Traffic Accident Visualization

Course	CSE 5520 Fall 2021, Data Visualization of Communication
Author Name:	Lynn Pepin
Date:	December the 19th, 2021

Introduction

> 1 page. What is your project about? What are you trying to communicate by publishing your Dashboard? What data analytics visualization are you offering?

> **In short:** This is a project about traffic accidents in the United States. The purpose is to provide an enviornment for data-exploration. This dashboard provides several visualizations helping people view the distribution of traffic accidents, and how the severity and location of accidents differ by common traffic controls (such as stop signs or crossings) and by weather conditions (such as temperature or visibility.)

This is a project focusing on traffic accidents within the United States. Specifically, traffic accidents from 48 states (Alaska and Hawaii excluded), from February 2016 to December 2020. The goal is to provide the user with tools for *data exploration*.

Most people are aware that being near car roads or driving on car roads is, statistically, one of the most dangerous things we do. Many of us have had people close to us injured or lost to car accidents, or have been injured ourselves. Car collisions are sudden by nature, making them different than other leading causes of death such as heart disease or cancer. This is all the more personal for all of us after the recent and tragic loss of one of our students, Nhuong Nguyen, to a traffic collision on November 30th.

There is not a central thesis to this dashboard, Rather than lead users to specific conclusions, the goal of this dashboard is data-exploration, to allow users to learn by 'playing' with data. This kind of approach helps us understand our personal experiences and contextualize them within the greater world.

The dashboard allows people to explore the location and incidence of traffic accidents, and how the severity of accidents differ with certain key variables. Key variables that are considered are the presence of traffic controls (such as stop signs, roundabouts, crossing, 'traffic calming', etc.) as well as weather conditions (such as temperature, humidity, precipitation, visibility, etc.)

This dashboard offers dashboards to visualize the distribution of accidents by these key variables (that is, the incidence of them through box-plot/violin-plot/KDE/histogram) and the locations (per latitude-longitude. The impact of accidents are measured by severity (in how much the collision delays traffic) and by the distance of road impacted.

There are two modals included to allow the user to make concrete conclusions: Hypothesis testing and regression studio. The hypothesis testing allows users to determine whether certain traffic control measures provide a statistically significant difference on the severity and distance of collisions. The regression studio allows users to determine if there is a meaningful correlation between any two key variables, as well as perform polynomial regression against these two variables. The regression studio also provides the Pearson correlation coefficient. This allows the user to quantify suspected correlations and visualize them easily.

Dataset Description

> 0.5 pages. What are the specifics of data sets you are using? What real world (public/private) data or artificial (fake) data are you using? What are the reasons you choose them?

> In short: This is a spreadsheet containing information about 1.5 million accidents over 4 years in the US. I used only this one real-world dataset. I chose this because I think it is important to help visualize and communicate one part of the large negative impact vehicles play in our life.

The dataset is "US Accidents (updated)" by Sobhan Moosavi on Kaggle, described as "A Countrywide Traffic Accident Dataset (2016 - 2020)".

This dataset provides a CSV containing 1.5 million rows corresponding to traffic accidents in the United States between February 2016 to December 2020, compiled using various state-provided APIs. Each accident has the start and end geocoordinates provided, the date and time, the severity measure (of impact on traffic flow), the distance of road impacted, as well as measurements of various road conditions during the accident (such as whether it occured near certain traffic controls) and various weather and meteorological conditions (such as temperature, precipitation, humidity, and what time of day it occured at.)

I chose this dataset because the issue is important to me. In the United States, law requires learning motorists (often young) spend 40 hours driving on the road before obtaining a license. No tests later in life are ncessary. Owning a vehicle is so vital that judicial system is often hesitant to rescind these licenses permanently. More recently, factors such as LED lights and the 2020 post-lockdown increase in driving speeds may be accelerating these inherent problems.

Even ignoring the pressing climate issues, I have very deeply-held concerns regarding the vehicle culture in the United States. I could not find datasets specifically about costs, injuries, deaths, and secondary effects of traffic accidents, so I chose this dataset.

System Functionality

> 1.5 pages. What specific data visualization does your website offer? In writing, make references to figures with proper numbers that you will include in Section 4. In your writing, make sure you include what "data analytics" your Dashboard addresses with plots/graphs? You are not showing the graphs and plots for the sake of pretty figures. What are the analytic (computation) goals behind the figures?

This project offers six main, key modules. I'll describe each.

Part 1: Box/violin plots.

This module provides box and violin plots allowing the users to estimate how the 'Distance' of an accident (as measured in miles-of-road-impacted) differs according to different traffic controls (such as stop signs, traffic calming techniques, etc.) The X-axis has "false" or "true" (for the presence of the control) while the Y-axis measures the distance impacted.

Part 2: Histogram and KDE

This module provides a histogram-plus-KDE plot (with margin visualization) of the distribution of weather conditions during traffic accidents. Users can set different bin-sizes. For example, one might use this module to see the distribution of visibility during traffic accidents. Most accidents happen at 10-mile visibility (the max, and the norm), but there is a notable bump in the KDE (and histogram) around the 0.0 mile mark.

Part 3: Hypothesis testing

This module performs hypothesis testing. It plots the t-distribution of the two populations, as well as accompanyinh box-plot and p-score. The magic of this module is that it is modular: The user can specific one of four different independent variables (traffic accidents with or without stop signs, with or without crossings, with or without calming, and those with or without traffic signals) against one of two independent variables (distance or severity).

Effectively, this visualizes, tests, and answers eight hypotheses.

Part 4: Accidents map

This modal plots accidents on a map according to specific subsets. That is to say, it allows users to plot traffic stops with certain severity levels and certain traffic controls. This

effectively ties-back to the first module, but now allowing users to physically and tangibly *see* where accidents happen.

Part 5: Clustering accidents

To be honest, this part I included just to fulfill the rubrik, and because it is a bit pretty. However, there is significant technical depth for the *presentation* of coloring. A relevant stackexchange is here: https://gamedev.stackexchange.com/questions/46463/how-can-i-find-an-optimum-set-of-colors-for-10-players

We found $k = 12$ to be a reasonable and optimal number of clusters during our homework. Choosing colors for an arbitrary number of k points is difficult. The most reasonable choice is to generate colors where, for arbitrary and small k, the hues will be reasonably evenly distributed across the circle. This is accomplished using the equidistribution theorem with a modulus of the golden ratio (i.e., 1.618034.)

Unfortunately, plotly provides only HSV colorspace and not HSL or more modern color-spaces, so the perceptual brightness varies across hues on most monitors (peaking at green.) Implementing code for alternative color-spaces would be too much work for this project, as would wrangling color-calibration in sRGB across everyones monitors.)

Part 6: Regression studio

This part gives the users hard quantitative tools just like in Part 3. This part is the most complex and so it is saved for last, for when users are more experienced with the data. The regression studio prompts users to pick an X axis variable (such as precipitation) and a Y axis variable (such as severity of an accident) for regression. The graph outputs a regression between these two variables, printing out the plotting formula and the Pearson correlation coefficient.

In addition to this, the regression is polynomial. The user can choose linear, quadratic, cubic, or higher polynomial regression, to explore more complex relationships in the data. The full polynomial is plotted (albeit with rounded coefficients.)

Screenshot Attachment

7 pages max.

You can include multiple pages of screenshots as needed up to 7 pages. Each screenshot should have a figure number (e.g., Figure 1, Figure 2, etc.) and each should have a short annotation explaining what the plot/graph is about (i.e., Figure caption). The figure numbers should be referred to in the body of the paper as needed.

Accomplishment Summary

Use the 1 page Excel sheet.

Categories	Visualization Methods	Basic	Medium	Intensive
Clustering:	Hierarchical	X		
	K-means	X		
Classification:	Gaussian mixed model	X		
Network Analysis:	Network Visualization	X		
Correlation Analysis:	Linear regression			X
	Pearson correlation			X
	Kernel Density Estimate			X
Hypothesis Testing:	t-test		X	
	p-value		X	
Statistics:	Boxplots			X
	Violin plots			X
	Histogram			X
Geospatial Analysis:	Cartogram map		X	
NLP/Text Mining:	WordCloud	X		
	Barplot	X		
Basics:	Line graphs			X
Other:				

Notes: Hierarchical and K-means clustering were skipped for being redundant with GMM and computationally unnecessary. The linear-regression, t-test, p-value, and hypothesis testing categories include combinations of variables due to the interactive nature. The NLP section is not applicable, and the line-graphs are enhanced as part of the regression studio.

Extension summary

> 0.5 pages. You can summarize any "substantial" development effort since your presentation if you have done so

---

This is not a summary of included features, but a description of what I would add if I spent more time on this.

This dashboard is lacking due to time constraints I faced this semester. In particular, this dashboard does not provide exploration for temporal data, and does not provide tools to help users contextualize accidents across other variables. For example, the traffic accident map might as well be a population density map. A more polished dashboard would process this into accidents-per-100K-people.

Furthermore, only 10 thousand of the 1.5 million points are provided, and if I had the time and knowledge, I would "pre-bake" the visualizations of each module. Good visualization includes relative priors that help further contextualize these accidents. For example, an accident-per-population cartogram would be much more meaningful than an accident map.

Finally, this project should be expanded to include relative climate data, and perhaps visualizations relating to the prevalence and opportunity for public transport. This would establish a meaningfu narrative of "problem-problem-solution".

And, on a technical note, I would explore more-efficient WSGI servers for deployment (on my personal site), or even more-efficienct server alternatives to Dash (such as Rust Plotters).