Data Visualization lab

First practical work

Introduction

The goal of the first visualization project is to create a visualization that compares the traffic crashes happened in New York City during the months of April 2019, 2020, and 2021, which correspond to pre- during, and post-COVID measures. Data can be obtained from the NYC Open Data website. You need to download the data, extract the desired months and clean the data: https://data.cityofnewyork.us/Public-Safety/Motor-Vehicle-Collisions-Crashes/h9gi-nx95.

The dataset contains information regarding several years that includes location, time, victims, vehicles involved, etc. Your task is to create a visualization that is able to answer the following questions:

- Have the accidents changed during stay-at-home period?
- Do accidents in 2021 look similar to 2019?
- What about the number of victims, how has this behaved?
- Are there any areas with larger number of accidents?
- What are the main reasons of accidents?

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.

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. Therefore, just deleting the non-desired months before dealing with the data with Open Refine will be useful.

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 set of three months 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. Describe your cleaning decisions too (e.g., if you drop records for some reason, argument why).

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 will 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 evolution of victims… 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 has to be implemented individually.

You have to provide the raw data as well as the clean data. Be sure not to include the original file, just the necessary months.

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: Have the accidents changed during stay-at-home period?
- Answer to Q1 could be: In chart C1 you can see the accidents evolution along the three different months. Note how on April 2020, the number of accidents was lower.

The delivery must consist on a single ZIP file with a name that corresponds to the username of the author, that contains the datasets (raw and clean), the Colab file(s) (*ipnyb*) and optional extra documents if required. Of course, the Colab document need also be formatted properly. It should start with the name of the author, for example.

The deadline for the delivery of this lab project is the 13th of March.

Important remarks

The final grade will take into account the number of variables included (e.g., number of accidents, areas with accidents, gender of victims, number of victims…). 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 may be 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 whether the current work is enough or needs more effort.