Mid-Semester Progress Report

DSA5900 – Spring 2023

Eduardo Cerqueira

03/20/2023

Introduction

The International Syndromic Surveillance project focuses on monitoring the COVID-19 virus. It is crucial to track the number of COVID cases, deaths related to COVID, and hospitalization rates. Early access to projected rates would greatly benefit public health agencies, epidemiologists, governmental organizations, and decision-makers. With access to these surveillance systems, decision-makers responsible for health-related issues can respond more quickly and effectively in their efforts to contain the spread of viruses like COVID-19.

PanViz 2.0 is an interactive web-based tool that utilizes multiple data streams to simulate and model disease spread metrics across the United States. These metrics include positive cases, deaths, hospitalizations, and infection rates at different spatial hierarchical regions, such as country, state, and county. The system was developed by researchers at OU DISC in collaboration with other departments across campuses of OU and OUHSC. It can be used to project COVID-19 cases, deaths, and hospitalization rates, and the project aims to expand the system's regional capacity to include countries like Peru, which have a different geographical hierarchy.

Objectives

Peru's hierarchy has four levels - country, department, province, and district - and the project will explore the country's current COVID-19 data streams to find optimal ways to process the information into a meaningful aggregated form similar to the data provided by the New York Times. The data will be evaluated at different temporal levels, including daily, weekly, biweekly, monthly, and annually. Several machine learning algorithms will be run to project cases, deaths, and hospitalization rates, and the results will be integrated into the PanViz system.

The project aims to provide knowledge of Peru's COVID-19 data and the current PanViz system for the US. Automated scripts will be created to perform data aggregation for Peru and run machine learning models for integration.

Data

The Peru government website (https://covid19.minsa.gob.pe/sala_situacional.asp) is the source of data for this project, including the number of positive COVID-19 cases, deaths, and hospitalizations. Although the data is primarily numeric, there are also qualitative data such as location, yes/no questions, and vaccine types, which need to be converted to quantitative data. My current responsibility is to collect data, which involves inserting new data into the Peru database file each day while removing duplicate data. To accomplish this task, I have generated a Python script that performs web scraping on Peru's National Database, cleans and converts the data to a CSV file, and uploads it into a database. The problem arises when removing duplicate data from the next day, as each following day includes previous data that was already uploaded into the

database for the PanViz system's usage. However, I overcame this obstacle by incorporating a merging process into the code using each patient's UUID, which is a unique ID for each individual. Furthermore, the PanViz user's computer is configured to run the Python script every day automatically, ensuring that new data is available for later data aggregation and processing to understand Peru's database and geographical hierarchy.

Methodology

Techniques, Results and Analysis

To scrape and filter data from the Peru website, I used a combination of Python scripts provided by my sponsors. I blended the scripts into one and defined functions to collect data into a DataFrame, clean it, and create a data quality report for analysis. Using the SQLite and Pandas packages, the code reads SQLite query results into a Pandas DataFrame and then created and inserted the data into a table. To avoid duplication, the code performed a merging process on the DataFrame before sorting the data by the date of results. This process was repeated for positive cases and hospitalizations, with deaths being the first table created. Outliers were identified, and tasks were automated to create reliable, accurate, and easily accessible models that will provide clear and concise information for future generations.

The final deliverable will be to integrate the Peru data into the PanViz system. The existing functionalities of the PanViz system will be expanded to communicate findings using standard plotting, such as line and bar charts, to predict COVID cases for future dates. The project is expected to be completed by April 28th, 2023. During the following three months of this practicum study, I will partition my time by understanding the codebase, converting the Peru data into more structured data, running the existing machine learning models on the Peru data, and storing the results, and integrating all features into the PanViz system for final analysis.

Deliverables:

- Evaluating the performance of several machine learning models in doing predictive analysis.
- Integrating the final results into the Peru version of the PanViz.

References

https://covid19.minsa.gob.pe/sala_situacional.asp

Appendix

Figures

	metodo *TEXT*	edad *INTEGER*	sexo *TEXT*	criterio_fallecido *TEXT*	fecha_resultado *TEXT* ↓	UBIGEO *INTEGER*	UUID *INTEGER*
1		25	FEMENINO	NULL	2023-03-20	150131	13914260
2		26	MASCULINO	NULL	2023-03-20	110101	14469495
3		46	MASCULINO	NULL	2023-03-20	110110	14516358
4		54	FEMENINO	NULL	2023-03-20	150122	14908588
5		62	FEMENINO	NULL	2023-03-20	80201	15375959
6		58	FEMENINO	NULL	2023-03-20	40103	15819273
7		73	MASCULINO	NULL	2023-03-20	150101	15908931
8		24	MASCULINO	NULL	2023-03-20	150133	16407990
9		77	FEMENINO	NULL	2023-03-20	120401	17221492
10		49	MASCULINO	NULL	2023-03-20	40101	18385293
11		17	FEMENINO	NULL	2023-03-20	150108	18436340
12		47	MASCULINO	NULL	2023-03-20	150122	20091673
13		28	FEMENINO	NULL	2023-03-20	150133	20899494
14		31	FEMENINO	NULL	2023-03-20	200115	21020158
15		28	MASCULINO	NULL	2023-03-20	150113	21058434
16		34	FEMENINO	NULL	2023-03-20	150132	21657899
17		20	MASCULINO	NULL	2023-03-20	150133	21303986
18		67	MASCULINO	NULL	2023-03-20	60101	21587680
19		29	MASCULINO	NULL	2023-03-20	150140	21833379
20		65	MASCULINO	NULL	2023-03-20	150136	22799727

F.1- Table example in database file for positive cases, generated by the python script.

	UUID *INTEGER*	fecha_recopilacion *TEXT*	fecha_resultado *TEXT*	edad *INTEGER*	sexo *TEXT*	criterio_fallecido *TEXT*
1	4746095	2022-12-13	18/08/2022	58	F	Serological
2	3064284	2022-12-13	12/04/2021	50	M	Virological
3	36730475	2022-12-13	13/02/2021	70	M	SINADEF
4	13111055	2022-12-13	17/07/2021	44	M	SINADEF
5	11651032	2022-12-13	7/06/2021	61	F	SINADEF
6	321058	2022-12-13	20/09/2022	66	F	Clinical
7	443319	2022-12-13	16/03/2021	66	F	Serological
8	36855263	2022-12-13	16/02/2021	63	M	Virological
9	15813827	2022-12-13	29/06/2022	55	F	Serological
10	404083	2022-12-13	14/05/2021	72	M	SINADEF
11	9900866	2022-12-13	25/04/2021	41	M	SINADEF
12	11129415	2022-12-13	26/04/2021	74	M	SINADEF
13	20581563	2022-12-13	6/05/2021	87	F	SINADEF
14	10671471	2022-12-13	17/09/2022	78	M	Clinical
15	6728415	2022-12-13	24/08/2022	83	M	Serological
16	649856	2022-12-13	13/09/2022	70	M	Serological
17	1020608	2022-12-13	2/04/2021	90	M	SINADEF
18	12734548	2022-12-13	21/11/2021	65	M	SINADEF
19	18525	2022-12-13	27/03/2021	76	M	SINADEF
20	3000706	2022-12-13	9/12/2022	90	M	Serological

F.2- Table example in database file representing hospitalization rates generated by the python script.

	fecha_recopilacion *TEXT*	fecha_resultado *TEXT* ↓	edad *INTEGER*	sexo *TEXT*	criterio_fallecido *TEXT*	departamento *TEXT*
7	2023-03-20	2023-03-19	82	FEMENINO	Virological	LIMA
8	2023-03-20	2023-03-19	89	MASCULINO	Virological	LIMA
9	2023-03-20	2023-03-19	63	MASCULINO	Virological	LIMA
10	2023-03-20	2023-03-19	83	FEMENINO	Virological	LIMA
11	2023-03-20	2023-03-19	84	FEMENINO	Virological	LIMA
12	2023-03-20	2023-03-19	56	MASCULINO	Virological	LIMA
13	2023-03-20	2023-03-19	82	MASCULINO	Virological	LIMA
14	2023-03-20	2023-03-19	85	FEMENINO	Virological	LIMA
15	2023-03-20	2023-03-19	98	MASCULINO	Virological	LIMA
16	2023-03-20	2023-03-19	69	FEMENINO	Virological	LIMA
17	2023-03-20	2023-03-19	66	FEMENINO	Virological	LIMA
18	2023-03-20	2023-03-19	85	FEMENINO	Virological	LIMA
19	2023-03-20	2023-03-18	24	FEMENINO	Virological	ICA
20	2023-03-20	2023-03-18	64	FEMENINO	Virological	PUNO
21	2023-03-20	2023-03-18	88	FEMENINO	Virological	LIMA
22	2023-03-20	2023-03-18	64	MASCULINO	Virological	LIMA
23	2023-03-20	2023-03-18	77	MASCULINO	Virological	ICA
24	2023-03-20	2023-03-18	88	FEMENINO	Virological	LIMA
25	2023-03-20	2023-03-18	79	FEMENINO	Virological	JUNIN
26	2023-03-20	2023-03-17	86	FEMENINO	Virological	LIMA

F.3- Table example in database file for death cases, generated by the python script.