Hayley Baek

Car Accidents in Brazil

Abstract

With cars being the main mode of transportation in many parts of the world, car accidents are a regular occurrence and can lead to injuries and even loss of life. They continue to be a growing concern for public safety issues. Identifying trends in causes and types of car accidents can help to illuminate the growing body knowledge of road safety and provide actionable insights for city planners, insurance companies, and even the general public.

Brazil is one of many countries that have a strong car presence. Brazil ranks as the seventh highest populated country in the world [1] with 47.12 million motor vehicles in circulation in 2023 [2]. This study analyzes car accident data in Brazil to visualize potential factors and key trends and propose a potential solution to reducing the rate of car accidents.

Introduction

Road traffic accidents are a significant public safety issue worldwide, and Brazil is no exception. As one of the largest and most populous countries in the world, Brazil experiences a high volume of road traffic, resulting in a concerning number of car accidents each year and leading to injuries and loss of life. This paper aims to leverage several visualizations to explore car accident data from Brazil and identify key risk factors and trends.

By examining accident characteristics such as location, time of occurrence, weather conditions, and road conditions, this study seeks to provide a comprehensive understanding of the factors that contribute to road traffic incidents. The types and causes of car accidents that are considered highly dangerous can occur in higher frequency or reportedly more severe outcomes, measured by the number of people injured in a given accident instance.

The visualizations are motivated by a series of tasks defined in a later section and crucial tools for analyzing the data and translating them into actionable insights. The insights gathered from this analysis can inform policy recommendations aimed at reducing the frequency and severity of accidents, ultimately leading to safer roads and improved public health outcomes. Understanding the causes and patterns of car accidents in Brazil is crucial for developing effective prevention strategies.

Dataset

The dataset for car accidents in Brazil is sourced from Kaggle user Mlippo who in turn sourced the data from the Brazilian government's website [3]. There are almost half a million rows of data that's relatively clean, meaning there is very little preprocessing to be done for handling null values, dropping meaningless columns, or enrichment by adding columns from another dataset. While the dataset originally published on the Brazilian government website is in Portuguese, the

English translated version of the data is generously, supplementally published in the Kaggle dataset.

The dataset covers years 2017-2023 and includes a number of meaningful columns like exact location and time, number of injuries and deaths from the accident, types and causes of the accident, road type, and weather. Some columns included but are largely disregarded are weather timestamp, which indicates whether the weather is sunrise or sunset, victims condition, which can instead be observed from the different number of victims columns, and two other columns (‘regional’ and ‘police station’) that include the state abbreviation but nothing more about the conditions of the car accident.

The location data is provided in latitude and longitude coordinates as well as city and state names. This richness of the data is part of why the dataset doesn’t require joining other datasets to add meaningful context. The locations are able to be plotted on a map to both small and large scales without losing information, and this ease of use is helpful for analysis through geographic visualizations.

The causes of accidents attribute has around a hundred unique categorical values that have been grouped into larger categories of causes: aggressive driving, negligent driving, environment, DUI, mechanical failure of the car, and pedestrian involvement. These causes categories are defined solely for the purpose of visualizing a larger set of data that will be discussed in a later section of this paper.

Tasks

This study also sets out to analyze what external factors might contribute to car accidents. The data includes weather conditions and times of the day, so the creation of a visualization between these two variables is necessary by the total number of people injured. One hypothesis is that rainy or snowy weather causes more car accidents and that there is also a higher number of accidents around the nighttime than in the daytime.

Similarly, road type is another external factor considered in the trend analysis of the car accident data. Whether a road is one lane, double lanes, or multiple lanes, the next hypothesis is that there is a higher frequency of accidents across roads that have more lanes for more cars.

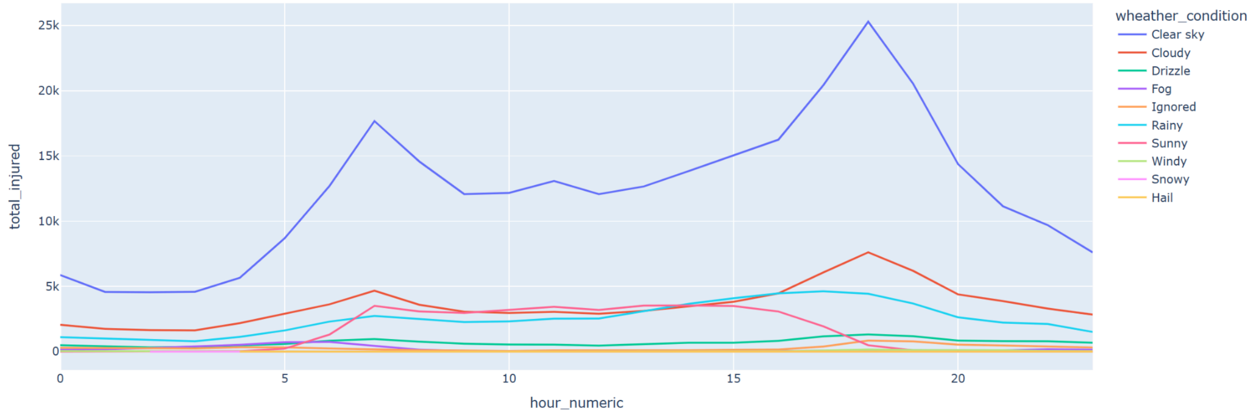
One large problem to be solved by the data is to visualize trends in the causes and types of accidents as well as any potential factor in causing car accidents. Causes of accidents include but are not limited to using phones while driving, objects in the road, or any vehicular mechanical failures. Types of accidents include but are not limited to vehicle collisions, vehicle fire, or pedestrian collisions. There is determined to be a high likelihood of accidents being caused by drivers more than the environment or mechanical failures resulting in collisions to other vehicles.

Accident types are further analyzed through potential differences in frequency between the different states in Brazil. This is to uncover whether there are significant differences between higher and lower population states having a similar frequency of accident causes. The hypothesis for this is that there are still more accidents caused by other drivers even across a large spread of population sizes.

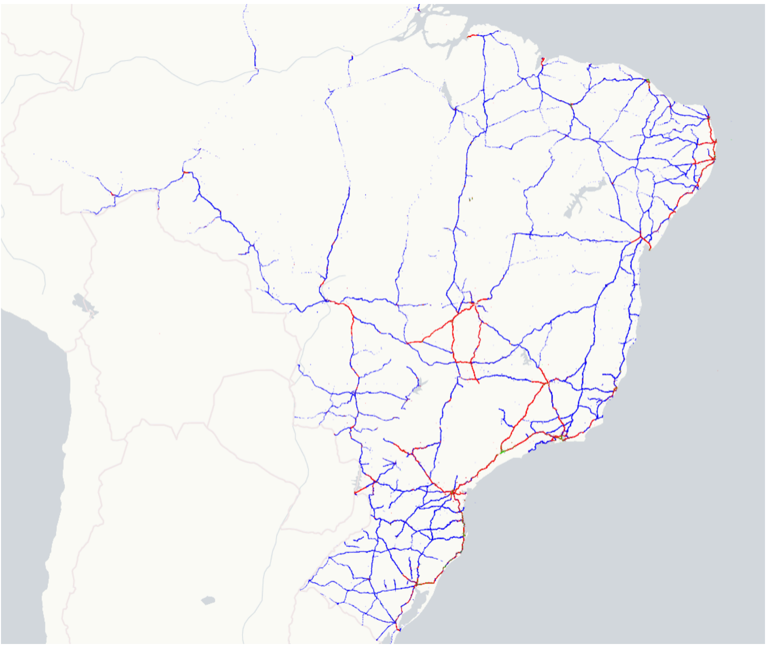
Solution

The visualizations were implemented using the Python libraries Pandas, Matplotlib, Plotly, Geopandas, and Lonboard. Pandas allows for highly flexible manipulation of the data into dataframe objects that further allow for tasks like creating meaningful visualizations. Matplotlib and Plotly are visualization tools that can create a multitude of different types of charts and graphs that handle different types of data. Geopandas and Lonboard are packages that specialize in visualizing a large volume of geospatial data and can create high level maps.

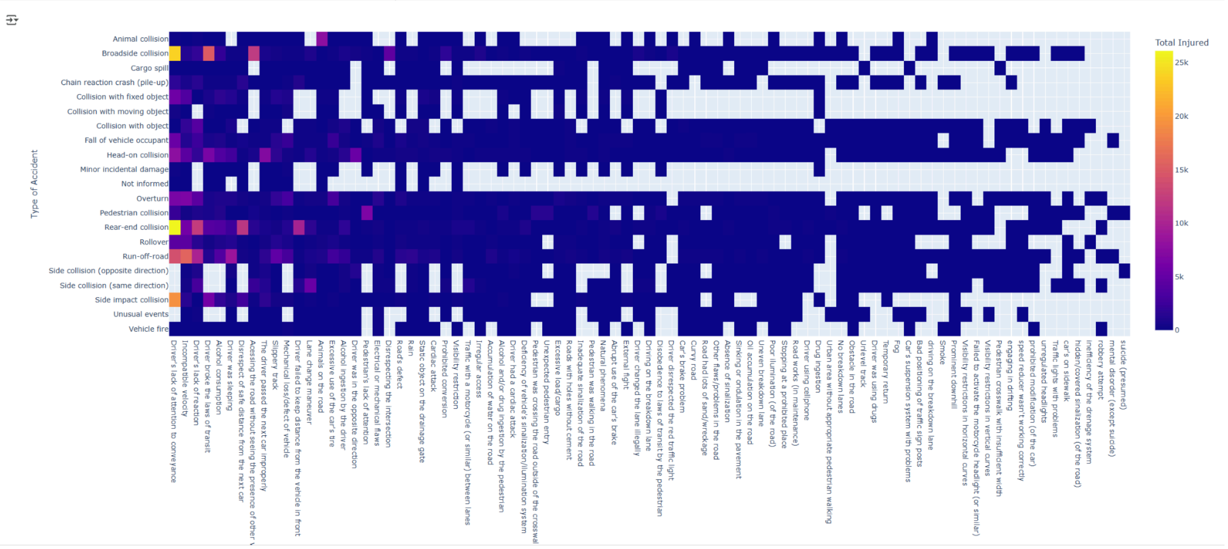
Four visualizations were created to answer the question of what trends exist in the Brazilian car accident data, each visualization uncovering a new insight for what risk factors contribute to high consequence accidents.



The first visualization is a line graph of the total number of injured persons by time of day. The lines in the graph are segmented by the weather conditions (misspelled as wheather from the Portuguese translation). The x-axis range is from 0 to 24 to represent the hour of day, and the y-axis ranges from 0 to 25,000 The line graph is the reliable go-to visualization choice for plotting time data because of its ability to communicate time-sensitive changes across any numeric attribute. The colors are arbitrarily assigned to the weather categories by Plotly's default categorical colormap, so they might not be the most effective for plotting a large scale category than the weather attribute which has a definitively small number of allowed values.



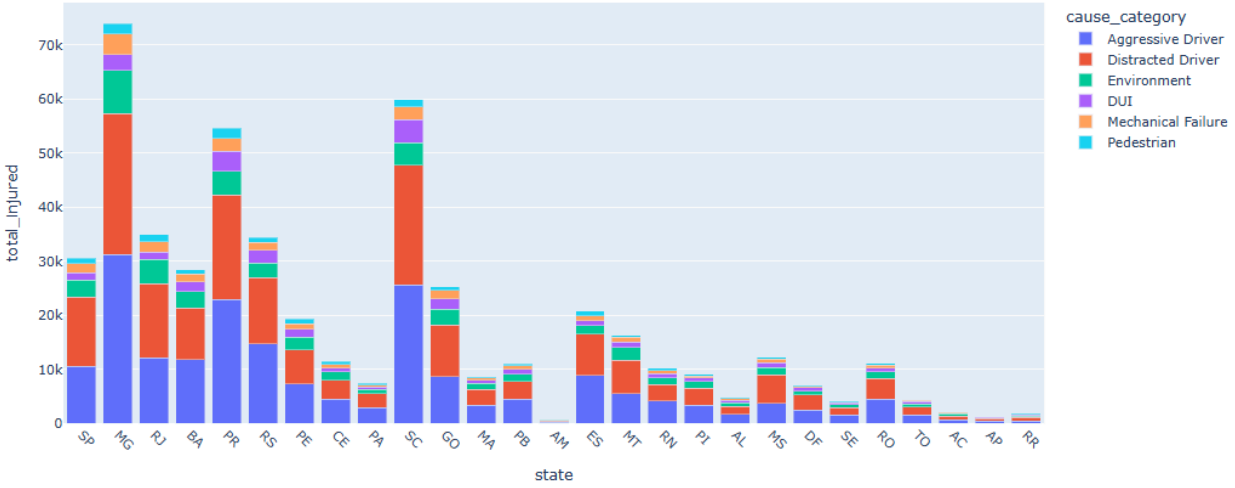
The second visualization is a map view of the accidents colored by road type. There are three types of roads determined by the dataset: blue (single lane), red (double lane), and green (multi lane, not visible in the visualization). This map was created using Geopandas and Lonboard, which are Python packages specifically for creating geospatial visualizations. Lonboard was utilized over other packages because of its ability to handle the large volume of data, over 400k rows in the entire dataset, and display individual points in a much quicker way from other geospatial visualization packages like Geopandas or Folium. Lonboard also leverages zooming as a crucial interactivity component, as the user can see the exact density of data points in any given view.



The third visualization is a heatmap of the types of accidents by causes of accidents. The x-axis is ordered by its count of the total number of injured persons so that the distribution is relatively skewed to one side. This is opposed to the x-axis being ordered alphabetically and having the heatmap having a horizontally even distribution of colors. This ordering draws the viewer to one side and allows for a faster visual representation of what specific combination of accident causes and types appear more frequently.

The colormap is a hue gradient with a completely different color representing nonexistent accident cause and type combinations. While it might be difficult to view a group of cause and type pairs that have similar values of total number of injured persons, this colormap gradient allows the view to juxtapose more extreme ends of the range and makes outliers stand out much more clearly and effectively.

Because this visualization is so large, the user is encouraged to interact with the graph by scrolling horizontally and hovering over each cell to access the tooltip. The tooltip window shows the exact count of each pair of cause and type accident pairs, which is useful for a future case of adding more values of either categorical attribute.



The fourth and last visualization is a stacked bar graph of categorized causes of accidents by state. The numeric measure for this graph is the total number of people injured. There is a fixed number of states in Brazil that is unlikely to change regularly, so there should not be an issue with plotting all or most of the 26 Brazilian states. The x-axis is ordered by the states' populations [4] so the user can see an implicit correlation between population and number of total injured people.

There are almost a hundred different values found in the dataset for the causes of accidents, and this visualization would not have benefitted from having even a filter of the most common type of accidents because it would have misrepresented the total count that's available in the data. The causes of accidents therefore are grouped into six categories specifically for this visualization in the study: aggressive driver, negligent driver, environment, DUI, mechanical failure (of the vehicle), and pedestrian. The colors are arbitrarily assigned by Plotly's default categorical colormap, which might not be the most effective but are distinct enough to differentiate the small number of categorical values.

Results

The line graph of external factors measuring the rate of car accidents reveals most clearly that clear sky weather is actually the weather that single-handedly accounts for most of the total number of car accident injuries. This directly contrasts the hypothesis set forth in the Tasks section that rainy or snowy weather would correlate to a higher rate of car accidents. While if all of these weather conditions were grouped together to a single "not clear" category, it might compare to the number of accidents with clear sky weather, but it still illustrates the next most significant observation illustrated by the graph that time of day is consistently peaked in rate of car accidents around rush hour. This is also a different result from the hypothesis in the task that proposed that a higher number of accidents occurred in the nighttime than in the daytime.

The key takeaway from this line graph visualization is that there are more likely to be accidents when there are more people out on the roads at the same time, mainly around rush hour when people are commuting to and from work.

The map view of Brazil laying out the exact geographic location of the accidents in the entire dataset reveals that there are double roads that connect major cities in Brazil and single roads around most of the other roads in the country. Zooming into the map also reveals that there is a lower density of accidents in the single roads and a higher density of accidents in the double roads. Interchanges and any road leading towards city centers have a higher density of accidents of any road type. There are more accidents that occur on double roads, and the country's layout of double roads is mainly to connect major cities.

It's not necessary to suggest that having more single roads that connect major cities would result in fewer accidents, but one consideration is that more car accidents are likely to occur with any higher volume or density of cars.

The heatmap visualization reveals that the top causes of car accidents are: distracted driving, driving too quickly or too slowly, and delayed reaction time. This was easily found because the columns or x-axis values were ordered in descending order by the total number of injured people per cause. The most frequent pair of accident types and causes is distracted driving leading to rear-end collisions followed by distracted driving leading to side-impact collisions. This was also easily found because of the frequency of types-causes accident pairs being so unevenly distributed in a small handful of pairs. The hue color gradient makes any outliers stand out more noticeably from the trend, so the top frequency of types-causes accident pairs were visually noticeable.

After hovering over the cells to find approximate counts of the total number of injured persons, careless driving alone is found to comprise over 17% of all car accident causes. This heatmap served as inspiration to delve further into the causes of accidents to see approximately how much of all car accidents could be attributed to specifically drivers versus other types of causes.

The stacked bar graph reveals that more than half of all accidents across any state can be attributed to other drivers, whether they be aggressive drivers or negligent drivers. With the x-axis being ordered by states' populations, this visualization also illustrates a slight correlation between accident rates and population sizes. For some Brazilian states like MG (Minas Gerais) and SC (Santa Catarina), their higher accident rates could be explained due to their terrain being more rugged compared to other Brazilian states. Minais Gerais has some of the highest peaks in the entire country including the Pedra de Mina peak in the Mantiqueira Mountains [5], and Santa Catarina has the most mountainous terrain, where 52% of the territory is located above 600 metres [6].

The causes of accidents were grouped into larger categories of causes. Only four accident causes are attributed to distracted or negligent driving, but those four causes make up for about half of all causes of accidents by drivers. This is in contrast to twenty-three causes of accidents categorized into accidents by aggressive drivers that make up the other half of all causes of accidents by drivers. The key takeaway from the stacked bar graph is that there is a pattern of drivers accounting for half of all accidents across all states' population sizes.

What all four visualizations have in common is that high rates of car accidents can be explained by driving behavior. Whether from the line graph revealing that higher accident rates correlate with rush hour and more people being out on the road at the same time, the geographic map revealing that double roads that connect major cities have a higher frequency of accidents, the heatmap revealing that most frequently accidents are caused by driver negligence leading to some car-on-car impact, or the stacked bar graph revealing that drivers alone account for more than half of all car accidents across all states, drivers are the common denominator for high car accident rates.

Bibliography

[1] https://www.worldometers.info/world-population/brazil-population/

[2] https://www.statista.com/statistics/831145/motor-vehicle-fleet-size-units-brazil/

[3] https://www.kaggle.com/datasets/mlippo/car-accidents-in-brazil-2017-2023?select=accidents_2017_to_2023_english.csv

[4] https://www.worldatlas.com/articles/brazilian-states-by-population.html

[5] https://trilhaserrafina.com.br/trilha-pedra-da-mina/

[6] Garschagen, Donaldson M. (2000). "Santa Catarina". Nova Enciclopédia Barsa. Vol. 13. São Paulo: Encyclopædia Britannica do Brasil Publicações Ltda.