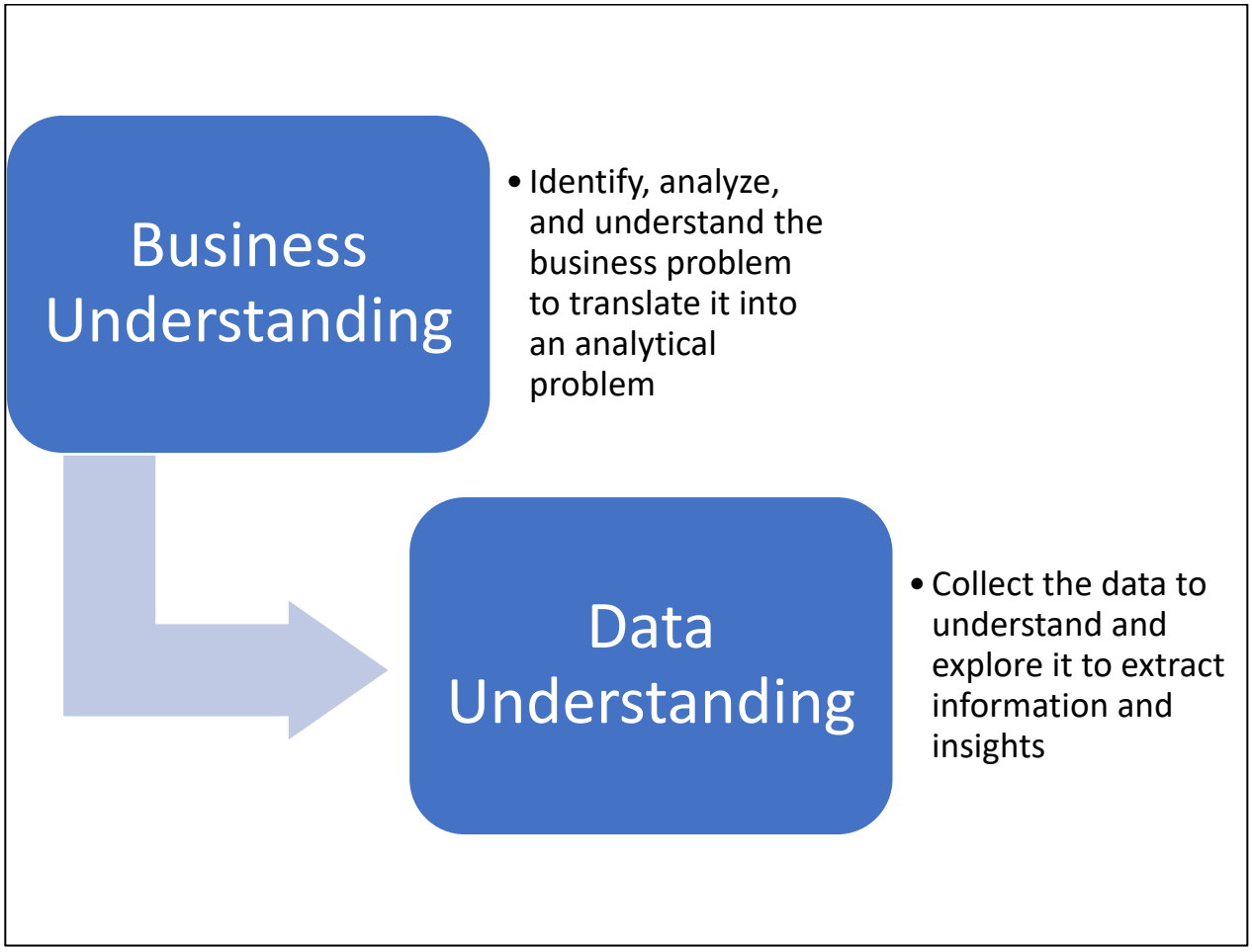
```
+---+--------+-----+--------+-------------------+-------------------+---------+----------+-------+-------+-----------
-+--------------------+------+------+----+------+----------+-----+-------+-------+----------+------------+-----------
--------+-------------+-------------+-----------+-----------+-----------+--------+-------------+-------------+--------
---------+-----------------+-------+-----+---------+---------+---------+-------+-------+----------+-------+-----+-------
--------+-------------+------------+-------------+-------------+---------------+--------------------+
| ID|  Source|  TMC|Severity|         Start_Time|           End_Time|Start_Lat| Start_Lng|End_Lat|End_Lng|Distance(m
i)|         Description|Number|Street|Side|  City|    County|State|Zipcode|Country|  Timezone|Airport_Code|  Weather_
Timestamp|Temperature(F)|Wind_Chill(F)|Humidity(%)|Pressure(in)|Visibility(mi)|Wind_Direction|Wind_Speed(mph)|Precipi
tation(in)|Weather_Condition|Amenity| Bump|Crossing|Give_Way|Junction|No_Exit|Railway|Roundabout|Station| Stop|Traffi
c_Calming|Traffic_Signal|Turning_Loop|Sunrise_Sunset|Civil_Twilight|Nautical_Twilight|Astronomical_Twilight|
+---+--------+-----+--------+-------------------+-------------------+---------+----------+-------+-------+-----------
-+--------------------+------+------+----+------+----------+-----+-------+-------+----------+------------+-----------
--------+-------------+-------------+-----------+-----------+-----------+--------+-------------+-------------+--------
---------+-----------------+-------+-----+---------+---------+---------+-------+-------+----------+-------+-----+-------
--------+-------------+------------+-------------+-------------+---------------+--------------------+
|A-1|MapQuest|201.0|       3|2016-02-08 05:46:00|2016-02-08 11:00:00|39.865147|-84.058723|   null|   null|        0.0
1|Right lane blocke...|  null|I-70 E|   R|Dayton|Montgomery|   OH|  45424|     US|US/Eastern|        KFFO|2016-02-08
05:58:00|          36.9|        null|       91.0|       29.68|         10.0|          Calm|           null|
0.02|       Light Rain|  false|false|   false|   false|   false|  false|  false|     false|  false|false|          fa
lse|         false|       false|          Night|          Night|          Night|               Night|
+---+--------+-----+--------+-------------------+-------------------+---------+----------+-------+-------+-----------
-+--------------------+------+------+----+------+----------+-----+-------+-------+----------+------------+-----------
--------+-------------+-------------+-----------+-----------+-----------+--------+-------------+-------------+--------
---------+-----------------+-------+-----+---------+---------+---------+-------+-------+----------+-------+-----+-------
--------+-------------+------------+-------------+-------------+---------------+--------------------+
only showing top 1 row
```

Spark is an excellent tool to use for this analysis given how large this dataset is and the numerous benefits of Spark:



CRISP-DM

To complete this project in an organized and methodological manner, I utilized CRISP-DM. CRISP-DM is a widely used and accepted methodology for data related projects that was developed by a consortium of over 200 organizations in 1996. Although CRISP-DM is a six-phase methodology, my project will not require the last four phases given the nature of the project. Thus, I only completed two phases: Business Understanding and Data Understanding:



Once I completed the Business Understanding and Data Understanding phase, I built tables to answer questions about the dataset, which will be discussed in the following section.

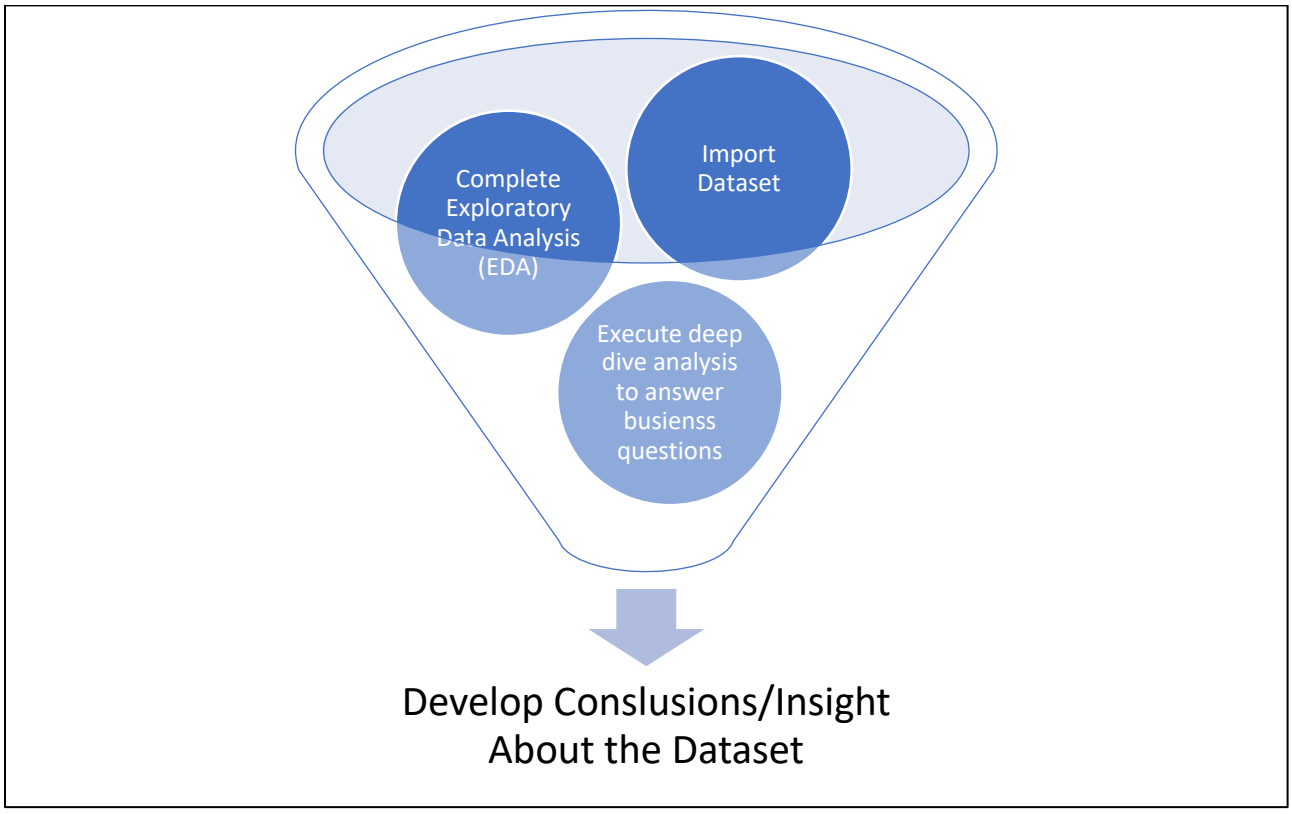
Goal of the Analysis

The goal of my analysis is to answer the following ten questions about my data:

Business Question 1	• In which states do most accidents occur?
Business Question 2	• Which one of the aforementioned states have the most severe accidents?
Busienss Question 3	• What are the top descriptions of the accidents?
Busienss Question 4	• Under what conditions do most of these accidents occur?
Busienss Question 5	• What is the landscape of accidents in the area in which I live?
Busienss Question 6	• What is the average wind speed for each of the severity levels?
Busienss Question 7	• What is the average precipitation of each of the severity levels?
Busienss Question 8	• What is the average temperature for each of the severity levels?
Busienss Question 9	• How severe are accidents at crossings?
Busienss Question 10	• How severe are accidents at railways?

Completed Steps

After determining which business questions, I completed the following steps to answer those questions:



Deep Dive Analysis

Leveraging the lectures and notes from the professor, I completed a deep dive analysis of the dataset to answer the aforementioned questions. As a result of my analysis, here are the answers to those questions:

In which states do most accidents occur?

- It seems that accidents occur the most in California (CA), Texas (TX), and Florida (FL).

Which one of the aforementioned states have the most severe accidents?

- Between CA, TX, and FL, it seems as if FL has the most severe accidents.

What are the top descriptions of the accidents?

- It seems as if the descriptions of the accidents include information regarding congestion on interstates, which may align with the intutiion we have about safety of interstates. Maybe public officials in these areas should reconsider these dangerous areas!

Under what conditions do most of these accidents occur?

- Contrary to intuition, it seems that most accidents occur while the weather conditions are clear, fair, or mostly cloudy. This most likely has to do with the fact that climates differ depending on which state you live in. For example, it may tend to rain more in Florida than in California.

What is the landscape of accidents in the area in which I live?

- I have lived in Charlottesville, VA since August of 2015. Although I am not familiar with the first accident listed above, I am very familiar with the roads/interstates mentioned in the description of these accidents. Fortunately, there haven't been very severe accidents in the area nor have I been involved in them!

What is the average wind speed for each of the serverity levels?

- It does seem as if wind and severity are positively correlated albeit marginally.

What is the average precipitation for each of the severity levels?

- It does seem as if rain and severity are positively correlated albeit marginally.

What is the average temperature for each of the severity levels?

- It does seem as if temp and severity are negatively correlated. The colder it is, or the lower the temperature, the more likely there is to be an accident.

How severe are accidents at crossings?

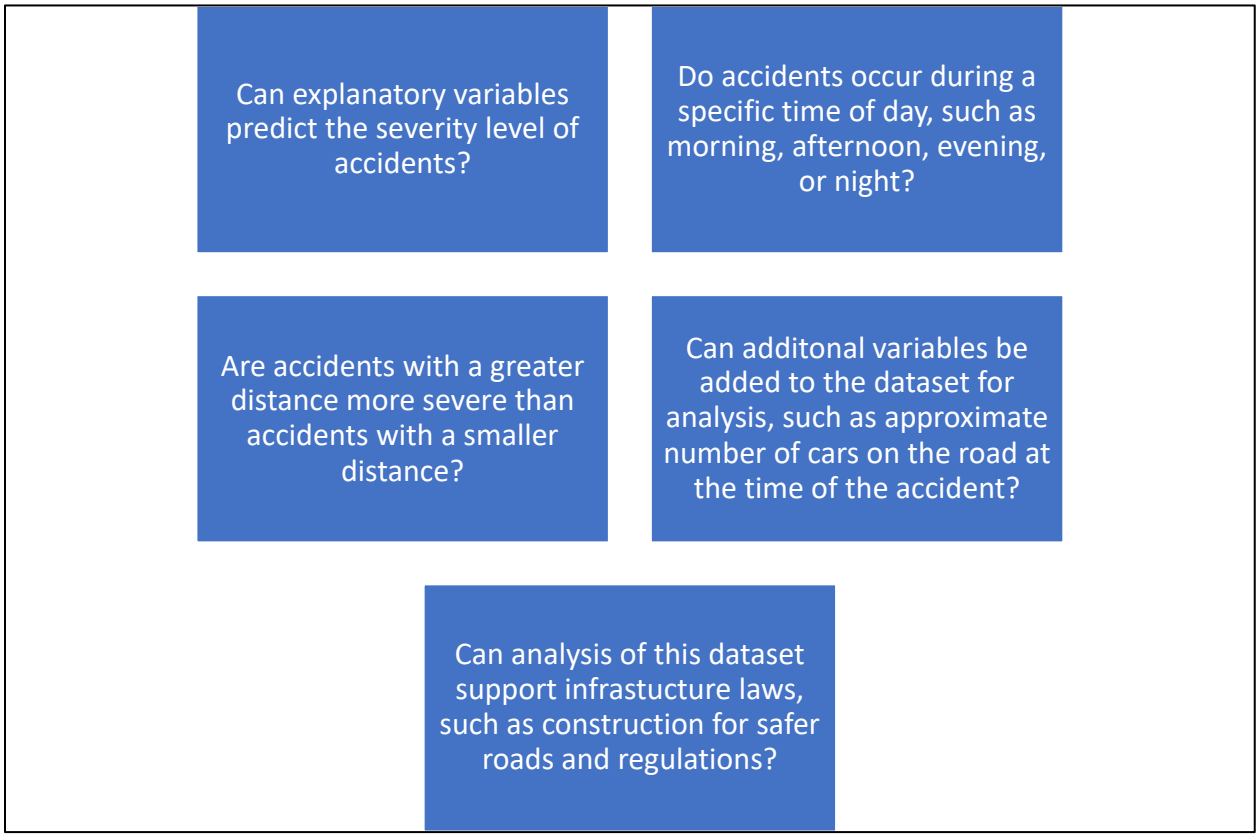
- It seems as if accidents at crossings are quite severe. Most accidents at crossings are either a 3 or 4 category.

How severe are accidents at railways?

- It seems as if accidents at railways are not as severe as they could be with most severe accidents at railways being a category 2. Relative to crossings, railways are a bit safer.

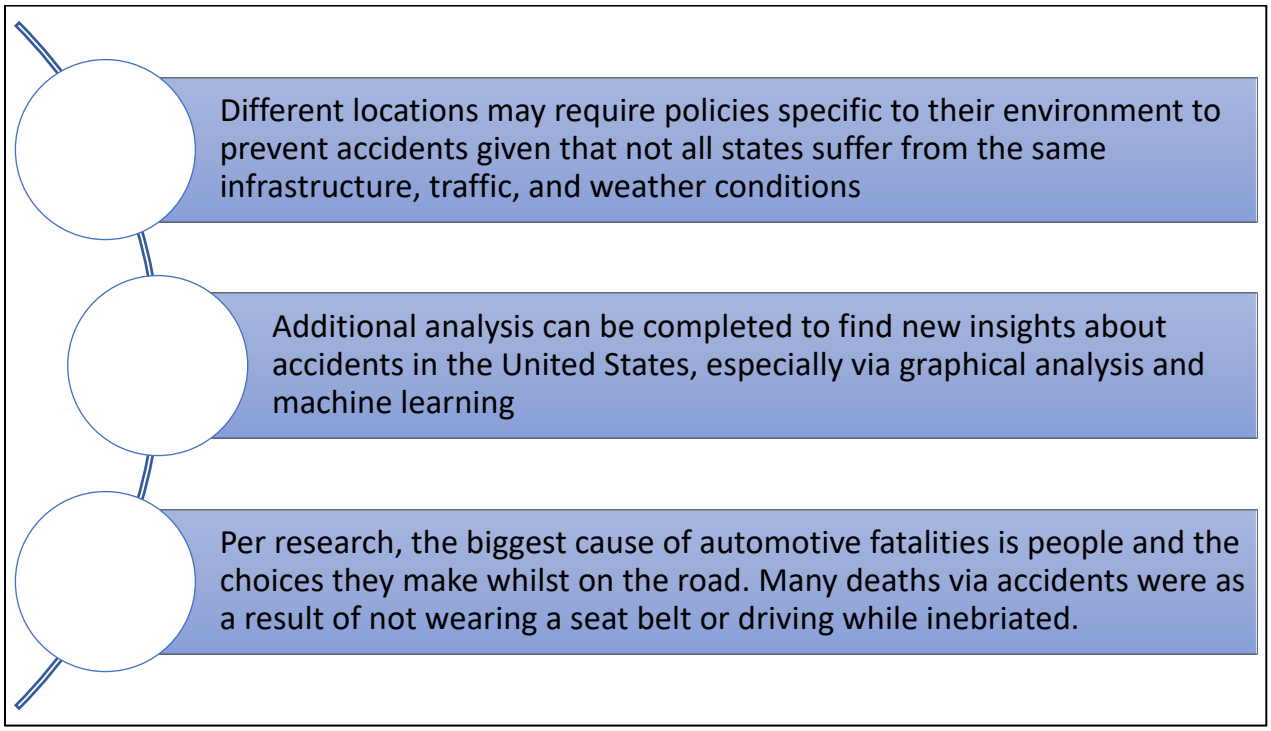
Future Analysis

Although my analysis answers ten interesting business questions about accidents in the United States, there is so much more analysis that can be done! Here are five additional questions I would like to answer once I develop my Spark skills to a greater level of expertise:



Conclusions/Insights

Although I have answered several business questions and could answer additional questions in the future, there will still be several elements about accidents in the United States that the data won't be able to answer. Nonetheless, there needs to be policies implemented in the United States intended to reduce the accident and death rate on the road. Thus, what else might we be able to take away from this analysis and what additional information can we learn about accidents in the United States?



Video Presentation

Here is the link to the video presentation that discusses this project in detail: Spark Individual Assignment (C. Nelson)