COVID-19 Data Sharing/BR

INSTRUCTIONS TO CREATE DATABASE AND UPLOAD DATA

Danilo Carlotti, João Eduardo Ferreira, Fátima L. S. Nunes

1. Introduction

This document presents a relational model, the necessary programs to create the database and upload the files available in the repository COVID-19 Data Sharing/BR, available in https://repositoriodatasharingfapesp.uspdigital.usp.br/.

The set of files in this folder contains:

- analises_descritvas.py – program that generates descriptive analysis (see section 6)
- connection_mysql.py – program to connect to mysql server (see section 5)
- create_db.sql – script in sql format that creates the database according to the data dictionary (see section 4)
- data_dictionary.pdf – data dictionary (see section 5)
- instruções.pdf – this document in portuguese
- instructions.pdf – this document in english
- map_server_bokeh_infections_per_city.py – program that creates html file with infections per city
- model_db.mwb – mysql workbench file with the model for the database (see section 2)
- parse_exams.py – program used to parse exam files (see section 5)
- standard_reference_value.py – program used to help with the transformation of data (see section 5)
- upload_dados.py – program used to upload files in the database (see section 5)
- validator.py – program used to check the integrity of files, their encoding, the columns and other properties before uploading

The files with anonymized data from COVID-19 patients can be found in https://metabuscador.uspdigital.usp.br/handle/doc/12344.

The necessary steps to install the server can be found in section 3.

2. Data model

Figure 1 presents the Entity Relation Diagram and it contains a description of the tables from the database and their relations according to the data dictionary. It is also reproduced below.

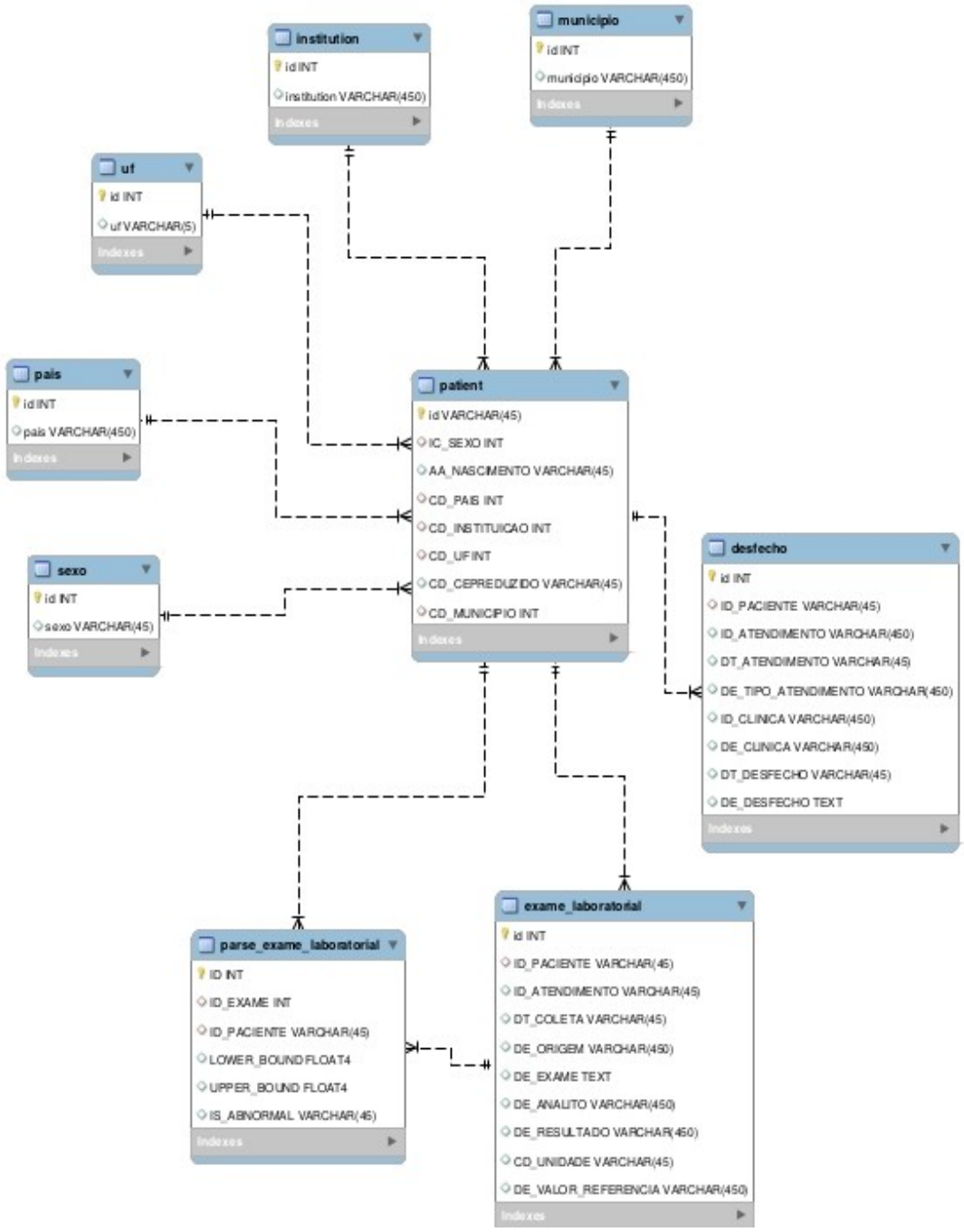

Figure 1 – Entity-Relation Diagram of the database

3. Installing the server

The database was designed to work with a SQL server, such as the free version of MySQL. This server can be downloaded from https://dev.mysql.com/downloads/mysql/.

A tutorial on how to install and configure the server can be found in: https://dev.mysql.com/doc/refman/8.0/en/windows-installation.html.

If the users wishes for a Graphic User Interface they can use MySQL Workbench: https://dev.mysql.com/doc/workbench/en/wb-installing-windows.html.

4. Script to create the database

The script used to create the database and all its tables can be found in "*create_db.sql*". Using the graphical user interface or through command line the user can create the database.

5. Upload and data processing

All programs were written in Python. The requirements to execute the programs are:

- python version 3.5 or higher
- pandas
- pymysql
- geopandas
- matplotlib
- bokeh

Python (version for Windows) can be installed with the following link: https://www.python.org/downloads/.

The libraries can be installed with pip. A tutorial is available in: https://docs.python.org/3/installing/index.html.

The file containing the necessary method for connecting to the database is "*connection_mysql.py*". The address for the server, user and password are, respectively: 'localhost', 'root', 'root'. This setting can be changed by the user.

The data can be uploaded only if in compliance with what is described in "data_dictionary.docx".

To run the program that uploads the data it is necessary to supply the program with three arguments:

python3 upload_dados.py "[PATH TO DATA FILE]" [TYPE OF FILE(exame,paciente ou desfecho)] "[INSTITUTION NAME]"

Example:

python3 upload_dados.py "hsl_small_dataset_fapesp/small/COVID_PATIENT.csv" paciente "HSL"

After uploading all data, it is necessary to run the program "parse_exams.py", without any arguments.

The results, after parsing, are stored in the table *parse_exame_laboratorial* and they indicate indicating if the exam was abnormal or not.

6. Preliminary visualizations

There are two programs that generate data visualization.

The first one is "*analises_descritivas.py*". This program has functions that create html files with pizza plots for: the cities of the patients, the outcomes, patients sex and the top 15 most abnormal exams. It is also possible to generate a histogram with patients age in intervals of 10 years. There is also a function used to create an excel file with a count for the cities, used in the next program described.

The second file is "*map_server_bokeh_infections_per_city.py*" and it uses the library bokeh to visualize the number of infected per city. The names of the cities, however, have to match the names in the shape file for the state of São Paulo (*35MUE250GC_SIR.shp*).

Examples of data visualization that were generated in 06.26.2020 can be found below in figures 2 through 7.

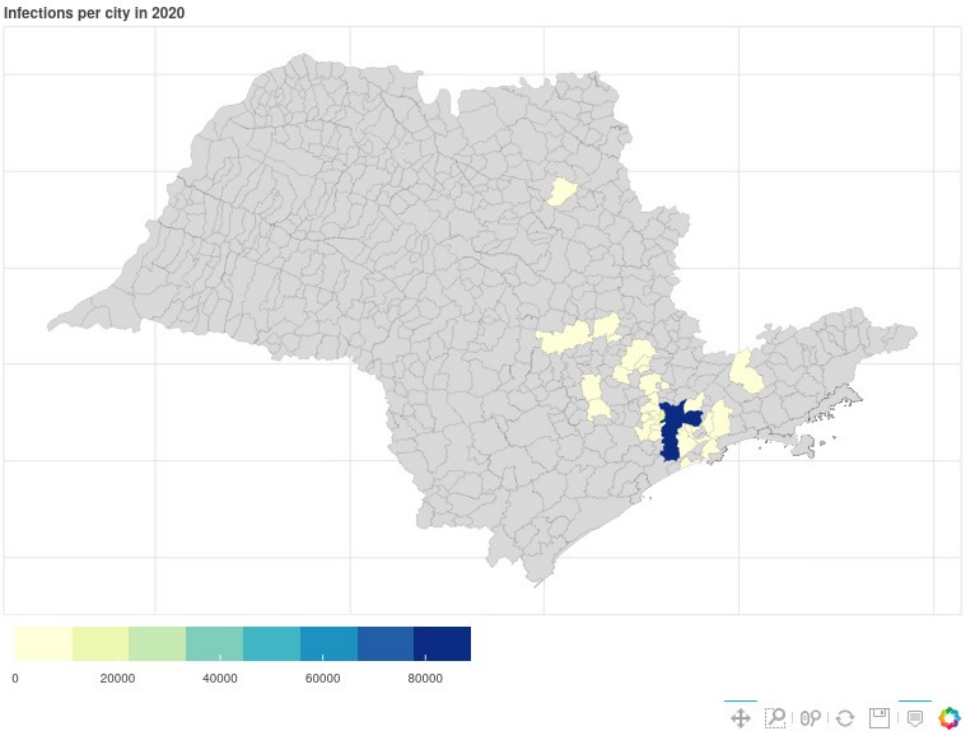

Figure 2 – Map of infections in cities in the State of São Paulo

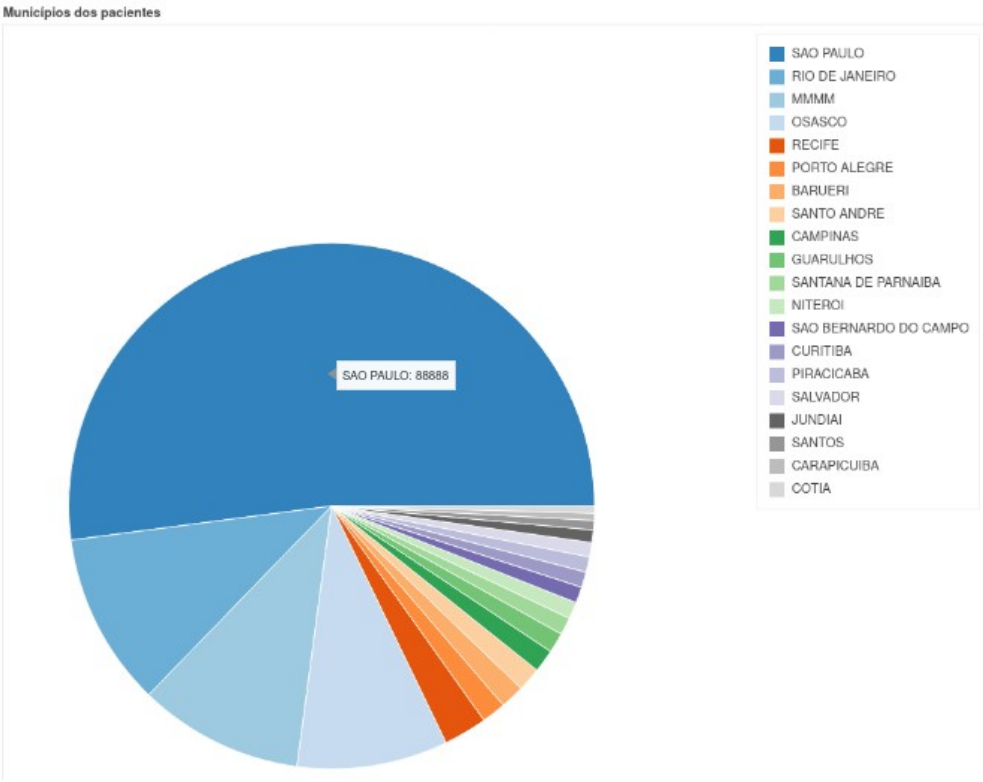

Figure 3 – Pie plot of infections in cities

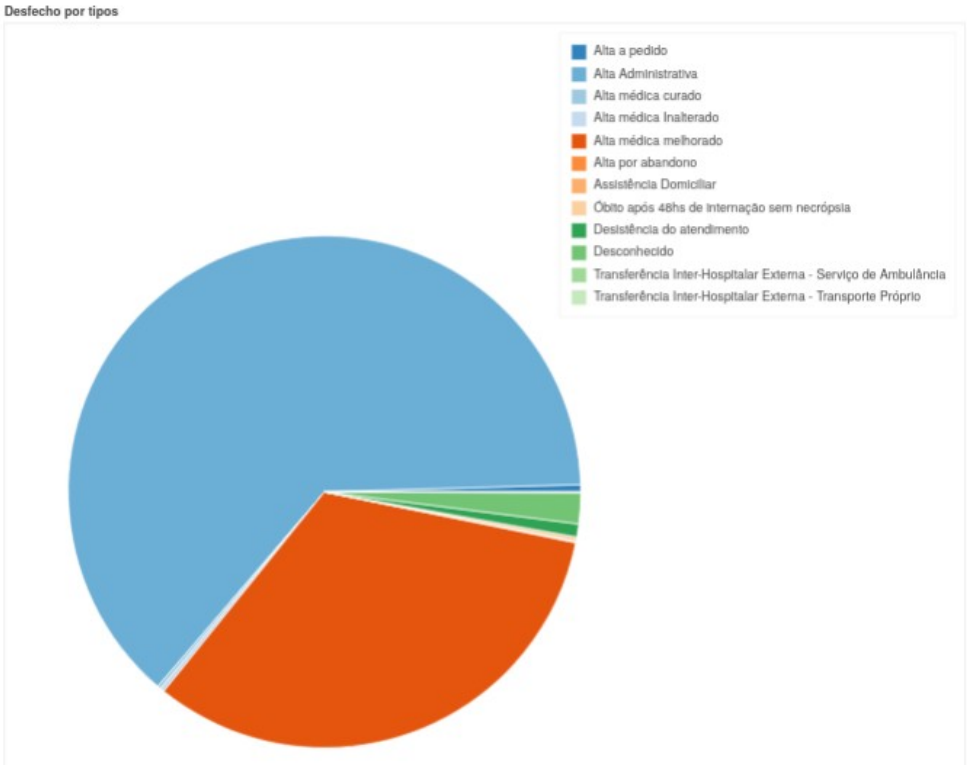

Figure 4 – Pie plot of outcomes

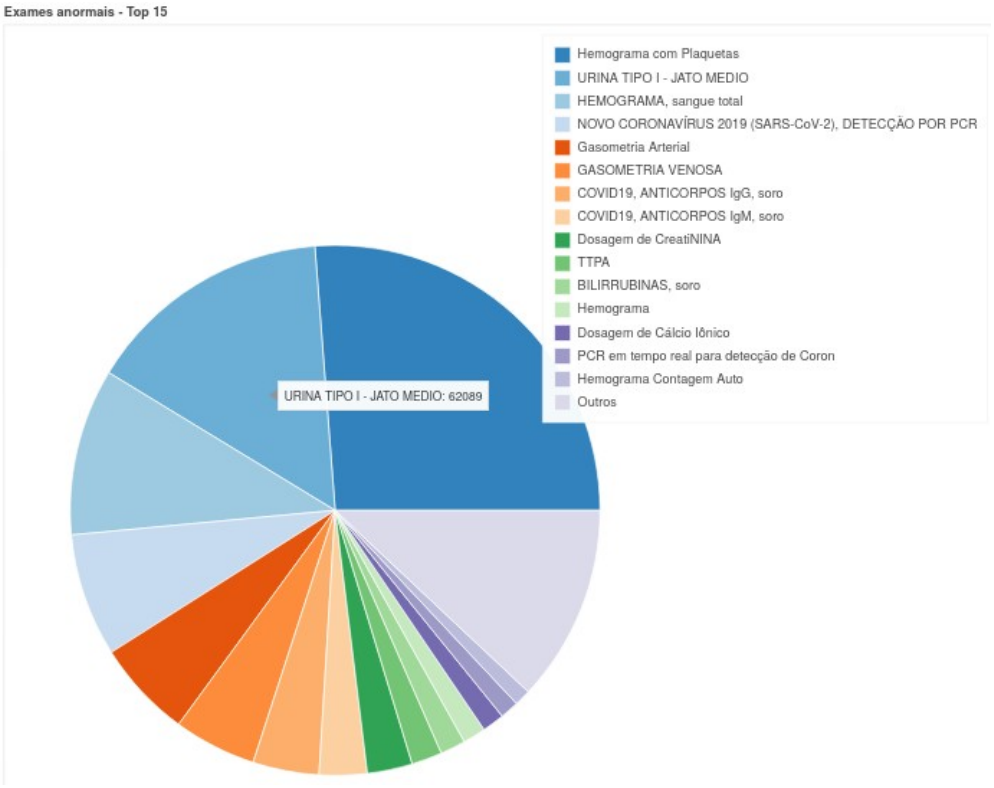

Figure 5 – Pie plot of most abnormal exams

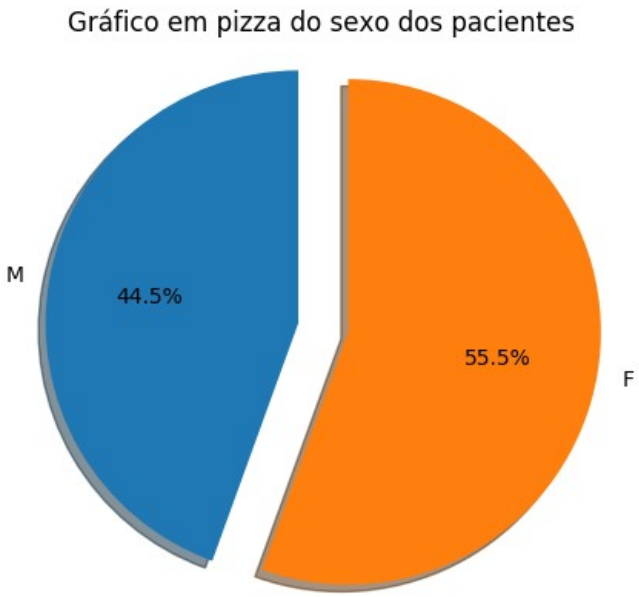

Figure 6 – Pie plot of patient's sex

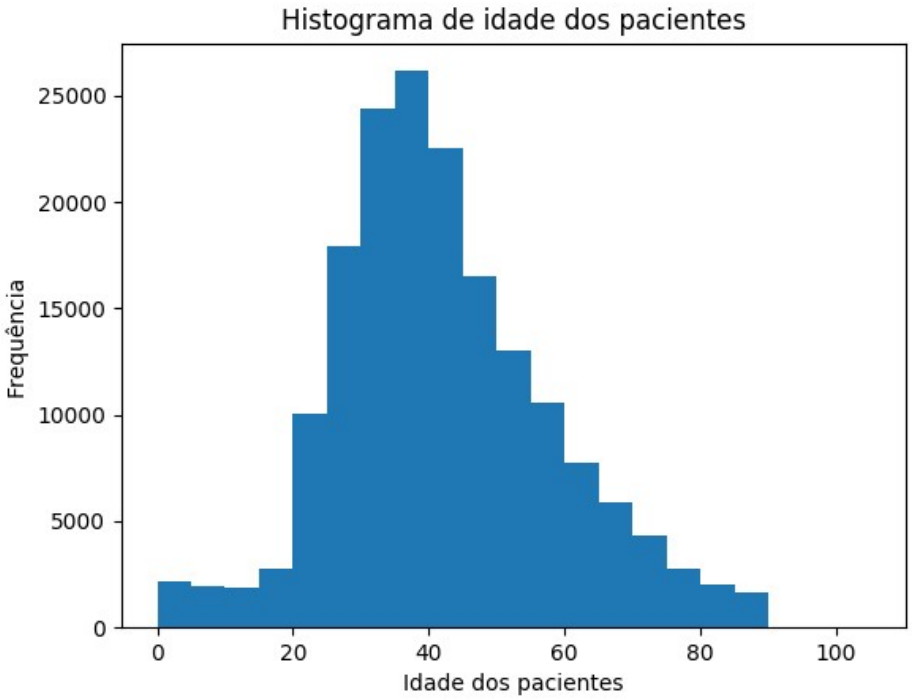

Figure 7 – Histogram of patient's age