Data Visualisations Project on US Road Accidents

Data Visualisation (STAT3011)

Jeevanpreet Singh Elizabeth Stinton

20153780 19136033

Background

The increasing usage of personal automobiles in the United States has led to at least 10 million purchases of vehicles a year since 1976 ([www.statista.com](http://www.statista.com), n.d.). As a consequence, vehicle accidents and fatalities have risen: In 2019 alone, 33 000 vehicle fatalities were recorded with the fatality rate ranging from 3.3 up to a startling 25.4 deaths per 100 000 people (www.iihs.ord, n.d.).

In the effort to lower these traffic fatalities and accidents, informative statistics play a critical role in informing prevention-aimed policy and decisions by transport regulators. These statistics often require the gathering of both locational and situational data to clarify the conditions of car crashes. In particular, geospatial data has paved the way in distinguishing how indeterminate factors, both environmental and social, may impact the rate of accidents in specific regions at specific times. This has allowed scientists and experts to mitigate traffic accidents with increasing precision (Ghadawala, 2019).

Our project aims to better leverage geospatial data (alongside other relevant data) through data visualisations. As modern decision-making is heavily reliant on data (Berinato, 2016), optimal visualisations are critical for supporting regulators in making effective and appropriate decisions. This is especially relevant as government and regulatory bodies begin to investigate how the COVID-19 pandemic has impacted driving and traffic behaviour.

Problem Statement

Our project will utilise the ‘US-Accidents’ dataset (2016-2019) to generate insights on variables that have significant impact on traffic accident possibility and occurrence, including an emphasis on geospatial and/or environmental variables.

In addition, the US-Accidents dataset will be used in conjunction with another dataset that includes relevant traffic data such as traffic violations. This will assist in determining the impact of socially-driven variables of driving (poor driving, laws that disincentivise reckless driving, etc.).

Furthermore, the recent addition of 2020 data in the US-Accidents dataset will allow an examination of how driving behaviour may have changed during the pandemic. This may help inform accident prevention policy and decision-making going forwards as the impact of COVID-19 continues to evolve.

In summary, our project will aim to answer three questions:

1. How does geospatial data relate to traffic accident occurrence?
2. How does relevant social and environmental data relate to traffic accident occurrence?
3. Has driving behaviour and accident occurrence significantly changed with the onset of the pandemic?

These questions will be answered through a variety of informative data visualisations and summarised in a dashboard, enabling a comprehensive assessment of traffic accidents in the US for transport regulatory and government bodies.

Data Sources

The US-Accidents dataset originates from a Cornell University Machine Learning research paper and consists of traffic accident data spanning 2016 to 2020 (Moosavi et al., 2019). The purpose of this dataset was to succeed the sparse, limited, and lacklustre accident datasets that preceded it.

The second dataset originates from an open-source domain and consists of traffic violations recorded by law enforcement departments from 2012 to 2016. The purpose of this data is to comprehensively collate traffic violations for regulators to implement stricter laws in road networks (Gutierrez, 2017).

Finally, a Census dataset will be used to determine population totals of each US state and assist in determining accident rate, as opposed to relying on total count which is heavily influenced on the populations of states in the US (Vikas, 2018).

Literature Review

Prior research on traffic accident analysis and reduction ‘have used small-scale datasets with limited coverage’ (Moosavi et al., 2019) or relied on outdated larger or incomplete datasets. The dataset we have chosen (US-Accidents) addresses these issues, providing millions of points of data for traffic accidents nation-wide. This has provided insight into various traffic behaviour, including:

* the distribution of accidents across periods of time,
* the distribution of the location of accidents, and
* the distribution of the road-type where accidents occurred.

Specifically, the authors of the US-Accidents dataset have leveraged data visualisations to provide these insights across 3 years (2016-2019). Our project will utilise further datasets to build upon these data visualisations and provide further insight into the 1st and 2nd project questions.

As stated previously, the US-Accidents dataset has been recently updated with 2020 data for traffic accidents nationwide. This has provided the opportunity to determine how the pandemic has affected traffic accidents in the US compared to prior years (3rd project question).

Prior research by the United States Department of Transportation’s National Highway Safety Administration (NHTSA) in 2020 has showed an increase in fatal traffic accidents throughout the pandemic alongside a significant shift in driving behaviour (www.nhtsa.gov, n.d.).

However, NHTSA’s research has focused exclusively on the impact of the pandemic on traffic fatalities: In their 2020 fatality projections report, it highlighted the percentage change of fatalities in 2020 as compared to 2019 and the percentage change from each quarter to its corresponding quarter in the previous year (see below).

Diagram

Description automatically generated

Chart, application

Description automatically generated

The US-Accidents dataset will allow our project to supplement NHTSA’s insight into traffic fatalities during the pandemic with further insight into non-fatal traffic accidents.

Furthermore, in NHTSA’s report, ‘Update to Special Reports on Traffic Safety During the COVID-19 Public Health Emergency: Fourth Quarter Data’, periods of lock-down across the US were examined in terms of their relation to traffic behaviour and fatalities. Information in this report on lockdowns will enable us to investigate how non-fatal traffic accidents and behaviour were affected throughout periods of lockdown across the US.

Methodology

The methodology of this project is subdivided into three separate components to specifically answer each individual project question.

1. Firstly, R Studio and packages including tidyverse will be used to generate visualisations highlighting the most significant geospatial conditions of accidents. Proposed visualisations include different types of map plots and bar plots. These visualisations adhere to the objectives being completed, utilising multiple form of visualisation based on the nature of the variables within the dataset.
2. Secondly, data remediation will be applied to our two datasets and comparative R visualisations will be used to strategically identify any relationships between both datasets. The violations dataset, although not overlapping with the US-Accidents dataset (2012 - 2016 as compared to 2016 - 2020), will provide sufficient insights into traffic-related violations and the encompassing insights it provides in regards to the social circumstances of poor driving. Proposed visualisations include histograms or dot plots and mapping plots.
3. Thirdly, visualisations will be used to highlight the effect of the pandemic on traffic accidents.

Proposed visualisations include grouped bar plot or map plots.

Objective 1: Generating Insights on Accident Dataset

Rationale and Visualisation Choices

Before attempting analysis on the accident dataset, it’s first important to consider the nature of variables to distinguish the types of visualisations required to fully observe patterns within the data and any significant points of interest. The main objective is to observe highly identifiable trends in accidents across the US and the environmental and/or locational circumstantial factors that influence the rate of accidents occurring. The dataset includes 1 500 000 observations across 45 variables. Within these variables are categorical and numerical factors that are assessed by a multitude of streaming traffic APIs that dictate a variety of circumstances that are present during the accident occurrence. It’s first important to consider which plots should be utilised for specific variables within the dataset.

Time variables will be assessed through bar plots and facet grids to generate insight of multiple time variables in regards to accident occurrence. Within the date format, the yearly, monthly, weekly and daily values will be considered to fully understand trends that are exhibited with the time of accidents. Locational variables will be initially explored through bar plots to identify hotspots and locations in which most accidents occur. The census population dataset (REFERENCE THIS) will be used to determine the rate of accidents, acknowledging the influence of densely-populated areas on proportionately higher accident counts, and thus further geo spatial data will be used in comparison with accident count to see the locations with highest accident rate. When identified, these locations will undergo pie chart analysis to identify the most significant buildings and/or road structures that impact rate of accidents.

Environmental stimuli will be visualised through bar plots and interactive scatter plots to identify the relationships between certain environmental conditions and the impact of each variable on accident occurrence. This will have to be done on a subset of the data considering that 1 500 000 would provide an inefficient analysis. Relationships are most significant as they identify the cumulative degree of influence that the environment has on accidents occurring. Boolean variables such as road structure and amenity presence will be visualised through labelled dot plots to identify the amount of accidents happening at these places, placing emphasis on whether these factors impel drivers to be affected by their surroundings.

Data Remediation

The pre-processing of data is an important step in ensuring that missing values are omitted from the data while ensuring the mutation of all character values to effectively plot categorical variables where colour can become an aesthetic that can be altered according to levels.

Timeline

Description automatically generated

Figure 1.1

Figure 1.1, an initial plotting of missing values implored the removal of the Number column, a unique ID for each accident. This is unimportant for analysis, thus omitted. Other columns were also removed, taking into consideration that the only variables of interest are date-time, locational and environmental stimuli. Moreover, throughout the analysis, functions were implemented to omit NA values where necessary, in which entire rows did not require removal due to single missing values but rather were only called and removed during the function. Piping the dataset allowed for this, simplifying the process of visualisation without consistently regenerating values that were omitted in the data.

Progressing through the data, it was evident to split the date time format column in Start Time to it’s time components. This was done with a sub string function, in which all values (YYYY-MM-DD HH-MM-SS) were split and placed in separate columns. A factor conversion was then applied to ensure all variables within the dataset remained either numeric or factor.

Time Visualisations

Having formatted and converted the date time values into separate variables, it is now possible to analyse trends of accidents according to when they occur. This is especially important when determining which periods of driving cause the highest occurrence of accidents and whether precautions can be imposed based on these insights.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 1.1 Figure 1.2

The bar plots above provide insightful comparisons between yearly and monthly accident trends throughout the four year range of the dataset. Figure 1.1 exhibits a significant increase of accidents, adding to more accidents than all previous years accumulated together. This provides concerns regarding the social climate and stigma of driving under pretences of imposed social restrictions and limited travel, an observation that will be explored in objective 3. Figure 1.2 demonstrates that most accidents occur during December, an unsurprising outcome considering a large amount of travelling takes place during holiday periods. As it appears, accidents increase steadily and reach a drop-off at July where lowest accidents occur at around 50 000, but then increases rapidly until the aforementioned peak at December with over 250 000 accidents.

A picture containing graphical user interface

Description automatically generated

Figure 1.3

Figure 1.3 verifies the alarming range of accident values between July and December. Having previously seen the large amount of accidents in 2020, it’s obvious that the data is a reflection of accidents during 2020, in which July and August rates were extremely low due to travel restrictions and possible unavailability of data during that period. Previous years follow however a slow increase towards December, with a low point occurring during the midyear. Having accounted for seasonal trends, we can now find observations for daily entities.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 1.4 Figure 1.5

Figure 1.4 demonstrates that accident occurrence follows a weekly trend, in which a the values within a 7 day range correspond to its adjacence periods, accumulating to a set of 4 identical patterns of movement during the month, with an evident decrease on the 31st considering most months only consist of 30 days. Figure 1.5 verifies this claim, indicating that most accidents occur between Monday and Friday while weekends consist of merely half of any workday value. This follows the logic that travelling due to work influences rate of accident rather than weekend travel, though it would be compelling to identify the influence of intoxication or socially-driven factors on the remaining observations on the weekend.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure 1.6 Figure 1.7

Figure 1.6 demonstrates hourly trends on accident rates and justifies notions of accidents occurring due to work travel; in which most accidents occur during 8am-10am (travelling to work) and 3pm-5pm (evening rush home). It corresponds that afternoon values are quite low, as well as periods before sunrise and hours approaching midnight. As such, there are many revelations to be considered here, in which alternative options of travel may be discussed to decrease accident occurrence during these peak times. Figure 1.7, a facet grid facetted upon weekdays, confirms that every day of the week follows a similar trend in regards to accident totals. Let’s now consider how the time of accidents may correspond to the severity of the accident occurrences.

Chart, bar chart

Description automatically generatedChart, bar chart, histogram

Description automatically generated

Figure 1.8 Figure 1.9

Severity categorised accidents into levels 1 to 4, from least severe to most severe. It can be assumed that they are considered to be low severity, medium severity, high severity and extreme severity. Figure 1.8 demonstrates that a large amount of accidents occur with a Severity level of 2, followed by 3, 4 and 1. Figure 1.9, a reimplementation of the yearly bar plot, indicates that all severities are similar in proportion for each year, except severity 1 which does not exist before 2020. It can be assumed that this has been implemented as a new level.

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Chart, histogram

Description automatically generatedChart, histogram

Description automatically generated

Figure 1.10

Figure 1.10 is a generic bar plot with backgrounded grid lines, split into severity and their corresponding monthly trends. Severity 2 appears to follow most closely to the overall monthly trend of the data due to its large proportion, however severity 3 and severity 4 visualisations seem to indicate a high accident occurrence in June, at over 22000 and 10500 accidents respectively. This demonstrates that drivers are more likely to encounter extreme crashes during the mid-year, unlike what was represented previously in the monthly trends. Therefore, it’s important to consider the same trends in the daily period while also analysing the proportion of severities at the year level to ensure not all data is eclipsed by a single level.

Chart, bar chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Figure 1.11 Figure 1.12

Figure 1.11 indicates that there are similar trends of accidents regardless of severity, with highest accidents occurring during the time 8-10am and 4-6pm time slots, indicating that severity cannot be determined on the hour of occurrence. Moreover, Figure 1.12, an interactive dodge bar plot, conveys the notion that all years have the largest proportion of Severity 2 cases, followed by 3 and 4. Proportion is useful in ensuring that the large value of a single severity does not entirely reflect the trends of the entire dataset, and hence is a useful visualisation to reuse when appropriate.

Chart, bar chart

Description automatically generatedChart, treemap chart

Description automatically generated

Figure 1.13 Figure 1.14

A simplified classification of accident times can be found in the Sunrise/Sunset variable, in which the accident may occur at levels Day or Night, reliant on the system of the sun rising and setting. Figure 1.13 Demonstrates an extremely similar proportion of accidents respective of their severities, however a larger proportion of extremely severe accidents occur at night, as well as a decrease in low severity. Figure 1.14 demonstrates however that most accidents occur during the day, which is indicative of why there may be a larger spread of severity in Day observations (more values hence more chance of variation). However, as this is not the only time classification of interest, it’s important to compare and evaluate whether most other time systems justify similar trends within the daily and nightly counts.

Chart, treemap chart

Description automatically generatedChart, treemap chart

Description automatically generated

Figure 1.15 Figure 1.16

The confusion matrix visualisations above demonstrate the comparative frequencies of matched values between sun and twilight classification systems. Figure 1.15 demonstrates that the sunrise/sunset system is similar to civil twilight, observed through the diagonals being filled by Frequency of similar values, while Figure 1.16 demonstrates similar values as well. These comparative visualisations indicate that most accidents occur during the day and night and are observed similarly as such, but nearly 60 000 accidents considered daytime in civil twilight is night time in the sunrise/sunset system. As such, we can assume that many accidents occur during the transition between day and night, specifically periods of twilight and sun movements. This may pose a bearing on driver visibility, though it can’t be assured considering these are merely classifications and do not justify similar weather conditions every day.

Environmental Conditions Visualisations

Environmental factors are a major objective of this analysis, in which certain environmental factors will be evaluated together to justify any trends in accidents that may be impacted by natural effects. Considering the environmental variables are mostly continuous numerical values, bar plots would not efficiently convey the relationships between these factors. Interactive scatterplots, boxplots and density plots will be produced.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 1.17 Figure 1.18

Figure 1.17 demonstrates relationship between humidity and temperature in relation to severity. It appears that though severity is random, most accidents occur at high temperature and low humidity and low temperature and high humidity. As such, a weak negative relationship may be assigned next to these two conditions, these will be further explored in regards to weather conditions. Figure 1.18 demonstrates that high severity accidents occur when it is spread across a larger distance, though requires better axes.

Chart, scatter chart

Description automatically generated

Figure 1.19

Figure 1.19, applying a log10 scale transformation, implores that high severity accidents occur mostly at a higher humidity, as well as most accidents only covering a distance of 2 miles in total, while many occur at near 0 miles.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 1.20 Figure 1.21

Figure 1.20 demonstrates that there is usually no precipitation occurring during an accident, while visibility appears to range between 0 and 10 mainly (in terms of miles). This is quite regular for a driver’s range but anything that impedes a driver’s visibility must be considered as point of interest. Figure 1.21 merely confirms that temperature is strongly consistent with wind chill, those the values appear to spread more under 50 degrees. It is an interesting phenomena, though it confirms enough that only temperature can be used when considering variables regarding warmth.

Chart, bar chart

Description automatically generated

Figure 1.22

Figure 1.22 demonstrates the most common weather conditions within the data, selecting the highest 10 from a possible 117 conditions. It is evident and logical to have most accidents occur in fair conditions considering the weather is stable for the majority of the day, however it’s interesting to observe that cloudy conditions and light rain appear to have at least 50 000 accidents occurrences attributed to them, indicating partial need for greater examination of driving under these conditions. The top 5 will be used for comparisons below.

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated

Figure 1.23 Figure 1.24

Figure 1.23 and 1.24 both demonstrate the distribution of temperature according to weather conditions. It appears that, on average, cloudy and fair conditions have higher temperatures while light rain has a lower temperature and much tighter range, hence a tighter distribution. As it follows, mostly and partly cloudy distributions are nearly identical, while cloudy and fair conditions appear to occur mostly at slightly lower temperatures.

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated

Figure 1.25 Figure 1.26

Figure 1.25 and 1.26 demonstrate that most accidents occur at a higher humidity. Light rain conditions appear to also have a tighter distribution yet a significantly higher average humidity than all other conditions. Moreover, mostly cloudy conditions appear to have a slightly higher distribution, while partly cloudy and fair conditions are most identical. This may indicate that higher humidity may lead to more adverse weather conditions, hence a point of interest for locations with these environmental factors.

Chart

Description automatically generatedChart

Description automatically generated

Figure 1.27 Figure 1.28

Figure 1.27 displays extremely tights and identical distributions of pressure according to wind condition, though it appears that cloudy conditions have higher distribution peaks at pressures around 24-28. This may indicate that adverse conditions mostly occur at lower pressures. Figure 1.8, though similar to adjacent distribution exhibits that most visibility occurs at 10 miles. This is the standard figure and, characteristically, cannot be determined accurately through observation. However, most conditions appear to cause the same visibility concerns.

Chart, box and whisker chart

Description automatically generatedChart

Description automatically generated with medium confidence

Figure 1.29 Figure 1.30

Finally, Figure 1.29 and 1.30 demonstrate very similar distribution patterns for wind sped. It appears that fair conditions occur at near 0 wind speed, hence it is evident that higher wind speeds may lead to more adverse conditions.

Locational Visualisations

Locational visualisations will now be applied to complete the objectives of the analysis; in which hotspots are identified and geospatial visualisations will identify areas of highest accident occurrence.

Chart, histogram

Description automatically generatedChart, bar chart

Description automatically generated

Figure 1.31 Figure 1.32

Figure 1.31 demonstrates that the highest number of accidents occur in California, followed by Florida, Orlando, etc. Moreover, Figure 1.32 shows that Los Angeles is the city where most accidents occur, followed closely by Miami. It’s important not to consider these as relatively important figures, considering accident counts do not consider rate of occurrence, hence census data must be utilised to ensure states are effectively ordered by rate.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 1.33 Figure 1.34

Figure 1.33 displays that Los Angeles is the County with the most accidents, though it suffers the same fate as the previous bar plots in not considering populations within these areas. Figure 1.34, however, is less effected considering population cannot be calculated, though the use of streets will be significant. Regardless, the I-5 and I-95 streets appear to cause the most accidents, hence should be looked at for better safety regulations and or precautions.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 1.35 Figure 1.36

Having accounted for population, the dataset was then merged with 2018 census population information according to state and the rate of accidents per 100 000 were calculated. It appears now that Oregon is the state with the highest accident rate, while California is the third highest though it covers most accident observations in the dataset. Oregon, thus, is a significant point of interest as the accident rate is substantially higher than other states. In Oregon, Portland is the city with most accidents (Figure 1.36), thus this will be further explored through analysing road structures and their proportions through labelled pie charts.

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Figure 1.37 Figure 1.38

Chart, pie chart

Description automatically generatedChart, pie chart

Description automatically generated

Figure 1.39 Figure 1.40 (BAD LABEL, ALL STATES NOT OREGON)

Figure 1.37 displays the proportion of accidents that occur with some observable structure present, and identifies that junctions are the location in which 28% of accidents occur, followed by crossings (25%) and traffic lights (19%). Considering the extremely high rate of accident occurrence, the junction and crossing systems in Portland must be heavily flawed considering it accounts for most accidents in terms of rate. We can cross validate observations in Portland with the top 10 cities in Oregon (Figure 1.38) in terms of accidents and it confirms that the percentage remains the same, with junctions increasing to 35%. Further cross validation with all states in Oregon (Figure 1.39) shows nearly identical percentages to Portland, while cross-validating with the entire dataset (Figure 1.40) demonstrates that traffic signals and junctions should account for 70% of accidents. This is evident proof that the crossing structure in Oregon is heavily flawed, considering it causes many more accidents in proportion when compared to other structures in every other US state. As such, this structure should be re-evaluated or examined further to enhance safety and/or limit accident occurrence.

Having considered all locational data, taking into account proportional rates of accidents, it’s now important to identify, through geospatial visualisation, which states and regions throughout the US appear to have the highest and lowest accident rates. To do this, it’s first important to analyse geospatial visualisations in regards to highest accidents and populations to effectively determine highest accident rates.

Map

Description automatically generatedMap

Description automatically generated

Figure 1.41 Figure 1.42

Figure 1.41 demonstrates the states with most accidents by count while Figure 1.42 displays states with highest populations, with both gradient fills progressively getting darker for higher counts and blue highlighting for top 10 states in each regard (inherently, red underlining is lowest 10). It’s fairly obvious, through comparison to see that states with highest accident counts also have the highest population. In fact, many regions within the East and West coast, as well as Southern states appear to be densely populated, while lowest populations and accident occurrences exist near the centre. It’s important to consider that Oregon (above California which is darkest brown on the West coast), is not highlighted as a highly populated state in Figure 1.42, hence it is fairly obvious that there are issues there regarding driving safety and accident precautions

Map

Description automatically generated

Figure 1.43

Figure 1.43 displays the final geospatial visualisation, adjusted for accident rate. It appears that highest accidents occur at the coastal states, while lowest accident rates occur in the central regions of the US, including two smaller states on the East Coast. This may largely be due to the overall structure of roads and differences between each state, considering coastal areas with higher population will generally consist of heavier traffic, much tighter and narrower roads as well as a larger number of people attempting to travel to different locations in a day. However, Oregon does require the most re-evaluation, determinately on its crossing structures, considering it is less densely populated than most states but still accounts for this highest accident rate.

Objective 2: Generating Insights on Traffic Violations Dataset

Rationale and Visualisation Choices

Considering the accident dataset is the data of significance, this objective will be brief and simply identify and/or consider the socially-driven impacts behind road violations. In brief, it’s an extension of the initial dataset and gives insight on human error and situational factors in addition to the environmental, time-related and locational variables previously investigated. Considering this dataset spans from 2013-2016, it will not be merged to the initial dataset, though certain visualisations can be made in comparison to identify how violations relate to accidents.

As the dataset is considerably smaller and involves citation occurrence rather than accident occurrence, remediation will not be used and significant statistical insights can be made from bar plots and geospatial visualisations.

Visualisations

By observing data in terms of violations, we can attest to whether laws and regulations do a satisfactory job in decreasing accident occurrence. As such, we should begin by analysis location again to see any changes in trend.

Chart, bar chart

Description automatically generatedTimeline

Description automatically generated

Figure 2.1 Figure 2.2

Figure 2.1, a bar plot with a log 10 scale transformations, demonstrates fairly interesting observations, in which Florida is the only State that remains in the top 10 count of both accidents and violations. As such, there’s a possibility that Florida regulations are well implemented yet not effective in decreasing accident rate. Moreover, Maryland appears to have the highest number of violations, indicating most offences are committed in that area, yet many do not lead to accidents as it was not displayed in the accident bar plots.

Out of curiosity, Figure 2.2 was constructed to see the most violations on any given street and it appears that Northampton Drive sees around 40 violations between 2013-2016. Considering that we are focusing on human impact, we should identify the origins of each driver and consider how these influences could impact driving ability.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 2.3 Figure 2.4

Figure 2.3 demonstrates that most violations occur from drivers that originate from Silver Spring, which surprisingly has a street with around 30 violations in Figure 2.2. It appears that the quality of teaching driving in that area is lacking, and requires serious overhaul to ensure that drivers are more competent, or are less inclined to be influenced by intoxication and commit any offences. However, a driver’s city of origin does not indicate whether they learnt to drive in that area. Figure 2.4 confirms that drivers from the state of Maryland, as labelled on their driver’s license, commit the most violations. As such, Maryland requires its constituents to undertake stricter tests and increase the difficulty in attaining a license in that state.

Chart, bar chart

Description automatically generatedChart, bar chart

Description automatically generated

Figure 2.5 Figure 2.6

Figure 2.5 shows that most violations are committed by driver’s using vehicle makes of Hondas, reaching nearly 1500, and with a 4S model (Figure 2.6). This, inherently is not a useful observation as Hondas and Toyotas are extremely common and 4S is a model enlisted for multiple different makes, however it may indicated that individuals driving these cars are more prone to committing mistakes than other makes, regardless of car manufacturing rarity and/or the amount of driver’s using them. It would be logical to apply larger premiums on these makes and models to deter any driver’s from committing any violations.

Chart, bar chart

Description automatically generated

Figure 2.7

Figure 2.7 displays statistics regarding the race of offenders, effectively demonstrating that most driver’s that violate laws are white, followed by Hispanic and black individuals who are extremely similar. Though an interesting insight, it’s fairly inconsequential in regards to committing driving offences considering that the locational factors will have more influence than ethnic background. We can, however consider alcohol use and gender proportions in regards to violations being committed.

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

Figure 2.8 Figure 2.9

Figure 2.8 displays that men commit double the amount of traffic offences than women, indicating that males are inherently more likely to drive recklessly than women and thus gender may possibly be considered when issuing premiums in terms of insurance to deter individuals. Figure 2.9 evaluates that alcohol use, out of nearly 12000 observations, is only present in around 40 offences, indicating that most violations can’t be attributed to intoxicated driving, but rather driving competency. Still, regulators should aim to minimise driving while intoxicated and impose heavy fines if the laws are violated.

Finally, to objectively compare the violations to accident occurrence, a geospatial map visualisation will account for violations according to state and generate an overview of road regulations being implemented across the US.

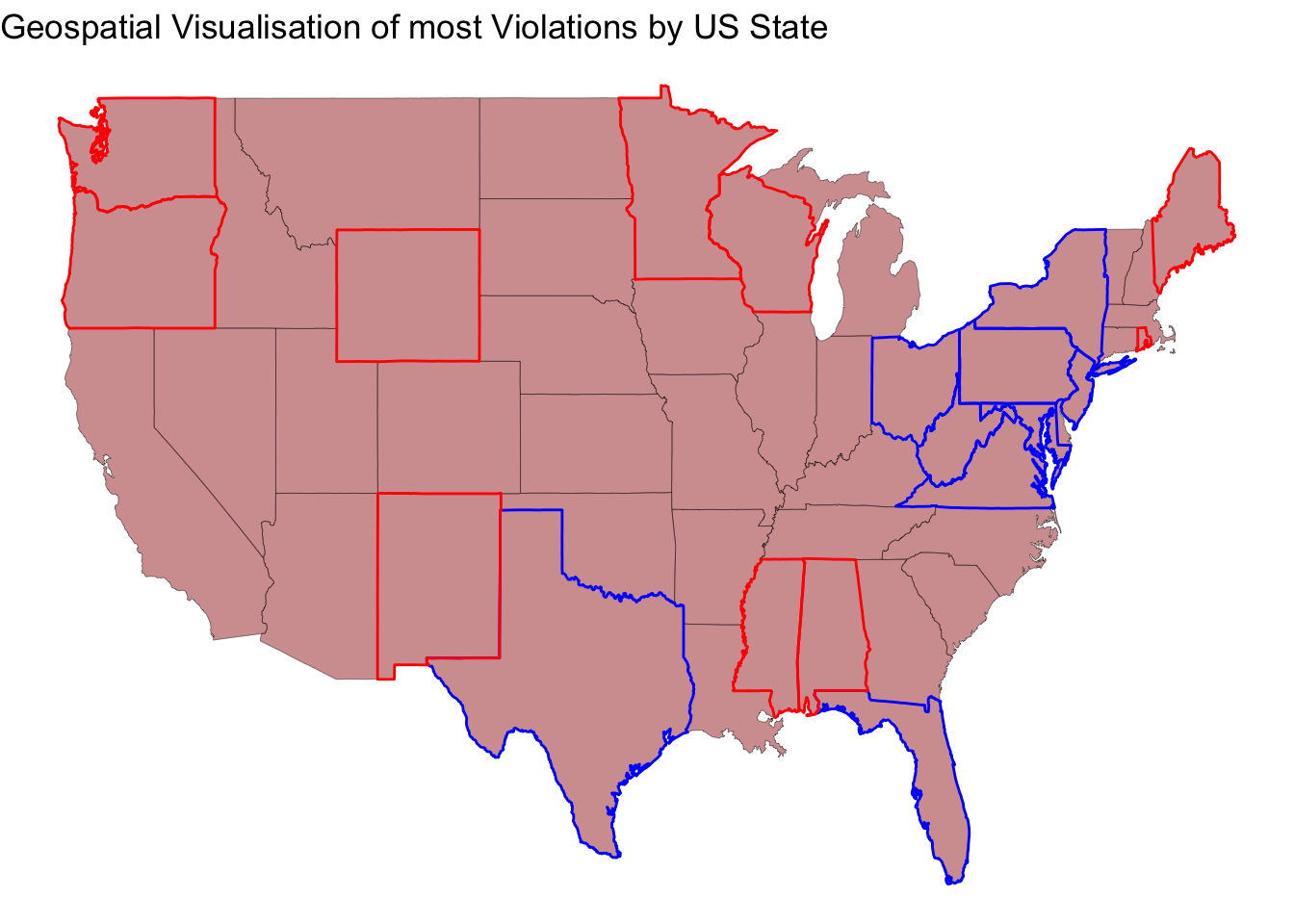


Figure 2.10

Figure 2.10 is a geospatial map highlighting highest violation counts in blue and lowest in red. This demonstrates that many states on the Eastern Coast have the highest violation occurrences, as well as a few Southern states. In fact, a large cluster of violations occurs in the upper Eastern region, while surprisingly, there are comparatively lower violations occurring on the West Coast. It is evident that there needs to be better enforcement of regulations in these areas of high violation occurrence as driver’s are currently unequipped with the adequate skills to drive well. Moreover, it may be considered a ‘social culture’ in the Eastern region, in which driving poorly is not as penalising as it is in other states. A less lenient approach and better approach to regulating road laws will indubitably decrease violation states.

An interesting insight is that Oregon, in the upper west region of USA, appears to be highlighted in red, and thus is a state in the 10 lowest violation occurrences in the US. This is a major concern considering it is the state with the highest accident rate. This obviously indicates that regulations in Oregon are too lenient and it is imperative that the road laws there receive a massive rehaul to ensure that drivers are penalised heavier and are less likely to drive recklessly.

References

# Berinato, S. 2016. Visualizations That Really Work. Harvard Business Review. <https://hbr.org/2016/06/visualizations-that-really-work>

# Ghadawala, M. (2019, 24-9). Top 7 Benefits of Geospatial Data in Driving Analytics. Geospatial World. <https://www.geospatialworld.net/blogs/top-7-benefits-of-geospatial-data-in-driving-analytics/>

Gutierrez, F. 2017. Traffic Violations in USA. <https://www.kaggle.com/felix4guti/traffic-violations-in-usa/>

Moosavi, S., Samavatian, M. H., Parthasarathy, S., & Ramnath, R. (2019). A Countrywide Traffic Accident Dataset. ArXiv:1906.05409 [Cs]. <https://arxiv.org/abs/1906.05409>

Vikas. 2018. US State Populations – 2018 <https://www.kaggle.com/lucasvictor/us-state-populations-2018>

[www.iihs.ord](http://www.iihs.ord). (n.d.) 2021. Fatality Facts 2019State by state. IIHSHLDI. <https://www.iihs.org/topics/fatality-statistics/detail/state-by-state>

www.nhtsa.gov. (n.d.). 2020 Fatality Data Show Increased Traffic Fatalities During Pandemic | NHTSA. [online] Available at: <https://www.nhtsa.gov/press-releases/2020-fatality-data-show-increased-traffic-fatalities-during-pandemic>.

www.statista.com (n.d.). 2020. Light vehicle retail sales in the United States from 1976 to 2020(in 1,000 units). Statista.com <https://www.statista.com/statistics/199983/us-vehicle-sales-since-1951/>

Datasets

Violations Data - . <https://www.kaggle.com/felix4guti/traffic-violations-in-usa/>

Accident Data - <https://www.kaggle.com/sobhanmoosavi/us-accidents>

Population Data 2018 - <https://www.kaggle.com/lucasvictor/us-state-populations-2018>

Code