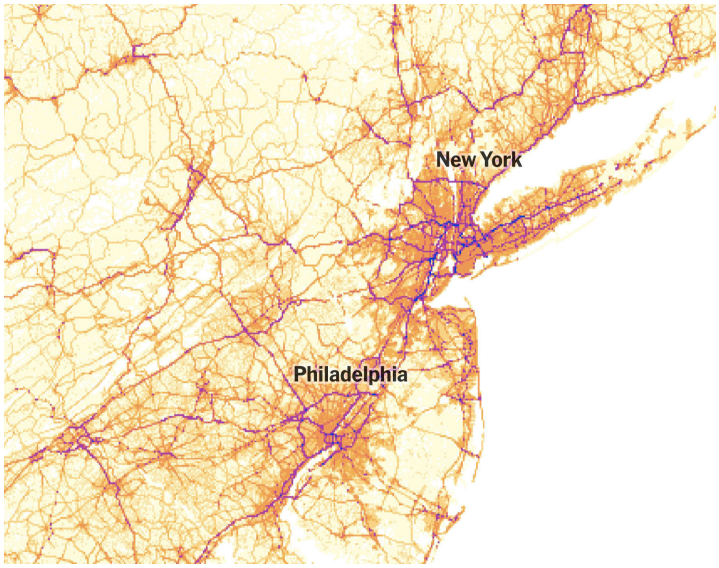
Drive-Safe USA



In-Depth Analysis of US Car Accidents

Table of Contents

1. Executive Summary

For this project we explored a dataset spanning US road accidents from 2016 to 2023. We developed a national statistics dashboard and an interactive map to present and visualize this data. The primary objective of the national statistics dashboard was to offer a detailed overview of accident distributions and patterns at the national level. On the other hand, our goal for the interactive map was to allow users to filter accident data based on locations, times, and driving conditions to drill down to their own backyard. The combined purpose of these two components was to empower users with an interactive platform to gain insights into accident hotspots and patterns. By leveraging our tool, users can explore accident distributions at the level of their interest. Ultimately, our hope was to create a user-friendly interface that contributes to informed decision-making, fostering a safer transportation environment for all.

Link to Visualization:
https://public.tableau.com/app/profile/himanshu.naidu/viz/interactive_map_desktop_third/LandingPage

2. Concept Background

2.1 Data and Concept Background

Traffic accidents pose significant challenges to road safety and public welfare, impacting communities and individuals across the United States. People caught in traffic accidents face serious injury and financial distress while those living around them can find their livelihoods thrown into disarray by the resulting disruption to local traffic. The general public, particularly commuters, would have much to gain from a comprehensive knowledge of the accidents in their area. Thus, understanding the underlying causes, geographical distribution, and trends associated with these accidents is crucial for devising effective preventive measures and enhancing road safety initiatives. This work endeavors to provide this knowledge, with a focus towards aiding the average U.S. commuter.

To perform our analysis, our work draws from the dataset first established and defined in the 2019 paper "A Countrywide Traffic Accident Dataset" by Moosavi et al. This dataset contains over 7 million records of traffic accidents from across the U.S. spanning 2016 to 2023. As with most accident datasets, this collection contains basic information on the time, place, and duration of each accident. But unlike most datasets, it also contains contextual, environmental, and geographic data for each entry as follows:

- Contextual Data:
  - Accident severity, duration, description, etc.
- Environmental Data:
  - Weather conditions, sunlight, visibility, etc.
- Geographic Data:
  - Points of interest such as ramps, bumps, and roundabouts

With this wealth of information, it becomes possible for us to visualize correlations between these additional features and the severity or location of accidents across the country. However, before we can confidently utilize this data, we must first understand how it was obtained and process it in preparation for its use in our visualizations.

In order to obtain such a multifaceted dataset, Moosavi et al. chose to combine data from multiple different sources. For each accident in the list, the basic and contextual information was pulled from a combination of state agencies, MapQuest, and Bing Map Traffic data. In this way, a large number of accidents could be cataloged. However, any overlap or repeats between each source needed to be removed *by hand*. By the authors' own admission, this process was likely incomplete, and our own work will need to take this into account when preparing the data for use. The environmental data was pulled from weather stations across the U.S. and reflects the consistency and priorities of these sources. This data is *extremely* thorough with relatively few NaNs on account of weather stations placing a high importance on uninterrupted service. However, the use of often outdated and somewhat imprecise machinery means that we can expect no guarantee of precise, error-free, or user-friendly data in these categories. Significant cleaning needs to be done in this area. And lastly, the geographic data was pulled almost entirely from OpenStreetMap, an open-source community-run project that

attempts to document road features from across the world and record them for use in future analysis. This project spans multiple continents but is largely run by smaller local chapters of the organization that attempt to follow the group's best practices but otherwise have no direct obligations. By its very nature, OpenStreetMap is not exhaustive, and its volunteers are not evenly dispersed across the United States. Despite this, OpenStreetMap is the most rigorous and comprehensive resource on roadway features to date and is impressively clean for what it is. Moosavi et al. took the locations of each accident and added additional columns documenting whether or not OpenStreetMap listed a particular point-of-interest in the vicinity of it. It is worth noting that for any accident where the presence of a feature was ambiguous and not mentioned in the description, Moosavi et al. chose to record that the feature *was not there*. As such, this dataset is strictly *under*estimating the presence of these roadway features and any correlations between them and accident severity are likely being *over*estimated on account of such features only being recorded when the record shows that it was relevant to the accident. All these factors taken together were used to guide both our future visualization design and our immediate data exploration. With the source of data now thoroughly understood, we continued on to inspecting the details of the data itself.

2.2 Data Exploration

Before our dataset can be used to inform our audience, we first need to ensure that it is in good condition. Nearly all real-world data contains issues such as missing values, inaccuracies, or poor formatting. As such, the data must be explored in order to determine its condition and identify cleaning targets. Following standard practice, we began by inspecting the distribution of each column in the dataset through histograms. When doing so, we made note of any columns that had anomalously high/low values or anomalously frequent values. In addition, the prevalence and the *meaning* of missing values in each column was investigated and noted down. Any columns with categorical information also had its unique values inspected in order to determine the format and consistency of their contents.

From this process, several key issues were identified. As was expected, any issues present in a given column were also present in any other columns that shared came from the same source (e.g. MapQuest). Within the contextual data, there was a negligible amount of missing data and the distributions of each column followed either normal or exponential distributions with reasonable values and no notable outliers. The only exception to this was the "Distance" column that encoded the amount of roadway whose traffic was affected by the accident. This column contained values in the hundreds and was thus noted down as a candidate for cleaning. Within the environmental data, very few missing values were found but nearly every column contained alarmingly high outliers. All environmental columns from "Visibility" to "Weather" were noted down as needing cleaning. The "Wind Speed" column in particular was given a high priority due to containing wind speeds of 1060 mph, which would exceed peak hurricane conditions *on Jupiter* and were thus guaranteed to be incorrect values. And within the geographic data, no anomalous values or unexpected categories were found but the majority of columns contained at least 80% missing values. A closer inspection of Moosavi et al.'s methodology reveals that this is because "NaN" values represent not a *missing* value, but an absence of the point-of-interest. As such, these values should not be altered. From this exploration, the "Distance" column and *all* environmental columns were identified as targets for data cleaning.

2.3 Data Cleaning and Preparation

Beginning first with the environmental columns: the "Wind Speed", "Precipitation, "Visibility, and "Air Pressure" columns all contained anomalously high or low values. For "Wind Speed", several points reported values in the hundreds, despite 70mph winds being sufficient to announce a tornado. Manual inspection of these points' accident descriptions and weather supported the notion that these values were the result of malfunctioning weather equipment since none of their descriptions mentioned poor weather and none of their weather entries were "Windy". Entries with tornado-level winds were trimmed from the dataset after confirming that these errors were distributed completely at random throughout the dataset. This check for a completely random distribution of outliers was also performed for all subsequent cleaning steps. For "Precipitation", many points reported daily rainfall of 20 inches or more. The U.S. record for daily rainfall is 49.6 inches for a specific region in Hawaii in 2018. While these values are thus possible, they are not plausible since none of these points have "Rain" as their associated weather condition. Such points are clearly anomalous. Entries with greater than 20 inches of rain were removed and entries with greater than 5 inches of rain, but no associated rainy weather condition were also removed. For "Visibility" several points had a visibility of 140 miles. Most of the dataset had a 10-mile visibility listed (due to rounding on the part of weather stations) so these values are anomalously high. Notably, these values ARE still possible. In the absence of any air pollution, 140 miles is considered the upper limit of visibility for states on the west coast. However, all of these high visibility points were reported in the vicinity of major cities such as San Francisco whose air pollution prevents more than 15 miles of visibility. As such, any points in a city with over 60 miles visibility were removed, as no U.S. city area has ever reported a visibility above this. Finally, for "Air Pressure" a surprisingly large number of values well above or below the mean of 30 inHg could be found. The maximum reported value was double atmospheric pressure while the minimum was the vacuum of space. Values outside the range of 29 to 31 can cause symptoms varying from migraines to instant death. As such, these values were clearly anomalous. The all-time low and high barometric pressures ever observed were 25.9 inHg and 32.01 inHg respectively so points outside this range were cleaned from the dataset. With this, the anomalous environmental columns were cleaned, and "Distance" could be more closely inspected. "Distance" had anomalously high values of over 400 miles listed for several of their accidents. But incredibly, an investigation of these points found that they were correct. Several accidents in the dataset resulted in the closure of hundreds of miles of highway and traffic disruptions that spanned over 400 miles. After consulting with external news sources to confirm the veracity of these points, the column was ultimately left unchanged. In all these cleaning steps listed above, care was taken to remove only entries that were guaranteed to be anomalous since individual outliers had a minor effect on the overall dataset and thus the preservation of the data could be prioritized.

Following these cleaning steps, there remained only one major alteration that needed to be made. The "Weather" column gave the categorical weather condition at the time of the accident for each data point. However, it used U.S. meteorological classifications for each weather condition and thus had over 80 different possible values and several instances of obscure jargon. Thus, each related weather category was binned into one of 10 simple-to-understand bins such as "rainy" to increase the column's accessibility to our intended layman user base. With this step completed, the dataset was now fully cleaned of anomalous, missing, or poorly formatted data and could finally be used to assemble our visualizations.

3. Process Description

3.1 Testing Procedures

Our testing was conducted in three phases. An in-class trial involving a paper prototype (Fig 7.1) was conducted, followed by an external evaluation of a prototype Tableau dashboard. Lastly, a presentation of our adapted dashboard served as an occasion for additional feedback.

The in-class paper test (Fig 7.1), managed collectively by the team, proceeded with one team member acting as the computer and operator, two members acting as prompters, and two others as feedback collectors. Users from our class evaluated our paper prototype's state comparison pane, while we, as testers, focused on user interactions, filter usage, and refining tools. Feedback from users highlighted issues with slider labeling, map interaction, and clarity in the visual elements, prompting suggestions for improvement, such as clearer labeling and better visual cues. The subsequent discussion led to a re-evaluation of the project's goals, shifting the focus away from state comparison. This strategic pivot aimed to restructure our project to better fit what can be supported by our data.

The prototype Tableau dashboard (Fig 3.1) was tested by three tech professionals in tests conducted by individual members of our team, outside of class. Each test proceeded in an analogous way as to the in-class tests, but with one member of the team playing all of the testing roles. These tests proved extremely useful in inspiring our group to simplify our design, and proved a key component of moving beyond paper towards a viable product.

Our last phase of testing was a product demo (Fig 3.2) intended to gather feedback from our peers. For this test, we introduced our project and walked through a sample workflow of our dashboard. While half of our group presented, we had the rest moderate our group's slack channel to take questions and criticisms.

3.2 The Interactive Map

Our main goal in creating our interactive map was to allow users to explore accident hotspots in their local area at whatever scale they wish to set their focus. What made this a difficult task is that accident hotspots are not static. They change over time depending on a huge number of potentially overlapping factors including season, driving conditions, and more. For our purposes, we chose to focus on time and weather to inform our users.

To understand our initial intentions, here is a sample user story. A commuter takes I-5N to work and wishes to learn about where and when they should be vigilant during their commute today. Using our tool, they navigate to their hometown, find, and select I-5N, the weather, and the time of day, and then view the map to look at the distribution of accidents on I-5N that happened on similar days. This data would then serve as a proxy for where the user should be cautious on their way to work.

The implementation of our interactive map occurred in three stages. First, we made a Tableau prototype (Fig 3.1) and tested it for our C3 submission. Second, we incorporated feedback from these tests into a presentation draft (Fig 3.2). Third, we refined the map using the feedback we received from our peers during the in-class review.

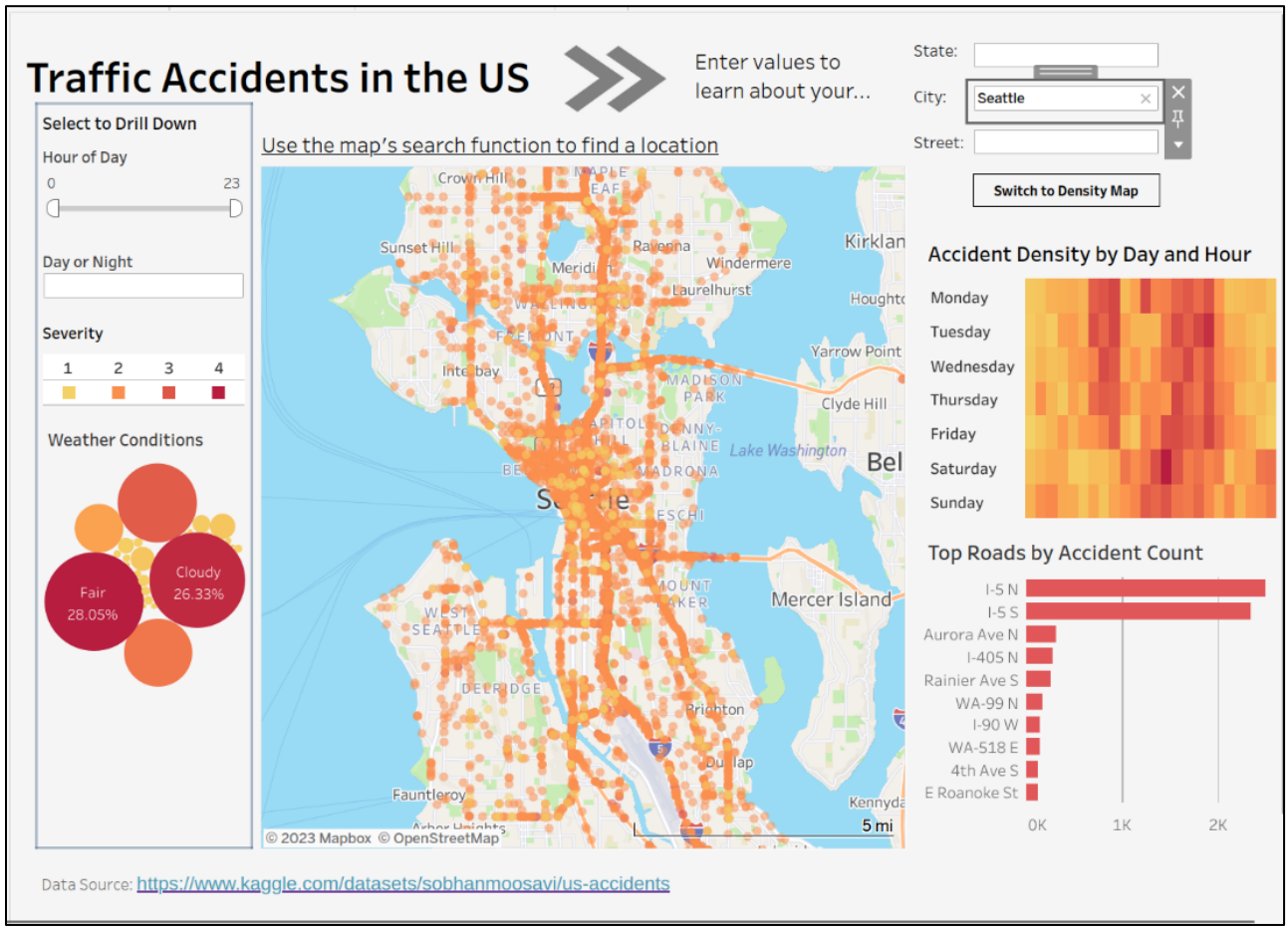


Fig 3.1 – First Tableau Prototype

Above is our first iteration on the interactive map. In this prototype, we attempted to save space by using a heatmap to concisely present both the weekly and daily distribution of accidents, and a bubble map to concisely present the distribution of weather conditions. The ability to switch between a point and density map was also implemented for flexibility, as in low-density areas a point map was believed to be more effective than a density map.

Unfortunately, this first draft of our interactive map was poorly received. The feedback from our product testers highlighted several issues. They found the controls cumbersome, the color scheme inaccessible (one of our testers was colorblind) and inconsistent (we used the same colors for severity and our distribution charts), and the charts fragile at high levels of "zoom." Additionally, users encountered challenges in coordinating the filters and expressed a need for clearer instructions and guidance. In essence, the primary problems centered around the separation of data and view selection, the overuse and misuse of color, chart malfunction upon zoom, filter and control difficulties, and sluggish map performance. This feedback proved extremely useful and was incorporated into our presentation draft, given below in Fig 3.2:

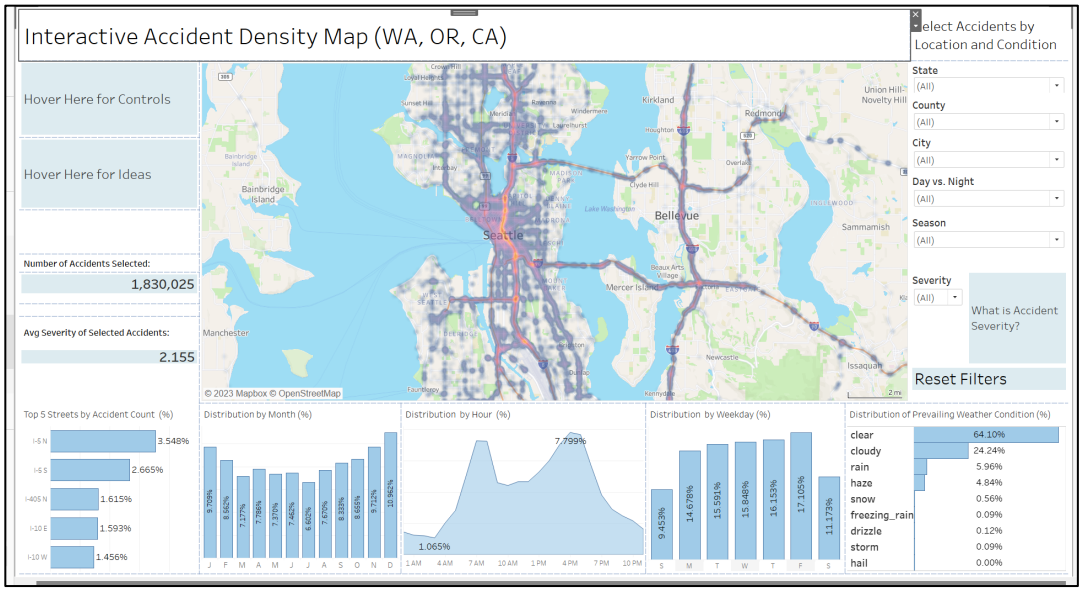


Fig 3.2 Presentation Draft of the Interactive Map

In response to our testers, we made several adjustments to our map. First, we removed the point view on the map and adopted a darker color low-density areas to try to compromise between the two views. Second, we removed all unnecessary uses of color to avoid confusion. Third, we scoped down to the west coast states to improve load times for users. Fourth, we implemented a reset button as per user request, along with hover buttons to present instructions, introductions, and additional information to our users. Finally, we simplified our charts by removing elements such as the heatmap and map, opting instead for bars that maintained integrity at multiple levels of zoom, and importantly, remained selectable even when data for a desired category was sparse.

Though we believe this map was a definite improvement upon our prototype, it was far from perfect. Again, we received valuable criticism, this time from our peers during our presentation. In no particular order, we received criticism for: visual clutter, the map theme and lack of legend, not accounting for priors (to be discussed further in our critical evaluation), and the placement of our filters and hover buttons. Fig 3.3 shows our finalized dashboard that incorporates this final round of feedback.

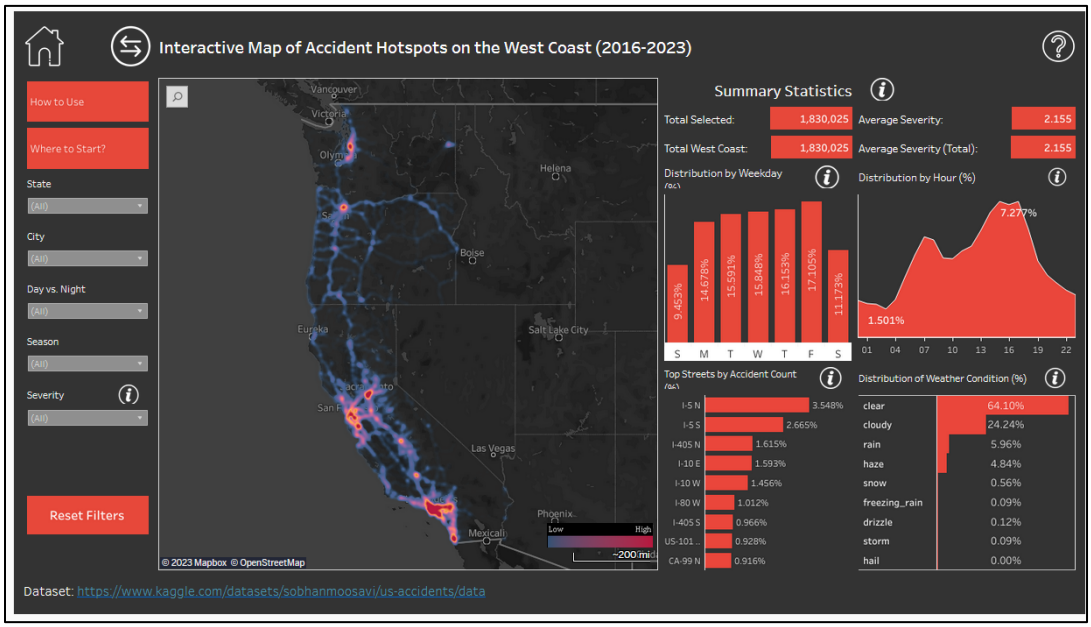


Fig3.3 - Interactive Map of Accident Hotspots on the West Coast

First, we managed to implement a dynamic zoom that snaps to a city or state when you select it in the filter, reducing the separation of data and view selection. Second, we removed the month histogram to declutter the interface, considering we already had a seasonal filter. Third, efforts were made to enhance visibility and usability by enlarging the map, fixing a better aspect ratio, and making and including a color legend. Fourth, to streamline user interaction, the filters and hover buttons were consolidated into a single container positioned on the left-hand side of the interface. Fifth, a shift to a sparse dark theme was made to improve the legibility of the map and supporting charts. Sixth, info buttons were also added to further document our dashboard. Lastly, dataset-wide averages were introduced to facilitate basic comparison across the dataset.

We'll conclude our review of the interactive map with a brief reflection. Most importantly, we found that testing is *extremely* effective. A cyclical process involving prototypes was key to figuring out which of our ideas were not worth implementing. Another key takeaway was that a high degree of interactivity requires a lot of data and extreme flexibility. Our project definitely would have been more successful, and we would have been able to incorporate more priors, if we had focused on a single metro area.

3.3 The National Statistics Dashboard

In response to invaluable feedback received during the C3 prototyping and feedback activity, we worked on elevating the National Statistics Dashboard into an insightful tool for understanding accident data across the entire United States. This endeavor was driven by the acknowledgment that the interactive density map fell short in delivering a holistic portrayal of nationwide trends. Our presentation draft is given in the figure below:

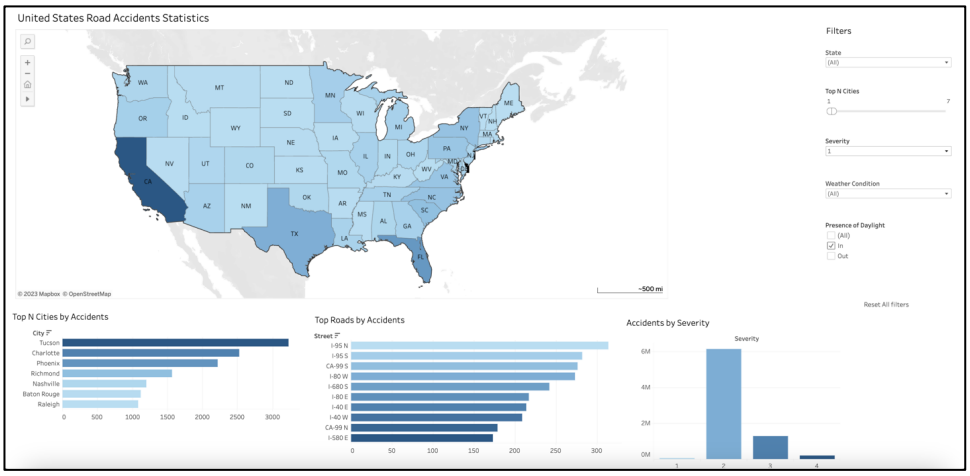


Fig 3.4 Presentation Draft of the National Statistics Dashboard

The presentation draft (Fig 3.4) included a choropleth map, top cities and roads by accident counts, and a severity distribution graph. Filters such as state selection, severity, weather conditions, top N cities, and presence of daylight were integrated to empower users with the flexibility to tailor their insights. However, feedback we received during the presentation prompted a reevaluation of our design choices.

Navigating the challenge posed by the non-uniform distribution of accidents across states, we made a strategic pivot from relying on raw accident counts to utilizing average severity as a more equitable metric. This decision was pivotal in ensuring fair comparisons between states with diverse population sizes and accident rates. The subsequent simplification of the dashboard, achieved by judiciously eliminating certain filters, was an intentional move towards a cleaner and more user-friendly interface.

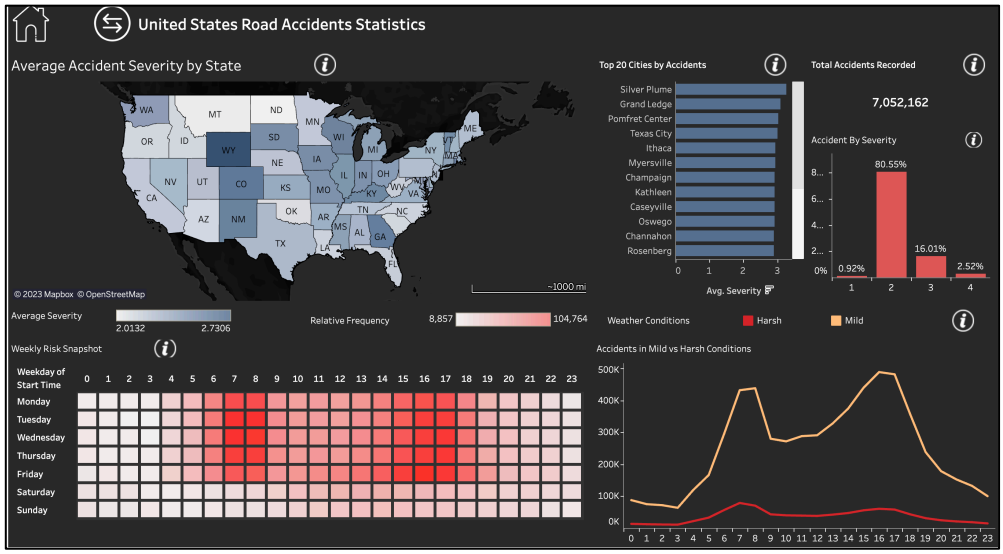


Fig 3.5 The National Statistics Dashboard

The choropleth map, originally driven by accident count, was redone to reflect the average severity of accidents in each state. Similarly, recalibrating the ranking of top cities using average severity, as opposed to raw accident counts, offered a more meaningful portrayal of the prevalence of accidents in certain cities.

To enhance temporal insights, the frequency of accidents based on days of the week and time of day was illustrated with a heatmap. This addition enabled users to discern patterns across both variables with greater clarity. Responding to feedback from the course instructor, a line graph depicting hourly trends in accidents based on weather conditions was introduced. The categorization of weather conditions into mild and harsh facilitated a more nuanced exploration, addressing the instructor's query on the impact of rare weather events on accident rates.

The National Statistics Dashboard has undergone a substantial transformation, shaped by user feedback, team analysis, and instructor input. The resulting dashboard (Fig 3.5) now prioritizes simplicity, intuitiveness, and visual appeal, offering users a powerful and user-friendly tool to explore and comprehend accident data across the United States.

4. Critical Evaluation

Originally, our group intended to make comparative analysis the focal point of our project. However, as we developed our interactive map, it became very clear that we needed more data to accomplish our task. As was noted in class by our peers, our professor, and our TA, comparison relies on knowledge of prior probabilities. For example, at the moment, our charts show peaks in accidents during the morning and evening commutes. Does this mean that your chances of getting into an accident are greatest at these times? The answer, of course, is not necessarily. To effectively compare locations, times, and conditions, we needed to account for the volume of traffic matching our filters. In other words, we needed to normalize our data to avoid misleading our users.

Realizing that these priors were required was the easy part, unfortunately. Given the scope of our project and the variety of levels of zoom we wanted to capture, the comprehensive traffic and demographics data we needed we found either too difficult to find, or too difficult to incorporate. Our dataset was already extremely large, and our dashboard was already relatively slow for our purposes. In the end this was a difficult roadblock for us that we did not quite overcome.

Besides this fundamental weakness of our project, there were other issues with our data that indicated we were out of our depth. For example, numerous reporting issues complicated our analysis. Notably, the lack of uniform standards and practices across states led to nonrepresentative gaps in time and space in certain locations, and response bias resulted in less severe accidents going mostly unreported due to their minimal impact on traffic. Our data did not stand up, in all cases, to the amount of zooming we wished to allow, and was not, in fact, a representative sample of the reality we wanted to explore.

Lastly, our goal was also unfeasible from the perspective of Tableau. The map plugin we relied on proved inadequate for our needs and struggled to handle the data we attempted to visualize, even after filtering down to the level of a single city or sampling down to a minimal subset of the data. We should also have anticipated the difference between local performance and the capabilities of Tableau Public. In the future, before embarking on our project, we needed to perform an initial feasibility test of the concept. We considered performance only too late, assuming we would be able to make it work in any case.

In retrospect, instead of pursuing the goals that we did, our strategy should have centered solely on King County, or even solely on Seattle. This focused approach would have yielded more valuable insights, as **our data would have been more comprehensive, and we would have been able to incorporate complimentary supporting data**. Most of our problems would have been manageable had we taken this approach. Our data would remain small enough to allow for quick interactivity, and could be supplemented without becoming inconsistent, as it would be constrained to one state with uniform standards.

Fatal flaws notwithstanding, we are happy to say because of this project we have learned a lot about Tableau, including its strengths and weaknesses, and have produced a usable final product. In addition, our process of testing has reiterated many of the concepts we learned about in class. The feedback we received consistently focused on aspects of our visualization that interfered with its **pre-attentive effectiveness [3]**, a key concept from lecture. Our testers told us to simplify, consolidate, clean, and reduce the amount of coordination required to interact with our dashboard. They told us to make our tool easier to use, even at the cost of

functionality. In the end, we took their advice, and pared our dashboard down to a few carefully selected distribution charts, focusing on maintaining good UI/UX. This involved **avoiding the arbitrary use of color [2]**, much like Jock Mackinlay mentioned in his talk to our class. As he spoke about, color, as an encoding, is a natural but sometimes ineffective choice for certain common tasks. Our testing confirmed this, suggesting that it is best to avoid using color as an encoding where it is unnecessary. For example, we originally used color *and* size as overlapping encodings in the weather condition bubble map from our prototype. This was not only unnecessary, but it left our testers scratching their heads. Lastly, the feedback we received caused us to look back to Heer and Schneiderman's, **"taxonomy of interactive dynamics," [1]** to understand the issues our interactive visualization faced. In particular, the separation of data specification and view specification in our first two dashboards crystallized as a problem thanks to their terms.

5. Acknowledgements

We would like to extend our sincere gratitude to Professor Nathan Mannheimer for his invaluable guidance, support, and mentorship throughout the duration of this project. His expertise and encouragement have been instrumental in shaping the direction and quality of this Data Visualization project. A special thanks also goes to our dedicated Teaching Assistants, Tejal Kolte and Nizan Howard, whose assistance and insightful feedback greatly contributed to the development and refinement of this project. We express our heartfelt appreciation to our diligent testers, whose invaluable feedback and rigorous testing significantly improved the robustness and functionality of our project. Furthermore, we are grateful to our classmates at University of Washington for their encouragement, engaging discussions, and collaborative spirit, which have been instrental in our learning and development throughout this quarter.

6. References

1. Heer, J. & Shneiderman, Ben & Park, C.. (2012). *A taxonomy of tools that support the fluent and flexible use of visualizations.* Interact. Dyn. Vis. Anal. 10. 1-26.

2. Mackinlay, Jock. *Talk for DATA 511: Data Visualization for Data Scientists*, 30 November 2023, University of Washington, 45th Street Plaza, 1100 NE 45th St, Seattle, WA 98105, United States.

3. Few, Stephen. *Now You See It: Simple Visualization Techniques for Quantitative Analysis*. Analytics Press, 2009.

7. Appendices

Appendix 1 – Paper Prototype

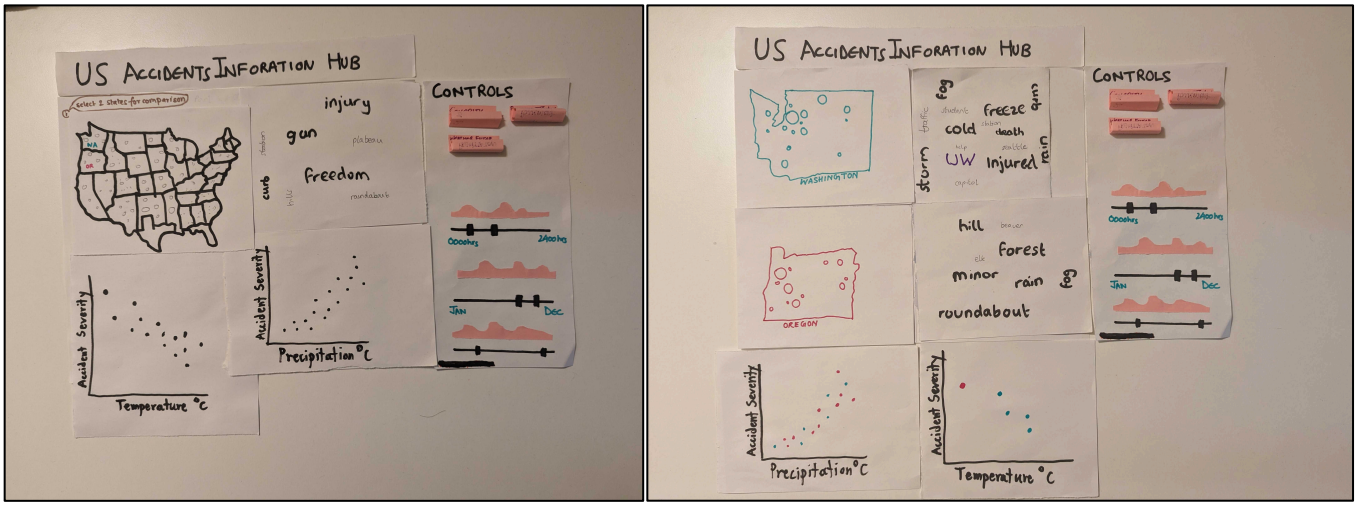


Fig 7.1 Paper Prototype

Fig 7.2 – Landing page

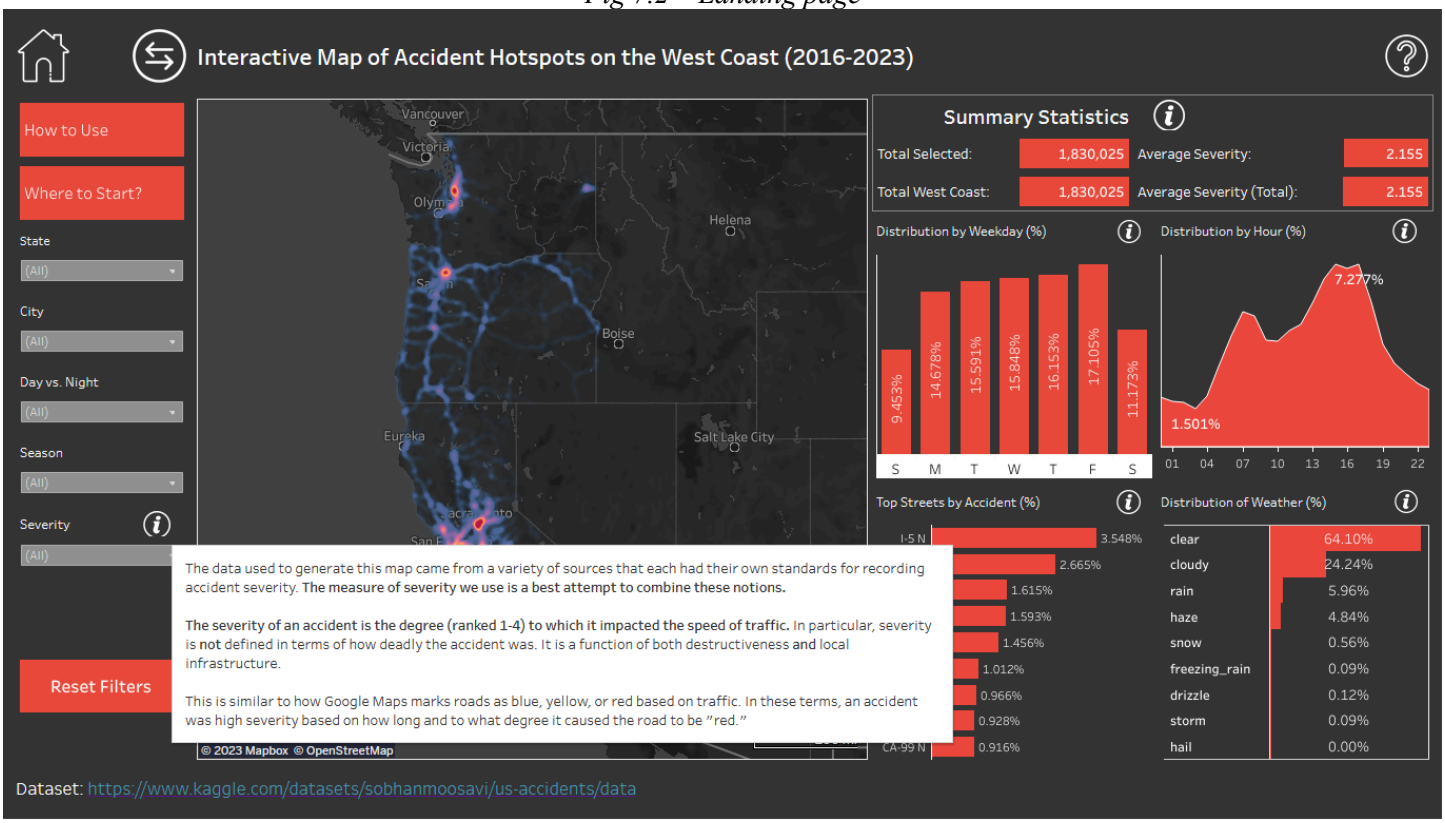
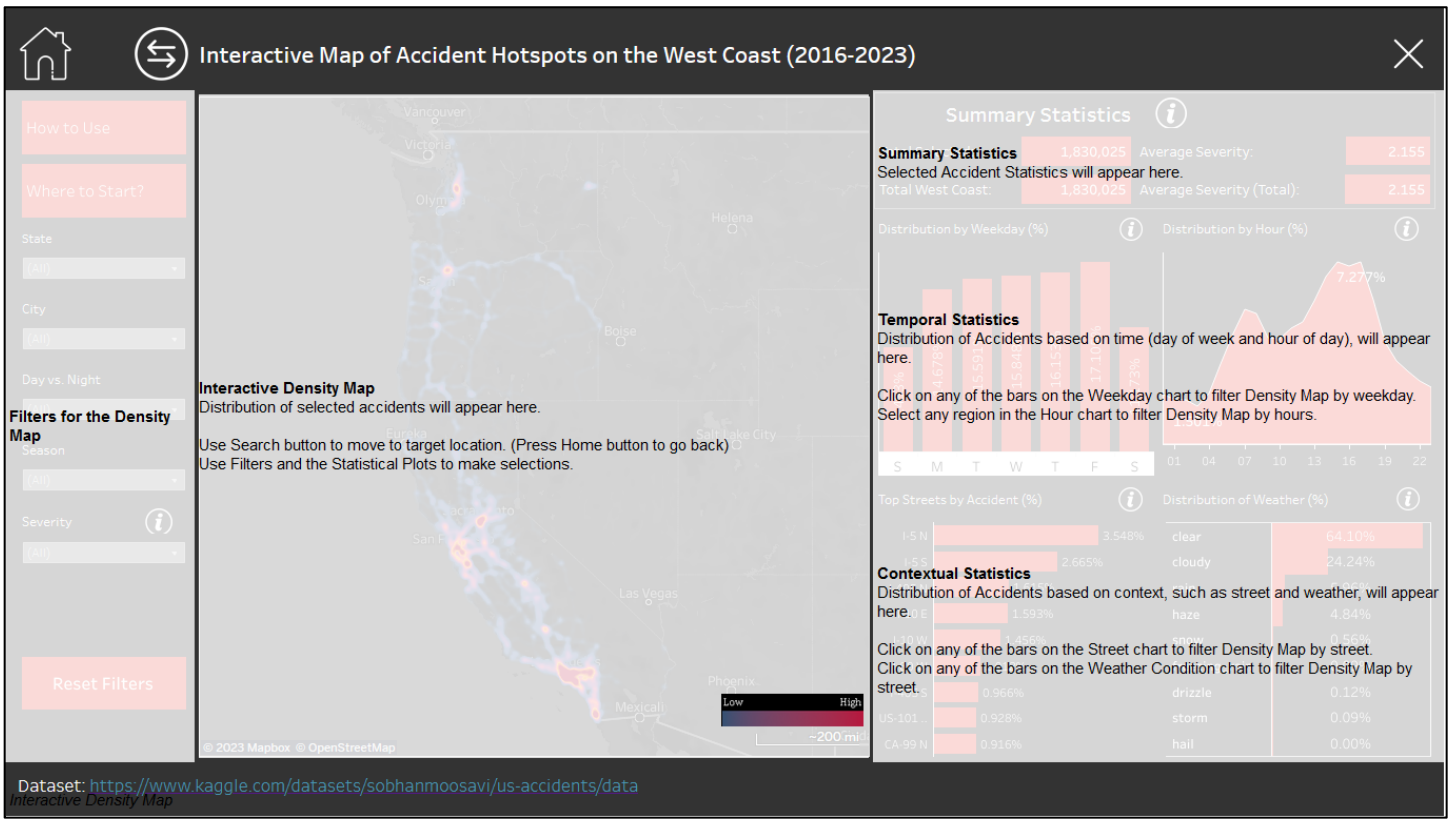


Fig 7.4 – Interactive Map with tooltip



Interactive Density Map

Fig 7.5 – Button Functionality