World Data League 2022

Challenge

Challenge Semi-Finals - Predicting a safety score for Women in Costa Rica

Team Name

Data Crocodiles

Authors

- Diogo Pessoa
- Fábio Lopes
- Francisco Valente
- Júlio Medeiros

Introduction (250 words max)

According to recent data, around 74% of people report they feel comfortable walking alone at night in the OECD countries. However, there are significant variances between countries when looking deeper. In some countries (e.g., Austria, Denmark, Finland), around 85% of people say they feel safe, but the scenario is just under 50% for other countries like Mexico, Brazil, Costa Rica (our case study) and goes below 40% in South Africa. Despite the increase of the efforts and awareness over the past few years regarding the topic of the reduced safety when concerning women and girls and the complaints reported by them as victims, some cities' policies and security still lack to respond to the community's demands. The case study of this challenge, Costa Rica, is ranked 32 out of 163 countries when it comes to overall peace in the 2020 Global Peace Index. Police statistics have shown that 70.6% of the complaints of street sexual harassment in Costa Rica in 2019 were submitted by women. This challenge aims to study how several factors are linked, negatively or positively, to safety issues and how they can be used to create a safety index. This index can be used to inform women regarding the safest districts, fulfilling their right to enjoy public space without feeling danger. Moreover, the safety index may also help inspect the key elements that can impact and improve the safety of a given city/district, i.e., which policies can be addressed to improve citizens' protection from harm.

Data (250 words max)

Our analysis was based mainly on four datasets provided and/or suggested by the WDL organization:

1. Crime records for Costa Rica, helpful for the security categories considered in our final safety index;
2. the different arcGIS data available, useful for the education, healthcare and public areas categories considered in our final safety index;
3. Districts data (demographics, Social Development Index) from Costa Rica, used for the education, healthcare, economics, and security categories considered in our final safety index;
4. Openstreetmap (the geolocation of Points of Interest), helpful for the public transportation category used as another factor on our final safety index.

We looked for additional data online, but no significant information was found. At this point, one of the limitations was regarding the data containing the COVID-19 period, which impacted the crime reports. Therefore there will be a clear break on the safety trend over the years before COVID-19 and during/after it. Given that, additional data, post COVID-19, could help better estimate and predict the trend in the next trimesters/years.

Methods and Techniques (250 words max)

First of all, we searched the literature to get a better understanding of the safety indexes around the world and how they are usually computed. Second, we extracted new features from the date-times. Third, we performed an extensive exploratory data analysis (EDA). It included the study of demographics and crime reports data. This EDA was strongly based on visual analysis for an easier and better understanding. Based on the research of safety indexes, we created our index taking into account the available data. That index is based on six categories that reflect different safety areas: security, education, healthcare, economical, public areas, and transports. For each category, several indicators were considered and computed. As we have some temporal data (crime reports), the index was obtained for different trimesters and years. Then, we developed forecasting models using ARIMA and linear regression algorithms. We made it for trimesters and years independently. To train them we used data from 2010 to 2017. To evaluate them we used data from 2018 and 2019. We removed data from 2020 to 2022 because we found that these data are different from the previous years as a consequence of the COVID-19 period. We evaluated the models using the relative root mean squared error and by comparing visually the expected values with the predicted ones. Finally, using a geographic library (folium), we created interactive plots to observe the value of the safety index (and its categories) over the years and the trimesters for each district of San José Canton.

Main Insights (300 words max)

As expected, we found, either by year or trimester, a significant downtrend in the crime reports from 2017/2018 onwards, independent of the gender of the victim. The COVID-19 lockdowns and measurements adopted during it and even after could probably explain this downtrend of reported crime (e.g. due to remote work potential victims were less exposed to dangerous scenarios). However, in general, the number of crime reports increased over the years before COVID-19. Furthermore, we verified that there are not many reported street harassment cases. This could be explained by the fact that people only started being aware of this crime after 2020 (only 3 districts had harassment cases reported).

Regarding the safety index, it is important to have a final score that is computed based on sub-indexes that consider different safety categories, as usually done by safety indexes in the literature. It also allows having a particular and global analysis. For example, Carmen district presents a high number of criminal reports but is one of the districts with better education, economics, and infrastructures. In this district, the overall safety index will be high, despite the higher number of criminal reports. Therefore, the product also offers the user an additional option besides the overall final index: the opportunity to get information on different factors the user might be especially interested in. It is also interesting to observe that some neighbor districts have very different scores for a given category (e.g. Pavas has much better healthcare infrastructures than its neighbors).

We also found a curious fact during exploratory data analysis: male people suffer more crimes during the 6 pm to 12 am period while for female people the worst hours are between 12 pm and 6 pm.

Product

Definition

A dashboard that provides a geographical representation with the safety index/metric per district based on the data of San José Canton (Costa Rica) regarding several safety-related categories (security, education, economics, healthcare, public areas and public transportation) and the global safety index. The dashboard informs women citizens regarding the safety index/metric per geographic zone by trimester and by year and can also assist the governmental decision-makers in order to create policies that improve the safety of citizens. Being a geographical representation, with score number and respective easy-interpretable colour, makes it easy to understand and to be used by anyone, independent of their age and level of literacy. This dashboard can be extended to other cantons upon the availability of related data.

Users

Both women and girls can use the dashboard for a safe accomplishment of daily life activities, both in terms of overall safety or a specific category of interest from the six ones being considered in our index. The dashboard is useful for both the resident population and tourists. For example, a San José citizen may be more interested in education or healthcare safety categories than a tourist, which, for instance, may be more interested to know where they can walk and visit without feeling the danger of being the victim of a crime. The fact of being a dashboard with a geographical representation of the safety index by numbers and respective easy-interpretable colours contributes to a friendly use application that any person can easily use, independent of their age and level of literacy.

Activities

- Suggests the safest regions for women and girls, overall and for each safety category. It also allows the police makers to know which areas they can address/improve to enhance safety.
- Predicts the trend of safety index/metric per geographic zone by trimester and by year.

Output

The main output is a geographical representation presenting a safety score/index of each district in the San Jose canton, overall and for several safety-related categories. Being a geographical representation, it is easy to understand and to be used by anyone, independent of their age and level of literacy.

🌍 Social Impact Measurement

Outcomes

- Increase of the perceived safety level, in particular by the female population.
- Decrease in the global number of crimes and improvement of safety-related issues (healthcare, education, etc.), as the local management and policy-makers can be more aware of which places require new measures to make them safer, and how to do it (which areas must be addressed).

Impact Metrics

- Women and girls' perceptions of safety and how comfortable/free of danger they feel (survey studies)
- Number of crimes reported by women and girls
- Variation of category indexes (e.g. if a district has a low public transportation score, we want that to be improved by, for example, creating more bus stations and bus routes).

Impact Measurement

The demographic and infrastructure sub-indexes (education, healthcare, public transportation, etc.) tend to improve, which will lead to an increase in the safety index. However, based on model predictions, the crime trend is upward, leading to a lower safety index value. Nevertheless, we expect that our product can better inform its users to avoid the regions most prone to crimes as well as help decision-makers reduce dangerous situations and then balancing that upward trend. Overall, we expect an improvement of the safety index due to a fruitful use of the tools and information provided by our product.