Traffic Collision Data Visualization in Los Angeles County

Bufan Zeng, Wing Wa Leung, Yunwei Zhang

University of Southern California

Abstract. Traffic collision data of Los Angeles county are analyzed and visualized using JavaScript and D3.js.

1 Introduction

Our project is designed to explore the traffic collisions data from *Los Angeles Open Data* [1] website. The design principle is mainly focused on providing charts and stories that are straightforward and easy to comprehend to the audiences. In a broad view, the trend of traffic collisions and some demographic information are visualized.

2 Data

The dataset is "Traffic Collision Data from 2010 to Present" from *Los Angeles Open Data* [1]. It has 18 attributes, which include geographic, time-series and demographic information about all traffic collisions occurred in Los Angeles county starting year 2010. There are over 400,000 rows of data where each row presents an accident record.

To visualize the geographic data, a GeoJSON file of Los Angeles county is needed. We obtained this file named "L.A. County Neighborhoods (V6)" from *Los Angeles Times* [2] website.

3 Related work and originality

Not many works have been found on similar topics about Los Angeles. We are largely inspired by the *Exploring NYC Vehicle Crash Data in Tableau* [3] website which explains the distribution of traffic collision data in New York City. However, the website uses Tableau to implement and lack of interactive features.

In our practice, we used JavaScript with D3.js to create all the charts and effects. Although not as easy to use as tools like Tableau and Google Charts, they are more powerful and able to provide more flexibility and diversity in visualization and interaction forms. The logic flow of our design (overview $\rightarrow$ trend $\rightarrow$ demographic information visualization) mimics story-telling process, which makes our audiences easier to feel engaged.

4 Design methodology and features

4.1 Framework and collaboration

We chose Angular as our framework. Each chart is an individual component in Angular. All of the graph plotting is implemented using JavaScript and D3.js. GitHub is used for us to collaborate with each other and manage version control.

4.2 Data cleaning

The most frequently used data cleaning technique is to check the distribution of the attributes of interest and eliminate outliers. For certain charts, pivot table is used and noise is further removed (for example, the X level in victim sex).

Since the data size is rather big for D3.js to load as a whole, and the project aims to visualize data through different perspectives, data cleaning is specifically performed for each chart.

4.3 Overview of data

Line chart To provide audiences with general idea of our dataset, we designed a line chart with the monthly cumulative amount of accidents. Each line represents a year. Data is processed by extracting the month and year from each record, then calculate the total count of each month. Categorical color scale is used to help audiences differentiate the lines.

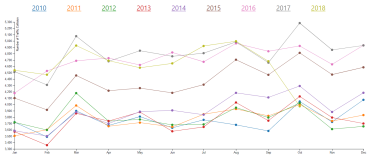

Fig. 1. Line chart

To make the chart interactive, we added events such that when the users click on the data points, specific values would appear. Furthermore, when the text at the top is clicked, the corresponding line will be turned on/off in the case that some users just wish to compare the results of a subset group of years.

There is also a mobile-friendly responsive feature in the chart that when the screen width changes, the chart will re-scale accordingly; if the screen is too small, the text at the top will disappear.

4.4 Trend of data

Following the overview, we implemented choropleth maps and slopegraphs to visualize the trend of the traffic collisions.

Choropleth maps Given latitude and longitude of each accident, the Python library "uszipcode" is used to query the name of the city where it occurred. From the rich information returned (area, population, wealth, etc.), we specifically extracted the "major_city" attribute. Together with the GeoJSON file of Los

Angeles county, results are aggregated and mapped to each city area in the choropleth maps.

The choropleth maps are consisted of 2 parts, the left panel shows the cumulative accident frequency in each sub-city in Los Angeles. The data is then divided into 4 time periods (morning peak, evening peak, daytime, nighttime) and the hourly frequency by rank is presented using "small multiples" visualization technique on the right panel. Since the variation of total number of accidents is large, "quantile scale" is preferred to encode color scheme over "quantize scale".

Moreover, a feature of highlighting particular city area upon mouse hover is implemented, along with a tooltip showing the detailed information so that users can explore the cities of interest.

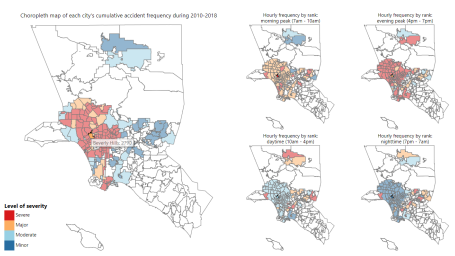

Fig. 2. Choropleth maps

Fig. 3. Slopegraphs

Slopegraphs The trend visualized in the small multiples of choropleth maps is qualitative. To have a quantitative view, slopegraphs are implemented to show the differences in hourly cumulative accident frequency. The 4 time periods are separated into 2 groups (morning peak - evening peak and daytime - nighttime) as comparing changes within each group are more meaningful.

The interactive feature for slopegraphs is also highlighting. When the mouse hovers on any piece of information about a city, the fonts would enlarge and the slope line will be highlighted in red.

4.5 Demographic information visualization

Bar chart After observing some data trend, another perspective of the data, demographic information, is presented to the audiences. The bar chat integrates information about victim gender and age as well as the total accident count in each group. As there are 3 dimensions of data in this visualization, a stacked bar chart is used. The age attribute has too many values in the original dataset. Therefore, we recoded it into age groups and used pivot table to get the ideal data structure (age group - gender - count) for this chart.

The button group at the top of the chart is used to associate on-click events with transition of chart contents. For groups beyond age 70, y-axis would also be changed because the accident counts are minimal as compared to others. Using

single y-axis for all age groups would make these data unreadable. Furthermore, tooltips are added for the bars and familiar color scheme is used to represent the gender.

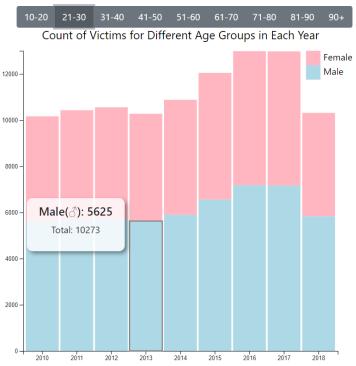

Fig. 4. Bar chart

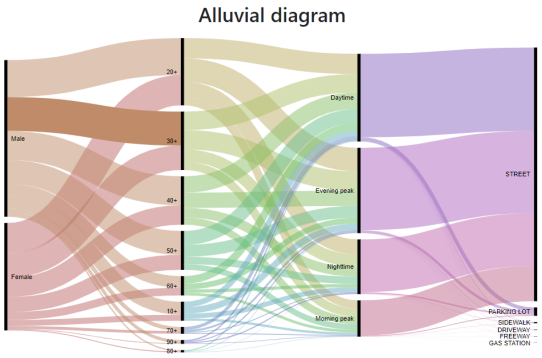

Fig. 5. Alluvial diagram

Bubble cloud To further explore other attributes, we designed a bubble cloud generated using D3.js circle packing layout. Data is the count of each victims descent, which is proportional to the radii size.

Alluvial diagram Having generated those features like age groups and time periods, an alluvial diagram could help audiences understand the relations between these attributes. Yet this diagram form is much more complicated, so we sourced an online application named *RAWGraphs* [4] to generate it. On top of the chart, we added mouse-over and mouse-out interactions using D3.js.

5 Evaluation

The infographics are evaluated against the visualization wheel (see Fig. 6). Data are represented using basic shapes such as points, bars and lines which are quite abstract. All charts are highly functional without any decoration. They are very dense in terms of the amount of information presented, and lean toward the multidimensionality side of the wheel. A wide variety of visualization forms have been used to encode the data.

All charts are interactive, features like highlighting, tooltips and transitions are implemented to better deliver the visualization findings. Apart from alluvial diagram in which some

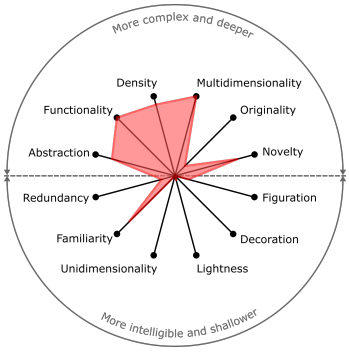

Fig. 6. Visualization wheel

users may find challenging to read, the rest are
easy to understand (high familiarity). There is little redundancy in the choropleth maps that the data are divided into 4 time periods and shown as small multiples.

6 Findings

From the visualization, the audiences can obtain some interesting and straightforward findings:

- The total amount of traffic collisions in Los Angeles is increasing each year.
- Although February has 2-3 days less than other months, the difference in the amount of collisions is definitely more than the length effect - it is always the month with the fewest accidents.
- Every March and October will have obviously more collisions than other months. The line chart also captures the sharp increase in amount from February to March every year.
- The cities which are farther away from downtown tend to have fewer collisions.
- Evening peak is the time where most collisions happen.
- Age group of 21-30 has significantly more chance of being collision victims than others. After 30, the chance of getting involved in collisions continuously drops.
- Male victims always involve more in traffic collisions than female victims do, regardless of age.
- The Hispanic and white are the top descents for the victims.
- Besides on street, a lot of people get into collisions in parking lots, too.

7 Conclusion

Besides the techniques we used to accomplish the project, we also gained insights about the situations of traffic collisions in Los Angeles county and successfully visualized them. If there are more time for future improvements, we may consider adding more layers or data sources to the visualizations to make them more informative.

References

1. "DataLA: Information, Insights, and Analysis from the City of Angels — Los Angeles - Open Data Portal." Data.lacity.org, data.lacity.org/.
2. "Mapping L.A. Boundaries API." Los Angeles Times, Los Angeles Times, boundaries.latimes.com/sets/.
3. "Exploring NYC Vehicle Crash Data in Tableau." InterWorks, 20 Apr. 2018, interworks.com/blog/modonnell/2015/08/26/exploring-nyc-vehicle-crash-data-tableu/.
4. "RAWGraphs." How to Make a Scatterplot RAWGraphs, rawgraphs.io/.