US MOTOR VEHICLE CRASHES

Scuola universitari professionale

della Svizzera italiana.

SUPSI

OCTOBER 12, 2022 - JANUARY 30, 2023

AUTHORS: GABRIEL TABACARU & MANUEL ROMANELLI
PROFESSION: STUDENTS

Index

Abstract

Road traffic crashes are a leading cause of death in the United States, but that is not the only place where the number of vehicle crashes is high, and because of that we decided to do a deeper analysis on those crashes to find where exactly the most crashes happen and the reason that could generate a vehicle accident.

Introduction

In our analysis we decided to study the phenomenon of road accidents.

We decided to choose this topic to underline their great frequency and importance in the impact of human life.

The main part of this project is to show the location where the accidents happened and the causes of the accidents.

The purpose of this project is to inform the population about the accidents frequency and to convince people to pay more attention to some aspects of that could cause an accident.

Our analysis also aims at making people understand how many mistakes humans can make, in this case while driving, and therefore the importance of finding a solution to reduce the number of annual accidents.

Data Sources

Being the continent with the most information available, we decided to choose the American continent as the subject of our analysis.

We got our data from:

https://data.ny.gov/Transportation/Motor-Vehicle-Crashes-Case-Information-Three-Year-/e8ky-4vqe

https://data.world/data-ny-gov/e8ky-4vqe

This dataset was published by the Crash Records Center of the State of New York, the coverage of the accidents is Statewide but there are also some data about other states. Since the territory of the United States is so vast and out data is mainly about one state, we decided to concentrate most of our analysis in the are of New York and nearby states.

Our data is from the year 2014 to the year 2021 and on their website the owners of the dataset post annually so it is possible to get new data every year to make more accurate analysis and check the growth or decline of the number of accidents.

For more specific analysis we also used another dataset containing all the counties with their specific position and their population.

We got a free version of the dataset from https://simplemaps.com/data/us-counties , which also offers some more detailed datasets, containing more fields, for a specific price.

Data pre-processing

To be able to have a more reliable analysis as close to reality as possible we decided to use more datasets, one from the years 2014 to 2016 and one from the year 2017 to 2021, luckily the Crash Records Center recently updated the second dataset, adding the years 2017, 2018 and 2021 that were missing when we started our project.

Our data preprocessing started with the concatenation of those two datasets and all the next steps were done on the concatenated dataset.

To be able to generate our data visualizations with specific data and to have the population of each county we used another dataset named uscounties.

In the United States there are currently 3143 counties but unfortunately our datasets contain the data about only 64 of them which are mainly in the state of New York and the ones close to it.

Because of that we decided to select only the state of New York and those close to it.

To be able to have more specify information about the number of crashes we use the population of each county to be able to create a new column containing the accidents per population of each county.

To find the conditions when the most crashes happened we used the information we had in our dataset about the: Road Descriptor, Road Surface Conditions, Lighting Conditions.

For the rest our dataset was pretty clean and we didn't have to handle any problems of data missing or other things.

Interface design

For the interface design we create an HTML page that we next uploaded to github, from there we were able to host our personal project page as a site.

The link to the web page is: https://taba013.github.io/Data_Visualization/

We divided the page in four main sections, one under the other.

In the first section we put the title, subtitle and the authors of the page, right below that it is possible to find our dataset that you can download and the Visualization Protocol related to our project.

In the second section we showed our first map of the US that shows the number of accidents for each country and below a fast introduction.

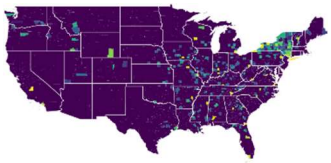
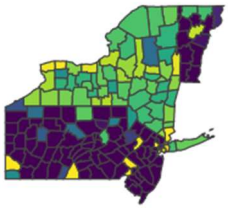
In the third section there is a dropdown that allows to select between two maps of the state of New York and the nearby ones. On the right of the map there is a bar plot that shows the counties with more than 10000 accidents.

In the last section of the page we used a Sankey diagram to show the conditions that caused the accidents and based on that we gave some tips to avoid as many accidents as possible.

Data visualizations

The main data visualizations that we decided to focus on are about the location of the accidents to find the counties with most crashes and some visualizations to find the conditions when most accidents happen.

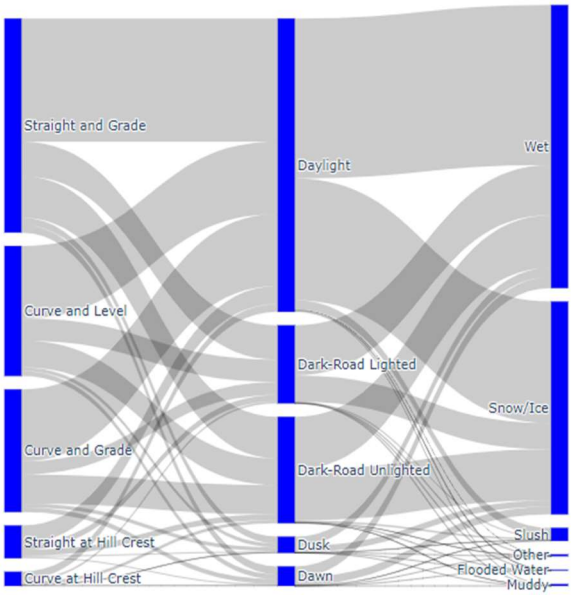
For the locations we decided to use two different maps:

- One map of the United States
- 
- One map of the State of New York and the states close to it.
- 

From the first map we found out where our data is centered and based on that we were able to find a more specific map which is the second.

The second map is focused on the state of New York and the ones close to it which are Pennsylvania, New Jersey, Connecticut, Massachusetts, Vermont. This map helped visualize counties with most of the accidents.

As our second visualization we decided to focus on the condition that caused the accident. This helped us understand how the road, the light and the weather conditions were when the accident happened.



To make this visualization we used a Sankey diagram to display the flow between the road descriptor, the lighting conditions and the road surface conditions.

As we can notice by the graph, it is possible to find when most accidents happened.

Thanks to that we can give some advice to when to pay more attention to the driving:

- When the road is wet and grade.

- When there is snow or ice on the road.

- When is dark and there are not lights.

- When there is a curve.

Next Steps

As next steps or to make the analysis more specific it is possible to find another dataset containing the weather conditions of each county.

That would allow to generate more accurate visualizations on the accidents conditions because our analysis are influenced by the fact that the road is dry, straight and level most of the times and that caused the visualization to seem like that's when it is most likely for an accident to happen, which is not true.