Visual Analytics - Final Project Due Date: 02/03/2020

Engineering in Computer Science
Prof. Giuseppe Santucci Alessio Fiorenza - 1661504
TA: Marco Angelini Federica Di Marco - 1631365

Road accidents in the United States

Introduction US-Accidents can be used for numerous applications such as real-time accident prediction, studying accident hotspot locations, casualty analysis and extracting cause and effect rules to predict accidents, or studying the impact of precipitation or other environmental stimuli on accident occurrence. In this work we concentrate only on accidents of a single year in the United States. Even considering a single nation and a single year the number of accidents is really high, so in this work we aim at providing a visualization of this data that helps in better understanding patterns in the number of accidents in the different states that were recorded across time.

Data We got our data from the *https://www.kaggle.com* site where there are collected, using several data providers, including two APIs which provide streaming traffic event data, data from February 2016 to March 2019. These APIs broadcast traffic events captured by multiple events, such as the US and state departments of transportation, law enforcement agencies, traffic cameras and traffic sensors within the road- networks.
The dataset covers the accidents in 49 US states and contains 2.25 million accident records and 49 attributes. Because of the considerable amount of data, we could not just load them via d3 or javascript because the time to load was lust too much. Therefore we decided to consider just one year, 2018. In particular the data we considered are about 800.000 tuples.

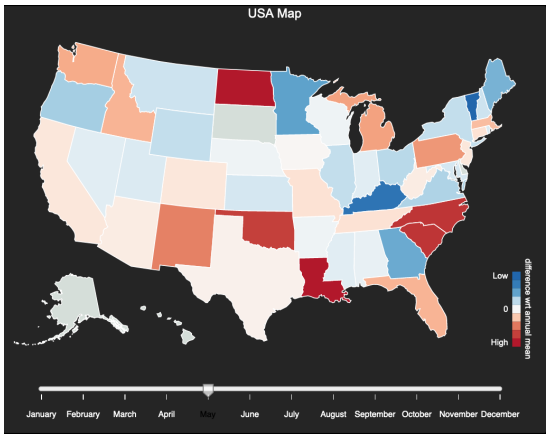
Moreover, the features we considered for the analysis and visualization will have the following symbols and meanings:

- `SEVERITY`: Shows severity of the accident (number between 1 and 4).
- `START_TIME`: Shows start time of the accident in local time zone.
- `STATE`: Shows the state in address field.
- `TEMPERATURE (F)`: Shows the temperature (in Fahrenheit).
- `WIND_CHILL(F)`: Shows the wind chill (in Fahrenheit).
- `PRESSURE(in)`: Shows the air pressure (in inches).
- `VISIBILITY (mi)`: Shows the visibility (in miles).
- `WIND_SPEED(mph)`: Shows the wind speed (in miles per hour).
- `PRECIPITATION(in)`: Shows precipitation amount in inches.

- `WEATHER_CONDITION`: shows the weather condition (in particular we used rain, snow, fog, clear, overcast, freezing).

Because of the big amount of data, we decided to do some preprocessing, before loading them. Accidents were grouped by `month, state` and counted to use those values in the **USA map**. The same happened for the **Stacked plot**. In the other plots, such as the Scatter, parallel coordinates and the PCA, we have a overall view. In addiction to that, we randomly sampled about a thousand records, in order to plot the single instances on the **parallel coordinates**, and perform the **PCA** on them.

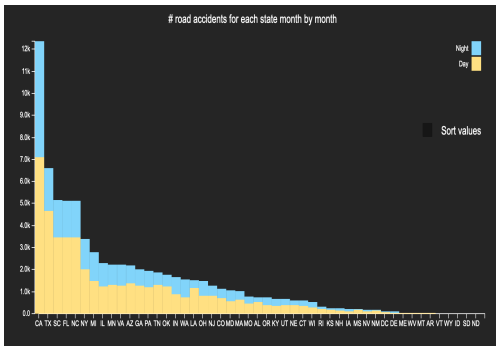
State Patterns The first visualization elements we are going to describe have the goal to display for each state the amount accidents in 2018. In particular we want to visualize how many accidents occour in each state during the different months of the year and how these numbers vary from state to state. For the first task we drew a map of the USA and colored each state according to the number of accidents for the selected month with a heat map. The color of each state in this case is independent from the color of the other states, it only depends on the mean of accidents for that state during the year. A hot color tells that in that month the number of accidents was higher than the average while a cold color that the number of accidents was respectively low. In particular, the domain of the color scale goes from *-2stdDev(state)* to *2stdDev(state)*.



A slider under the map allows the user to change the month under consideration. On the other hand, in order to visualize how the number of accidents differs from state to state in a quantitative fashion we display a stacked bar chart with a bar for each state whose height is proportional to the number accidents in that state.

We found interesting that some patterns appear really clear in the map, for example in June and July almost everywhere the number of accidents is below the average probably highlighting the departures due to holidays, on the contrary in October and November this number is far over the average maybe for the Black Friday and the Thanksgiving day in which more people take the car to move. An exception to the June month is Washington, probably because it is one of the destination for holiday travels and local population usually stays in the country. Everytime a month is chosen, the stacked plot shows the

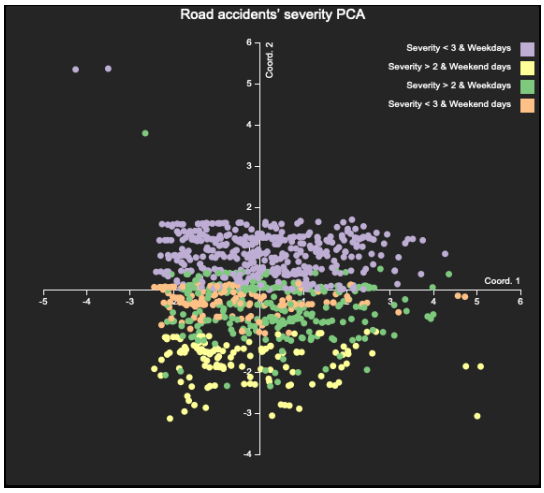
exact number of accidents for each state from the one with the highest number to the one with the lowest. Moreover is shown the division between day and night accidents which are shown in light blue for night and yellow for day. It is possible to highlight the daytime accidents we prefer selecting "night" or "day" and sort the values again pushing the "sort values" button to let them more clear. The plot shows that the number of day accidents is higher than the night ones for every month. This can be motivated since during the day there are more traffic because more people have to use a car, for example to go to the work or to go to airport for a travel. Below there is an example of the plotter usage.



Accidents The other visualization elements we deviced concern the severity of the accidents and their causes. These data are displayed in three views:

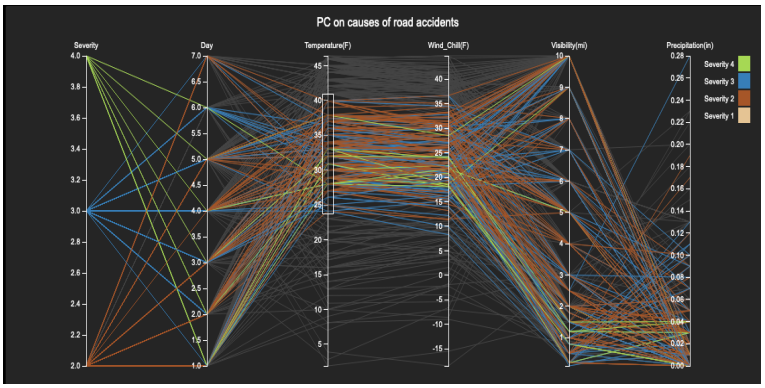
- A plot representing the results of a PCA projection if accidents' severities divided into weekend-accidents and not.
- Parallel Coordinates with some weather's attribute mentioned above.
- A scatter plot which shows some novel/peculiar characteristics on the total amount of accident based on the weather condition.

PCA plot This plot displays in 2D the records sampled in the dataset projected along the attributes with greater variance according to a PCA. We did not plot the entire dataset but we picked uniformly at random for each month a certain amount of data and used them for the visualization. Plotting all the accidents would have been too dense and it would have been difficult to distinguish points in the plot. Moreover the size of the data would have required a considerable amount of computation time on a Javascript engine and would have result in a bad delay experience for the user. We were concentrating on accidents' severity and their correlation with the weekends.



We identified clusters grouping the accidents according to different characteristics and we later observed on the plot that they were actually close together in the final space as you can see in the figure. So we divided the accidents in 4 groups: the ones that experienced during weekends and the ones that experienced during the rest of the week focusing on the severity (greater or smaller than 2). As you can notice accidents groups form clusters, and thus there is a good chance that they are close each other also in higher dimensions. On mouse click on a point of the plot, the correspondent cluster will be shown on the parallel coordinates plot.

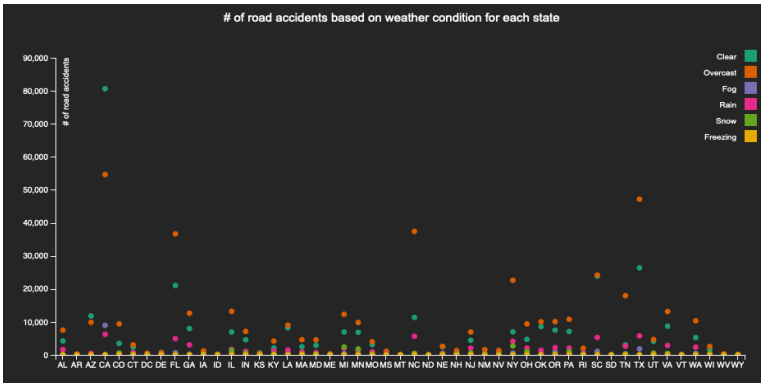
Parallel coordinates The same data used in the PCA plot are displayed in all of their dimensions in a parallel coordinates view. Parallel Coordinates are a standard method to visualize high-dimensional data on a 2D environment. We decided to display weather's possible conditions which cause accidents.



To give a better user-experience, axes can be moved around and reordered while maintaining a constant distance. To provide further user interaction in each of the axes the user can brush with the mouse over an interval of values for that axis, as a results the lines in that interval are put in foreground and the rest of the accidents changes to a light grey color and is put in background. This view is also coordinated with the USA map.

Indeed the user can click on a state and as a consequence in the parallel coordinated view only the accidents in that state are highlighted, in an orange color, and put in foreground.

Scatter Plot The last view in the second group is the scatter plot. It displays relationships between accidents in each state and the causes related to the conditions of the weather, such as clear, overcast so on and so forth. We decided to use a scatter plot where on the x-axis we have states, while on the y-axis we have intervals for the number of accidents. Each point represents the number of accidents for each state. The color of each circle encodes a cause of accident for each point and allows to see quickly which is the major cause of accidents (such as rain, snow..) giving the user possibly interesting insights. Moreover, to give further details to the user, a group of points can be zoomed, and a more precise number are shown, together with the cause of all accidents that the selection includes is displayed. This is useful in order to "query" the dataset. For example, thanks to this view, we noticed that during the whole year of 2018, the number of accidents during sunny and cloudy days were more that the accidents occurred during snowy and freezing days. Moreover, is possible to notice that the number of road accidents caused by snow or freezing are greater in northern states rather than in the southern states, this derives because the weather in those states are colder.



Why using this kind of plot? We opted for this type of scatter plot, over a `Radar Chart` because we wanted to display a general review of the different weather causes over the total amount of states and not focus on a single state, to find some correlation joining the weather causes of accidents between the states.

Discarded designs During the development of this application, many other views have been taken into consideration, but at the end they were dropped because of the lack of significant insights provided by the data. One of the first view to be dropped has been a `Time Series Chart` showing for each month of the year, the total amount of accidents. This because we focused our attention state by state on the USA map and we also used a stacked plot to see the total amount of accidents per month. It did not make much sense to include this static view.

Lastly, we dropped the idea of implementing a radar chart: in fact initially we thought

about the implementation of a radar chart in which each spoke contained a weather condition and a line was drawn to connect the data values of each spoke. So, clicking on a particular state of the map, the star should show the causes of the accidents in that state. The problem was that there was not a overview on the different states but just a focus on each state at a time.

Conclusion In this work we devised a simple visual environment with the goal of displaying the number of accidents and the entity of the causes in the different states in the USA in a time interval of a year. We used a map of the States coordinated with a stacked bar chart to visualize across the various months the amounts of accidents. As for the accidents we considered different attributes especially regarding the severity and the different causes and displayed the data qualitatively in a PCA plot and in the detail in a parallel coordinates view. Finally we devised a scatter plot with colored points in order to visualize the correlation between states based on the total amount of accidents with the same weather conditions. As for future work we think it may be useful to add the possibility of loading the data for different years if the user requests it; then it might be interesting to highlight clusters in the PCA projection with some cluster detection algorithm besides highlighting the information we already provided. At the same time it would be useful to ask users some feedback in order to add the small details that would make the interaction more intuitive and smooth. (Slides for this project are available here).