Utah Accident Visualization
Process Book

Alan Bird, Eric Sims

{u1136767, u1247999}@utah.edu

Fall 2022

Table of Contents

Table of Contents

Overview and Motivation

Related Work

Questions

Data
- Data Source
- Data Cleanup

Exploratory Data Analysis

Design Evolution
- Initial Designs
- Final Designs
  - Map
  - Distributions
  - Attribute Network
  - Filter
  - Bubble Chart
  - Heat Map

Implementation
- The Filter
- The Map
- The Attribute Network
- The Distributions
- The Heat Map
- The Bubble Chart

Evolution

Appendix A: Project Peer Feedback with Kyle Woods and Kasidy Fernandes

Overview and Motivation

It seems to be a common conversation topic of how terrible “Utah Drivers” are. How true is that? We believe that Utah roads are plagued with people going too fast, running red lights, disobeying traffic laws or being just plain ignorant of their surroundings on the road. We think it is worth considering looking at the accidents that have happened in Utah in a given year to determine what are the most common types of crashes and when and where they tend to happen.

We have heard that it is best to avoid the roads at night, but is that true? When do most DUI’s happen? Are teenage-involved accidents more common in urban areas than rural areas? Do elderly drivers become more involved in accidents during times of inclement weather?

Related Work

We were inspired by this work by just finding the data in the first place. We did not find any similar visualizations that had a similar story than what our plan has. While implementing the system we did find some accident related work including one visualization for shipwreck data.

Pie chart clustering visualization: http://bl.ocks.org/gisminister/10001728
Shipwreck visualization: https://bl.ocks.org/Andrew-Reid/21ff4b57267fa91dacc57ef1ccb7afb3

Questions

With this data visualization we are trying to answer many questions about diving and accidents in Utah in general, but specifically we want to show the correlation between different types of accidents, when they are likely to happen and what the causes of them are. With this data law enforcement will be able to better predict optimal times to be watching the streets and citizens using the Utah roadways will be more prepared and aware of potentially dangerous times to be on the road and locations with higher than average accident rates.

Data

Data Source

The data we are using is from this website:
https://dev.socrata.com/foundry/opendata.utah.gov/herb-zqda. Socrata is an open source data site that has hundreds of sets of collected data. We are focusing on the Utah data crash info that contains information of crashes from thousands of car accidents across a few years spanning from 2017 to to 2019.

Data Cleanup

The data arrived in an extremely usable format, the CSV file provided easily converts to a large array of JSON objects that is quite straightforward to use. The one data issue we found is that the location data is stored in UTM and our implementation through the google maps api required latitude and longitude. We created a python script to take the initial csv and convert the UTM coordinates to new "lat" and "lng" columns in a new CSV file.

We eventually created a script to map the relationships of crash attributes to each other, and how often they occur. We found that d3 and javascript couldn't handle this processing of 250,000 data points in an O(n^3) algorithm, so we created a python script to create a json object of the relationships. The python program took ten minutes to complete. The results of this are in "data/links.json" in the project.

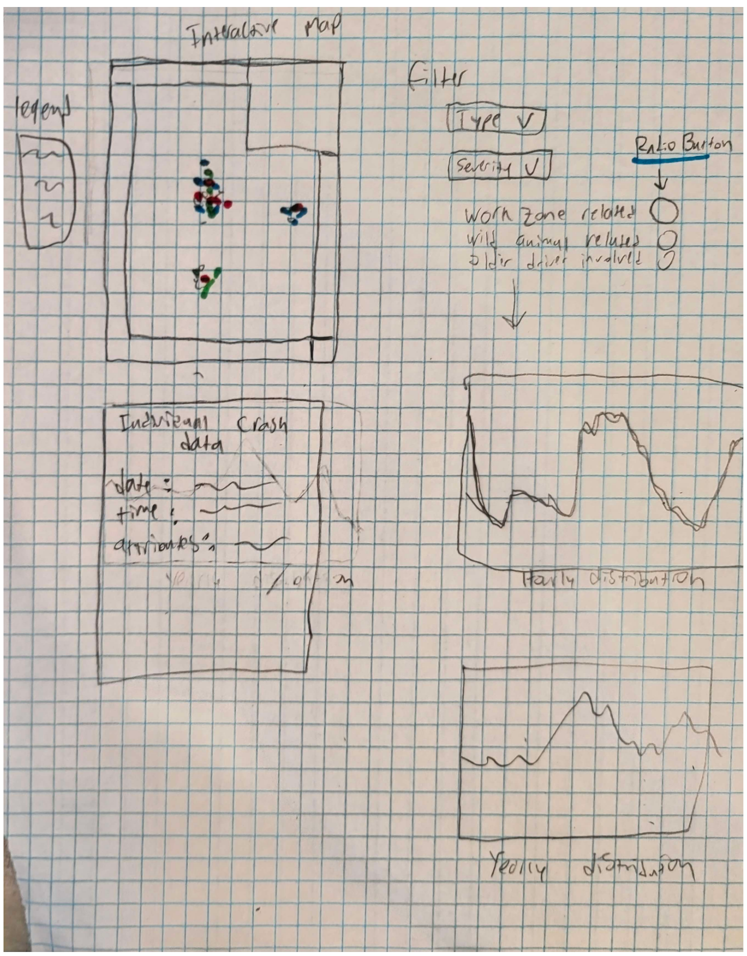
Exploratory Data Analysis

The data we found had an online tool showing the number of car accidents happening per month over time. So we could see that there was somewhat a pattern concerning the seasons and the number of crashes. We did some manual filtering of the data to see the relationships of different accident attributes, and we found certain answers like "teenage driver involved accidents happen frequently at intersections". This gained our confidence to move forward with the project, knowing that helpful insights could be gained.

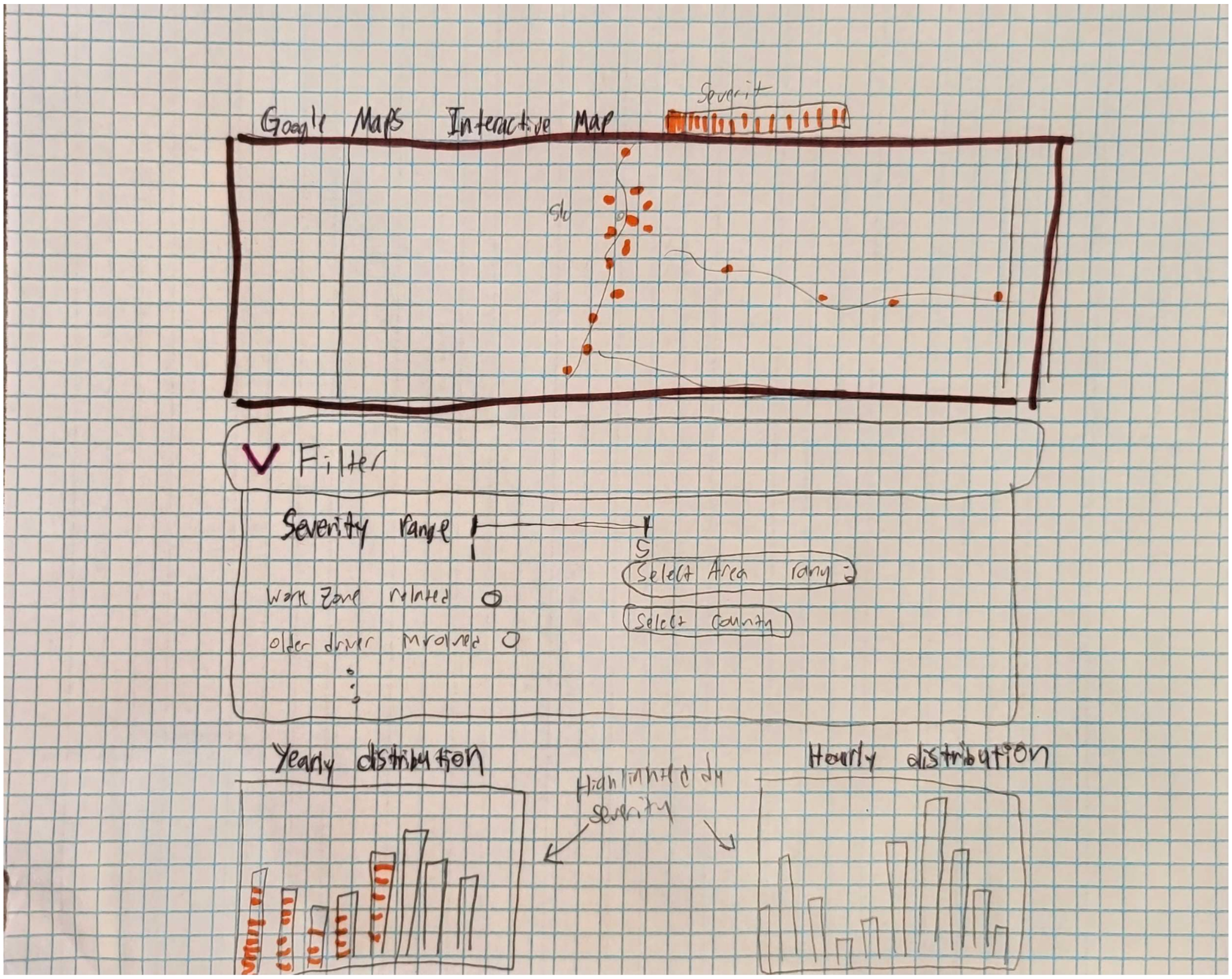
Design Evolution

Initial Designs

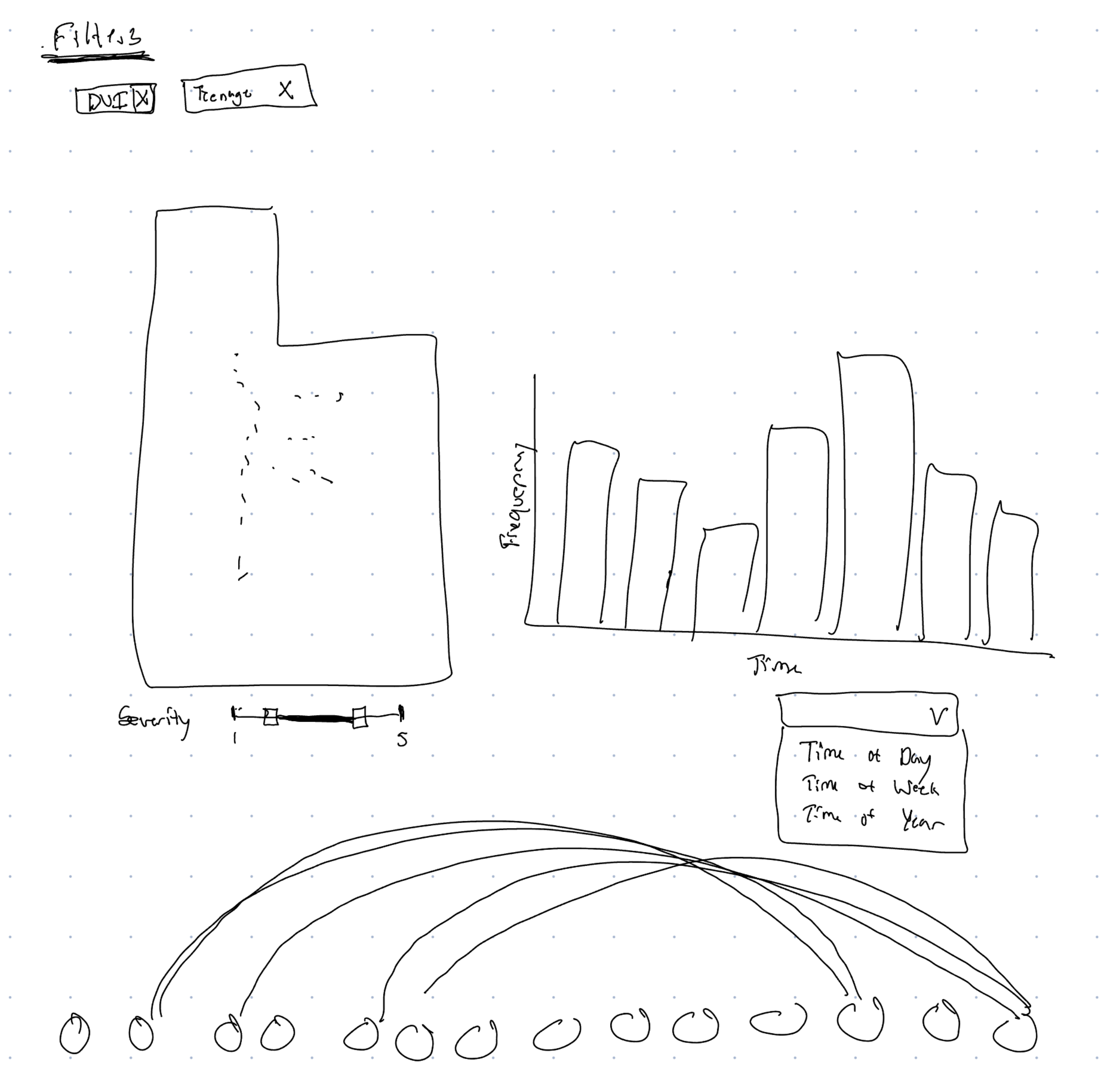
We knew that we wanted to have a few key features in our visualization. First, we wanted to have an interactive map that plots where the crashes have been. We also wanted a filtering area that users could focus on certain types of crashes. Furthermore, we wanted a detailed page for a selected crash and finally we wanted bar charts showing the most common times that crashes tend to occur.



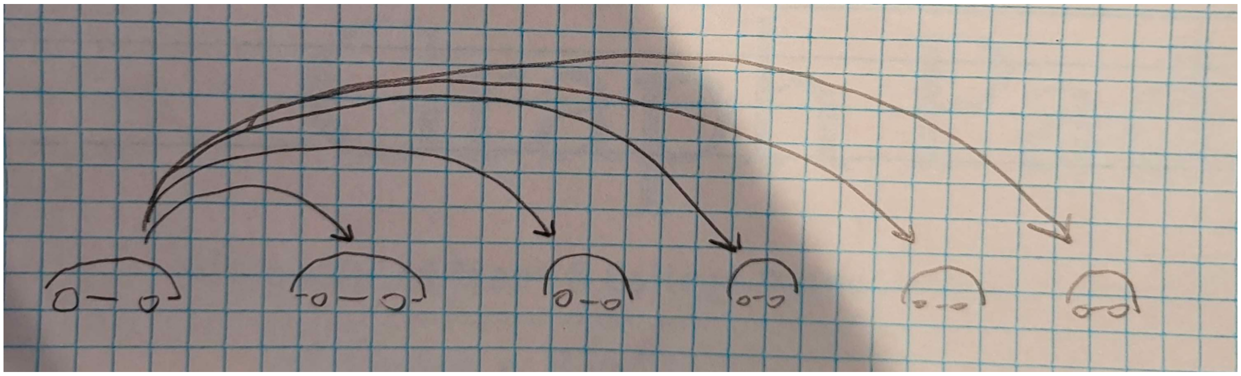
Prototype #1 - This was a pretty good home run of what our final design is. Although it did not consider a data derivation.

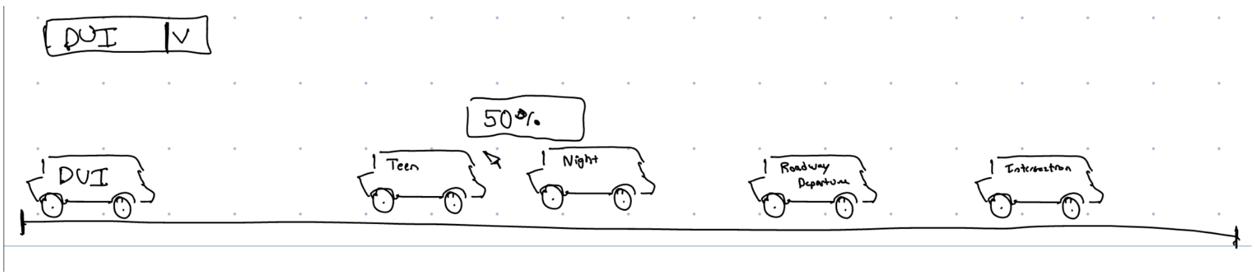


Prototype #2 - Experimenting with using a severity scale on the map to differentiate the accident plots

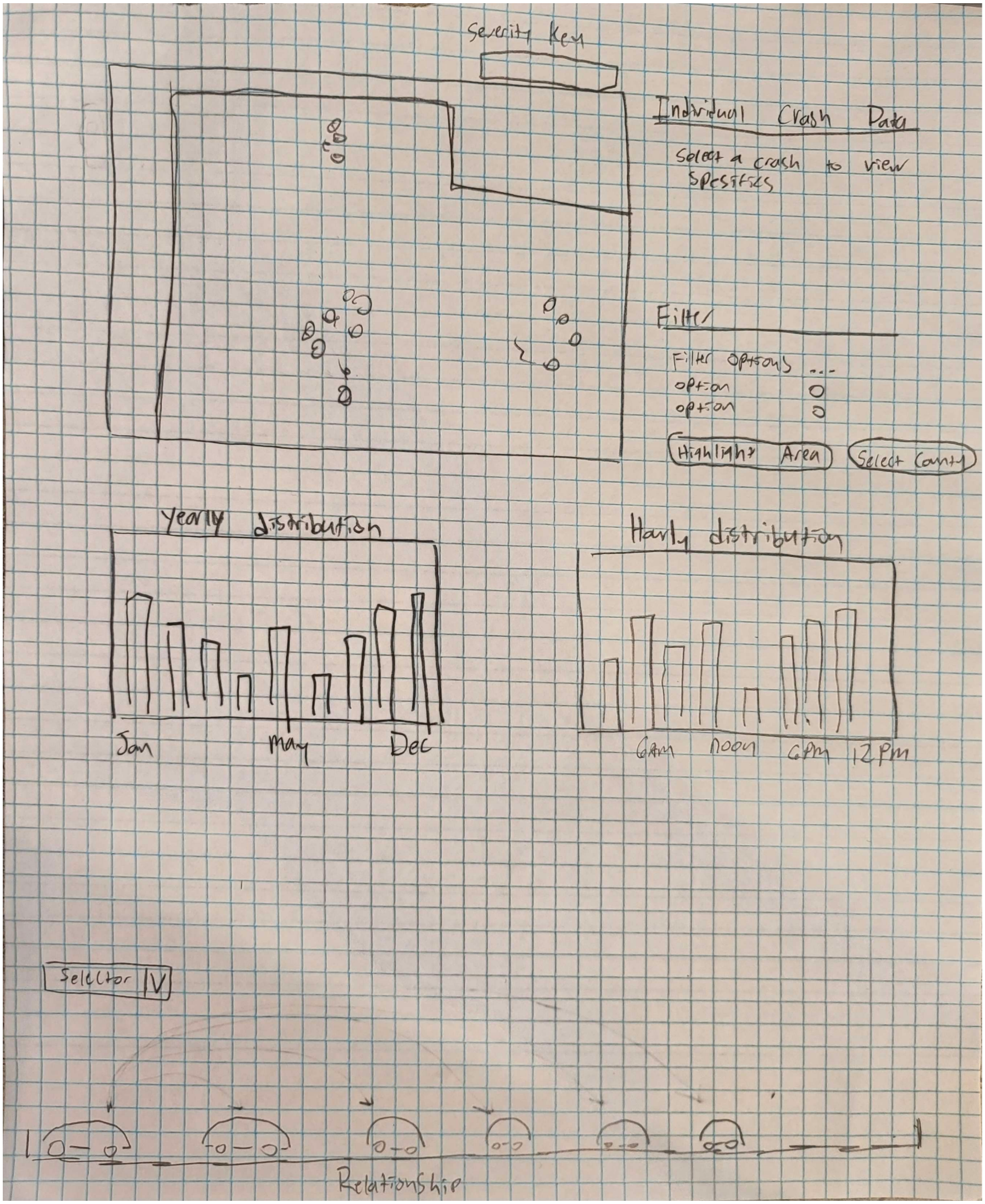


Prototype #3 - General layout design being finalized, brainstorm idea of data derivation





Required Feature 5 prototyping: Relationship Correlations, mouse hover to show percentage correlation between selected feature

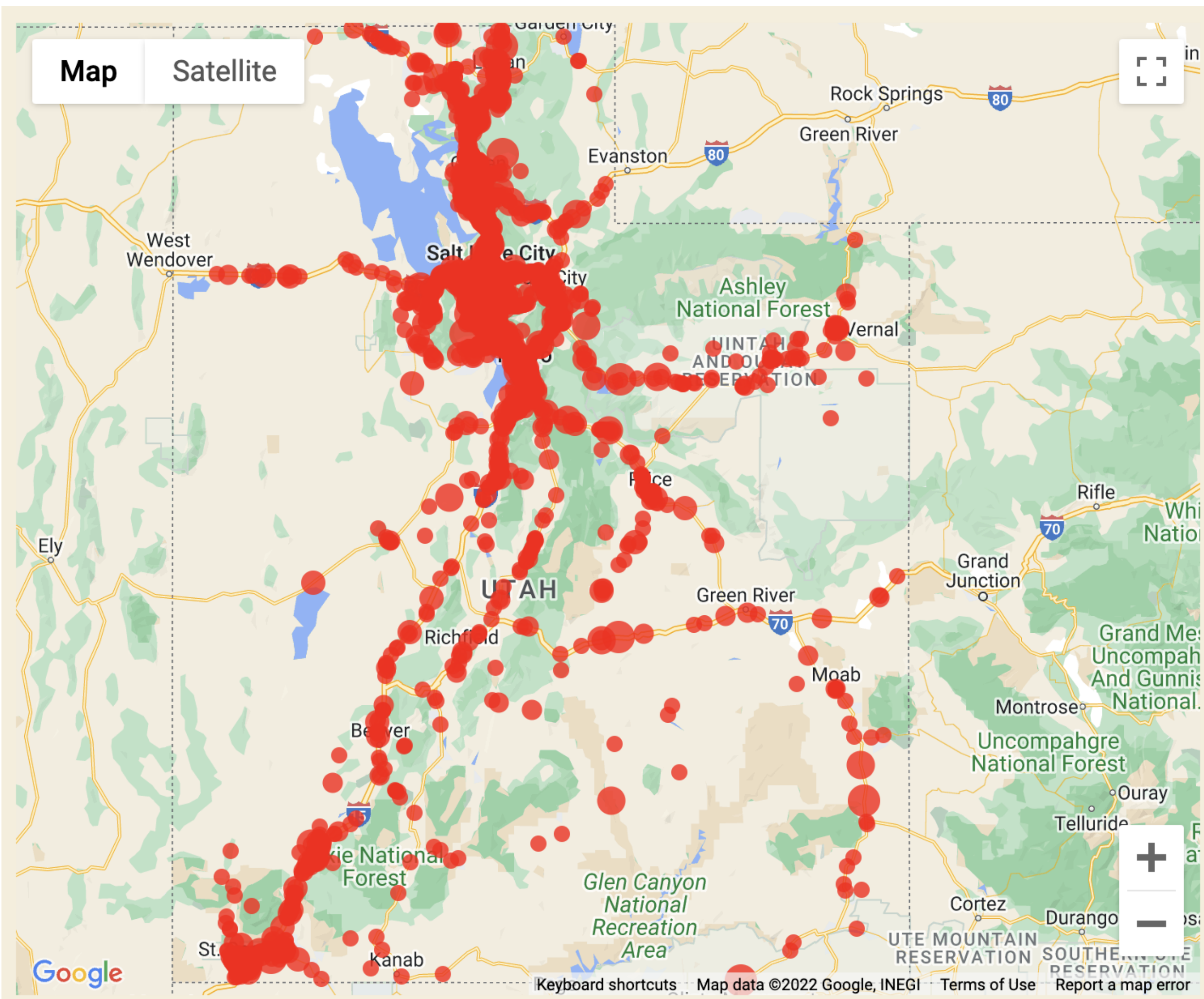


Final Design- relationship correlation display at bottom, ⅔ screen interactive Google map with overlaid accident data and individual crash data and filter options next to it and above Yearly and hourly distribution plots.

Final Designs

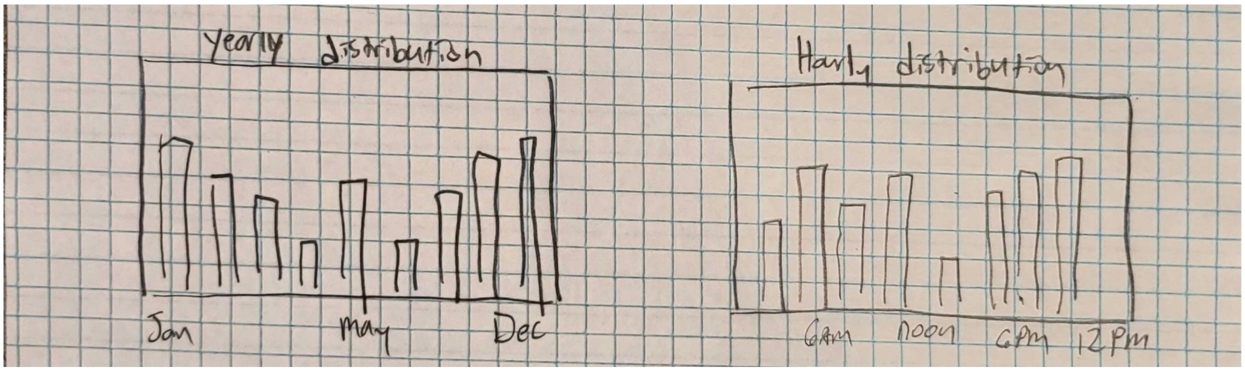
Map

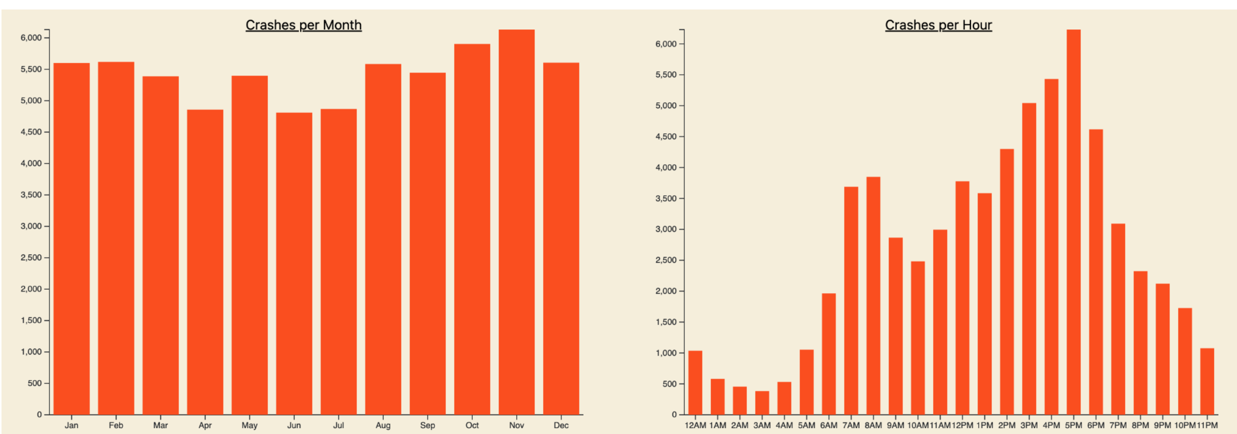
The map was the first item we implemented because we expected it to be the most difficult to finish. The full functionality was quite straightforward in actuality and it changed very little from beginning to end. We did try to tweak the design multiple times but we found that changes like data clustering or using other channels for the accidents was not as effective in showing the data as our initial design.



Distributions

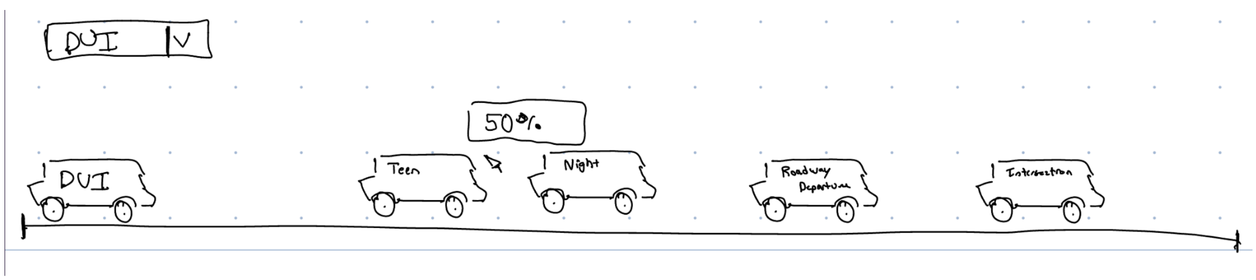
Between the final implementation and the initial design the hourly and monthly distribution charts did not change, we found this to be the most effective and interesting way to show this data.

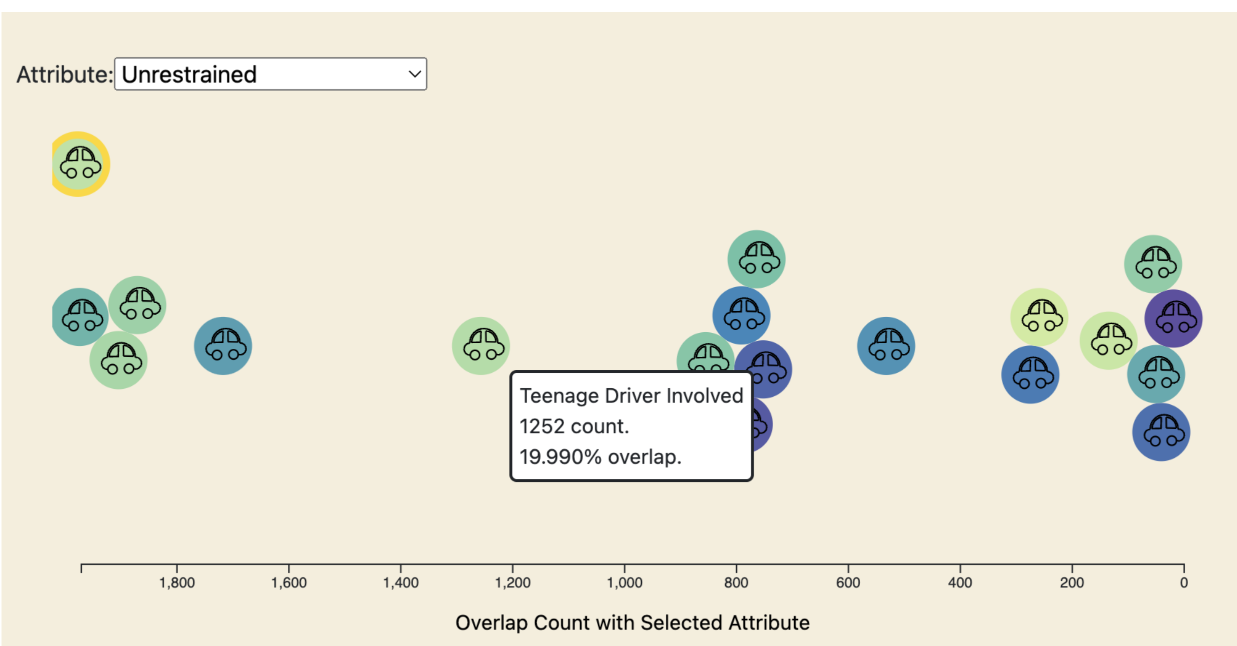




Attribute Network

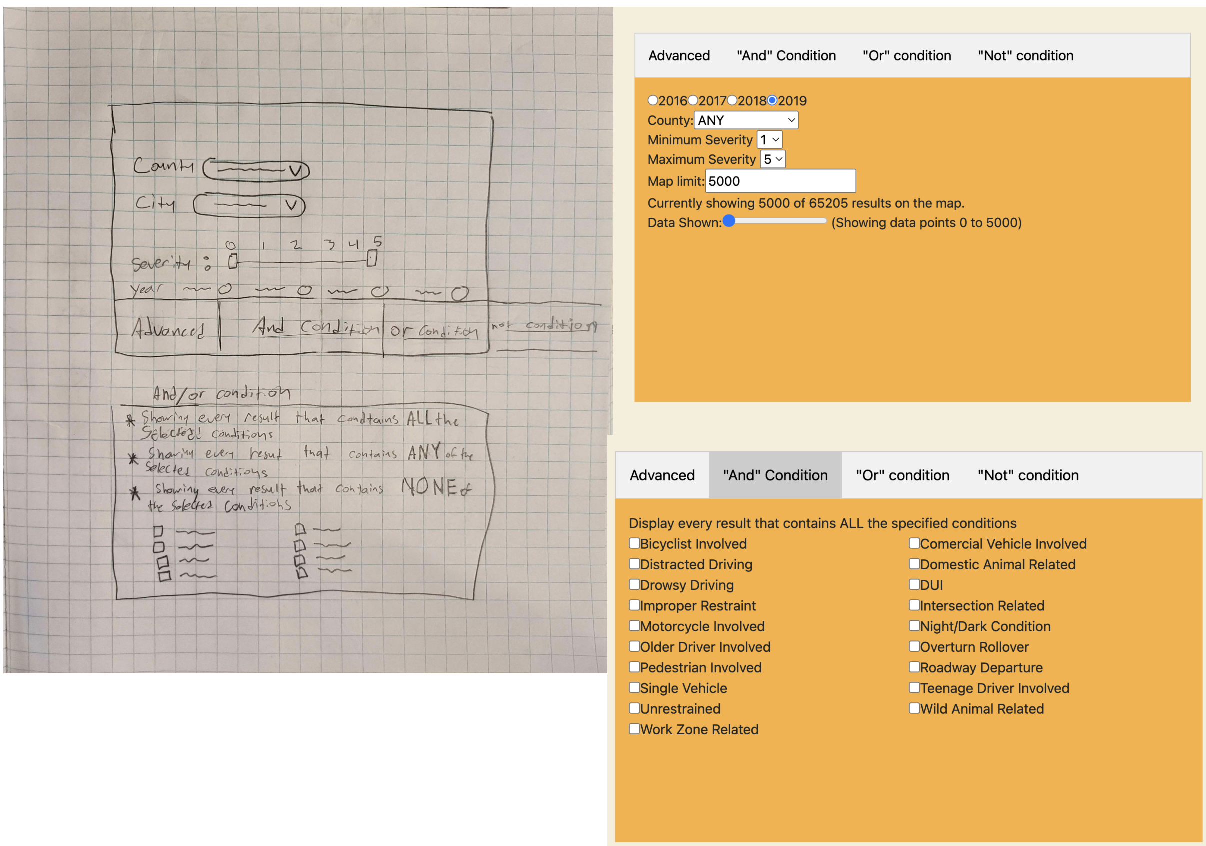
The attribute network ended up being the item that took the most work to look similar to our design. We tried multiple other designs to portray the data including an arc diagram and a heatmap and while the arc diagram did not make it to the final product the heat map we found to be a great way to see the whole picture all at once, but still included this as a detailed view of each individual attribute. The design of having the marks look like little cartoon cars ended up being much more difficult than expected but does add a lot to the overall aesthetic of the project. The colors of the cars in the final design have no correlation to the data, they are just a scale that looks good with the background. The filtering does not interfere with this visualization. It takes into calculation all of the accidents in the dataset.





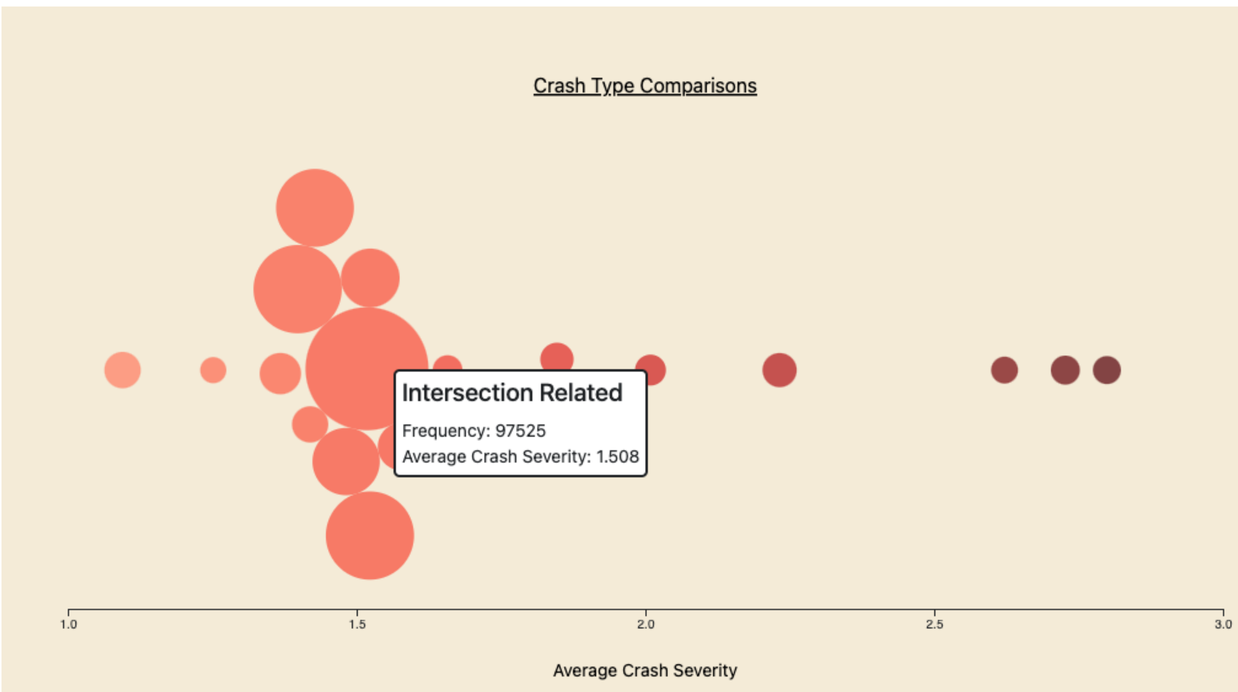
Filter

The filter ended up remaining similar in function and design from the initial plan of the project to the final implementation. The following image is of the final paper draft done for the filter before implementing the full functionality, the addition of the map limit was not something we had foreseen but ended up being a very simple addition to the filter. (See The Map in the implementation for more on the map limit)



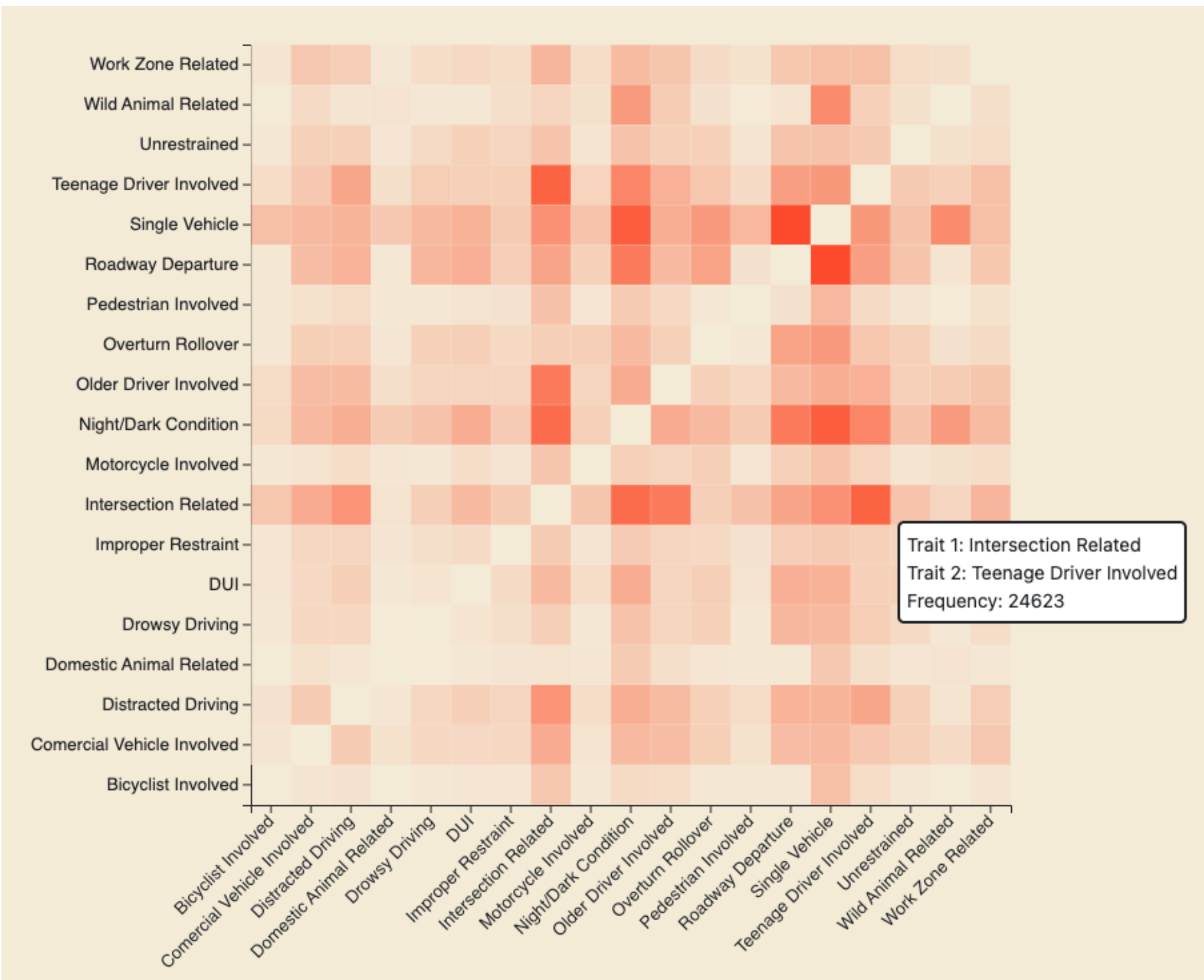
Bubble Chart

The bubble chart was a design decision decided later in development. We thought of a question that had not been answered in any of the other visualizations: “Which accident attributes averaged higher severities than others?” We also wanted to consider how often these accidents occur as well. Therefore we made a bubble chart. The bottom axis represented the severity of the accident, while the radius of the bubble represented how often they occur. This is an addition to our original project proposal, but one we thought was helpful in users gaining more insight.The filtering does not interfere with this visualization. It takes into calculation all of the accidents in the dataset.



Heat Map

The heat map is a two dimensional 30,000 view of the same data represented in the attribute network. Through this heatmap, we are able to compare the frequencies of any pairs of accidents at once. The darker the color, the more often the attributes are associated with one another in the whole data set. The filtering does not interfere with this visualization. It takes into calculation all of the accidents in the dataset.



Implementation

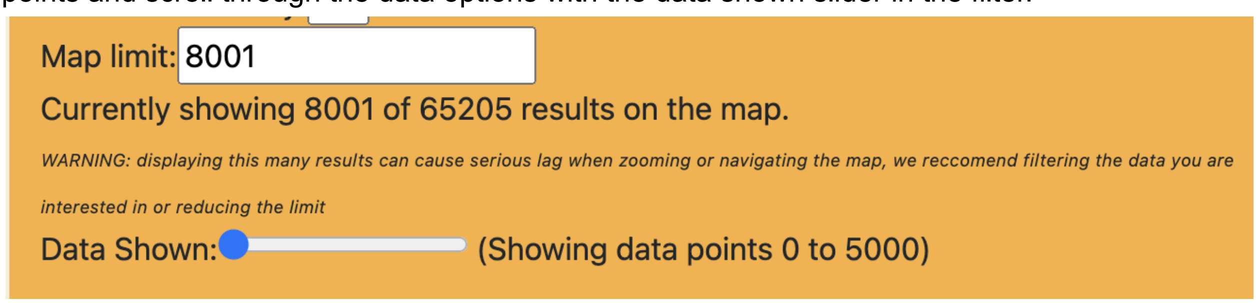
The Filter

The filter performs all of the functionality for changing which data is shown on the map and on the yearly and hourly distributions, the options for filtering include: severity, county, year as well as ‘or’, ‘and’ or ‘not’ filter on any of the boolean attributes recorded for the accidents meaning you can show only accidents in 2018 in Salt Lake county that involved a wild animal but did not happen during night or dark conditions.

The Map

On the map each circle represents an accident, and clicking on that circle will pull up the information for that accident above the filter.

The data that we are using has a total of 250,000 accident data points. Even split into 4 different years drawing this many points on a map is unreasonable. We had to find a way to show this data in a way that still allows us to draw conclusions from the data without leaving too much out. The first thought was to cluster the accidents into larger bubbles as the map was zoomed out and split into smaller bubbles as you zoom in, this may work for some other distributions but we want to show exactly what roads these accidents are happening on and just having large bubbles over each city in Utah was not useful data. The solution that we landed on was to simply allow the user to select how many data points are displayed on the map at a time, thus creating the ‘map limit’. If the map is stationary you ‘can’ draw every data point so we allow the viewer to do this, then if they want to move around they can reduce the number of drawn data points and scroll through the data options with the data shown slider in the filter.



The Attribute Network

The attribute network is a dynamic bubble chart that orders the attributes of car accidents according to how commonly they happen together. This allows you to select an attribute like DUI and see how often an accident involving a DUI also involves an underage driver or an unrestrained person. The bubbles representing the non selected attributes will be distributed across the y axis while the selected attribute will move to the top right corner and be highlighted in gold. Hovering over the individual attributes will show you the count of accidents that contain both attributes and the percentage of that overlap.

The Distributions

The distribution charts show the car accident data related to the time of day or the time of the year (in hours and months respectively). This reveals patterns about when accidents are most likely to occur. When an accident attribute (the “and”, “or” and “not”) filters are selected, the distributions are updated to show those respectively. The data is not limited by the map limit, and therefore is not affected by it. The bars can be clicked on and corresponding marks on the map are highlighted. A tooltip shows the number of accidents when hovering over a bar. See screenshot in design evolution.

The Heat Map

The heat map shows a two dimensional view of the same information from the attribute network. From here, a user will be able to know what pairs of accident attributes happen most often. A mouse tooltip helps the user to see what box represents which relationship and how often they happen in the data. The darker the box is, the more often it happens in the data. See screenshot in design evolution.

The Bubble Chart

The bubble chart represents both the average intensity of a crash type as well as how often that individual type of accident has occurred in the data. None of the averages were higher than 3. The bottom axis represents the average intensity while the radius of the circle represents how often it occurred. A tooltip shows information about what it is, the frequency and the average intensity. See screenshot in design evolution.

Evolution

We learned a few interesting patterns with our data. Here are a few examples:

- Car accidents happen most often during rush hour times (6AM-8AM and 4PM-6PM).
- Intersection related accidents are most often associated with night/dark conditions, teenage drivers or elderly drivers.
- Often, night/dark condition accidents are road departures.
- On average, pedestrian-related accidents are the most severe, but thankfully they are not the most common accidents.
- The most common accidents are intersection related accidents.

- In 2019 exactly 2 incidents involved a pedestrian and a domestic animal while across all of the data none involved a cyclist and a domestic animal

For a Utah driver, their questions on “when should I avoid the roads” or “where should I avoid driving” are definitely answered in this visualization. Furthermore, we believe that Utah driver educators could gain insights on what new drivers could be struggling with (i.e. intersectional accidents). Furthermore, law enforcement could use this visualization to determine what parts of roads should be monitored in order to create safer roads and issue more tickets.

If we had time to improve our visualization, we would work on having it be able to handle even more data. We struggled to get our map to display the data points that we wanted and we could not find a solution to cluster the data points together in time that we felt presented the data the way we wanted it. Furthermore, if we wanted to push this project further, we could possibly find a data feed of Utah car accidents that gets continually updated, rather than a report that captures accidents from 2016-2019. In addition the map limit and the data shown slider could dynamically update the map as the selection is moved but we did not have the time to implement or fully design that functionality.

Appendix A: Project Peer Feedback with Kyle Woods and Kasidy Fernandes

General observations from Kyle and Kassidy

-The project looks good and interesting

-The split between Must-Have features and Optional features looks balanced

-The data looks simple and good to work with, but the ideas you have should be scalable with more data quite easily

Suggested Changes (and our responses)

-There should be more data, getting data from multiple years will help with your time distributions, not to mention 2020 will likely be an outlier for driving/accident data because of the pandemic.

-Our assessment: This is a good suggestion, the place we found our data in only has the data from 2020 but as a part of our goals for our first week we want to get the data for multiple years as well as 2020 so that our averages better reflect and can predict current driving patterns.

(Update: It turns out that our data contains crashes from 2016 to 2019, even though it is called “Utah 2020 Crash Data”. What is interesting is that it seems that only 2019 has a full dump of crash data with 2016-2018 being sparse.

-Is there an ability to dynamically add new accidents as they happen?

-Our assessment: While this would be an interesting feature it lies outside the scope of what we believe our assignment should cover. The data we have has been curated to specifically be in the format we need it and the original source of the data may not be even available for a certain period of time. We will be increasing our data set but not to this extent.

Feedback analysis

Kasidy and Kyle did a great job with their feedback, they were insightful with their suggestions and their general observations were positive and encouraging. They could have been a bit more critical of our design look or had more suggestions for us to work with but their contributions were sufficient.