Information Visualization lab

First practical work

Introduction

The goal of the first visualization project is to create a visualization that analyzes the traffic collisions happened in Los Angeles during a month of the period 2010-present, that can be obtained from the Kaggle dataset: https://www.kaggle.com/cityofLA/los-angeles-traffic-collision-data.

The dataset contains information regarding several years that includes location, time, etc. You must select a month, drop the rest of the data (you can do this pretty easily using Python), and create a visualization that is able to answer the following questions:

- Are accidents more frequent during weekdays or weekends?
- How have the accidents evolved along the month?
- What time of the day are accidents more common?
- Are there any areas with larger amount of accidents?
- Is there a correlation between the weather conditions and accidents?

To answer the last question, you will require to find historical weather data, which can be easily obtained if you look for it a little bit.

You can add extra questions to these.

Data processing

You can process the data using either Open Refine, or another tool. Note that the initial file may be very large, so you might need to assign some more memory to Open Refine to process it properly. However, you can drop the unnecessary rows using Python before any other data processing. If you need more memory assigned to Open Refine, it is not necessarily simple in Windows (you might assign it more memory but the browser refuse to open a tab with such a big heap), but it is quite straightforward in Linux or Mac.

You can also process the data programmatically.

In any case, the initial cleaning procedures may take some time, and tools like Open Refine keep the data for being able to undoing changes.

Independently of the cleaning tool and process, you must describe your cleaning steps in your Google Colab document. If these are using pandas, for example, include the code in the document. We must be able to reproduce the steps and go from the raw data to the clean version.

For the delivery, you must provide the raw and clean version of the data. For example, if you follow our advice, the raw data would be a single month after dropping the unnecessary months, and the clean version will be the one without the columns you will not use, as well as the other modifications you did to the data.

Design and implementation

For the visualization, we need you to describe the design process also in the Google Colab document. This means that you may include all the steps that led you to the final visualization. You can remove (or group) some steps in the final document if you think it is better. But we need to see the design process, we want to understand how did you reach to the final visualization.

Before you start coding anything, you need to think on what visualizations will be provided. Note that the user needs to be able to answer the questions above with a single visualization, that may include multiple views.

For example, you might include a line chart to depict the accidents along the month, together with some charts that let you see the relationship with the accidents and weather conditions. You might also be willing to include some charts that depict the accidents per area…

This is just an idea, and the charts that you use in each view must be properly designed. Consider all sorts of charts that might be useful: line charts, bar charts, heat maps… Some views will contain several variables, so use visual cues, proper palettes to ensure they are understood properly.

Delivery instructions

The work can be implemented in pairs or individually.

You have to provide the raw data as well as the clean data. Be sure not to include the original file, just a single month.

You have to describe the cleaning procedure, so that I can generate the clean data from the raw data following your steps. This description must go in the Colab document.

You must include a step-by-step description on **how to solve tasks**. These can go in the Collab document. For example, one might have:

- Question 1: Is there any relationship between accidents and weather conditions?
- Answer to Q1 could be: In chart C1 you can see the accidents evolution along the month, and in chart C2, the weather conditions. It can be seen that the same days that weather conditions are bad, the number of accidents is higher.

The delivery must consist on a single ZIP file with a name that includes the authors, that contains the datasets (raw and clean), the Colab file(s) (*ipnyb*) and optional extra documents if required.

The deadline for the delivery of this lab project is the 15th of November.

Important remarks

The final grade will take into account the number of variables included (e.g., number of accidents, victim age, gender, weather condition…). Additionally, we will value the number of non-trivial tasks (adequately described in the documentation) that can be properly solved with your visualization tool. In this sense, adding other data sources is a good point (e.g., you could add information regarding whether a certain day has been a holiday to understand that there were less accidents during the week…).

Don't leave the project for the last day or do the minimum amount of work. In case of doubt, ask us whether the current work is enough or needs more effort.