How would you visualize your data?

Ruochen Pi
University of Sydney
Shanghai, China
ruochenpi@gmail.com

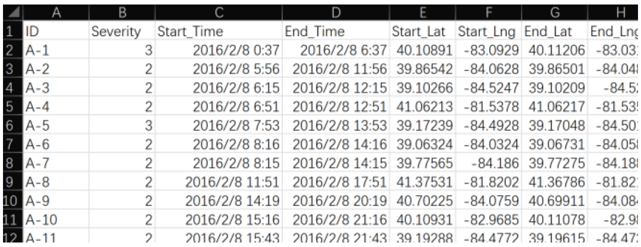
***Abstract*—This paper focuses on analyzing car accidents in the United States from February 2016 to December 2021. Through preprocessing such as attribute selection of the raw dataset, we select attributes of five different data types of car accidents for analysis and visualize them. For further analysis, we will show examples of visualization by symbolic representation and reintegrate the visualized data according to the changed symbols. The above data visualization work will help our target, the U.S. Department of Transportation, to analyze the data more effectively and reduce the number of future U.S. car accidents.**

Keywords—dataset, car accident, visualization, symbols, USA, typical consumer

I. DataSet

A. Dataset Introduction

The dataset is about U.S. car accidents covering 49 states with accident details from February 2016 to December 2021. The dataset is large, with 2.8 million pieces of data, including 47 attributes such as the time and place of the crash and the weather.

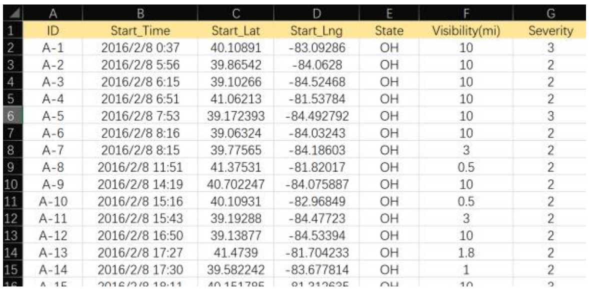


	A	B	C	D	E	F	G	H
1	ID	Severity	Start_Time	End_Time	Start_Lat	Start_Lng	End_Lat	End_Lng
2	A-1	3	2016/2/8 0:37	2016/2/8 6:37	40.10891	-83.0929	40.11206	-83.03
3	A-2	2	2016/2/8 5:56	2016/2/8 11:56	39.86542	-84.0628	39.86501	-84.04
4	A-3	2	2016/2/8 6:15	2016/2/8 12:15	39.10266	-84.5247	39.10209	-84.5
5	A-4	2	2016/2/8 6:51	2016/2/8 12:51	41.06213	-81.5378	41.06217	-81.53
6	A-5	3	2016/2/8 7:53	2016/2/8 13:53	39.17239	-84.4928	39.17048	-84.50
7	A-6	2	2016/2/8 8:16	2016/2/8 14:16	39.06324	-84.0324	39.06731	-84.05
8	A-7	2	2016/2/8 8:15	2016/2/8 14:15	39.77565	-84.186	39.77275	-84.18
9	A-8	2	2016/2/8 11:51	2016/2/8 17:51	41.37531	-81.8202	41.36786	-81.82
10	A-9	2	2016/2/8 14:19	2016/2/8 20:19	40.70225	-84.0759	40.69911	-84.08
11	A-10	2	2016/2/8 15:16	2016/2/8 21:16	40.10931	-82.9685	40.11078	-82.9
12	A-11	2	2016/2/8 15:43	2016/2/8 21:43	39.19288	-84.4772	39.19615	-84.47

Figure 1.1 Original dataset

B. Data Collection and Processing

This dataset was collected by the US government using data captured by traffic cameras and road network traffic sensors through an API interface. The data was published on Kaggle and I downloaded that data from the Kaggle website. I took 5 Colum (attributes) from this dataset, which are time, state, latitude and longitude, visibility, and severity.



	A	B	C	D	E	F	G
1	ID	Start_Time	Start_Lat	Start_Lng	State	Visibility(mi)	Severity
2	A-1	2016/2/8 0:37	40.10891	-83.09286	OH	10	3
3	A-2	2016/2/8 5:56	39.86542	-84.0628	OH	10	2
4	A-3	2016/2/8 6:15	39.10266	-84.52468	OH	10	2
5	A-4	2016/2/8 6:51	41.06213	-81.53784	OH	10	2
6	A-5	2016/2/8 7:53	39.172393	-84.492792	OH	10	3
7	A-6	2016/2/8 8:16	39.06324	-84.03243	OH	10	2
8	A-7	2016/2/8 8:15	39.77565	-84.18603	OH	3	2
9	A-8	2016/2/8 11:51	41.37531	-81.82017	OH	0.5	2
10	A-9	2016/2/8 14:19	40.702247	-84.075887	OH	10	2
11	A-10	2016/2/8 15:16	40.10931	-82.96849	OH	0.5	2
12	A-11	2016/2/8 15:43	39.19288	-84.47723	OH	3	2
13	A-12	2016/2/8 16:50	39.13877	-84.53394	OH	10	2
14	A-13	2016/2/8 17:27	41.4739	-81.704233	OH	1.8	2
15	A-14	2016/2/8 17:30	39.582242	-83.677814	OH	1	2
16	A-15	2016/2/8 18:11	40.151785	-81.312635	OH	10	2

Figure 1.2 Processed data set

C. Data details

- The time represents the time of the crash, in years, months, days and hours.
- State represents the state of the location where the crash occurred.
- Latitude and longitude represent the specific location of the crash on the map.
- Visibility represents the distance in miles that the driver could see at the time of the crash.
- The severity of the crash is expressed from 1 to 4, with 1 indicating minor severity and 4 indicating the most severe, judged by the time of the traffic jam.

II. Consumers

A. Who usually cares about the data

This dataset applies to all people, including the government, businesses, and the general public. Among them, government departments are most concerned about this type of data. This dataset responds to details about car accidents in the United States from 2016 to 2021, including environmental factors, time of day, and region. Various types of people use or analyze this dataset for different purposes.

B. Data consumers

The typical consumer of the dataset is the U.S. Department of Transportation. Their main responsibilities are to monitor road traffic violations, prevent and handle traffic accidents, maintain road traffic order, etc. These data consumers will use the data set to analyze the current situation of auto accidents and predict the future trend of auto accidents, so as to reduce the accident rate by increasing traffic lights, deploying more police officers, and imposing speed limits.

III. Data type

These data are 2.8 million by 5 information elements, which can be difficult when typical consumers read them directly. The best way is to visualize the data.

- Time is interval because it has the same interval and no real 0 value. In order to visualize this data, you can use the visual variables of values. After processing this data, time is categorized as a week, month, or year, and after statistical processing, the different positions of the Y-axis can be used to represent the trend of crash occurrence for different weeks, months, or years.
- State is nominal because it is a variable with no natural order or ranking. To visualize this data, visual variables of colour and size can be used, and the crash statistics for the state can be expressed in a bar chart.
- The latitude and longitude are intervals. in order to visualize this data, visual variables of location can be used. As an example, this dataset can show the exact location of a crash in America.
- Visibility is the ratio. To visualize this data, visual variables of colour, size, and value can be used.
- Severity is ordinal. To visualize this data, visual variables of value can be used. As an example, histograms can use shades of colour to indicate different categories.

I will use line graphs, bar graphs, and maps to describe to the audience. For this study, first I have a statistical classification for the time of car accident occurrence, we can use bar graphs to express the number of car accidents in different years and months. For the statistics of the state of the occurrence of car accidents I also choose to use bar graphs. For latitude and longitude, I choose to use a map to express it. For crash visibility and severity use, I will use a one-dimensional scatter plot for the statistics.

IV. Ask and answer the questions

- For time, answer the question of how many car accidents in the U.S. have occurred in recent years and what the trends are. For example, how much of car accidents were in the U.S. in 2018?
- For state, answer the question of how many crashes have occurred in each state in the U.S. in recent years. For example, how many car accidents in LA?
- For latitude and longitude, answer the question of how is the distribution of car crashes in the U.S. by location. For example, are there any locations with high crash rates? Where are they located?
- For visibility, answer the question of how many crashes in the U.S. in different visibility. For example, what is the highest number of crashes at what visibility level?
- For crash severity, answer the question of what are the statistics of crash severity in the U.S.? For example, the number of minor crashes compared to major crashes.

V. Typical problem

A. *Why do we need visualization*

Unprocessed datasets are not intuitive enough and the information is scattered. When we present an unprocessed US car accident dataset to the user, it is boring, and it does not attract the user's attention. It is not intuitive enough to express trends, and in general, is not very functional, which is a pile of scattered data. This is the difference between visualization data and raw data.

B. *Typical mistakes*

At the same time, there are some typical problems with visualization.

First, when we use inappropriate expressions in the process of visualization. For example, if the attribute is latitude and longitude, then it is best if we use a map, but using a bar chart is an inappropriate representation.

Second, the amount of information is too large. This can make the user feel dazzled and unable to grasp the focus.

Finally, not all information needs to be visualized. Sometimes it is better to show just one aggregated number than to show a complex chart of data. Especially when we have a large amount of data, it may be better to simply aggregate this data than to display each data point. For example, in the dataset studied in this paper, we can clearly know the year, month, day, and hour of each car accident. But users sometimes want to see a more general categorization, such as the number of crashes per year, or the number of crashes per month, rather than a list of every hour and minute.

VI. First symbolic representation

Using symbolic representation in graphical semiotics, this chapter describes and explains the imposition of visualization, and how to assign each data type to various visual variables.

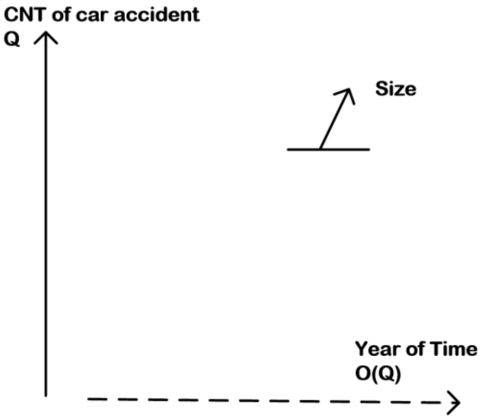


Figure 6.1 First symbolic of Time

The X-axis is the year of time, which is the dashed line. It is the heterogeneous dimension, which represents the categories are repeated several times. The Y-axis is the count of car accidents, which is the solid line. It is the homogeneous dimension, which represents the categories are established once and for all. The arrow size on the right side indicates that the visibility variable is size. Also, the symbol of O means Ordered, the data are ordered.

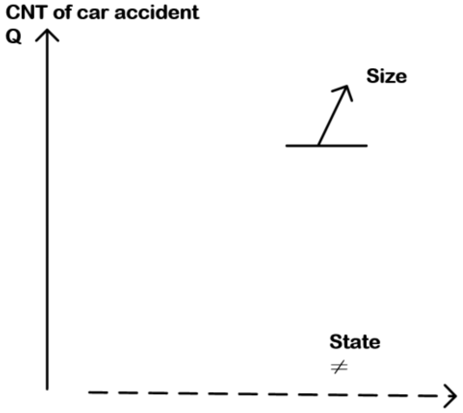


Figure 6.2 First symbolic of State

The X-axis is the state, which is the dashed line. It is the heterogeneous dimension, representing the categories are repeated several times. It is the homogeneous dimension representing the categories are established once and for all. The arrow 'size' on the right side indicates that the visual variable is size, and the symbol ≠ indicates the Selective perception that the data can be categorized.

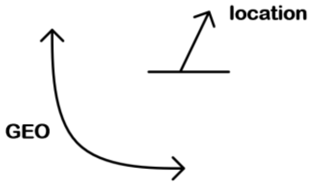


Figure 6.3 First symbolic of Longitude and Latitude

This is a map. The arrow 'location' indicates that the visual variable is location.

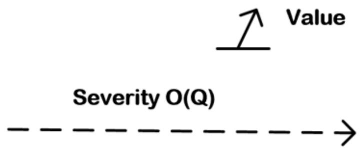


Figure 6.4 First symbolic of Visibility

The X-axis is visibility, which is a solid line. It is the homogeneous dimension representing the categories are established once and for all. Also, the symbol of O means Ordered, the data is ordered.

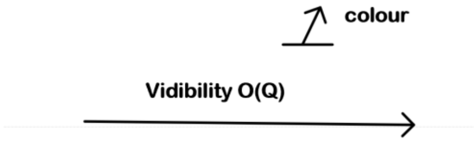


Figure 6.5 First symbolic of Severity

The X-axis is the severity, which is the dashed line. It is the heterogeneous dimension, representing The categories are repeated several times. Also, the symbol of O means Ordered, the data is ordered.

VII. First Visualization

This dataset is visualized through five different visual charts. Using the appropriate chart type relative to the attributes provides a better response to the data. Bar charts can represent data related to statistical counts, clearly expressing the magnitude of the values and helping to visualize the data for analysis. In addition, line charts can represent trends in the data, and maps can clearly represent the distribution of the data. The following visualization diagrams were obtained by analyzing Figures 6.1-6.5.

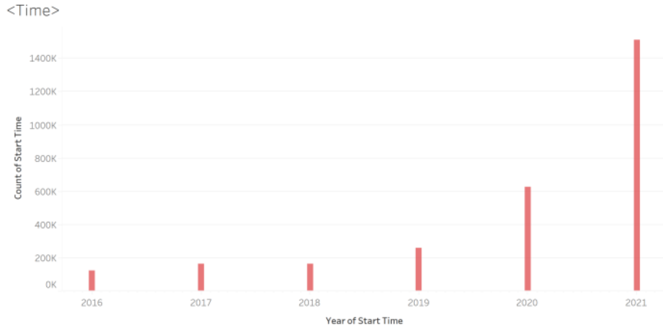


Figure 7.1 Visualization Chart of Time

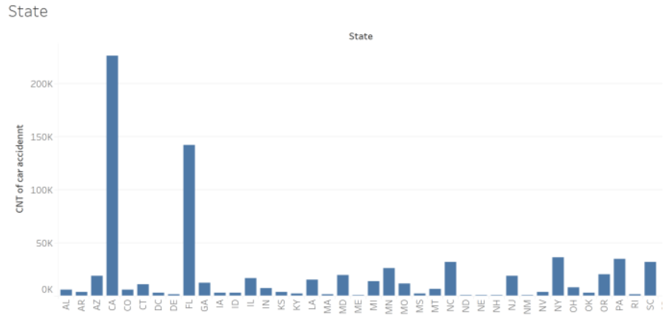


Figure 7.2 Visualization Chart of State

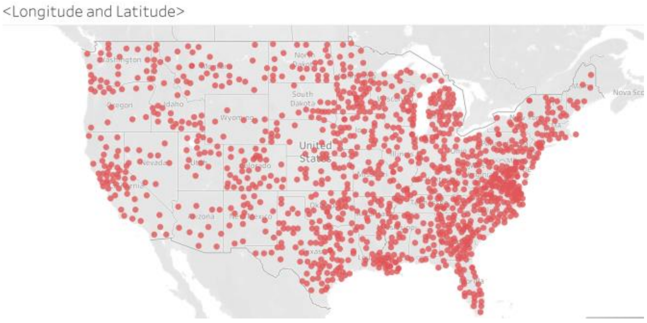


Figure 7.3 Visualization Chart of Longitude and Latitude

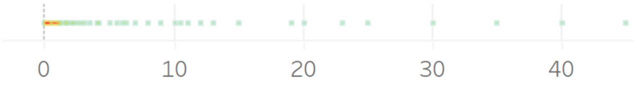


Figure 7.4 Visualization Chart of Visibility

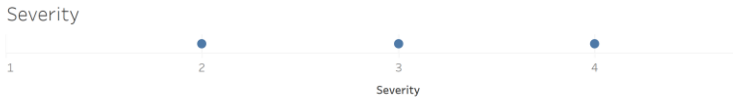


Figure 7.5 Visualization Chart of Severity

- The following are the visualization chart contents corresponding to the five attributes.
- The Time attribute uses a bar chart with the X-axis to reflect the change in time, and the X-axis ranges from February 2016 to December 2021.
- The State attribute uses the size of the bar area to represent the statistics of different states.
- The longitude and Latitude use circles to represent the distribution of crashes on the map.
- The visibility is represented by the circle.
- The severity is also represented by a circle, and the X-axis is severity, so you can see the distribution of severity as well.

VIII. Secondary Symbolic Representation

To better visualize this data, we optimize it. We try to represent Figure 6.1- Figure 6.5 with different symbols, the main way is merging as well as optimization. Most of the semantics in Chapter 8 is similar to Chapter 6.

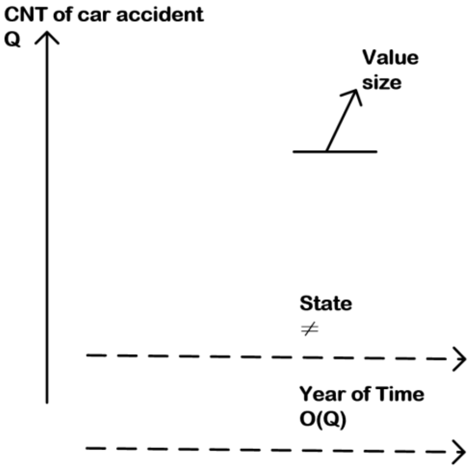


Figure8.1 Second symbolic of Time and State

Merge figure 6.1 and figure 6.2, the substance is the same as before. The X-axis is the year of time and state, both dashed, and the Y-axis is the count of car accidents, solid. The arrow size on the right indicates that the visible variable is size, and the symbol O indicates that year of time is ordered. ≠ indicates that the state is selective perception and the data can be classified.

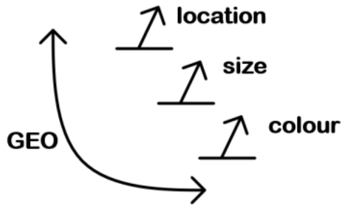


Figure8.2 Second symbolic of Longitude and Latitude

Optimize Figure 6.3 and add visual variables. Continue to use the map and select more visual variables, location, size and colour.

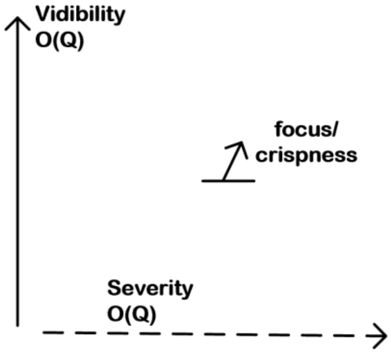


Figure8.3 Second symbolic of Visibility and Severity

Merge Figure 6.4 and Figure 6.5, the substance is the same as before. Change the X-axis of figure6.4 to Y-axis. X-axis is severity, which is a solid line. Y-axis is visibility, which is a solid line. Modify the visualization variable to focus/crispness. also, the symbol of O means Ordered and the data is ordered.

IX. SECONDARY VISUALIZATION

This chapter is an alternative visualization derived from the symbolic representation. In this chapter, we will describe how we assigned each data type to be used for the axes and various visual variables.

Finally, we visualize Figures 8.1-8.3 to obtain the following visualization charts.

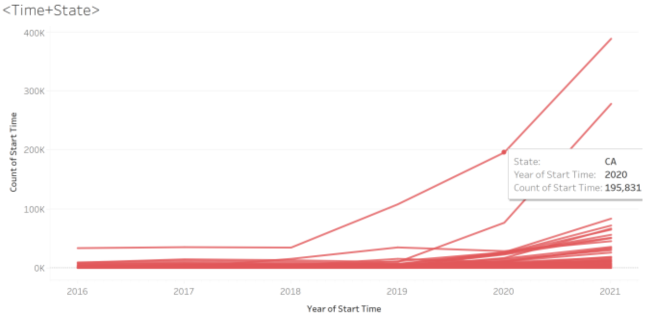


Figure 9.1 Visualization Chart of Time and State

This chart combines time and state in one visual chart, which is different from Figure 7.1 and Figure7.2. Users can analyze the crash situation of each state at different times. The data is processed from 2-D to 3-D. The X-axis represents time, and different folds represent the different states. Line chart instead of bar chart, which better reflects the trend of data size.

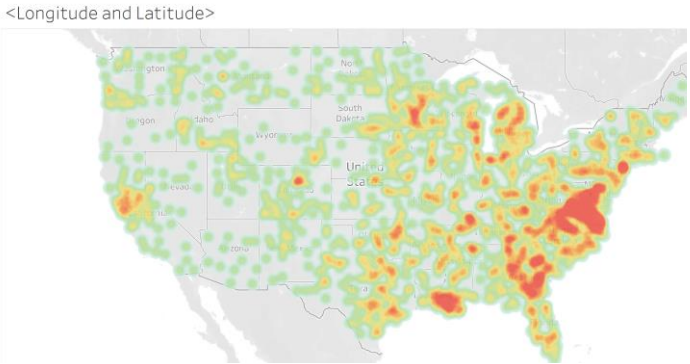


Figure 9.2 Visualization Chart of Longitude and Latitude

For the latitude and longitude attributes, the visualization chart is still presented in the form of a map. The difference is that the visualization variables are size, location, and color, which allows the user to get a clearer picture of where crashes occur, and the density of crashes in certain areas, making it easier to read which areas are the most frequent.

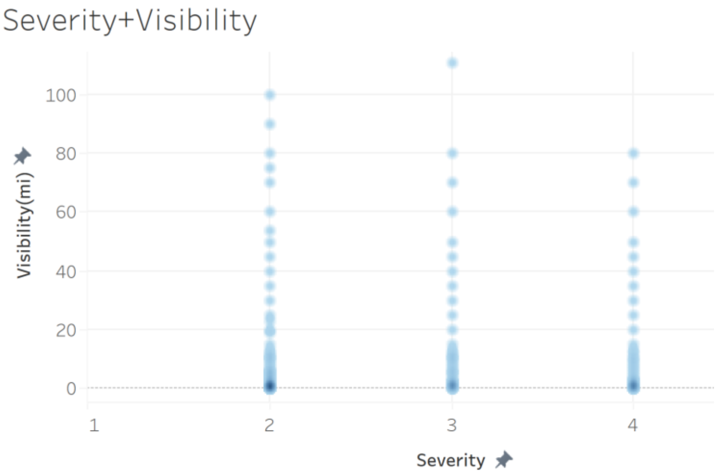


Figure 9.3 Visualization Chart of Severity + Visibility

For the attributes of severity and visibility, we combine these two attributes into a single visualization chart. The analysis of multiple dimensions allows more information to be learned. The visualization chart uses shades of colour to distinguish the distribution of the data. Darker colours mean that more crashes occur. For example, in this Figure 9.3, we can see that more crashes occur at the end of the visibility month.

REFERENCES

[1] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, and Rajiv Ramnath. “A Countrywide Traffic Accident Dataset.”, 2019.

[2] Moosavi, Sobhan, Mohammad Hossein Samavatian, Srinivasan Parthasarathy, Radu Teodorescu, and Rajiv Ramnath. "Accident Risk Prediction based on Heterogeneous Sparse Data: New Dataset and Insights." In proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, ACM, 2019.

[3] Masure, A. (2019). From Semiology of Graphics to Cultural Analytics: flaws in the mathematization of visible. *Abstracts of the ICA*.

[4] Dhieb, M. (2018). Translating Bertin into Arabic today: new hidden facets of Semiology of Graphics. *Cartography and Geographic Information Science*, *46*, 163 -