Supplementary document to the article “Primary Health Care Appointments and Hospital Stay: An Impact Analysis”

Joana Lopes[1] and Tiago Miranda[1]

[1] master’s students in Bioinformatics at the School of Engineering of the University of Minho

This document supports the article “Primary Health Care Appointments and Hospital Stay: An Impact Analysis” and aims to describe in detail the materials and methods used in this work. It should be noted that the original names of the dataset columns were maintained, which is why they are in Portuguese.

1 Materials

Both datasets were extracted directly from reliable sources, namely certified websites such as the Portuguese Public Administration's open data portal (dados.gov) and the “Transparency” area of the Portuguese National Health Service website, an Open Data initiative carried out by the Ministry of Health, in order to make available the vast set of data that underlie the operations and transactions that take place within the scope of the SNS's activities. This way we guarantee the reliability of the data.

1.1 Dataset 1: "atividade-de-internamento-hospitalar"

The dataset “atividade-de-internamento-hospitalar”[6], tracks the evolution of the number of discharged patients and hospitalization days by hospital institution and date (year and month) from January 2015 to December 2023. This dataset contains 14,215 records and 7 distinct columns that organize the data by period, region, institution, geographic location, type of specialty, discharged patients, and hospitalization days.

1.1.1 Column Description

The “Período” column stores data-type data in the year-month format. The “Região” column stores text data, associating each record with a specific region: “Região de Saúde do Alentejo”, “Região de Saúde do Algarve”, “Região de Saúde do Centro”, “Região de Saúde Lisboa e Vale do Tejo (LVT)” and “Região de Saúde Norte”. Additionally, the “Instituição” column corresponds to the respective hospital institution for each entry, and the “Localização Geográfica” column identifies the coordinates of each institution. The “Tipo de especialidade” column has values such as “Especialidade Cirúrgica”, “Especialidade Médica”, and “Outras Camas”. The “Doentes saídos” column stores integer data and considers all patients discharged from a given institution (excluding internal transfers), thus approximately corresponding to the number of hospitalizations, and will be renamed to “Número de internamentos.” Finally, the “Dias de Internamento” column represents the sum of hospitalization days used by patients whose stay exceeds 24 hours (excluding the discharge day).

1.2 Dataset 2: "evolucao-das-consultas-medicas-nos-csp"

The dataset “evolucao-das-consultas-medicas-nos-csp” [7], tracks the monthly evolution of in-person, remote/non-specific, and home medical appointments in Primary Health Care by Health Center Groupings (ACES) from January 2014 to December 2023. This dataset contains 6,900 records and 7 distinct columns that organize the data by period, region, entity, geographic location, number of in-person medical appointments, number of remote or non-specific medical appointments, and number of home medical appointments.

1.2.1 Column Description

The “Período” column stores date data in the year-month format. The “Região” column stores text data, associating each record with a specific region: “Região de Saúde do Alentejo”, “Região de Saúde do Algarve”, “Região de Saúde do Centro”, “Região de Saúde LVT”, and “Região de Saúde Norte”. Additionally, the “Entidade” column corresponds to the respective ACES for each entry, and the “Localização Geográfica” column identifies the coordinates of each entity. The columns “Número de consultas médicas presenciais”, “Número de consultas médicas não presenciais ou inespecíficas” and “Número de consultas médicas ao domicílio” store integer data representing the number of medical appointments provided to users by region/entity.

2 Methods

2.1 Data Preprocessing (Pandas)

This work followed ETL (Extract-Transform-Load) principles, and the data extraction phase has already been explored in the previous section. At this stage we transform our data for subsequent loading and storage in the database.

To handle our data, we used the Pandas library in Python, which allowed us to efficiently transform and prepare the data for analysis.

In the first step, after verifying the structure of the datasets, it was necessary to remove columns that were not relevant to our study, including the "Localização Geográfica" column from both datasets and the "Tipo de Especialidade" column from Dataset 1.

After that, given the presence of strange characters in Dataset 2, it was necessary to rename columns and rows that contained unreadable characters. Another important transformation was renaming the "Doentes Saídos" column to "Número de internamentos" as previously explained.

The most important step in this preprocessing, which allowed us to correctly merge the two datasets, was creating a mapping for each dataset. Since we had two columns in common in both datasets, "Região" and "Período" the initial solution was to merge by these columns. However, since for the same "Período"/"Região" pair in Dataset 1 we had multiple rows corresponding to the "Entidade" column in Dataset 2, we were, at best, duplicating the values in the "Número de internamentos" column.

For example, if in Dataset 1, for January 2015, for the Norte region, and for the Braga hospital institution, we had 500 hospitalizations, merging with Dataset 2 by "Período"/"Região" would result in this row being multiplied by the number of entities in Dataset 2 belonging to the Norte region. Suppose Dataset 2 had three distinct entities in the Norte region, each with a total of 100 appointments in January 2015. The merge would produce three rows for January 2015, Norte region, Braga hospital institution, each with 500 in the "Número de internamentos" column. This would lead to incorrect analysis results, with a query for hospitalizations in January 2015 in Braga showing 1500 instead of the correct 500.

Thus, the mapping allowed us to assign the institutions from Dataset 1 and the entities from Dataset 2 to the corresponding district and add a "Distrito" column to each dataset. After that, we aggregated our raw data: in Dataset 1, aggregating the number of hospitalizations and hospitalization days by "Região", "Período" and "Distrito"; in Dataset 2, aggregating the number of in-person, remote/non-specific, and home medical appointments by the same columns. This way, we had a single entry for each period/region/district in each dataset and could merge the two datasets by these columns.

2.2 Merging the Datasets

Using a time metric to measure the dataset merge, we found that using Pandas, the elapsed time was approximately 0.04 seconds. In comparison, using PySpark, the elapsed time was approximately 0.5 seconds. This difference is likely because the final dataset was not very large and could fit easily into the available RAM. In such cases, Pandas can be faster as it avoids the overhead of distributing data across a cluster.

After merging the datasets, we removed any missing values. We confirmed that our final dataset contained no missing values. The final dataset dimensions were 1944 rows and 8 columns.

2.3 Loading Data into MongoDB

Regarding data loading, the processed and ready-for-analysis dataset was loaded into MongoDB using a Python script with PyMongo. Since our data is temporary, meaning updated monthly, the data volume at the national level is likely to increase.

Therefore, we chose a highly scalable database like MongoDB. Additionally, even though it is a NoSQL database, MongoDB supports queries that allow for effective data filtering and analysis using the MongoDB Query Language (MQL) if needed.

2.4 Data Visualization and Analysis in PowerBI

Finally, the tool used for data analysis and visualization was PowerBI, which enabled us to create valuable interactive dashboards for data interpretation. To load our data, a connection to MongoDB was established, requiring the installation of an ODBC driver to access the data.

In PowerBI, when we started analyzing the data, we encountered some difficulties. Specifically, the comparisons between districts regarding the number of medical appointments and hospital stays were biased by the population differences in each district. Thus, it was necessary to import a dataset, "População Residente" [8], which allowed us to adjust the data according to the population of each district.

Additionally, we noticed that the “Número de internamentos” column represented the cumulative monthly values of hospital stays. For example, the value for February included the hospital stays for both January and February. Therefore, to calculate the metric of hospital stays per inhabitant, we added a calculated column “Monthly Hospital Stays” where each month's value was the current month's value minus the previous month's value.

Other metrics were generated to build the dashboards:

- “Total number of appointments” which corresponds to the sum of the number of medical appointments (non-presential or unspecified, home visits, and presential).
- “Hospital Stays per inhabitant” which corresponds to the division of the total number of monthly hospital stays by the resident population.
- “Appointments per inhabitant” which corresponds to the division of the total number of medical appointments by the resident population.
- “Non-presential appointments” which corresponds to the sum of non-presential or unspecified medical appointments and home visits.
- “Appointments per Hospital Stay Ratio” which corresponds to the division of the total number of appointments by the total number of hospital stays. In districts where this ratio is higher, we are dealing with districts that are supposedly more effective in terms of prevention.
- “Monthly Hospital Stays” which corresponds to the sum of the monthly hospital stays.

Since the analysis was to focus mainly on the years rather than the specific months of each year, a calculated column “Year” was generated, while still allowing the user to filter the data by month.