EPSILON – European Platform for Data Science: Incubation, Learning, Operations and Network

Training Material for Teaching and Self-Learning

Selected Use Cases

Module 4/6

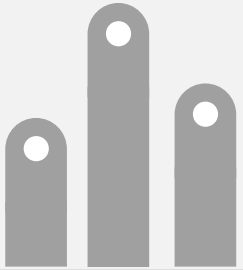
Hochschule Harz
Harz University of Applied Sciences
EPSILON
Funded by the European Union

Agenda

- Project 1: Where to build new bicycle parking spots in Paris?
- Project 2: Predicting Long-Term Unemployment in Portugal
- Project 3: COVID-19 mortality surveillance platform
- Project 4: Domestic Violence Data Observatory

Where to build new bicycle parking spots in Paris?

Supporting data-driven decision-making with open data



Where to build new bicycle parking spots in Paris?

Project Team:

Volunteers from CorrelAid

10 people divided into a Data and a Research Team

Stakeholder:

City of Paris Mobility Department

Problem Statement:

The City of Paris is currently in transition to “Green Mobility” which includes strengthening bicycle infrastructure. Until now, streets have been mainly used by cars as well as parking spots. At the same time multiple national and regional French governments started publishing public data.

In 2020, government officials published the “Plan Velo 2021-2026“, which includes plans for a massive improvement in cycling infrastructure.

The goal of this project is to support the Green Mobility Transition by providing a compelling visualization of where bicycle parking spots are needed.

Where to build new bicycle parking spots in Paris?

Detailed Information

DFG organization:
CorrelAid

Partner type:
Governmental agency

Partner name:
City of Paris Mobility Department

Sustainable Development Goal (SDG):
11

Type of interaction:
Short-term project

Type of analytics:
Modeling

Type of data:
Government data

All data is available here

Where to build new bicycle parking spots in Paris?

Internal & External Data

Most of the data was collected by the government and made available on open data platforms.

- Bicycle parking spaces
  - collected by Île-de-France Mobilités
- Visitor numbers at train stations
  - collected by SNCF (Société nationale des chemins de fer français)
- Incoming annual traffic volume per station of the rail network 2021
  - collected by RATP (Régie autonome des transports Parisiens)
- Green spaces and similar
  - surveyed by Mairine de Paris
- Parking on public roads - parking spaces
  - surveyed by Mairine de Paris

Where to build new bicycle parking spots in Paris?

Methods used for data processing

- Aggregation
- Normalization
- Formation of weighted average

Since the aim of the project was to provide graphical representation of parking space demand by geographical unit in Paris, a data processing process only took place to a limited extent. During data processing, the individual data frames from different data sources were aggregated and summarized into an overall data set. In addition, the individual variables considered and collected were **normalized** using the available bicycle parking spaces in order to finally determine the actual demand per level of the **smallest statistical spatial unit used in France (IRIS)**. In the subsequent presentation of the data, however, this was converted to the respective users or residents and the respective influence on the parking space requirements for bicycles and graphically displayed using a color scale.

Where to build new bicycle parking spots in Paris?

Data Sources

Bicycle Parking Spots
Data on already available parking spots in the city area and in train stations from two separate sources.

Census Data
Census provides current data on population density.

Green Spaces
Location and size of Parks and other green spaces.

Public Transport
Location and traffic through train, Metro and RER stations in Paris. Multiple data sources needed to be combined to obtain a full picture of public transport.

Museum Data
Location and attendance figures for museums in the city area.

Schools
Location and total capacity of schools.

Shop
Location of Shops.

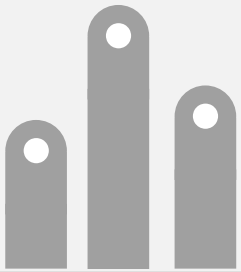
Where to build new bicycle parking spots in Paris?

Solution – Project Deliverables

1) Demo of a Dashboard displaying the various Supply/Demand Metrics
2) Publicly available code repository on GitHub

This dashboard provides an easy visual access to otherwise complex data

Predicting Long-Term Unemployment in Portugal



Predicting Long-Term Unemployment in Portugal

Project Team:

Data Science for Social Good Portugal (DSSG)

Data Science Knowledge Center @ Nova SBE

3 Data Scientists, 1 Technical Mentor, 1 Project Manager

Stakeholder:

IEFP, the institute of employment and vocational training in Portugal

Problem Statement:

The partner organization operates over 80 job centers nationwide. Within these centers, job counselors are assisting job seekers by suggesting interventions such as training courses and aiding in the job application process. Before this project, counselors were drafting personal action plans, picking from a multitude of available interventions.

This project had two main goals:

1. Improve on the process of identifying people at high risk of becoming long-term unemployed
2. Develop an intervention recommender system for job counselors, providing personalized recommendations to individual job seekers taking into account their personal profile

Detailed Information

DFG organization:
DSSG and Data Science Knowledge Center at NovaSBE

Partner type:
Governmental agency

Partner name:
IEFP

Sustainable Development Goal (SDG):
8

Type of interaction:
Short-term project

Type of analytics:
Modeling

Type of data:
Government data

All data is available here

Internal & External Data

The IEFP provided the project team with data on people registered in their unemployment programs.

This data consists of 12 years of transactional data with currently 3.1 million registered individuals.

Features:

- Demographic information
- Professional background
- Past history with IEFP
- Trainings attended, by type and outcome
- Job offers replied to, by outcome
- Numbers of times summoned by IEFP, by outcome

Methods used for data processing

1) Classification

2) Recommender System

1) Classification

Objective: Identify high-risk individuals for long-term unemployment

Method: Machine learning on historical data

Features Analyzed: Education, employment history, skills, demographics

Outcome: Proactively identify and prioritize high-risk individuals for job counseling

2) Recommender System

Objective: Improve suggested interventions using machine learning

Method: Predict success probability of interventions using client history

Features Analyzed: Predicted success vs. average success of a reference group (e.g. age)

Outcome: Provide error measure to aid counselors in drafting intervention plans

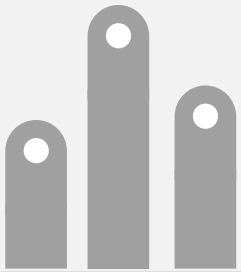
Solution

The project team improved job counselors' performance by developing two optimized models using historical unemployment data:

Risk Classification: A predictive model was created to identify clients at risk of long-term unemployment. It was tested, rolled out, and refined based on feedback from counselors.

Intervention Recommendation: A second model used data on past interventions to predict which actions are most likely to help individuals return to work. This provided personalized recommendations and a quantitative metric for counselors to assess success, reducing reliance on their experience alone.

COVID-19 mortality surveillance platform



COVID-19 mortality surveillance platform

Project Team:

Data Science for Social Good Portugal (DSSG)

3 data scientists

Stakeholder:

This project was initiated **without** being commissioned by a public institution

Problem Statement:

During the first years of the COVID-19 pandemic, mortality data played a central role in estimating the severity of the disease. As a result, it also proved to be a major factor in the decision-making process on measures taken to contain the disease.

The initiators of this project noticed problems with the publicly available data. Mainly, it was without structure and proper documentation.

The goal of this project was therefore to create a pipeline for improving the quality of the data and making it available in an accessible format.

Detailed Information

DFG organization:
DSSG PT

Partner type:
Project developed using open Protugues-government data

Partner name:
-

Sustainable Development Goal (SDG):
3, 17

Type of interaction:
Short-term project

Type of analytics:
Data consulting

Type of data:
Web scraping of open Portuguese-government data

All data is available here

COVID-19 mortality surveillance platform

Internal & External Data & Methods

Prior to the project mortality data was only publicly available on SICO – eVM, a portal documenting this data in Portugal.

On the website, multiple tables and graphs are provided, but the raw data is **not** readily available in a downloadable format.

Web Scraping:

To make data usable for visualization, machine learning, or other tasks, it is usually beneficial to provide it in a database or in an easy to handle file format like .csv. If interesting data is **only** available in form the of content on a website, **web scrapers** are a way to obtain the data in a usable format. When web scraping scripts are created to access the website of interest automatically and save the data in the format best suitable for further work.

Transforming Data:

Data in **JSON form** is provided by websites in a previously defined format. The project team analyzed the structure of the provided data and developed an automated script to format the data into an .csv.

Reporting:

Web scrapers are prone to changes made to the target website. Owners of a website may change the accessibility or the way information is displayed. In this case it is important to be aware of how data is provided by having a report system. This should include automated checks on whether or not data is still being obtained.

Solution

The project team was able to create an automatic system for accessing the mortality data on a daily basis.

This helped the following publicly available services:

Data repository

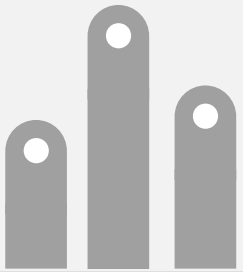
A well-documented and comprehensive repository containing historical data on mortality from 2014 up to today.

In addition to data in multiple formats, a data dictionary was created, explaining how the data is stored.

Web Scraping Example

The code for this project is publicly available and can be a starting point to create other web scrapers with automatic reporting.

Domestic Violence Data Observatory



Domestic Violence Data Observatory

Project Team:

Data Science for Social Good Portugal (DSSG)

3 data scientists

Stakeholder:

This project was initiated without being commissioned by a public institution

Problem Statement:

Domestic violence is the second most commonly recorded crime in Portugal. Although there is a multitude of data sources on the subject, there is a lack of easily accessible dash boards and data in simple formats (e.g. .csv)

This project has two main goals:

1. Create a data repository making data on domestic violence available in a simple format
2. Visualize the data in a dashboard

Detailed Information

DFG organization:
DSSG PT

Partner type:
Project developed using open Protugues-government data

Partner name:
-

Sustainable Development Goal (SDG):
3, 5, 16

Type of interaction:
Short-term project

Type of analytics:
Data consulting

Type of data:
Web scraping of open Portuguese-government data and NGOs data in Portugal

All data is available here

Data Sources & Methods

This project relied mainly on the following data sources. The data was only available in form of .pdf files. The aggregation of those files into an .csv file was handled individually.

Data Sources:

- APAV (Portuguese Association for Victim Support) Reports on Domestic Violence
- Quarterly Report on Domestic Violence provided by the Portuguese Government
- Yearly Report on Domestic Violence provided by Ministry of Internal Affairs
- Report on Victim Support Structures

Dashboard

- To provide a clear overview, data from multiple reports was aggregated geographically into a dashboard during this project.

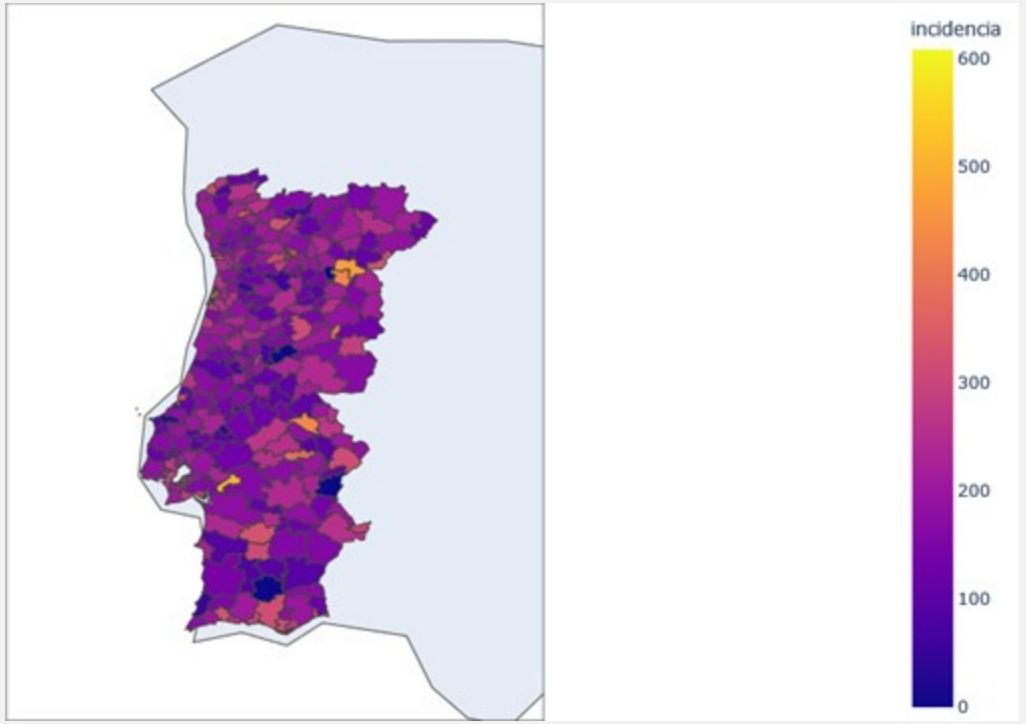
Solution

A dashboard displaying data on domestic violence in Portugal serves as a comprehensive and accessible tool for understanding key aspects of this social issue.

The dashboard integrates data from various sources to provide users with real-time insights, trends, and geographical patterns related to domestic violence.

A prototype of this project is displayed on the right.

The project is accessible here: https://www.dssg.pt/en/projects/domestic-violence-data-observatory/



Additional Projects

Identifying Fraud & Collusion in International Development Projects

Content from:
https://www.dssgfellowship.org/project/identifying-fraud-collusion-in- international-development-projects/

Paper:
https://www.dssgfellowship.org/wp-content/uploads/2016/12/world_bank_fraud.pdf

Surveying target groups for interest in a good life for the elderly in rural areas?

Content from:
https://www.correlaid.org/en/using-data/project-database/2020-03-DEN/

Identification of causes and optimization of waiting times for veterinary consultations at the AZP veterinary hospital

Content from:
https://www.dssg.pt/en/projects/identification-of-causes-and-optimization-of-waiting-times-for-veterinary-consultations-at-the-azp-veterinary-hospital/

Sources I

- Kloppenburg & Moura, K. & Dietrich, J. (2023) CorrelAid/paris-bikes: *Where to build new bicycle parking spots in Paris? Supporting data-driven decision making with open data*. Available online at https://github.com/CorrelAid/paris-bikes/ (last accessed on 04.12.2023).
- Data Gouv France: Where to build new bycicle parking spots in Paris supporting data driven decision making with open data. Available online at https://www.data.gouv.fr/en/reuses/where-to-build-new-bicycle-parking-spots-in-parissupporting-data-driven-decision-making-with-open-data/ (last accessed on 04.12.2023)
- DSSG Portugal (a): Predicting long-term unemployment in Portugal. Available online at https://www.dssgfellowship.org/project/predicting-long-term-unemployment-in-continental- portugal/ (last accessed on 11.11.2024)
- DSSG Portugal (b): Mortality Surveillance. Available online at https://www.dssg.pt/en/projects/mortality-surveillance/ (last accessed on 11.11.2024)
- DSSG Portugal (c): Domestic Violence Data Observatory. Available online at https://www.dssg.pt/en/projects/domestic-violence-data-observatory/ (last accessed on 11.11.2024)

Open Educational Resources

ATTRIBUTION 4.0 INTERNATIONAL - Deed

- **You are free to:**
- Share - copy and redistribute the material in any medium or format.
- Adapt - remix, transform, and build upon the material for any purpose, even commercially.
- **Under the following terms:**
- Attribution - You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use. If you wish to use this work in a way not covered by the license, please contact:

Harz University of Applied Science
Friedrichstraße 57 – 59
38855 Wernigerode
E-mail: info@hs-harz.de