**Car Accidents in Brazil Final Report**

**Osvaldo Morales, Carson Leung, Rezwan Rahman, Yuxuan Chen**

**Introduction**

A car accident is an undesired situation when a car makes physical contact with an object (another car, person, etc.). A car crash involves a user operating the vehicle, impact level(minor or major), takes place at a certain location (spatial), at a certain time(temporal), has a reason for crash (human error, external factors, etc.) and a concluding result(injured, uninjured, fatal). Car accidents are a prevalent issue that exists in societies that have many cars. It is an issue the modern world cannot avoid, which is why there is so much data existing today that pertains to car accidents. This topic is important to study because it affects people's everyday life, causing injuries and death.

For our project, we decided to study car accidents to see what areas of Brazil are dangerous when it pertains to car accidents. We collected a dataset from kaggle that contains 400,000 registers about traffic accidents that happened in 7 years(2017- 2023). We used this dataset to study patterns and trends in car accidents across Brazil, the relationship between factors like population and density on car accidents, what car accidents are like during big holiday events, and if external factors like weather and seasons have an effect on car accidents. Car accidents are not something that we can predict, however it is something that we can do a deep analysis of.

**====================================================================**

**Data Collection**

Main Data: <https://www.kaggle.com/datasets/mlippo/car-accidents-in-brazil-2017-2023/data>

Our dataset was originally in Portuguese but later translated into english. There are over 400,000 data points from the year 2017-2023. Every column contains important car accident factors, like date, time, location, cause of accident, number of vehicles involved, and result of accident.

*Data Integration*:

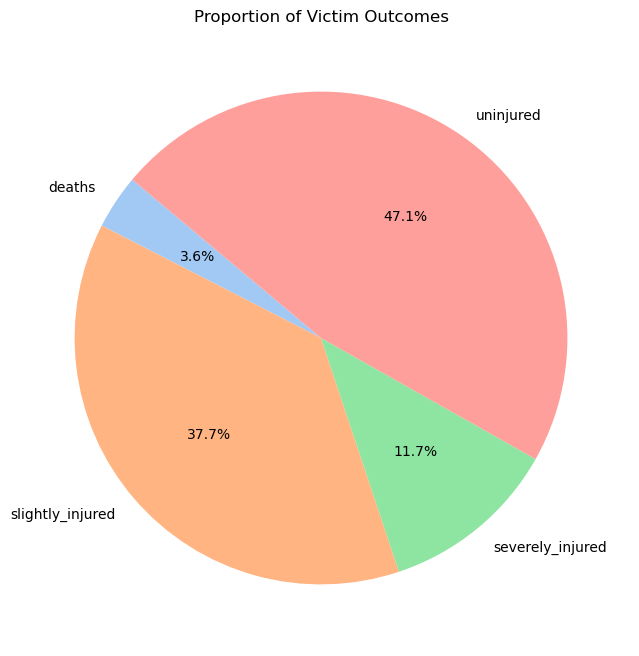
Integrated Data: https://www.worldatlas.com/articles/brazilian-states-by-population.html

In order to gain a better understanding behind our data and study accidents, we needed to insert additional data in order to get better results and understand the car accident situation in brazil. We attain population size, and density per km2 for every state. We added both state population and density as a column on our data frame representing population and density for every corresponding state. We also added a season column, representing what season the accident happened.

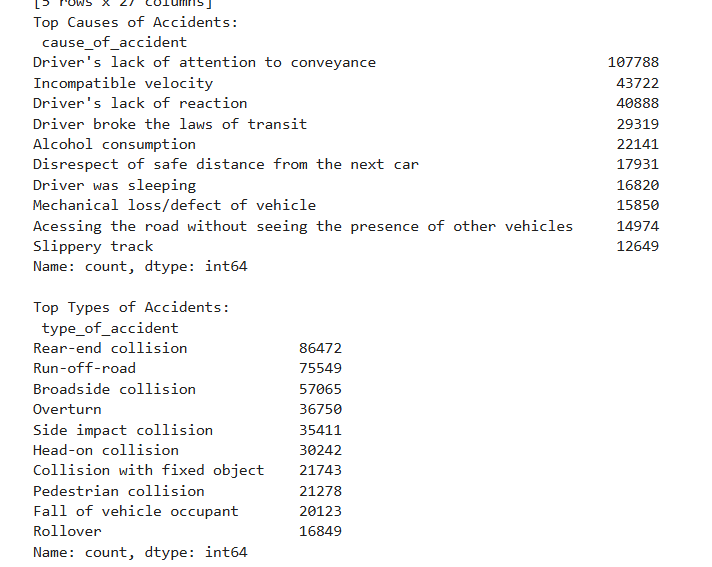
=====================================================================

**EDA**

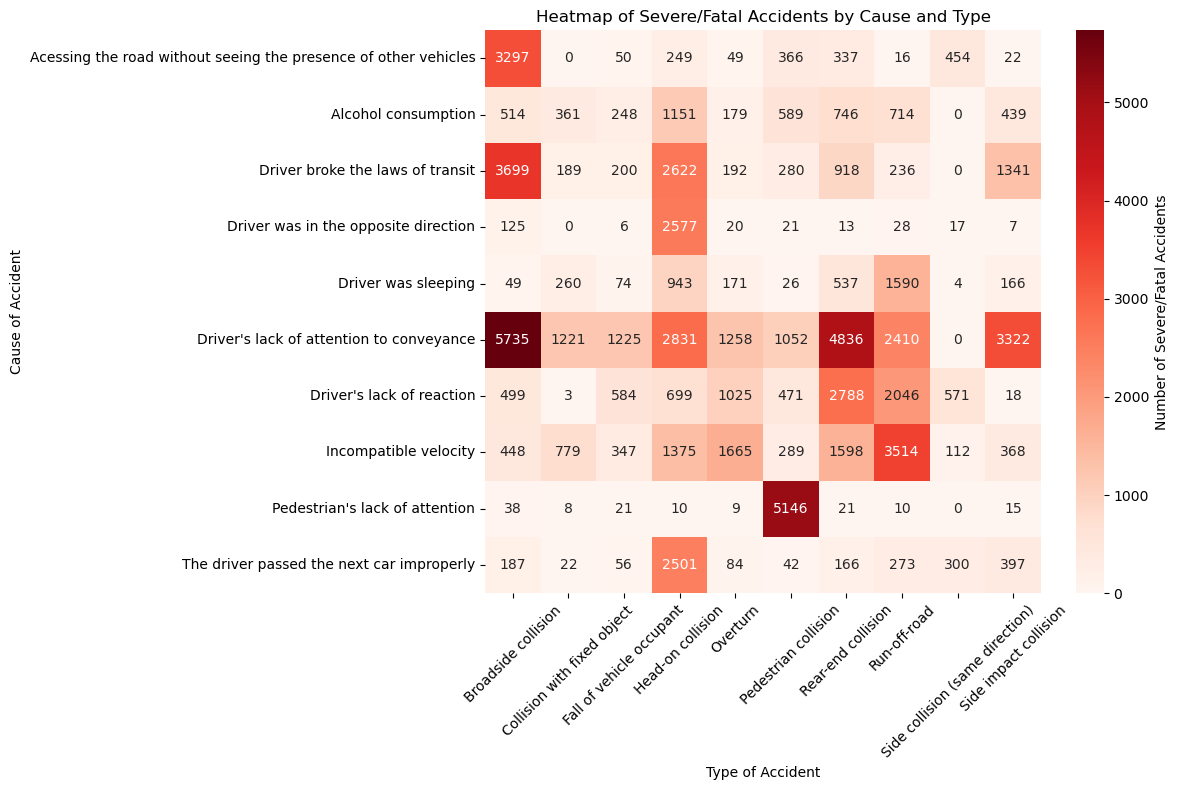
In our initial EDA of the dataset, one of the first things we looked into the severity of the accidents that occurred. From this, it resulted in the following pie chart:



Looking at the results of this pie chart, we saw that a majority of these accidents ended with people being slightly injured or completely unharmed. However, since our goal was to look at accidents that were potentially fatal or actually fatal, we decided to focus more on the smaller proportions of the overall accidents, which were the ones that had people severely injured (11.7%) and/or dead (3.6%).

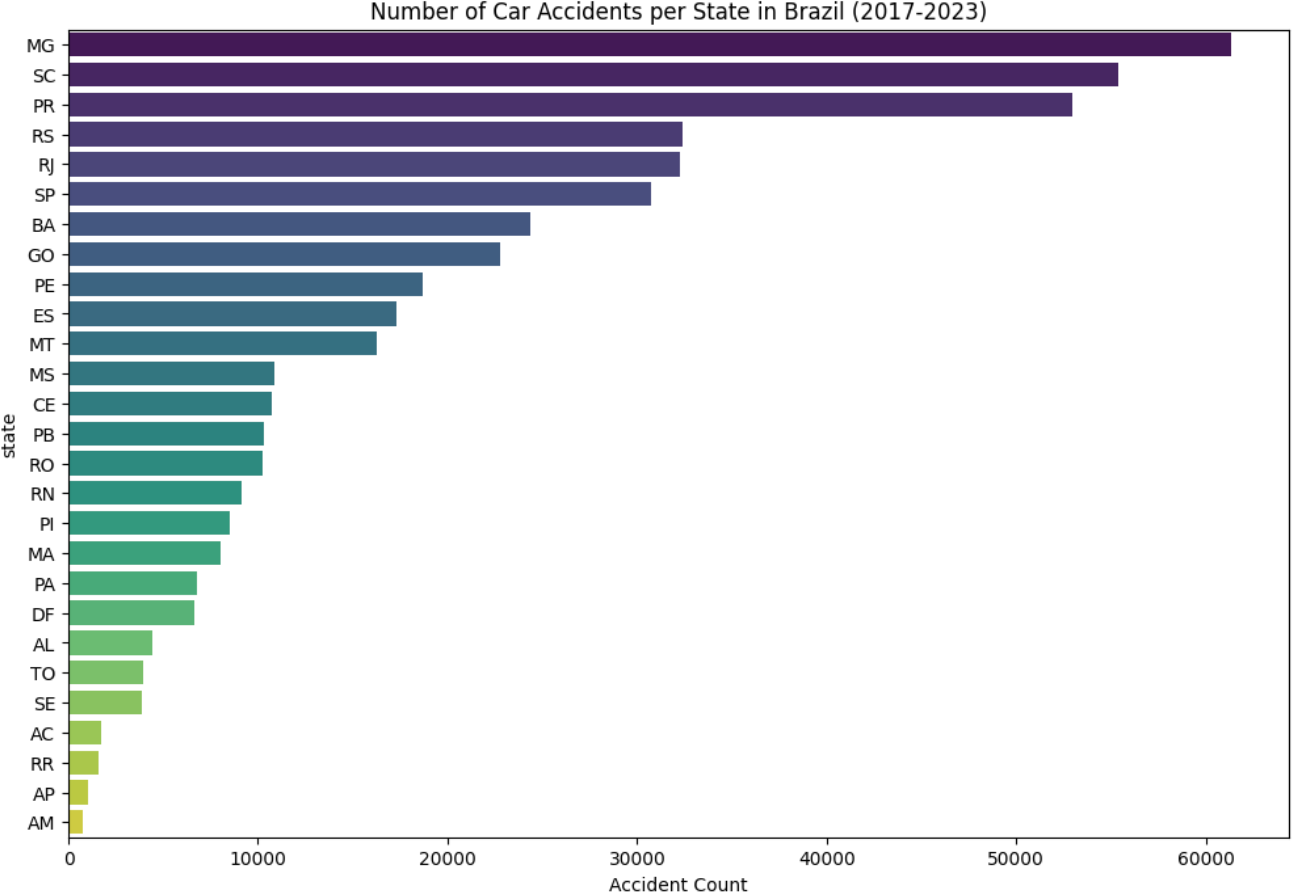


After our findings on the severities of the accidents, we looked into another couple of features in the dataset. One of them being *cause\_of\_accidents* and another being *type\_of\_accidents.* We printed out the top 10 causes and types of accidents based on the amount of them that appeared in the dataset. From this, we saw that the top causes of these accidents are due to human error. While we get an idea of what the top causes for these accidents are, we did not see the relationship between the causes and the types of accidents. So, to go a little further, we looked into what types of accidents are due to the causes. In the following two visualizations, you will see the relations of the causes and types of accidents(left) and the severity of them(right):



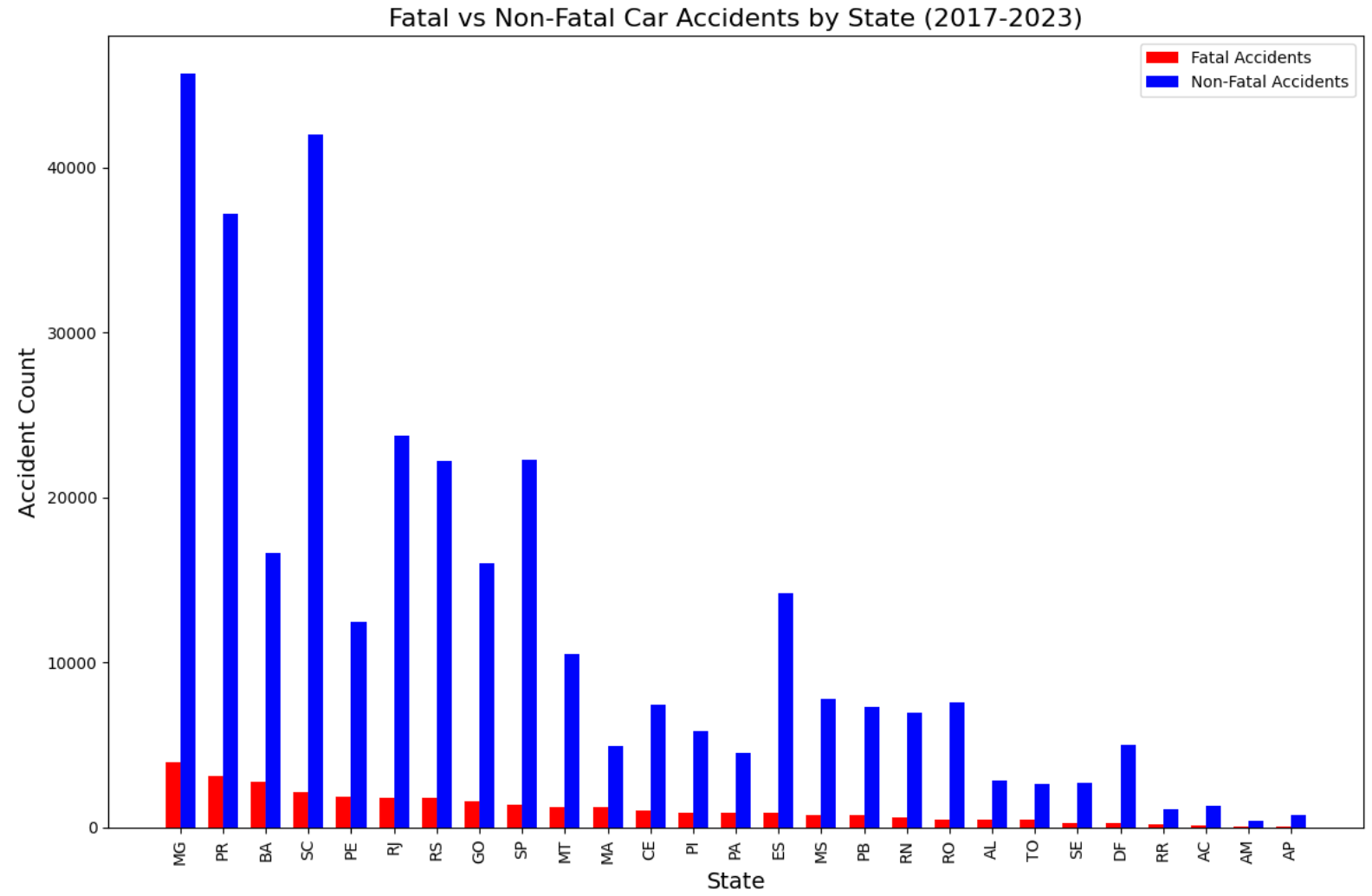
After taking a look at these visualizations, we were able to see which proportions of, for example, a “*driver’s lack of attention to conveyance*” was linked to a certain type of accident like “*Broadside Collision*” or “*Rear-end Collision*”. In the left visual, the darker the red color, the more severe the accident was. A few standouts we saw from this visualization were in a “*driver’s lack of attention to conveyance*” and “*Broadside Collision*”, as well as “*Rear-end Collision*”. These were two of the standout links, as they were the two of the darkest red in the visual, meaning that these accidents were the ones that are the most fatal.

We also made bar graphs showing what states had the highest total accident count per state:



We see that the states with the highest number of total accidents are “MG, SC, PR, RS, and SP”. This alone only shows us our total accidents per state. This may suggest that those states have more people.

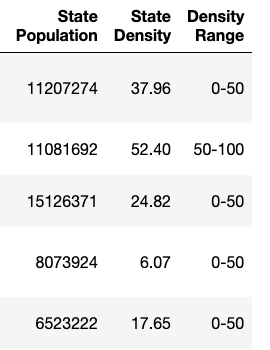
In order to take a deeper look and see the severity of states, we must compare how many of these accidents are fatal vs non-fatal:



We can see that some of the states with higher accident counts have less fatal accidents than some states with lower accident counts. The order goes “MG, PR, BA, SC, and PE”. It goes to show that knowing the accident counts of every state is not enough for us to determine which state is more dangerous in terms of car accidents. However, knowing the fatal vs nonfatal relationship for every state still does not give us all the information we need to know to answer this important question. Perhaps there are geographic and social factors which impact this. We must find more data in order to further examine this question.

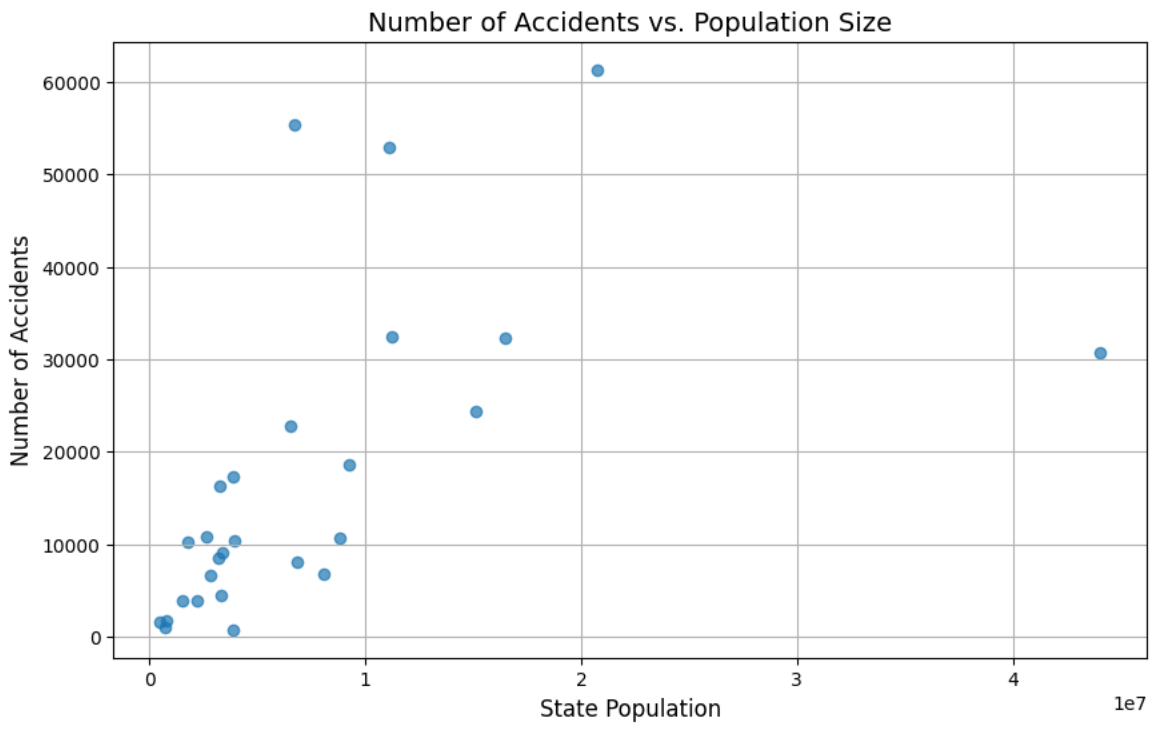
**Population and Density**

Two important data points that can help give us an important understanding behind which areas of Brazil are most dangerous in terms of car accidents is what the population of the state is, along with its density. We went to the website which is shown under the dataset section. It gives us the population size and density of each state, and we integrate those values into our original data.



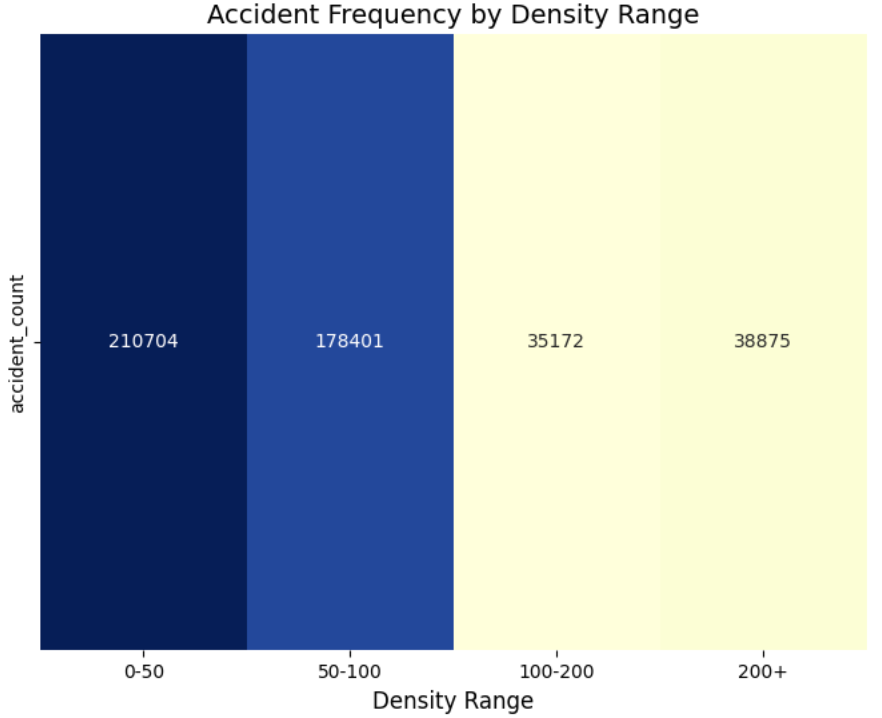
We created a column for state population, state density, and even made another column with density range. After inserting this into our data set we began to study what effects population and density has on car accidents in Brazil. The following visualizations will help in better understanding the relationship between accidents as it pertains to the states density size and population.

*Population vs Total accident count*:

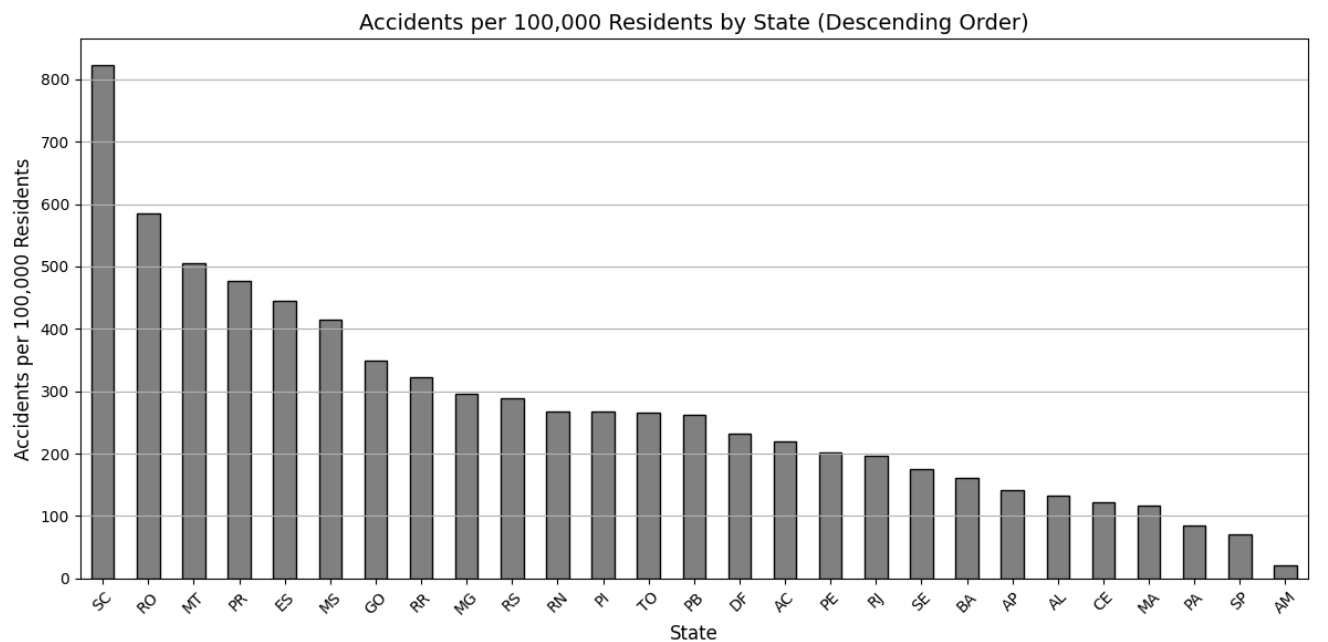


From this visualization, we can see that the lower the population size, the higher the number of total accidents. Many of the states we saw from earlier visualization with high total accident count. We clearly see a pattern, which indicates that a state's population size affects that state having a higher accident count. However, the state with the highest population does not have the highest accident count. Although their accident count is still pretty high. This just goes to show that although knowing a state's population can give you a good idea of what the states accident count will be, it is not the end all be all in getting a great accurate prediction.

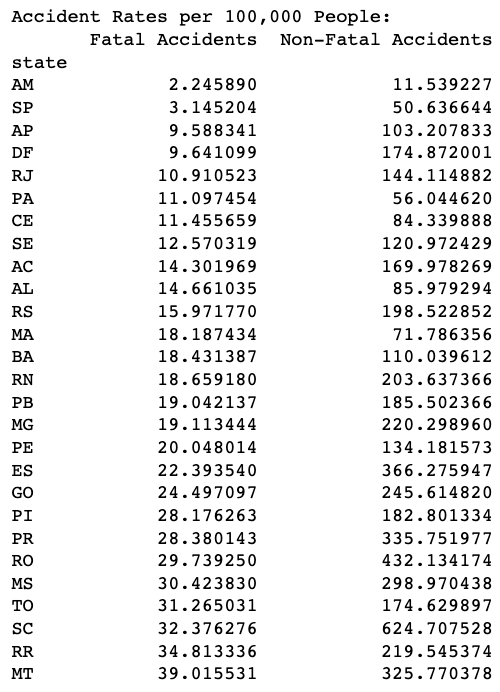
Next we will see what pattern state density has with accident count:



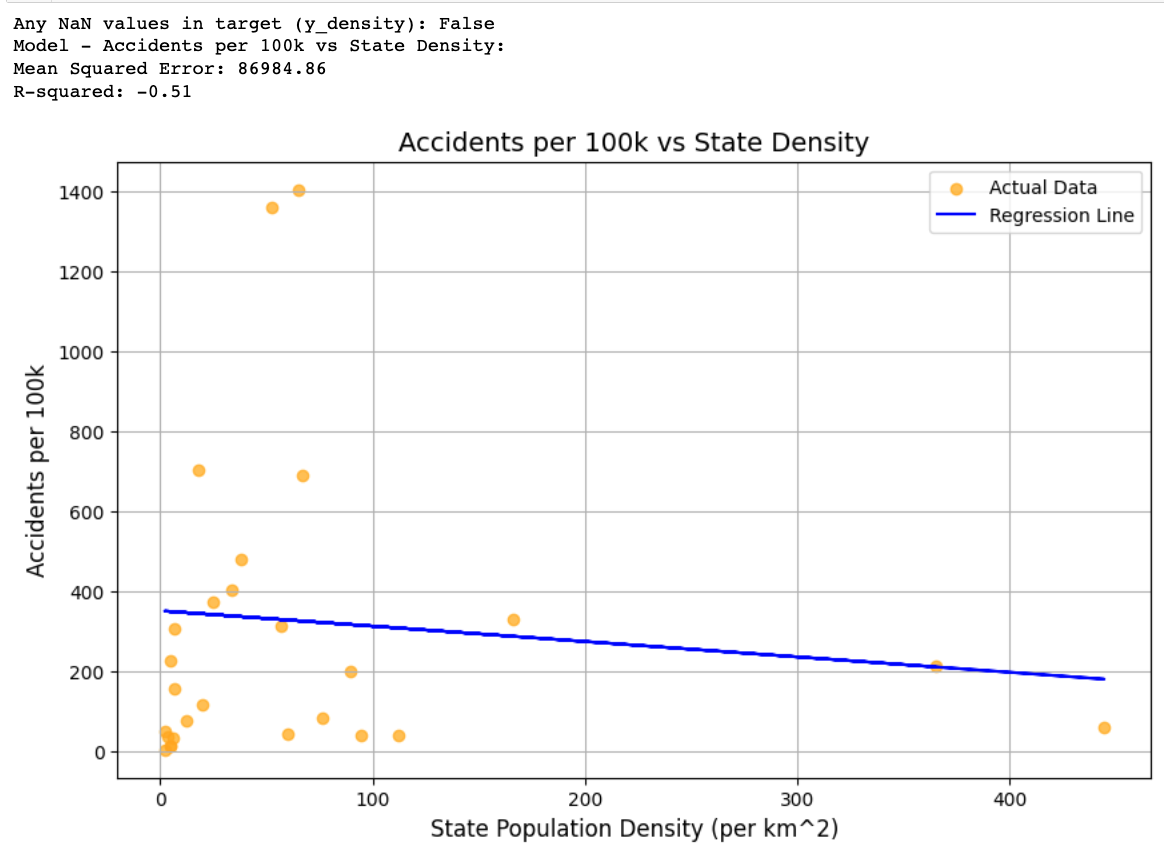
This is a heat map showing us what the accident count looks like based on a state's density. As we can see, the lower the density of a state, the higher the accident count. Similar to population, the density variable also shows us a pattern in the number of car accidents. However, after our density range goes above 100, we don't see a big difference between the amount of accidents. Although we cannot deny that both a state's population along with its density gives us a good indication or a good idea on how dangerous a state is, perhaps it is not enough to just focus on total accidents. Rather we should focus on the frequency of accidents per state. Let's use our population data that we integrated into our data set and see what the accident count per 100,000 people is per state.



Based on this visualization, we now have a better idea of which states are really more dangerous. Earlier we said that states “MG, SC, PR, RS, and SP” are most dangerous in that order based on total accidents. However, based on frequency states “SC, RO, MT, PR, and ES are more dangerous. This means that if we live in these states, we are more likely to experience a car accident than we are if we lived in other states. The only state which appears twice in these visuals is ‘PR’. Through the use of frequency of car accidents, we can now use this to see what relationship state population and state density has to the frequency of car accidents in Brazil. Now let's look at the frequency of fatal vs non-fatal car accidents in Brazil:

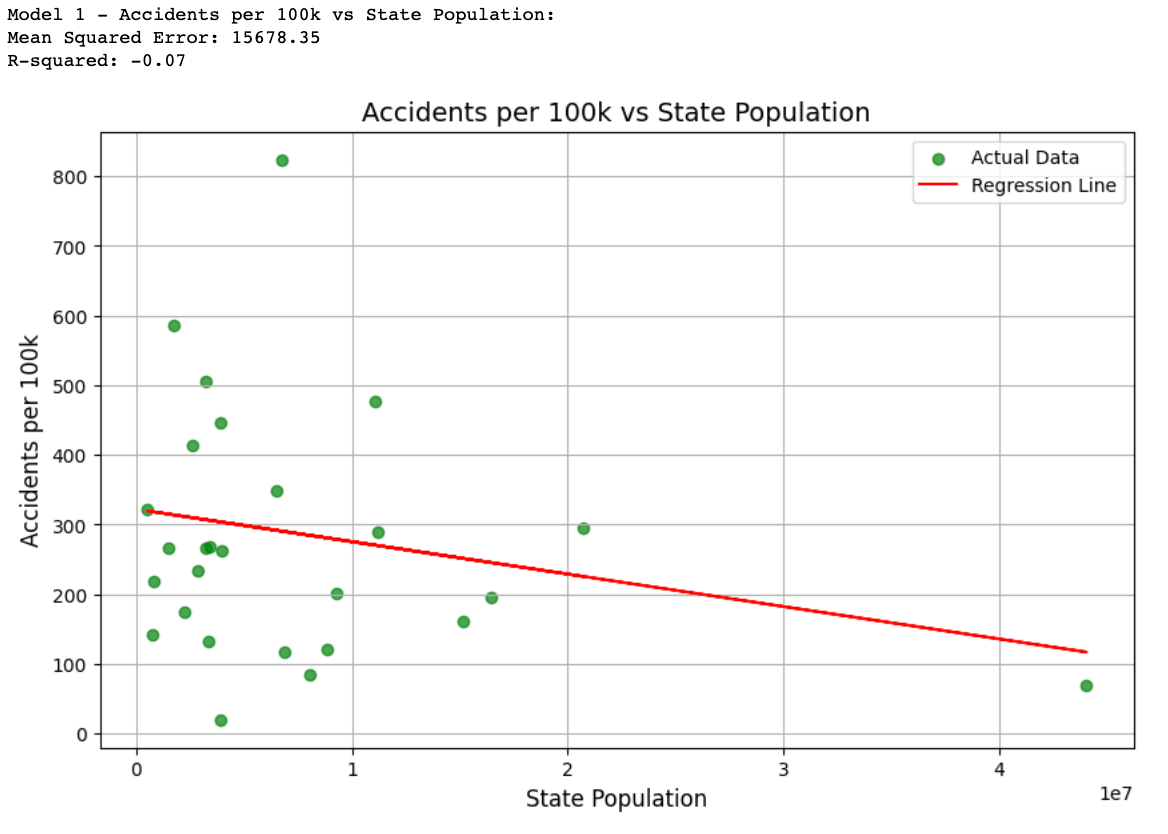


Our data is shown in descending order, so the highest is at the bottom. Our top 5 frequency of fatal accidents are “MT, RR, SC, TO, and MS”. Originally it was “MG, PR, BA, SC, and PE”. This helps us to conclude that the most dangerous state as it pertains to car accidents in Brazil is **SC** since it appears twice in both the frequency of car accidents, frequency of fatal car accidents, total fatal car accidents, and total car accidents.Now let's look into some linear regression models, showing us the relationship between frequency of car accidents as it pertains to state density and population. *Density vs Frequency of car accidents*:



This linear regression model shows us that as density increases, the frequency of accidents decreases. States that have higher density are much safer than states with low density. One idea is that densely populated areas have better infrastructure since the location is taken care of by more people, meaning more jobs, better roads, and possibly even transportation systems other than cars. Now let's see the relationship between population and frequency of accidents.

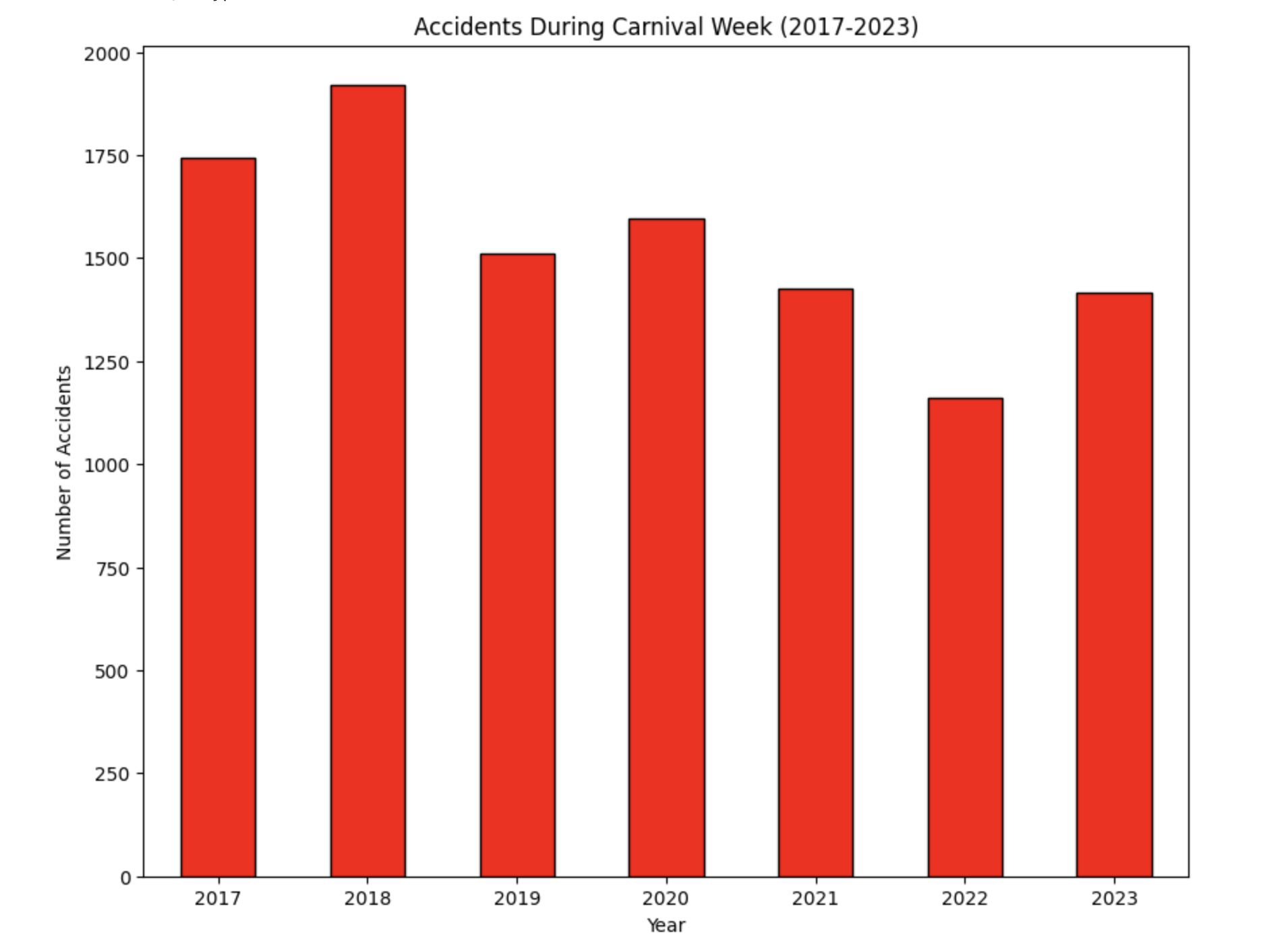
*Population vs Frequency of accidents*:

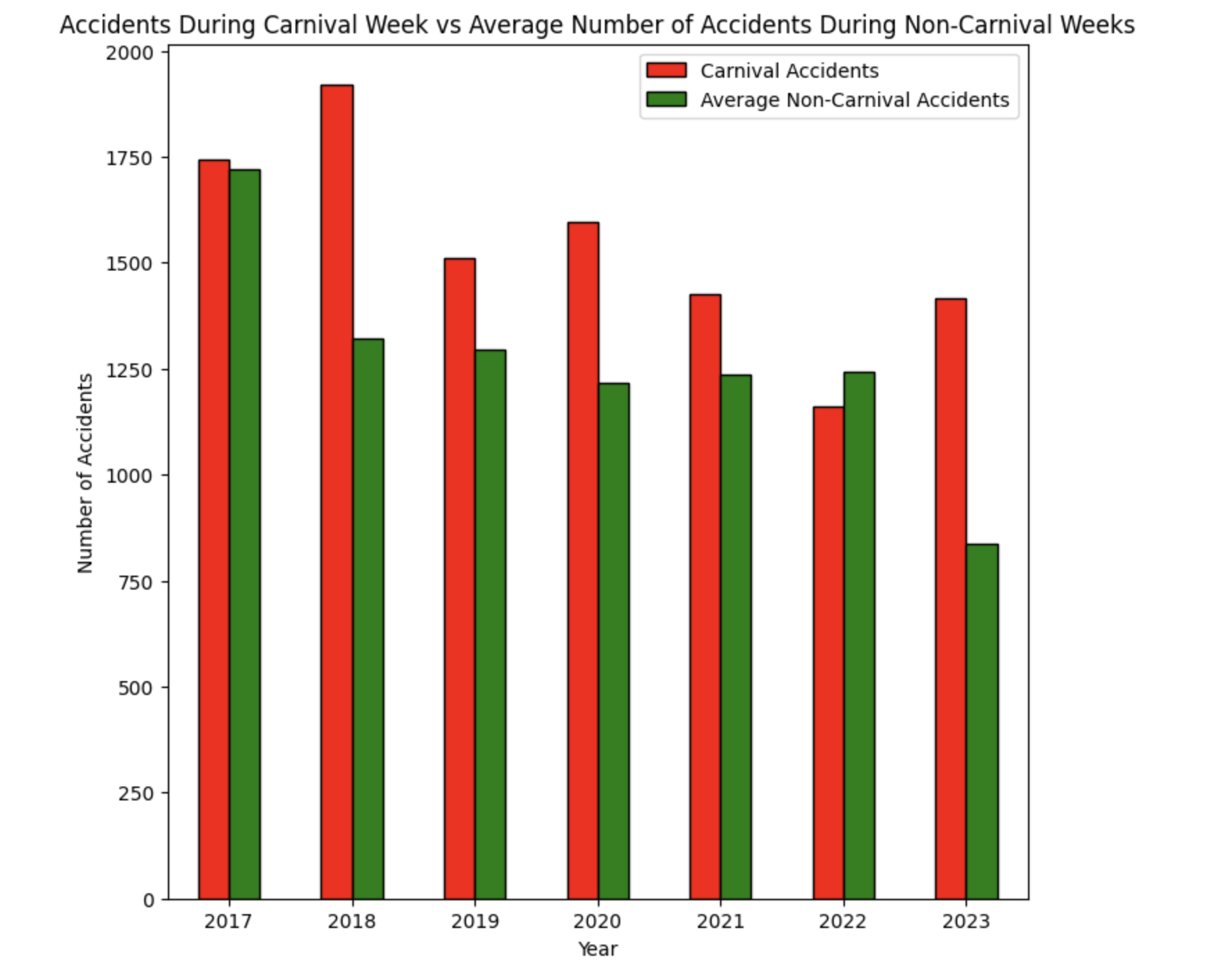
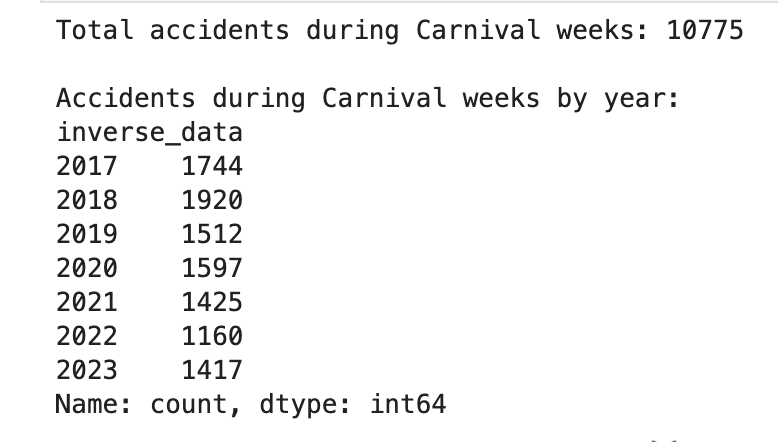


Although initially we can see that states with higher populations have higher total accident counts. We notice something a little different when we measure the number of accidents based on frequency of the accidents. We notice that there is one outlier where there is a high population, but low frequency of accident count. The highest accident count we can see on the graph, is the green dot at the very top, which represents the state ‘SC’. Overall, our result for population is similar to density when it comes down to frequency of accident count. This goes to show that the population has a higher total accident count, but the same is not reflected in the frequency of accidents. This is most likely because among the population count are babies, children, and people who are not in the proper age to drive. If we had the number of cars per state instead of population count, perhaps we would have gathered better and more accurate results.

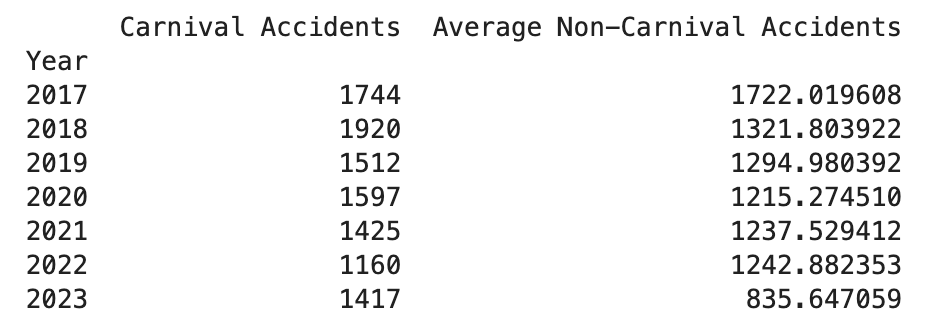
**Carnival**

We were interested in seeing if holidays had any impact on the number of accidents that occur during that time period and see how they compare to other times of the year. Carnival was chosen as the main focus for this topic because it is the most well known holiday in Brazil and attracts the most amount of tourists to the country. Road closures for the huge street parties and an increased police presence may also have an impact as a result of the celebration. Even though Carnival celebrations are not just limited to Brazil, the biggest one occurs every year in Rio de Janeiro starting the Friday before Ash Wednesday and ending on Ash Wednesday. During our research we realized that even though this is listed as the official start date in many news articles, the actual dates of celebrations appear to vary among the cities and states of Brazil. Carnival itself can occur as early as January and as late as March and many people may celebrate the holiday weeks before this and even past these dates. Therefore for the purposes of this project we defined Carnival week as the week ending on Ash Wednesday for each year in our dataset. This way we can have a consistent timeframe for analyzing the data.

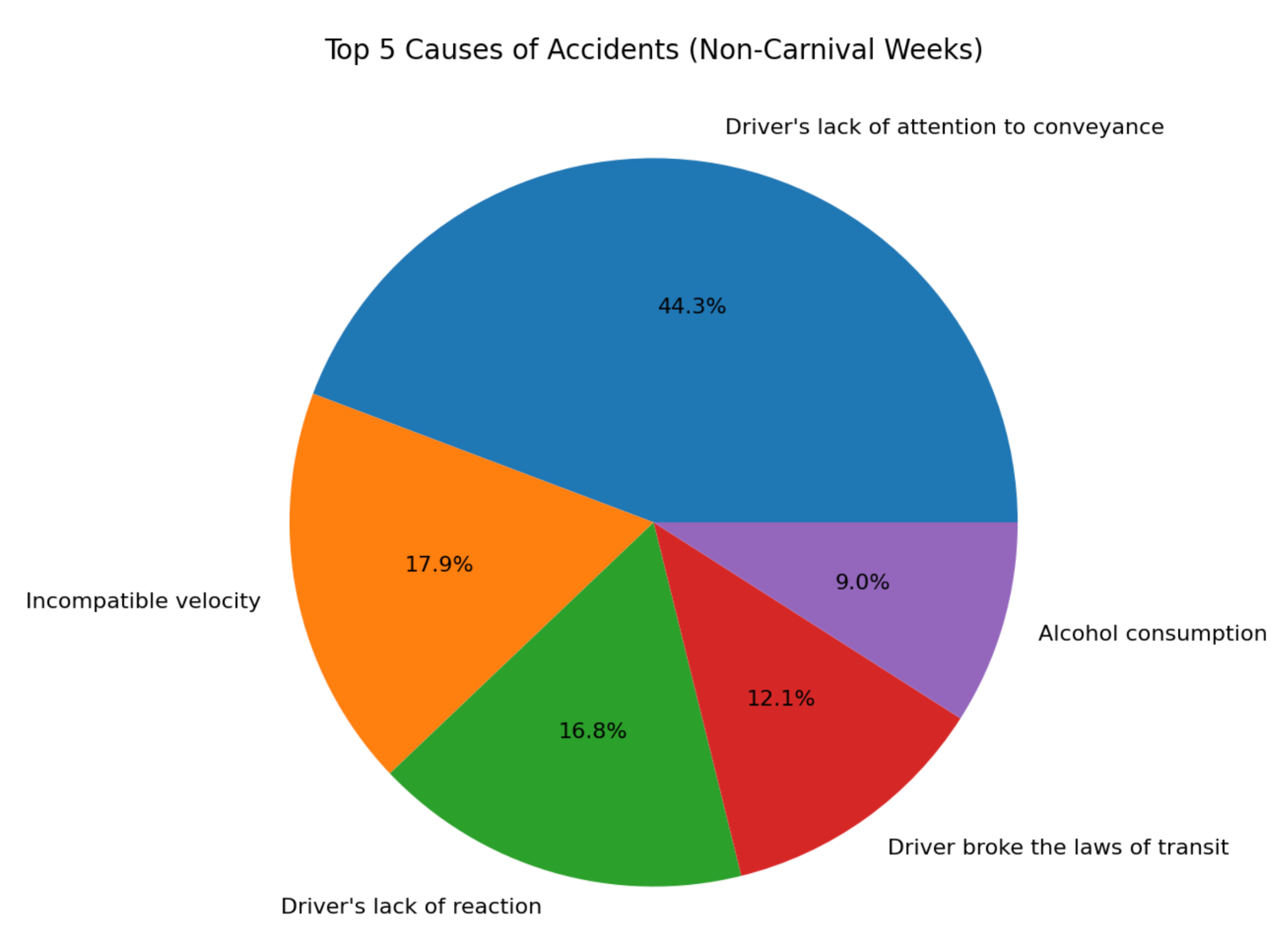
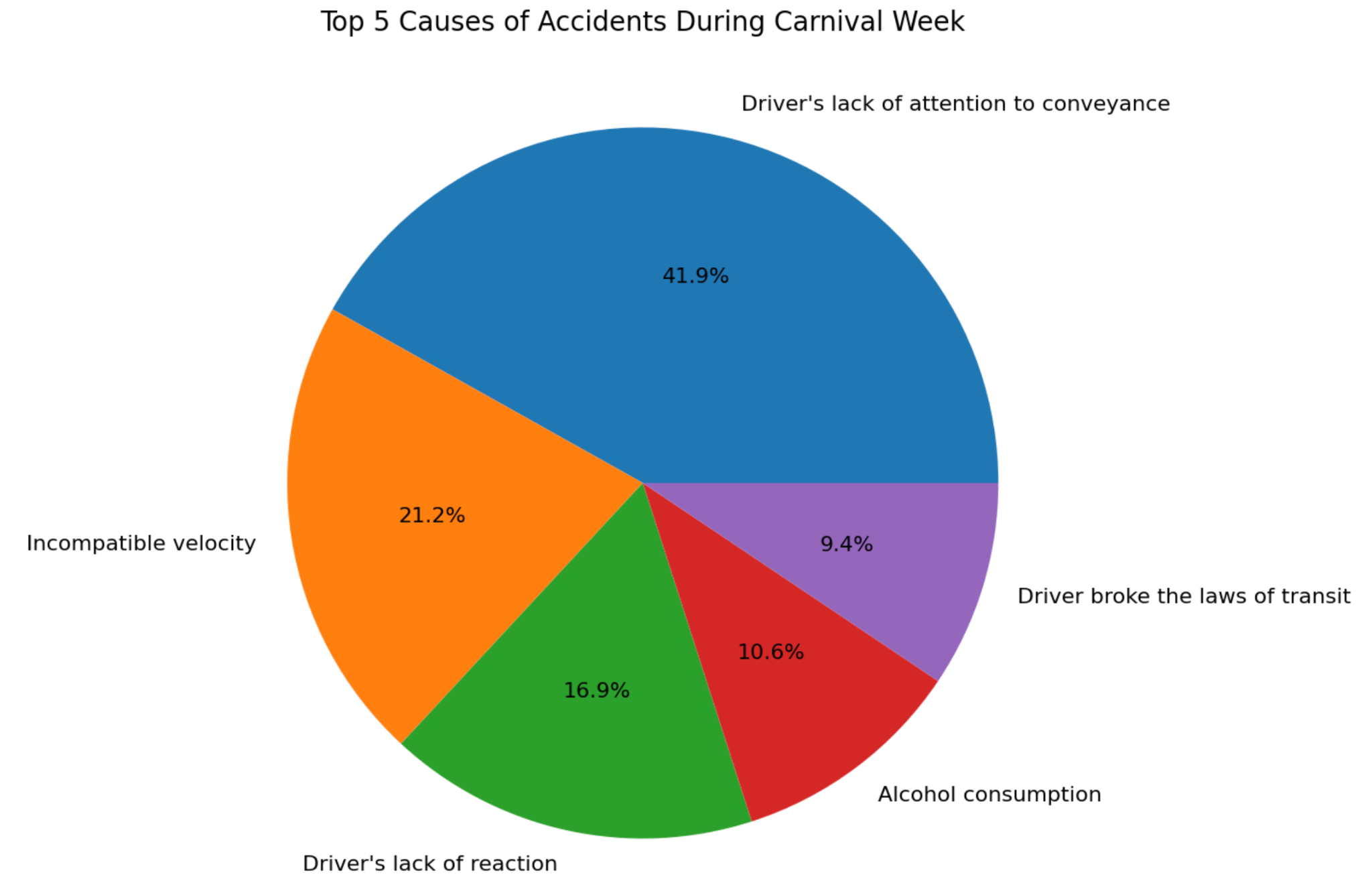


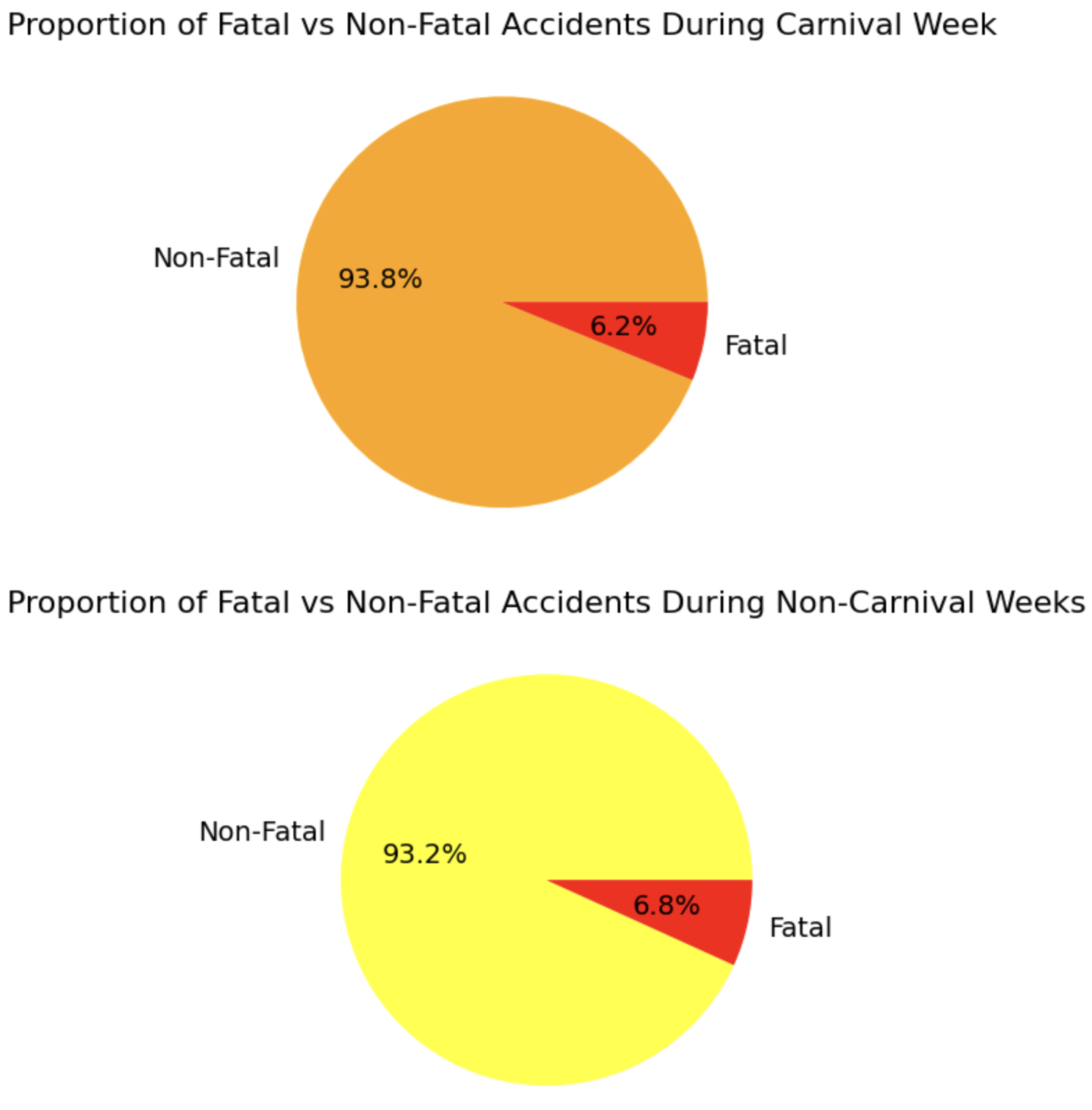
Above is the first visualization we created to see the number of car accidents that occurred during Carnival Week for each year in the dataset. This bar graph doesn’t take into account where the accidents occurred, only when they occurred. This is because with this dataset it was difficult to pin down the exact effect Carnival had on these accidents (i.e. road closures, redirected traffic, number of people out on the streets). One of the things we noticed is that cities like Rio de Janeiro (the city with the largest Carnival celebration) and São Paulo (the most populous city in Brazil) were surprisingly lacking in data points. Rio de Janeiro only had 1391 data points in the entire dataset and once we mapped the data points to a map using the latitude and longitude columns we could see that the big cities were relatively empty. This could be because the dataset only takes into account interstate highways, excluding city data. 1391 data points

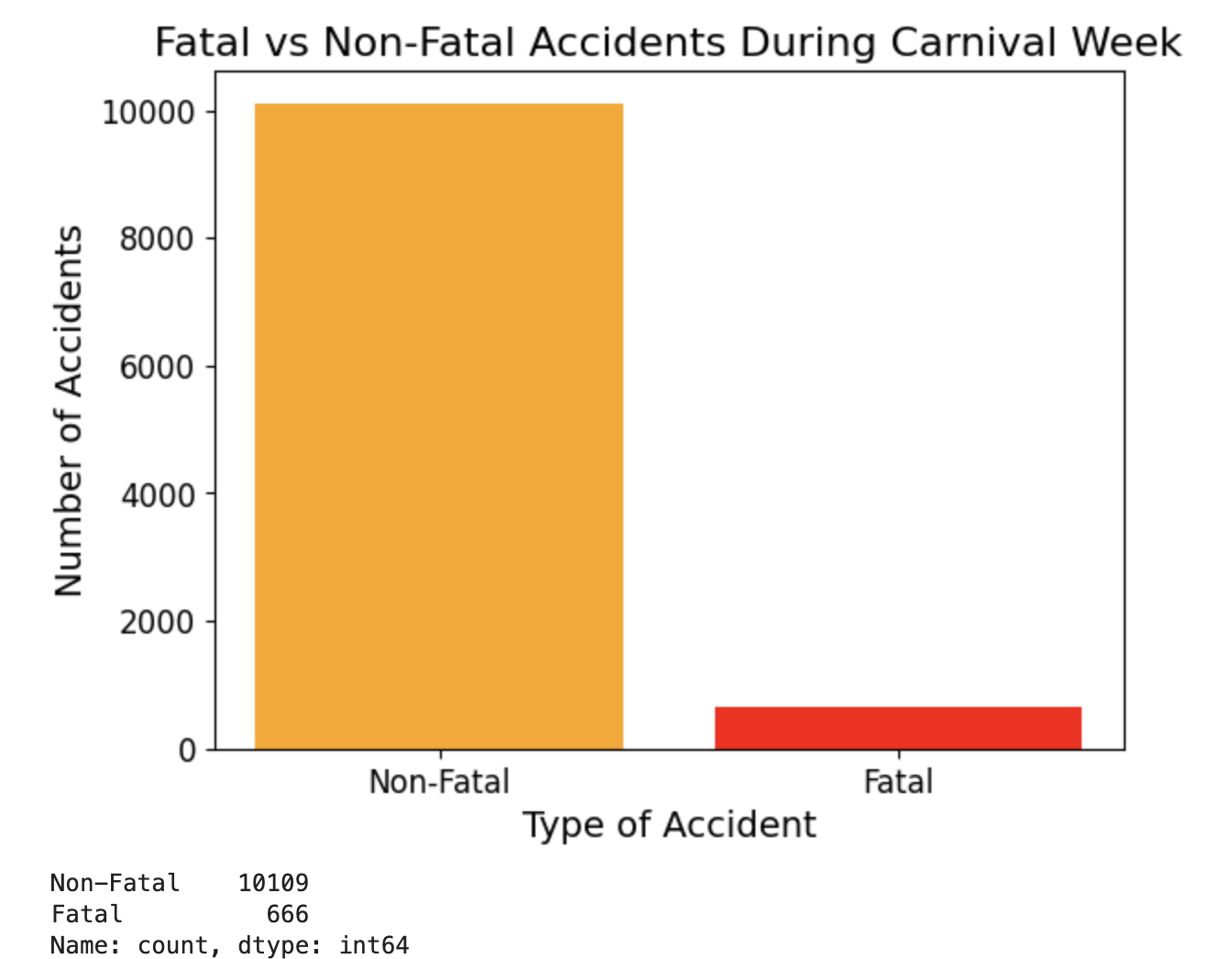
Next we compared the number of car accidents that occurred during the week of Carnival for each year and the average number of weekly accidents that occurred during the rest of the year. This average number was calculated by subtracting the number of accidents that occurred during Carnival week and subtracting it from the total number of accidents that occurred for each year and dividing that number by 51 ( average weekly accidents = total number of accident for a specific year - number of Carnival accidents for that same year / 51 remaining weeks of the year). We can see that for the most part the number of accidents that occur during Carnival and the average number of non-Carnival accidents are comparable, except for 2018 and 2023. It is difficult to pin down why there appears to be a spike in Carnival accidents for 2018 but 2023 is easy to explain because the dataset has no data points after August 2023.



Then we wanted to see if there was any difference in the top causes of accidents for Carnival compared to the rest of the year. Our first instinct was to suspect that alcohol would be a bigger factor in accidents given the celebratory nature. However as you can see below this was not the case, as “Alcohol consumption” and “Drive broke the laws of transit” merely switched places and the percentage of accidents that can be attributed to “alcohol consumption” appeared to have only risen by 1.6% .



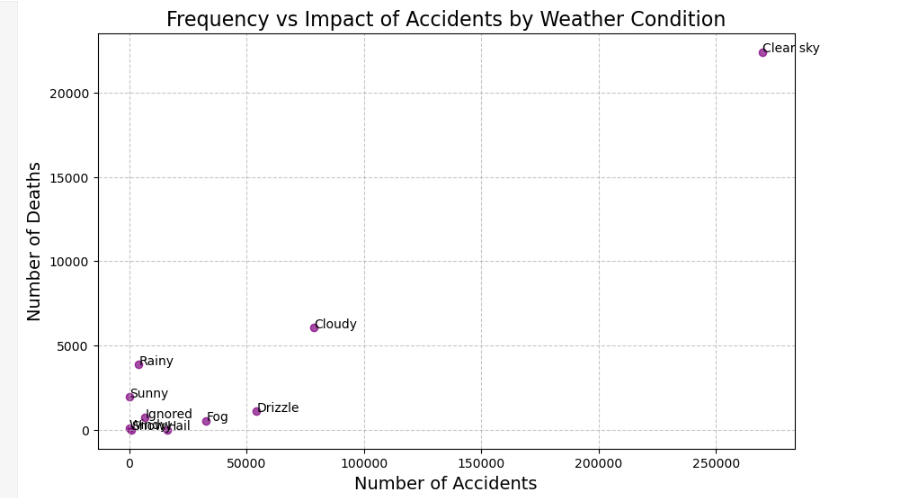




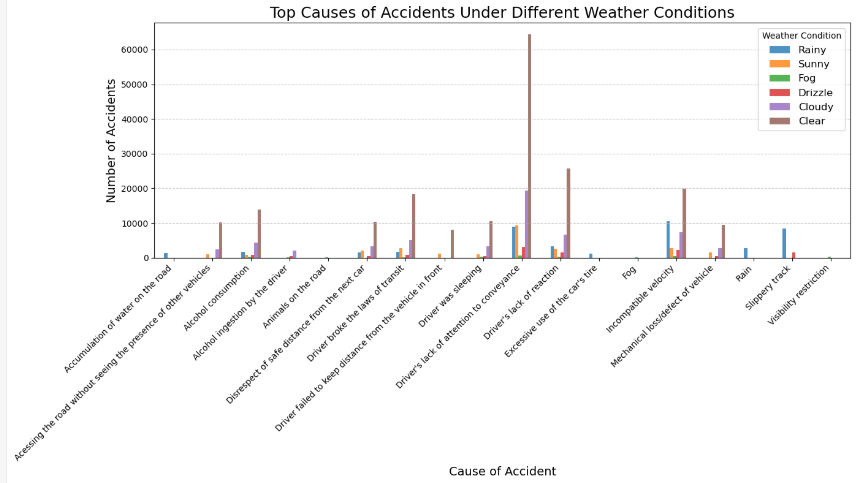
Finally we wanted to compare the number of fatal vs non-fatal accidents between Carnival and non-Carnival weeks. Here is the number of fatal accidents vs non-fatal accidents that happened during all the Carnival weeks combined (10775 data points). We didn’t think a bar graph was necessary to showcase the same comparison for non-Carnival weeks because it would obviously have way more data points and instead it would make more sense to look at the proportion of fatal vs non-fatal accidents for both types of weeks. But you can see there is barely any difference between the proportion of fatal vs non-fatal accidents. So after visualizing all the data for Carnival we can say that it appears to have little impact on the number of accidents that occur. This is surprising and not what we expected to see. Some ideas for what might be going on are that we are missing data for the big cities or perhaps mopeds and public transportation are more preferred methods of getting around. There’s a lot of factors to break down and not enough data to come to a firm conclusion.

**Weather**

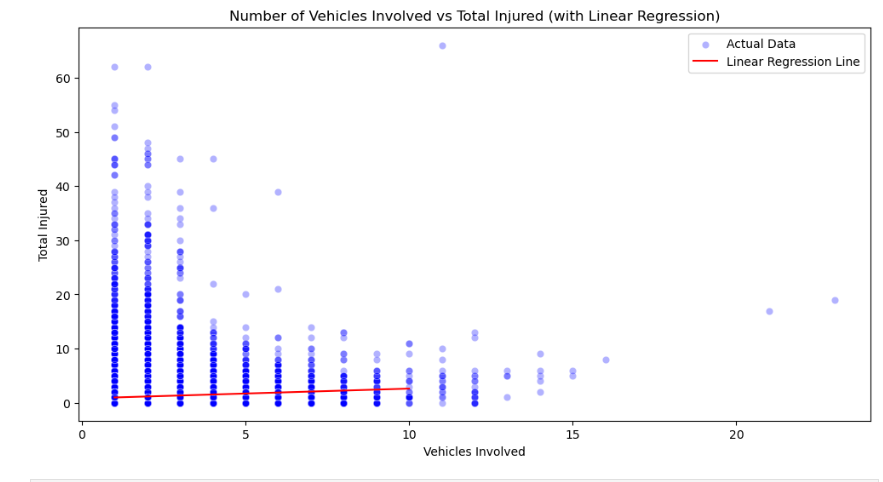
Being the most important part of the environment, weather conditions play a critical role in road safety, and it significantly affects the likelihood and severity of car accidents. We are trying to analyze the relationship between weather and the accidents. Aim to identify high risk conditions and provide insights that can enhance driver awareness and improve road safety.



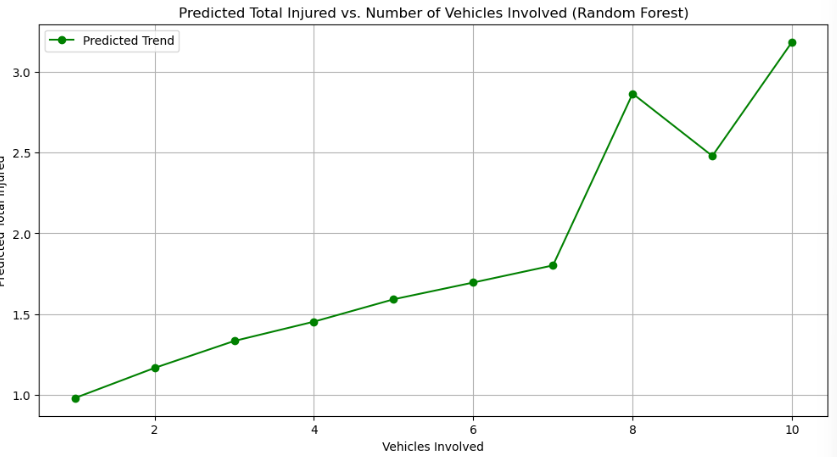
This graph is a basic showing of how many times an accident has happened and when it happens many people are dead and under what weather. In the graph it shows that most accidents have happened during the clear sky. However, this doesn't mean that clear skies are the most dangerous, likely due to more vehicles being on the road. However, it can also mean that most of the time Brazil's weather is clear sky. Cloudy weather has moderate accident rates compared to clear weather, and other weather conditions like rain and snow have fewer accidents but pose significant risks when they occur.



This is an overview of top causes of accidents under different weather conditions. Most accidents that occur during clear skies are likely due to human mistake. Another leading cause is “Excessive use of the car’s brake”. This also indicates that driver error might be the major factor when it is clear sky. Other weathers are different for example the rain and fog and drizzle in the diagram shows that “Slippery roads” and ”Visibility restrictions” are the majority. Although the first graph shows the clear sky has the most accidents, it does not really have a relationship between clear and accidents. The most effective weather is weather like rain and fog. Clear and sunny weather have the highest accident rates because of increased road usage and driver mistakes. Rain, drizzle, and fog highlight weather all mainly because of environmental reasons.

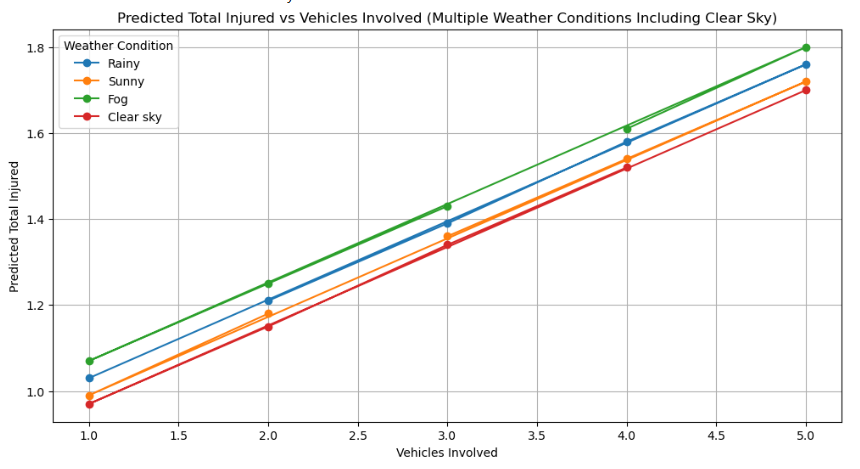


As we invest into using linear regression to predict the relationship between the number of people get injured and the number of vehicles involved in the accidents, we can see that the line has a positive slope. Even though it looks like the linear line is on the bottom left and it didn't really go through the whole map, the main reason is even though there is a plot that is above the linear line, it still counts as a minority when compared to the whole dataset. Most accidents in the dataset didn't have those high numbers of total injuries. After the calculation of prediction, it shows that as the number of vehicles increases, the number of injuries also increase. However, the model also shows its limitations for higher vehicle count because it is also rare to have a huge accident that has a lot of vehicles involved.



With one prediction it is hard to really show the relationship, we are also trying the random forest method on the model. The green line represents the predicted number of total injuries for different numbers of vehicles involved. Different from linear regression, random tree models capture nonlinear patterns, and reflect sudden increases in predictions as the number of vehicles increases above 7. Before 7 vehicles, the line increases slowly, but after 7 there is a huge increase and also decrease.

Both graphs above all show nearly identical MSE values, which might indicate that their average error in predicting total injuries is very similar. However, it might be too early to reach a conclusion. There are still many things that can cause car accidents.

The most important thing is to try to find the relationship between weather and accidents. After two different types of prediction, We also came up with linear regression with One-Hot Encoding. 

This graph shows distinct lines for different weather conditions. Each representing the predicted total injuries for a given number of vehicles involved.

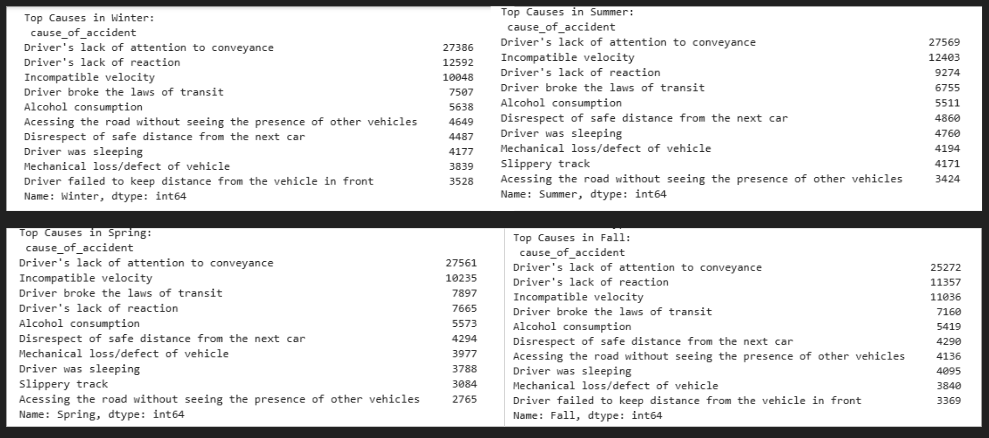
* Rainy weather : consistently predicts the highest injury rates among all weather conditions, shows the increased risks when it is raining.
* Clear sky: at the beginning it was outstanding among other weather, but here it shows lowest predicted injury rates, demonstrating better driving conditions despite high accident numbers.
* Sunnyand foggy conditions: falls in between the clear sky and rainy weather, with fog posing more risky than sunny.

These lines show a clear positive increase, indicating that as the number of vehicles increase, the total number of injuries also increase.  
 =====================================================================

**Data Preparation**

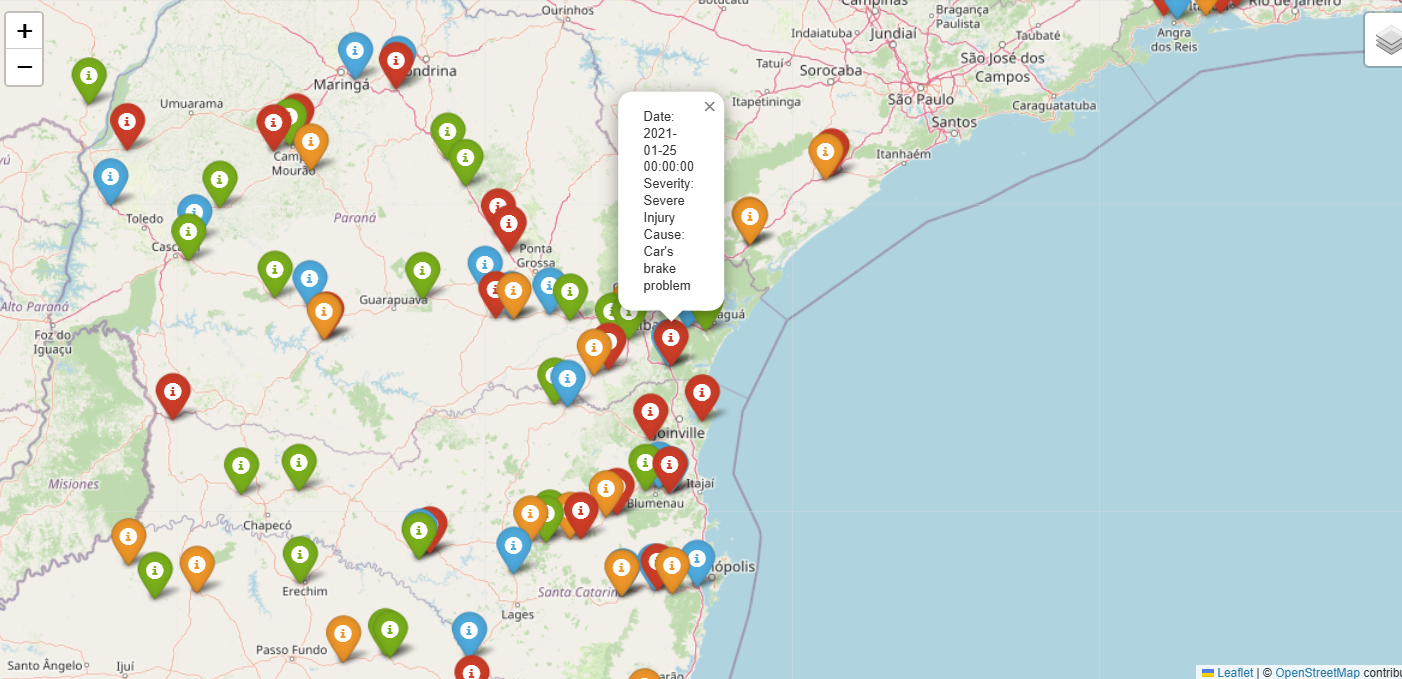
**Feature Engineering**

One of the things we did for data preparation is the use of feature engineering to make a new column for seasons. Since our goal was to find potentially dangerous areas in Brazil for accidents to occur, we were curious as to how seasons would play a role in the accidents. For instance, whether there would be more accidents during a certain season or how the causes and types of accidents might change for a certain season. As there was no season column in the dataset, we had to make a new column for the seasons using the temporal data that was given to us in the dataset like the time and dates of the accidents. In the process of doing so, we had realized/remembered that because Brazil is located in the Southern Hemisphere, their seasons are different from our seasons. In fact, their seasons are essentially the opposite of when we have our seasons. To give an example, our Summer(~June - August) is when Brazil would have their Winter.

After engineering the seasons column, we check the top causes and types of accidents for each season, which is shown in the screenshot below:

Having seen these results, we were not able to draw much information out of this because there is a lack of change in the top causes. For the most part, a majority of the causes are the same and are in the same order, meaning that despite the changes to the roads during each season the causes still remain the same.

In addition to looking at the top causes, we had also decided to map the locations of these accidents for each season in hopes that we can get an idea of where in Brazil is the most dangerous for car accidents. After making the season column, we filtered the accidents to have only accidents that resulted in severe injuries and/or deaths. Then, using the longitude and latitude given to us from the dataset, we successfully made a map with markers at specific locations that would be markers for accidents that were severe or fatal. The markers have four different colors to differentiate between the seasons(blue = winter, red = summer, orange = fall, green = spring). Here is a snippet of the map:



=====================================================================

**Evaluation of Results**

While the map helps us in identifying areas in Brazil that may be dangerous, the map that we have created still leaves us with a lack of an answer. As the markers are pretty evenly spread and not in areas that we expected based on results from our other analysis on population density and fatalities in different states of Brazil. For example, in the snippet of the map above, you can observe that Sao Paulo, one of the major big cities in Brazil, is without markers, implying that while there definitely many car accidents in the city, none of them are severe or fatal. This is, of course, a very surprising find, however, it should be taken with a grain of salt because not all markers were mapped because of rendering issues. Only the top 100 of the severe accidents were mapped due to this issue, and therefore, our conclusion may not be accurate.

There is a clear positive relationship between the number of vehicles involved in accidents and the total injuries. When the vehicle increased, injuries also increased. For predictive modeling, both linear regression and random forest models show similar accuracy, with random forest showing better flexibility in capturing non-linear relationships. However, there are still weaknesses. Linear regression models might oversimplify the relationship and random forest models may require careful tuning to avoid overfitting.

=====================================================================

**Improvements**

* Explore additional models, such as K- means and Naive Baye to improve predictions.
* Adding more features (example: time of day, road type, or reasons) to enhance model performance.
* Integrating more datasets such as number of cars per state, to see if this has a relationship they have with frequency of car accidents.
* Using city features to gather information on car accidents per city.
* Creating more and better folium visualizations of our data and results.

=====================================================================

**Conclusion**

Based on our findings, we can see that higher state population and lower density means there will be a higher number of car accidents. Our predictive model can highlight this, although there will be outliers due to factors we cannot measure. We know that the state **SC**(Santa Caterina) is the most dangerous as it pertains to the frequency of car accidents and showing up high on every car accident feature tested including fatal car accidents. Weather conditions such as fog and rain affected the severity of the car accidents, especially fog. Linear and Random Forest models provide similar predictions, however they fail to fully capture real world complexities such as non-linear relationships. Special events in Brazil do not show much of an impact on accident patterns, considering high amounts of alcohol consumption during that time. Perhaps this can be due to data on this not being readily available.

Our challenge when working on this project was that we had a limited number of data points meaning we had to actively search for more data to integrate in order to better and accurately study car accidents in Brazil. We also felt our work kept overlapping due to this issue. We were able to create predictive models, however they all fail to fully capture real world complexities.