Crime in Los Angeles – Project 2 Data Visualization Dashboard

Authors: António Pinto, Davide Farinati, Henrique Vaz, Philipp Metzger

https://losangeles-crime.herokuapp.com/

I. Introduction

This report describes the creation of a dashboard that interactively visualizes data on crimes that happened in Los Angeles between 2010 and 2019. This project was part of the course "Data Visualization" as part of the master's degree "Data Science and Advanced Analytics" at NOVA IMS in Lisbon in 2021.

II. Dataset Description

The main dataset in use in this project contains information on crimes that occurred in Los Angeles between 2010 and 2019. This dataset can be found on Kaggle.com at [1], where a CSV file with 2.12 million records is available for download.
The dataset contains 28 columns that contain information like dates, locations, crime descriptions and information about the crime victims, such as their ethnicity, age and gender. The features used for the analysis and visualizations of this projects are: 'DATE OCC' (date of crime's occurrence), 'AREA NAME' (the name of the area where it occurred), 'Crm Cd Desc' (a description of the occurred crime), 'Vict Descent' (the ethnicity of the victim), 'LAT', and 'LON' (the geographical location of the occurred crime).
Apart from this dataset, the authors also resorted to the Los Angeles Police Department's website [2] in order to retrieve each area's population and the square miles area of each neighborhood. Since for three areas of Los Angeles ('Olympic', 'West LA', and 'Topanga') this information was not available, the median of the respective variable was used for these three areas.

III. Visualizations and Interaction

1. Inspiration
   The content of the crime dataset described above served as the source of inspiration for the visualizations created for this project.

2. KPIs / Big Numbers
   In terms of visualizations, the first thing the reader will see are three numbers that will provide a very high-level perspective of data. Big numbers are a very common practice in dashboards and are used to give the viewer some context before diving into deeper analysis. The information this visualization provides to the viewer is where in Los Angeles the crime incidence is lowest, where it is highest and a comparison of the numbers of reported crimes between the selected year and the year before, in order to give the viewer an idea on how things evolved.
   Further below, the user also comes across two more big numbers. For the case being the user can see in which year a given neighborhood had more crime occurrences. In addition to that, to the right of that number is also shown the total population of that same neighborhood and to what percentage of the total Los Angeles population that corresponds.

3. Map of Los Angeles Crime in Selected Year - Choropleth / Map
   This graphic displays a spatial heat map, showing Los Angeles, in order to show which areas had more reported crimes. The values in each area can be compared by means of a color scale/intensity. This is a very handy graphic since it very concretely shows general areas of higher incidence, always considering a specific year selection.

4. Top 10 Most Common Crime Types – Bar Chart
   An also very common practice in dashboards, since there is limited space, is to present top analysis using the most significant categories. In this case, from 142 described crime types, the most frequent 10 are shown for a given year. Furthermore, the user is given the possibility of passing the mouse over the bars to find the exact values of occurrences.

5. Line Chart
   For the line chart, we decided to include also the top 10 crimes committed, where the user can visually observe the behavior of the number of occurrences of these crimes throughout the selected year in a specified neighborhood. Since displaying 10 lines, corresponding to the 10 categories, would be imperceptible and hard to visualize, by default the plot only shows the 5 most frequent crimes, while the others are only listed in the legend, but with the possibility to be activated by clicking on the respective legend elements. This way the user can select or deselect crimes according to his intentions, without being overwhelmed with a chaotic chart at a first glance.

6. Average Number of Reported Crimes per inhabitants per year for the respective area of Los Angeles - Bar Chart
   This bar chart combines information on the population, the square miles area and the yearly average crime count of each area in one plot. The user is able to select an area of Los Angeles in the selector above, which results in the respective data to be visualized in the left bar of the two-bar plot. The right bar with the label “Total Los Angeles” serves as an object of comparison to give the user an insight on the magnitude of the crime rate in the respective area compared to Los Angeles as a whole. A two-bar plot was chosen for this visualization, because it is the most effective way to provide the described comparison to the viewer. Note that this plot has no variability in relation to the year that is selected by the user.

7. Victim’s Ethnicities Distribution for the respective area of Los Angeles - Donut Plot
   This donut plot visualizes the ethnicity distribution of the crime victims of the area that is selected by the user, using the same selector as above. For a better visual result only victim’s ethnicities that make up at least 1% of the occurred crimes are displayed as a separate category. The others are aggregated in the category “Other”. A donut plot was chosen, because it is able to transport the same amount of information as a pie chart, whilst using significantly less ink due to its hole in the middle. The colors representing each category were chosen in the way they are displayed in order to ensure optimal visual separability of the different classes and in order to not imply any type of order between the categories. Note that this plot has no variability in relation to the year that is selected by the user.

IV. Technical aspects

Similarly to what was taught in the practical classes, the plots were created using the Python library Plotly, which enable the creation of dynamic dashboards for an enriched visualization of

data. Even though Dash, the platform taught on classes, is a compatible solution to create dashboards using the Plotly plots, for familiarity and simplicity reasons, we decided to use a library called Streamlit [3]. Streamlit, like Dash, is a library for building data dashboard web applications that has been gaining a lot of popularity due to its rapid prototyping capabilities. Lastly, on the implementation and deployment part, we hosted our web application using Heroku [4].

Regarding the dashboard content, our application offers some features to facilitate its visualization, which we will go through briefly.

All visualizations except the two lowermost ones are dynamically displayed according to the year selected in the draggable bar (2010-2019), presented in the top part of the web page. Below, are three indicators that present the best and worst neighborhoods in terms of crimes and also shows the total number of crimes for the selected year as well as a comparison with the previous year, except for 2010 (that is the first year of the analyzed timeframe).

Further down, we display a map with a heatmap for the crimes occurred in Los Angeles. Here, the user can observe what are the most dangerous areas, interacting with the plot by zooming in for a more precise analyses of certain regions. It is relevant to mention that the app is processing a large amount of data (all geographical positions of the occurred crimes) and due to this, refreshing its content can take quite some time.

Next, for this second part of our web page, there is a new level of granularity by enabling the user to filter by neighborhoods. In the dropdown menu, there are 21 neighborhoods of LA, where the upcoming plots will display information based on the selected option. The default value is the area of Newton.

Filtering by neighborhood is important because it gives the operator the opportunity to explore each of the areas analyzed previously in the map deeper. The part below the area selector provides information about the types of crimes that occurred in each area, how the areas' crime rates compare to Los Angeles as a whole and about the ethnical distribution of the victims of the crimes occurred.

The GitHub repository linked to Heroku for the deployment of the web app can be found at [5].

V. Discussion

In the course of the development of this project, we were able to not only produce meaningful plots in order to display the desired information, but also to learn about other concepts related with the composition of a dashboard and subsequent deployment of a web application.

For this second Data Visualization project, the challenges faced in the first project were faced again, more specifically: The processing of the data to match our needs and expectations as well as finding suitable plots for the right visualizations and interactions. Regarding the dashboard and the deployment, the documentation provided by the official websites is very complete and precise, with the most challenging part being the execution of call-backs by the filters and the optimization of the web app for faster loading times.

With the accomplishment of this project, we were able to produce a web page that could provide relevant insights for the city of Los Angeles. For instance, this could help users decide on where to look for their next apartment based on areas that are prominently placed above the other in terms of safety.

VI. References

[1] https://www.kaggle.com/chaitanyakck/crime-data-from-2020-to-present/version/5?select=Crime_Data_from_2010_to_2019.csv
[2] https://www.lapdonline.org/

[3] https://streamlit.io/
[4] https://devcenter.heroku.com/
[5] https://github.com/Fari98/datavis_dashboard