# Visualizing and Analyzing Traffic Accidents: USA's Perspective

Syed Mahbubul Huq
*Msc in Data Science*
*City, University of London*
syed.huq@city.ac.uk

***Abstract*— This paper analyzes and visualizes traffic accidents occurring in 49 states of the USA covering data from 2016 till 2021. In this report, we try to find if there is a relationship between different weather factors and the occurrence of traffic accidents by doing different visualization. Besides, the method of clustering and heatmaps are done to classify and differentiate dense areas of frequent and several fatalities in the occurrence of accidents. We also build a predictive model and take help from visualization to make the model reliable and efficient to predict the severity of the accident in a certain region.**

## I. Problem Statement

Around 1.35 million people are killed every year on roadways [1] and millions of others are fatally injured. In most cases, careless driving is considered a major reason for the accident. Our report goes beyond this and focuses on finding the relationship between traffic accidents with multiple external events and environments which is often neglected. We would also focus on finding a general trend of the occurrence of an accident. Lastly, a prediction model would be made by us which would predict the future severity of accidents based on the given factors. As a result, our report goes in the direction of answering the following:

1. Is there an effect of weather changes on accident occurrence?
2. Which geographical locations are considered the hotspot of accidents and in which parts severity of accidents is the highest?
3. What is the current trend in the occurrence of accidents?
4. How accurately can we predict the severity of an accident with machine learning model?

The chosen dataset [2] for our report covers 49 states of the USA in total. The selected dataset is suitable for answering our questions as it contains exact latitude and longitude information of the occurrence of the accident and also has vast environmental information like precipitation, temperature, and weather condition along with the exact time and date of occurrence. Besides, enough attributes and entities are provided in the dataset which would give us accuracy in visualizing and predicting.

## II. State of the Art

**Task 1:** In the journal [3], authors analyzed the effects and impact of weather on the intensity of vehicles on road. The dataset consists of weather and traffic data of Belgium from 2003 to 2004. They used a heatmap visualization technique after analyzing the map. In their visualization, they used legends and varied color densities to indicate traffic occurrence in a certain location. They concluded that there is a change of intensity in the vehicle on road due to weather and environmental change. This gives us an idea and indication that traffic on the road is changed due to the effect of the external environment. As the vehicle numbers vary, this gives us an indication of the busy nature of the street due to the change of weather. From this, we can assume that traffic accidents would also vary with the variation of weather.

In another paper [4] Leard and Roth directly visualized and analyzed the effect of weather change on traffic accident occurrence. They used colored heatmaps and compared the changes in accidents by change of color in their figures. They compared two maps in the occurrence of accidents, and it was found that there was some reduction in fatalities when the weather was shifted from snow to rainfall in a location and also there was an increase in fatalities with the increase of temperature. The approach that they followed is legitimate and goes totally with my first research question. I would apply a similar technique but in a different visualized map to conclude.

**Task 2:** In the paper [5], the authors explained the importance of heatmap in perfectly visualizing the geographical location and finding the exact location of incident occurrence. They used classification and clustering methods on their heatmap to efficiently process the spatial dataset. The author studies variation color, ranges, and radius in the map in getting an efficient result. I would also use this technique and change the properties of the graph to get more insights and knowledge about the map.

**Task 3:** In the paper [6] the author visualized and analyzed the trend of change accidents over time using time series graph. the graph beautifully represents the growing trend of accident and give details about the trend and occurrence. I would use this idea to establish the trend in my temporal data of the occurrence of accidents. Besides, I would also add and take factors like, day, month, and time into consideration while doing the visualization.

**Task 4:** In the paper [7] the author analyzed and visualized the occurrence of road accidents in 10 years of data for the UK. They also predicted monthly accidents using autoregression. They visualized the result in a time-series line graph and visualized about predicted and original number of accidents. I would use a similar methodology in visualizing my prediction and original data in a graph while predicting the severity of the accident.

## III. PROPERTIES OF THE DATA

The selected dataset [2] has traffic occurrence in 49 US states with data from February 2016 till December 2022. The initial dataset contained 28,45,342 rows and 47 columns. Different state and government entities were used to collect traffic accident occurrence incidents. Besides covering traffic incidents, our dataset also covers information regarding the environment and external factors, like exact latitude and longitude along with precipitation occurrence, weather conditions, exact time and date of the occurrence, and so on.

The major characteristics of our dataset are the characteristics of the entries that we have, all of the entries are suitable for doing spatiotemporal analysis. Temporal analysis is possible because of the availability of information on date and time in our dataset. A general trend can be established with these temporal characteristics. Besides, the exact latitude and longitude along with the County, City, and Zip code of the occurrence of the accident are mentioned in our dataset, this gives a new dimension and accuracy to our data.

After analyzing the initial dataset in Python, it was observed that all of the entries were not suitable for our analysis and report. Missing values along with unusual entries were observed in our data. In python isnull().sum() was used in finding the percentage of missing values of individual columns, it was observed that the column 'Number' which showed the street number of accident occurrences has a good number of values missing, which is 61.29%. The whole column was removed because the information was not that important for our report. After that, entries having any null occurrences were cleared using dropna() feature of pandas. Finally, our main dataset was created which contains 22,07,325 rows and 46 columns.

Feature engineering was carried out in our dataset to extract temporal features like the year, month, day, time, and weekday as separate columns, from the 'Start_Time' and 'End_Time' columns respectively. After outliers are checked for the time occurrence table to find if there are any negative values in the column, some negative values were found, which were then replaced with the mean time value.

In the below bar chart, an overview of some important features for our analysis is presented. We can observe the column 'Severity' which represents different severity of road accidents, with 1 being less severe and 4 being the most severe form of accident. Severity level 2 dominates in our dataset. Observing weather columns like 'Temperature(F)', 'Wind_Chill(F)', 'Humidity(%)', and 'Pressure(in)' we can say there are some outliers that are represented by the small-sided values. To confirm this, we plotted a box plot in python and found in the weather data columns, there were some noticeable outliers. It is decided not to remove the outliers. Outliers in weather data are the abnormal weather conditions that occurred in the timeframe that we are dealing with. Abnormal weather conditions should also be considered while analyzing because I believe this can be a crucial factor in road accident occurrence.

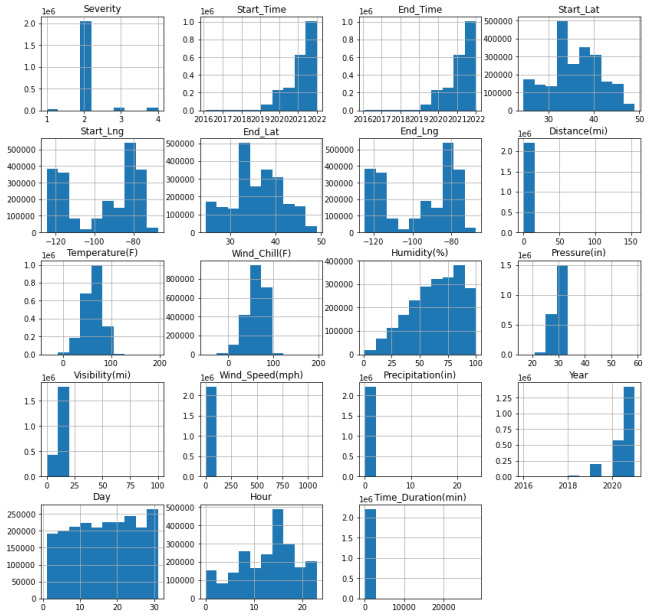


*Figure 1*

## IV. ANALYSIS

### *A. Analysis Approach*

For our analysis and visualization, the main tools that we would follow are python and MS Power BI.

For conducting our research and concluding with the report, we make an analysis plan and would follow it for getting our desired results. The following diagram represents our analysis approach:

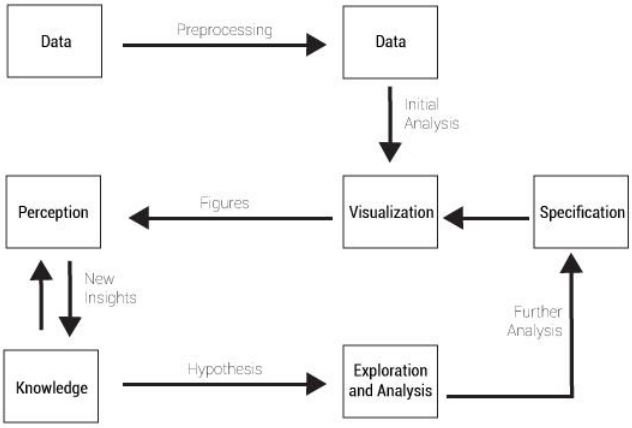


*Figure 2*

Initially, data is collected from our source. After doing a basic analysis of the data, we do data preprocessing and do the steps as described above. In the above step, the works involved are- analyzing data, getting perception, having insights and knowledge, repeating it to get more insights and knowledge, creating a hypothesis, doing EDA, and analyzing further to get specifications. Then the processes of getting perception are again repeated in an ongoing loop which would give us more knowledge, insights, and specification. The loop continues until we are totally satisfied with our visualization.

**Task 1:**
We first find the Humidity in the air percentage value count and then sort the data. Seaborn library of python is used to plot regplot which shows the relationship between the humidity of air and the occurrence of traffic accidents. From the graph, human thinking and knowledge would be used to see the visualization and answer about the relationship and the trend that humidity causes.
For the same task, we use a basic boxplot of seaborn, which calculates the count of accident occurrence based on weather conditions and gives us visualization of the occurrence of traffic accident counts.

**Task 2:**
To find the hotspot of traffic accident occurrence Power BI heat map is used. With dense color in the map, the frequency of occurrence in that particular area can be understood.
To find the severity of the occurrence of an accident in areas, we plot our graph and add visualize the graph, and divide and cluster them based on the severity of the occurrence of an accident.

**Task 3:**
To understand the trend of the occurrence of traffic accidents, a time series line graph is constructed, this would simplify everything and would give us insights into the trend by which traffic accidents are occurring over the years.
We also classify the occurrence of accidents based on months, and hours and show them in a simple boxplot which would help us understand the seasonality, month, and hour trends in which accidents are frequent

**Task 4:**
The K-NN algorithm is set and used by us to build a model which predicts the severity of the occurrence of an accident. For this, important features like location, city, county, time zone, state, temperature, pressure, wind direction, time, etc are considered.
After the prediction of accuracy, the model with training and testing score variation based on the number of selections of neighbors is visualized. This would show us the importance of choosing the correct N values while doing the prediction.

### *B. Analysis Process*

Our report focuses on answering the following question:

1. Is there an effect of weather changes on accident occurrence?
2. Which geographical locations are considered the hotspot of accidents and in which parts severity of accidents is the highest?
3. What is the current trend in the occurrence of accidents?
4. How accurately can we predict the severity of an accident with a machine learning model?

**Task 1:**
There is a close relationship between weather changes and the occurrence of traffic accidents. From the dataset, we decide to go with the humidity parameter and do a scatterplot. An article is also published stating the contribution of humidity in causing traffic accidents [8]. It is stated there that health condition deterioration like dizziness, confusion, etc can be caused to drivers which can lead to severe accidents. Vehicle damage can also occur due to this. This leads me in choosing Humidity as a parameter to understand its contribution to traffic accidents.

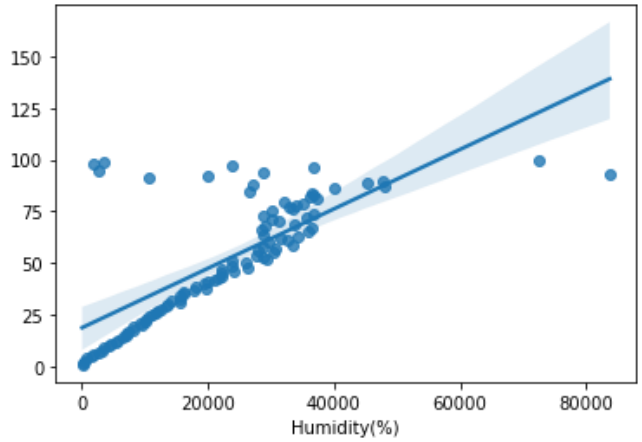


*Figure 3*

From the graph, we can observe that with the rise of humidity in the air, the number of traffic accidents also increased eventually. To understand the trend, we also draw a line, it is understood that the trend is upward and a gradual increase in accident cases is found with the gradual increase in humidity. From this, we can say that traffic accident occurrence is proportional to humidity. Besides, an assumption or idea of the occurrence of traffic accidents based on season can be made. During the summer season, the humidity of the air is the highest, it can be assumed from this discussion that traffic accidents occur frequently during summer. Human critical thinking can be applied to this and proper reasoning can be made by stating the problems we face during summer seasons, like health problems, vehicle problems, etc which is cause of the occurrence of accidents during that period.

It can be assumed that weather condition also has a part in causing a traffic accident. Keeping this consideration in mind, we plot a bar chart that represents the number of occurrences of accidents and ranked the weather condition in order.

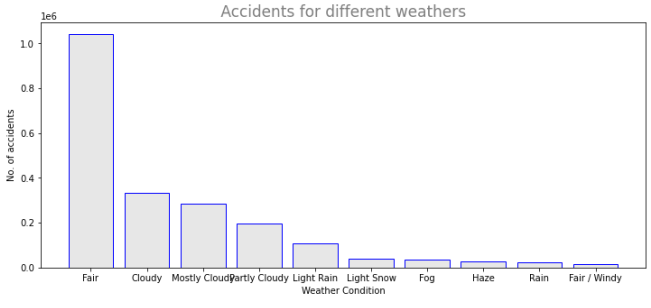


*Figure 4*

From the above graph, it is observed that most of the accidents occur when the weather condition is Fair. Which is a normal phenomenon. But from the general human

assumption, it can be assumed that weather conditions like rainy and snowy would have more impact in causing traffic accidents. As the season are not more in number compared to Fairweather, we cannot get a proper answer to this question. We can relate this with the task 3 question and try to make and come to another result.

**Task 2:**
To understand the geographical hotspot of occurrence of accidents, a heatmap is used and density-based bubbles are made to our map. We can make an assumption and reasoning that traffic accident occurrence would be the highest in urban areas and highways.

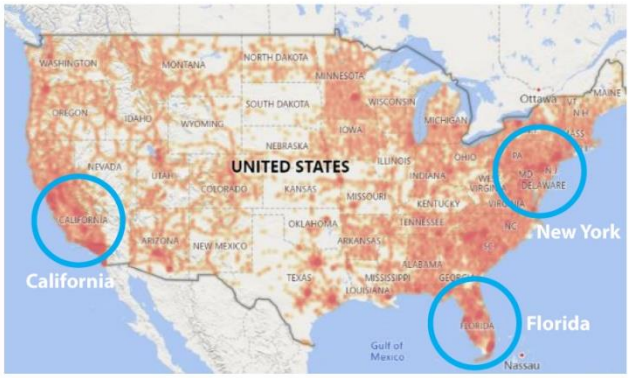


*Figure 5*

In the above graph, we can see the heatmap is dense in California, Florida, and the New York area which suggests that traffic accidents are very frequent in this region. A common pattern can be observed from this, all of the three places mentioned are sea-sided, generally, seaside are tourist spots and people usually gathers in those areas. As a result, a collision or accident occurrence may occur.

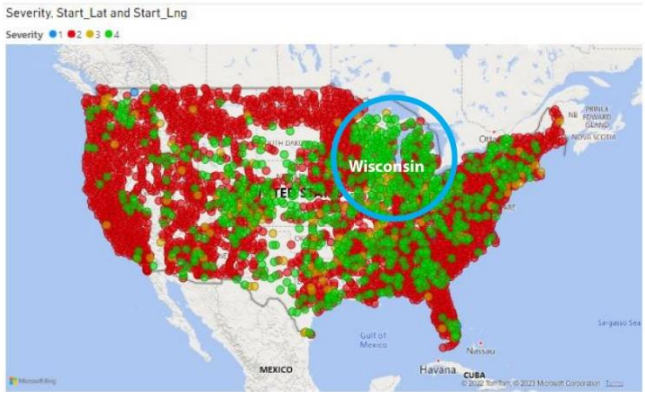


*Figure 6*

In the above figure, we can observe the distribution of accident occurrences in the form of bubbles on the map. Here all four severity levels are illustrated on the map.
The highest severity of accident occurrence is seen all over the state, but a major form of concern is for Wisconsin. Though the number of accidents in that area is not as dense as observed from the previous graph, almost all of the accidents that occurred are deadly. This suggests, the vehicles were at high speed and most of the accidents occurred on highways which are fatal. This suggests the weak infrastructure of traffic signs and signal maintenance in that area.

**Task 3:**
We find the general trend of what month of the year accident occurs by plotting a simple barplot.

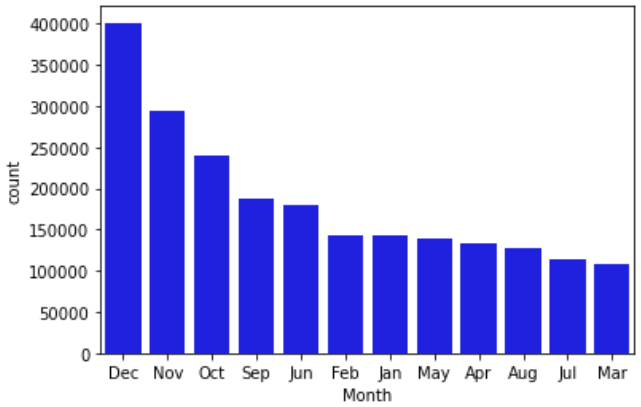


*Figure 7*

In the above graph, we can observe that December, November, and October are the top three months when an accident occurs. Generally, these three months mostly are the holiday and winter seasons. Besides during this period small or heavy snowfall is also experienced which can be a cause of making the last 3 months of the year accident-prone.
Linking this observation to the observation while analyzing task 1, we can add a piece of additional information to that saying that observing the trend of accident occurrence, bad weather condition like dull and snowy weather which is experienced in the last 3 months of the year has contribution in the occurrence of traffic accidents.

It is required to observe the trend of occurrence of accidents according to hours. It would give us insights and knowledge about the occurrence and would give us hints of occurrence.

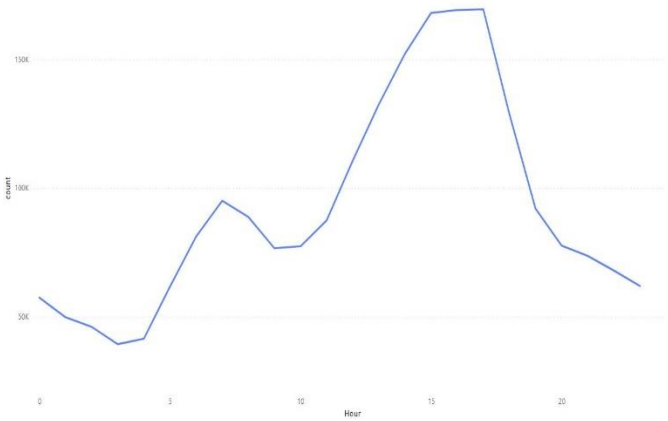


*Figure 8*

In the above graph, we can observe accident occurrence is at the peak in the evening time. That is the busiest time of the day as all offices get closed and a huge rush on the road is

seen which increases the chances of fatality and injuries. Besides, another observation can be made about this hour is on the peak. During this time, the sun sets or the sun is about to set which creates a hint of darkness around, as a result of this there is a tendency of accident occurrence. This factor can be related to environmental changes and their effect on accident factors.

**Task 4:**
Accurately predicting the severity of the occurrence of an accident is very much important. For a certain region, accurate models would predict the severity and would warn us about the consequences that we might face. Steps should be taken according to that.

We would be using the K-NN algorithm to do our prediction. Important features are selected and the model is run to predict the severity of the occurrence of an accident. It was observed changing the value of N neighbors, gave different results. We randomly used the value of N as 2 which gave us an accuracy of around 96%.

We plot the below graph to visualize and understand the accurate and efficient value of N that we should pick.

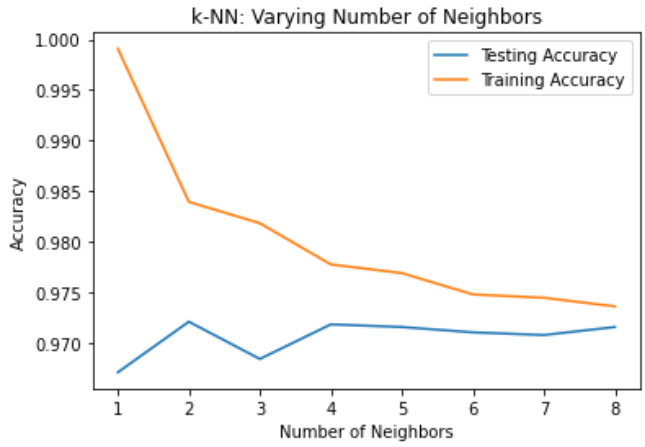


*Figure 9*

It is observed that training and testing accuracy are getting closed which reduces the overfitting and increases the accuracy of the model. From here, we can try with different values of N, the best value from the visualization seemed to be 8. After implementing N=8, we observe a change in accuracy. The accuracy of our model became 97.2% which improved our model accuracy and would help us to perfectly and efficiently predict the severity of an accident occurrence.

### *C. Analysis Result:*

From our observation, we found that there is an effect of weather in causing traffic accidents. It was found that humidity, the summer season, and snowy weather conditions were responsible for causing traffic accidents. Though we were not able to conclude and give this final statement without observing and answering all the questions, as few questions are related to each other. It was also observed the general trend of the occurrence of more accidents was seen mostly during the office ending hours and during the last 3 months of individual years. Most traffic accident-prone areas are California, Florida, and New York. I assume this is so mostly because of their geographical location and because of being modern, urban and busy areas. Surprisingly, Wisconsin does not have too many traffic accident cases compared to others, but still the numbers they have mostly caused several injuries in accidents. I assume, because of the highways and the absence of strict street rules and regulations, traffic accidents are occurring with severe fatalities. While predicting the severity of a fatality, it was observed, the appropriate N neighbor value was ensured by doing proper visualization which gave authentication and efficiency to our model.

## V. CRITICAL REFLECTION

Analyzing the graph which established the relation between humidity and traffic accidents gave an accurate result. Instead of using a traditional scatterplot, a scatterplot showing a linear relationship was plotted on the graph which showed the correlation between the two factors. Though the exact value of the factors was not determined by this graph a region of approx. value was understood.

Afterward plotting a bar chart showing the effect of weather did not show too many surprising results. But there was a limitation in this case, as per my assumption small and heavy snowfall had an effect on traffic accident changes but it was not shown in this case. This is because in the majority of the year, bad temperature weather does not prevail but as long as it prevails there is a certain effect of it in creating traffic accidents. This inspired me to study weather trends and we found weather conditions in the last 3 months of the year, which is snowy time had the greatest number of cases of traffic accidents. I related these two answers and graphs to know and understand the effect of the weather.

Though we were successful in detecting hotspot areas for traffic accidents and severity occurrences we did not study them more in-depth about the area. We could have gotten more insight and details about the locations. We also found a hotspot area for the occurrence of severe accidents but, we did not focus on other minor areas, there were many minor clustering and areas which were neglected by us. We also should have done clustering based on accident cases in areas which would have allowed us to divide the areas into sub-areas and zones. Various methods could have been used in this cased to determine clusters based on location and then divide the total area based on certain characteristics.

We only tested our dataset in a single machine-learning model. Though changing the neighboring values by seeing visualizations gave us good accuracy for K-NN, still, we could have explored more models and should have tried to understand why the score varied. More models with results and accuracy plotted in graphs would have given more opportunities to visually compare the models and get even better results or examine and know why the results are different from each other.

I worked with basic analysis, which would be better if multiple other features were grouped and then analyzed. In the future, a more detailed and broad analysis should be done. Like I found in which hour of the day accident happens the most, it should be expanded and should be linked and grouped with other factors like the day. Other important weather factors like precipitation, air pressure etc. should also be taken into consideration. While doing further research, the analysis should be done on multiple dimensions and the relationship of features together rather than considering one or two.

## VI. REFERENCES


[1] Control and Prevention, Dec. 14, 2020. https://www.cdc.gov/injury/features/global-road-safety/index.html#:~:text=Whether%20you

[2] "US Accidents (2016 - 2021)," *www.kaggle.com*. https://www.kaggle.com/datasets/sobhanmoosavi/us-accidents

[3] ]M. Cools, E. Moons, and G. Wets, "Assessing the Impact of Weather on Traffic Intensity," Weather, Climate, and Society, vol. 2, no. 1, pp. 60–68, Jan. 2010, doi: 10.1175/2009wcas1014.1.

[4] "Weather, Traffic Accidents, and Climate Change B e n j a m i n L e a r d a n d K e v i n R o t h." Accessed: Jan. 10, 2023. [Online]. Available: https://media.rff.org/archive/files/sharepoint/WorkImages/Download/RFF-DP-15-19.pdf

[5] https://www.degruyter.com/document/doi/10.1515/geo-2018-0029/html?lang=en

[6] https://www.researchgate.net/publication/341665300_Towards_Big_Data_Analytics_and_Mining_for_UK_Traffic_Accident_Analysis_Visualization_Prediction

[7] A. Tyagi, A. Kumar, A. Gandhi, and K. Mueller, "Road Accidents in the UK (Analysis and Visualization)." [Online]. Available: https://arxiv.org/pdf/1908.02122.pdf

[8] M. Brady, "How Heat, Humidity, and Summer Weather Can Lead to Car Accidents | Green Bay Car Accident Attorneys," Herrling Clark Law Firm, Jul. 21, 2021. https://herrlingclark.com/how-heat-humidity-and-summer-weather-can-lead-to-car-accidents/ (accessed Jan. 10, 2023).


| Paragraph | Word Count |
|---|---|
| Problem Statement | 240 |
| State of the art | 483 |
| Properties of data | 488 |
| Analysis Approach | 471 |
| Analysis Process | 1145 |
| Analysis Result | 197 |
| Critical Reflection | 483 |
| **Total** | 3507 |