Visualization Protocol

US Motor Vehicle Crashes

Gabriel Tabacaru, Manuel Romanelli

Timespan of the datasets:

First Dataset: from 2014 to 2016

Second Dataset: from 2017 to 2021

Data Source:

We got the 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 is also data about other states.

For some more information about each county, 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.

Metadata:

Column Name	Description	Type
Year	Calendar year of incident	Number
Crash Descriptor	Reported injury or damage outcome of crash: • Property Damage & Injury Accident • Fatal Accident • Injury Accident • Property Damage Accident	Plain Text
Time	Reported time of crash (hh:mm)	Plain Text
Date	Reported calendar date of crash (mm/dd/yyyy)	Date & Time
Day of Week	Reported day of the week the crash occurred	Plain Text
Police Report	Indicator of whether a police crash report is on file with NYS DMV. Y = “Yes” / N = “No”	Plain Text
Lighting Conditions	Reported lighting conditions at time of crash.	Plain Text
Municipality	Reported municipality of crash location	Plain Text
Collision Type Descriptor	Collision Manner Type Description	Plain Text
County Name	Reported New York State county where the crash occurred.	Plain Text
Road Descriptor	Reported road description where the crash occurred.	Plain Text

Column Name	Description	Type
Weather Conditions	Reported weather conditions when the crash occurred.	Plain Text
Traffic Control Device	Reported traffic control device present where the crash occurred.	Plain Text
Road Surface Conditions	Report road surface conditions when the crash occurred.	Plain Text
DOT Reference Marker Location	Department of Transportation reference marker present at location of crash.	Plain Text
Pedestrian Bicyclist Action	If applicable, the Pedestrian Bicyclist Action at the time of the crash.	Plain Text
Event Descriptor	The reported description of the crash event.	Plain Text
Number of Vehicles Involved	The reported number of vehicles in the crash	Number

Abstract

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

The main features that we focused on are the number of accidents per county and the conditions of the road where the accident happened.

To be able to get the number of accidents we decided to group them by the county where it happened.

To create the map of the accidents we combined the number of accidents of each county with the dataset of all the counties which also contained the position and population of each one.

We first generated a map of the US where we noticed, as expected, that the data is mainly focused on the state of New York and the ones around it, so we decided to just consider those states.

Thanks to the population we were able to generate a more specific map with the accidents per population of each county.

As last visualization we decided to do a Sankey diagram with someone of the conditions of the road: the road descriptor, the lighting conditions and the road surface conditions.

Data handling

First, we check that the columns of our two datasets are the same so we don't have any problem when we concatenate them.

After creating the concatenated dataset, we export it as CSV and that is the dataset that you can download from our page.

We then check how much data we have of each year by grouping the accidents by year and generating a bar plot.

We create a new dataset containing the accidents grouped by county.

We import the dataset containing all the counties and add a new column with the number of accidents of each county.

After that we create our first data visualization which is the map of the US with all the accidents per county.

Now we focus more on the states of New York, Pennsylvania, New Jersey, Connecticut, Massachusetts, Vermont.

Based on those states we create a new dataset containing all the counties in those states.

From that dataset we create our second data visualization which is a map of the accidents per county of those states and also create a bar plot to show all the counties with more than 10000 accidents.

To create the same data visualizations but more accurate we added a new column containing the number of accidents per population.

In the end we removed from our initial dataset all the unknown values to be able to create a Sanky diagram to display the flow between the road descriptor, the lighting conditions, and the road surface conditions.