A Visual Analysis on the Impact of Covid-19 on Hardest Hit Areas Using Machine Learning Algorithm

Jia Lin
Team C, DBAP CA2
National College of Ireland
Dublin, Ireland
x22117644@student.ncirl.ie

Min Chen
Team C, DBAP CA2
National College of Ireland
Dublin, Ireland
x19205040@student.ncirl.ie

***Abstract*—In the context of the Covid-19 pandemic, this project concentrates on a further analysis on factors which impact the status of the hardest hit areas in USA. The result has been determined by using a Machine Learning algorithm, decision tree. A Covid-19 dashboard has been created to visualise the pandemic situation of the entire world and USA, and identify the hardest hit areas in USA. This project, followed the CRISP-DM methodology flow, has built a ETL pipeline to implement transforming unstructured data to structured data. In this process, MongoDB is used to store the unstructured data, PostgreSQL is used to organise the structured data. Python programming language is used to complete the whole project.**

***Index Terms*—Covid-19, Machine Learning, Decision Tree, Dashboard Visualisation, ETL pipeline, MongoDB, PostgreSQL**

I. Introduction

The Covid-19 pandemic has caused enormous society and community crisis word widely. The impact of the Covid-19 has been studied, and research outcome has provided academic support and guidance for communities and societies to live with the pandemic. While the pandemic has been hitting harder on certain areas and communities. Obviously, those unequal impact implies that there are some other key factors behind which have impacted the community status differ from each other. This project is focusing on visualising the factors, which impact status of hardest hit areas apart from Covid-19 pandemic using decision tree machine learning algorithm.

Moreover, the impact of the Covid-19 on human society is unprecedented, and everyone's life has been greatly affected. People have to continue to pay attention to the changes of Covid-19 data. And the best way to exhibit such data changes is statistical visualisation which can be used to help people discover and identify the hardest hit areas in time.

II. Related Work

A. Dataset 1. Covid-19 Vulnerable Community Crosswalks

It is now the third year since early 2020 Covid-19 pandemic broke out and spread-out world widely, which has caused over 6.6 million global human life deaths by the end of November 2022 [1]. Studies on the pandemic impact on different territories, areas and communities has been explored for social support purposes [2], [3]. Gao [4] discovered Covid-19 has had immediate negative impact on poverty concentrating within specific groups as regards to their family status, hardship, education, and race or ethnicity. Islam et al(2022) states that communities in the United States with low income and minor racial or ethnic populations have been disproportionately hit by the pandemic [5]. However, previous research has not been found with any conclusions that if rural community or tribal community, low income, and poverty, as well as poverty percentage can really determine the status of the hardest hit area apart from the pandemic, which dataset 1 could be explored to answer this question.

B. Dataset 2. Covid-19 Statistics

The most direct way to know the latest news of Covid-19 is to search online. The most prominent data on the web are statistics from different organisations, governments, research institutes and universities. As one of the most authoritative data providers, World Health Organisation (WHO) [6] built it dashboard [7] to show the global pandemic. People can see the number of confirmed cases and deaths cases for the entire world everyday. And such two numbers of each regions and countries are displayed as well. The Johns Hopkins University (JHU) [8] has the most prestigious and top medical school, as famous as Harvard Medical School. Since its ranking, the School of Public Health has been ranked No.1in the United States of America (USA) over the years. Johns Hopkins Center for Systems Science and Engineering (CSSE) [9] provides a dashboard of Covid-19 [10], they presents more detailed information for Covid-19 data in USA. Ireland government combines the Covid-19 daily data and the vaccination information together [11], to demonstrate the Covid-19 circumstances.

III. Methodology

"The CRISP-DM methodology" [12] is applied for this project in order to achieve practical and realistic goals base on the two datasets chosen for the project. Fig. 1 presents the methodology flow of this project.

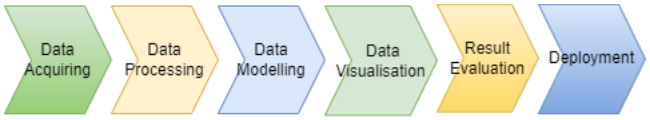


Fig. 1. Diagram of Methodology Flow

An ETL (Extract, Transform, Load) pipeline (see Fig. 2) has been deployed throughout this project.

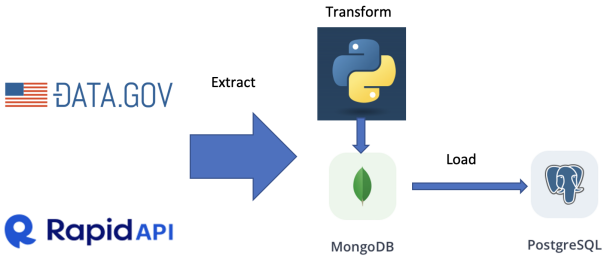


Fig. 2. ETL Pipeline.

A. Data Acquiring

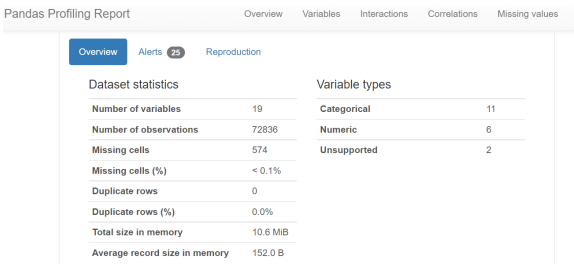
- Dataset 1. Covid-19 Vulnerable Community Crosswalks
  Dataset 1 was acquired at DATA.GOV website, which crosswalks tribal community, rural community, low-income area, poverty score and percentage on those census area with hardest hit area status affected by the Covid-19; all data values are weighted using FCC's scoring method for evaluation [13]. It gives the first academic reason to choose this dataset for studying and making analysis, apart from the motivation for choosing dataset 1 based on relative previous research and discoveries.

TABLE I
DATASET 1 VARIABLE DESCRIPTION

Variable Description
1.STATE: state in the u.s.
2.ST_ABBR: state abbreviation in the u.s.
3.STCNTY: county number of the state
4.COUNTY: name of the county
5.FIBS: fib numbers
6.COUNTY FIBS: county fib numbers
7.LOCATION: detailed census location
8.Total Score: total score for the census community
9.Max Possible score: maximum possible score from the census for the community
*10.Hardest Hit Area(HHA): status of the affected community by Covid-19 * the target variable for ML tree analysis*
11.HHA Score: scores for different hardest hitting areas
12.Low Income Area (LIA) County SAIPE - (Poverty Percentage): poverty percentage for low-income area county SAIPE
13.Low Income Area (LIA) County SAIPE- Score: census score for low-income area county SAIPE
14.Low Income Area (LIA) Census Tract (Poverty Percentage): poverty percentage for low-income area census tract
15.Low Income Area (LIA) Census Tract - Score: score for low-income area census tract
16.Tribal Community\n(1 if yes): tribal community
17.Tribal Community\nScore (Geographic Only): tribal community census score
18.Rural: rural area
19.Rural - Score: rural area score

The configuration for dataset 1 is downloaded by config.Json reproduction. The variables and overview of such dataset see Table I and II. It overviews 72836 observations with 19 attributes, suitable for applying python with machine learning algorithms to model and access

TABLE II
OVERVIEW OF DATASET 1



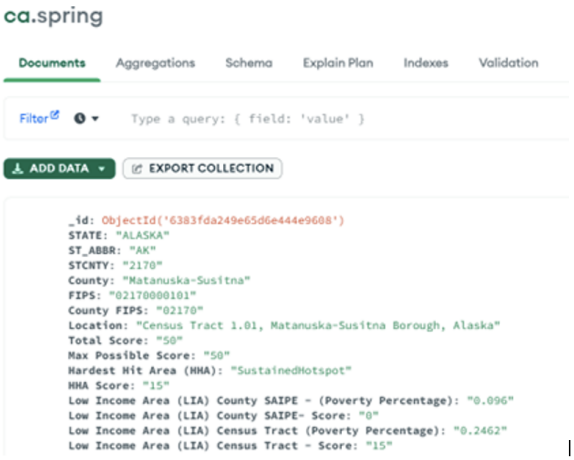
to achieve research purposes, which explains the second reason to choose this dataset 1.

- Dataset 2. Covid-19 Statistics
  Covid-19 Statistics [14] is a free API which is build based on public data by Johns Hopkins Center for Systems Science and Engineering (CSSE). The api provides the Covid-19 data for the entire world and statistics of data in USA by states as well. Data is updated everyday, the latest information of total number of confirmed cases, the daily new cases, the total number of deaths cases, the new deaths cases, the fatality rate and so on. In order to create a dashboard of Covid-19 for presenting the current spread of such disease, data has been extracted from this API. By sending specific request queries to the api, JavaScript Object Notation (JSON) objects have been returned.

B. Databases Management

- Dataset 1. Covid-19 Vulnerable Community Crosswalks
  MongoDB is applied to import semi-structured Json file dataset , because dataset 1.Json is larger than 16MB, which has exceeded the maximum data storage capacity in Json format at MongoDB, instead CSV format was imported to Mongodb with a data capacity of 10.6MB (see Table III).

TABLE III
DATASET 1 IMPORTED TO MONGODB



As a fully cloud-based data platform, MongoDB provides many advantages for dataset processing. It can provide bigger, safer, and more flexible database. The data for processing will be fetched from MongoDB by using Python programming languages. The environment of Jupyter in Anaconda will be applied to achieve the whole methodology flow. Cleaned data would be stored into PostgreSQL database for further analysis.

- Dataset 2. Covid-19 Statistics
  MongoDB: Data acquired from the source API is semi-structured organised in JSON format. As a NoSQL database, the schema-free feature of MongoDB, allows it can read and store JSON data with a high performance [15]. Data obtained from the API is stored in a specific MongoDB database as collections. The these collections are easy to be converted into CSV files. A MongoDB database called "jialin_Mongo_database" has been created to store these collections (see Fig. 3).

```
database names:
admin
config
jialin_Mongo_database
local
test_database


collection names:

['global_date_over_time_collection',
 'reports_US_data_over_time_collection',
 'provinces_of_US_collection',
 'reports_US_provinces_cities_collection']
```

Fig. 3. Databases in MongoDB and Collections In the database

PostgreSQL: When data has been cleaned and transformed into CSV files, a specific PostgreSQL database has been built to load such structured files as tables. By copying contents in CSV files to tables in a PostgreSQL database, they are ready for statistical analytics and visualisation. A PostgreSQL database named "jialin_postgresql_database" has been built to load all structured tables (see Fig. 4).

```
Tables in PostgreSQL database jialin_postgresql_database:
('global_data_over_time',)
('reports_us_data_over_time',)
('provinces',)
('reports_us_provinces_cities',)
('new_provinces_data',)
('final_provinces_data_hha',)
('new_provinces_with_state_code',)
('min_hha',)
```

Fig. 4. Tables in the PostgreSQL database

C. Programming Language and Machine Learning Algorithms

- Dataset 1. Covid-19 Vulnerable Community Crosswalks
  Python3, as a high-level, object-focused, and highly efficient programming languages working with machine learning(ML) algorithms, is used for data processing and decision tree ML analysis for dataset 1.
  Decision tree, as a supervised ML algorithm, is ideal for identifying predictions, outcomes, or decisions by a predictive modelling approach [16]. Decision tree ML algorithm will help to fulfil the goal of dataset 1 to see if status of hardest hit area by Covid-19 is also affected by low income, rural community, tribal community, and poverty percentage of those areas. A collection of libraries and packages is imported to suit for data overviewing, data fetching from MongoDB, missing data processing, variable rename, data modelling and decision tree ML analysis, as well as data visualisation.
  - pandas, pandas_profiling, and numpy are impoted for general operations.
  - Pymongo to connect Python with MongoDB for fetching data purposes.
  - psycopg2&sqlalchemy to connect python with PostgreSQL DB for fetching data.
  - KNeighborsClassifier is applied for decision tree model analysis.
  - Matplotlib and graphviz are applied for data visualization.
- Dataset 2. Covid-19 Statistics Python [17] was created by Guido van Rossum in 1991 [18]. As a high level programming languages, Python was designed to be readable, hence, it is understandable and reusable [19]. Python supports Object-oriented (OO) design and functional programming [19], so it can be accepted by most of programmers easily. Python is selected to do the programming part of this project, since it provides plenty of libraries. These libraries support to accomplish all requirements of this project and the entire process of implementation step by step. Firstly, data has been collected from the API, JSON objects have been returned. Secondly, data has been preprocessed and cleaned, by connecting to MongoDB, semi-structured data has been stored into a MongoDB database. Thirdly, data has been converted in CSV format and loaded into a PostgreSQL database. Eventually, statistical results have been visualised in a dashboard. The libraries used in this project to present the spread of Covid-19 are listed in Table IV.
  Jupyter Notebook [20] is an open source technology which allows to run many different popular programming languages on it, such as Python and R [21]. Jupyter notebook was developed in 2014, it is used to edit and run the code and show the visualisations, like charts, plots, maps and so on. Jupyter notebook has been chosen as the programming platform to implement this project and it worked very well[1].

[1]In order to make the execution time shorter, the author modified the "ioput rate" with terminal command:
jupyternotebook − −NotebookApp.iopub_data_rate_limit = 1.0e10

TABLE IV
LIBRARIES CHOSEN IN THIS PROJECT

No.	Library Name	Utilisation
1	request	Library request is used to fetch data from API.
2	json	Library json is used to load json objects.
3	datetime	Libary datetime is used to convert between datetime and string.
4	pymongo	Library pymongo is used to build connection with MongoDB.
5	pandas	Library pandas is used to create a ETL pipeline, It is used to implement visualisation as well.
6	psyconpg2	Library psycopg2 is used to make connection with PostgreSQL.
7	cvs	Library csv is used to store csv to PostgreSQL.
8	sqlalchemy	Library sqlalchemy is used to connect with PostgresSQL.
9	plotly.express	Library plotly.express is used to create charts, such as line, scatter, bar charts, etc.
10	plotly.subplots	Library plotly.subplots is used to organise plots in a single page.
11	plotly.graph	Library plotly.graph is used manage graphs, in particular to realise display charts in a dashboard.
12	dash	Library dash is used to create dashboard

D. Data Processing

- Dataset 1. Covid-19 Vulnerable Community Crosswalks (Data processing and data modelling)
  There are 3 main steps involved in data processing and data modelling.
  Step 1. Fetching Data from Mongo DB (see Fig. 5).

Fig. 5. Fetching dataset 1 from MongoDB

Step 2. Processing data types, missing values, and rename columns. Fig. 6 illustrates the frequency of the processed numerical data.
Fig. 7 outlined statistical description of processed numeric data for Dataset 1, which indicates that values underneath 0 represent processed missing values or NAs from selected variables.
Step 3. Data Modelling
Identifying target variables, featured variables through variable correlations and Cramér's V (ϕc) analysis.
Cramér's V coefficient measures the relational strength of associated variables, which "coefficient ranges from 0 to 1" indicates a perfect association [22].

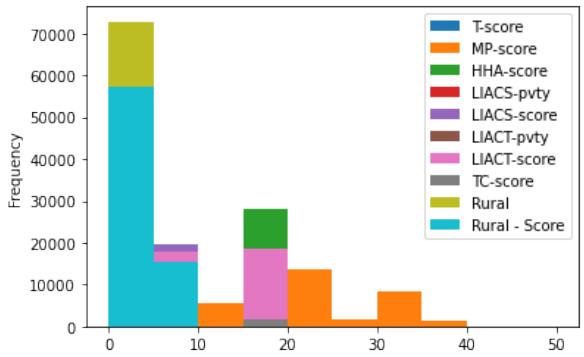


Fig. 6. Histogram of Processed Numeric Data (Dataset 1)

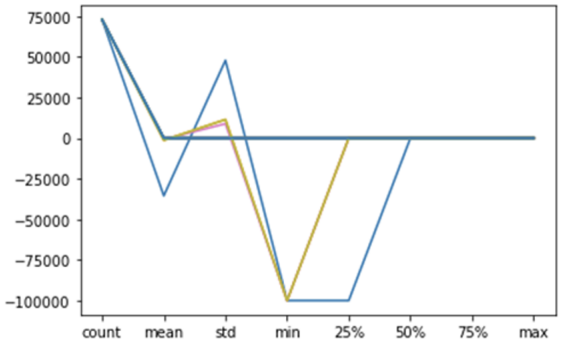


Fig. 7. Statistical Description of Numeric Data (Dataset 1)

The status of hardest hit area(HHA) will be selected as the target variable because HHA shows perfect association with many of the other variables like HHA score, low-income county area, rural community, and tribal community (see Fig. 8, 9). Featured variables are identified as:

- T-score, MP-score, HHA-score,
- LIACS-pvty, LIACS-score,
- LIACT-pvty, LIACT-score,

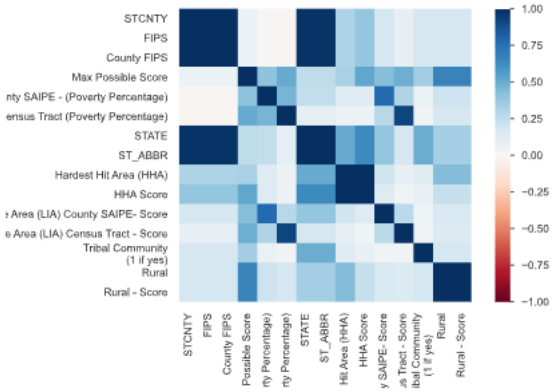


Fig. 8. Correlations between variables(Dataset 1)

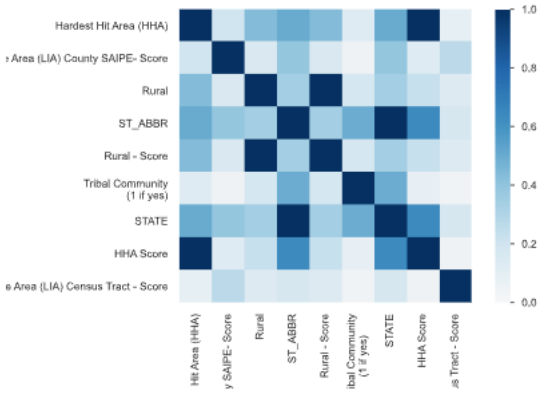


Fig. 9. Cramér's V (ϕc)- Dataset 1

– TC-score, Rural, Rural – Score.

- Dataset 1. Covid-19 Vulnerable Community Crosswalks (Visualising Model Analysis Flow)
  1. Generate Decision Tree Log for HHA. see Fig. 10

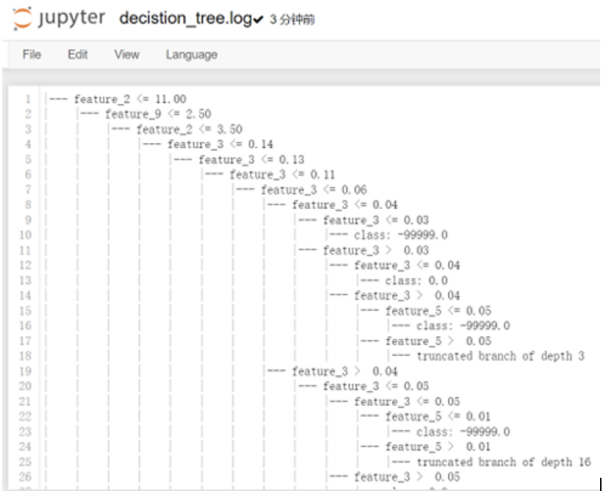


Fig. 10. Part of HHA Decision Tree Log (Dataset 1)

Featured variables contain different classifications belong to other class of features, which enables the tree to split from root node to end leaves [23].

2. Display HHA tree to Depth 7, see Fig. 11, the complete HHA tree see Fig. 17
3. Overview of the Whole HHA Tree, see Fig. 12
   A detailed closer look could help understand the nodes of the tree. Root node, decision node, and end leaf node are the main components of the HHA decision tree, see Fig. 13. Root node of HHA tree starts with the white box. It represents the main objective of the tree; the rest of the branches and leaves are all derived from this node.

– HHA-Score $\leqslant$ 11.0: The objective of the whole HHA tree is if HHA-Score is less than or equal to 11.0. True or False to follow.

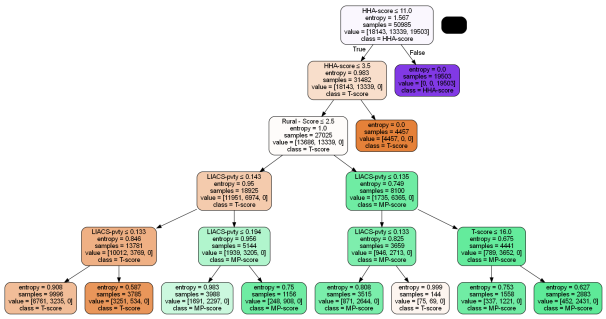


Fig. 11. HHA Tree to Depth 7 (Dataset 1)

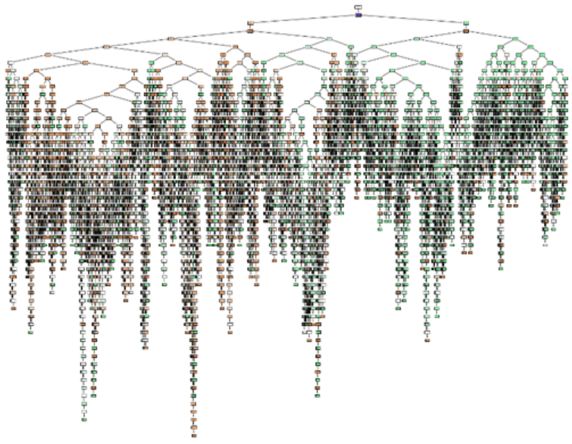


Fig. 12. Overview of the Whole Tree (Dataset 1)

– Gini $= 0.659$: Gini score larger than zero quantifies the contained samples in that box belongs to various classes. If gini score is zero means no samples belong to any more class, the node should be the end of terminal leaves.
– Samples $= 50985 : 70\%$ of the trained data from original.
– Value $= [18143, 1339, 19503]$: 18143 is the sample numbers of Rural-Score; 13339 is total samples of

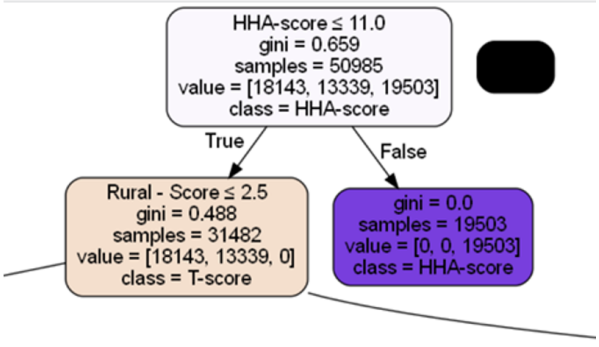


Fig. 13. Main Components of HHA Tree

T-Score; 19503 means number of samples of HHA-Score > 11.0.

- Class = HHA-Score: Class illustrates the prediction of HHA-Score are made by the value list, on the bases which HHA-Score $\leqslant 11.0$.

Decision node of HHA tree are series of nodes emerged from the root node, see Fig. 12 and 13. Each node questions a decision to be made or node splits at the point. Leaf node means the end of leaves with outcomes, like the purple box above implies samples with HHA-Score > 11.0 have no more decisions to make.

- Dataset 2. Covid-19 Statistics
  The "Covid-19 Statistics" API provides daily data for the entire world and USA. Only one day data cannot used to illustrate the spread of Coronavirus. A time series of data needs to be created to display the history and the trend of the Covid-19. By setting a specific number of days, a list of results returned to present a time series changes of Covid-19. Before analysing and visualising, the collections of data needs to be cleaned, any irrelevant data needs to be expunged and the missing data needs to be replenished or deleted. The final decision depends on whether such data affects the presentation of whole dataset or not. Before transforming to structured data, these data in JSON format have been preprocessed and cleaned.
  - The global data:
    By sending a specific date query, a JSON object has been returned. This JSON object contains 11 keys and their values in a format of key : value , The key names are date, last_update, confirmed, confirmed_diff, deaths, deaths_diff, recovered, recovered_diff, active, active_diff, fatality_rate. A key name with a _diff suffix indicate the new number of such case. For instance, the confirmed_diff indicates the number of new confirmed cases from a day before, the deaths_diff indicates the number of new deaths, the recovered_diff indicates the number of new recovered cases and the active_diff indicates the number of new active cases. Comparing with the accumulated number of confirmed, deaths, recovered and active cases, the "_diff" series can be used to present more intuitive change.
    By setting the number of days, a time series of global data has been obtained by sending a list of date queries. The first Covid-19 case was reported in Wuhan, China in December 2019 [24]. There are approximate 1078 days since then (until 13/12/22). We picked up 1000 days data to analysis and visualise. The global data cleaned, no duplicated record and no empty entry.
    - This dataset has been stored as a collection, called "global_data_over_time_collection" in the MongoDB database "jialin_Mongo_database".
    - "DBAP_CA2_global_data_over_time.csv" has been exported from the above collection.
    - A table named "global_data_over_time" has been created in PostgreSQL database "jialin_postgresql_database", contains all contents in "DBAP_CA2_global_data_over_time.csv" file. This structured table is ready for presenting. Then line chart of used to show the spread of Covid-19 within 7 days, 30 days and 1000 days in the first tab of a dashboard.
  - The USA federal data:
    By sending a specific date and the iso code "USA", a JSON object has been returned. This JSON object contains 12 key : value entries, the first 11 entries are the same as the global date. The last entry's key name is region whose value is another JSON object with province and cities informations of Covid-19. At this stage, it is not the time to dive into provinces or cities informations. Hence, the big "tail" region has to be cut off, only the federal information is required. However, a iso code key need to be maintained to indicate it is a data for federal. Hence, a new JSON object is created which is composed of 12 key : value, the last key name is iso. By importing json library and creating a dictionary object, the federal data without provinces information has been obtained.
    By sending a list of date queries with the iso code "USA", a time series of responses have been store into the MongoDB database, as a collection named "reports_US_data_over_time_collection". In this project, since the data structure of America information is very similar to the global data, although the first Covid-19 case was detected in USA in March of 2019, We set the number of days as 365 to show the spread of disease in last one year. In terms of observation, there is no duplicated records and no empty entity as well.
    - The "reports_US_data_over_time_collection" has been stored in "jialin_Mongo_database".
    - "DBAP_CA2_reports_US_data_over_time.csv" has been generated by the above collection.
    - Then the contents of such file have been inserted in a table "reports_US_data_over_time" in PostgreSQL database "jialin_postgresql_database".
  - The USA provinces and cities data:
    The "Covid-19 Statistics" API provides provinces of each country by sending the request of the corresponding country's iso code. By sending a query of iso code "USA", a list of US provinces has been returned. This provinces list is used to search detailed information for each province. This JSON object contains many duplicated and erroneous information. It was kept as an intermediate provinces list.
    Further more, by sending date, "USA" and a province name, more detailed information for cities in the

corresponding province returned in a JSON object format. This JSON object contains 12 key : value entities, the first 11 entries are the same as federal information and global data. The last key is region. This time, the first 11 entries are not considered, Covid-19 data of US provinces and cities are extracted from the value of key region. The value of region is a JSON object contains 6 items in key : value format. These 6 items are ″iso″ :″ USA″, ″name″ :″ US″, province whose value is state name, e.g. ″Washington″, lat whose value is a latitude number, long with its value is a longitude number, and cities whose value is a JSON object contains Covid-19 cities information of the corresponding province. The geographical information, like latitude and longitude of a province, are not used in this project. From the response information, there is no directed Covid-19 data for each state. But the state data can be calculated from the cities information. By exploring the value of cities, there are 10 key : value items. Their keys are, name indicates a city name, date indicates the query date, fips indicates FIPS code [25] in USA, lat, long, confirmed, deaths, confirmed_diff, deaths_diff and last_update. Note that these information only for one province.

In order to fetch Covid-19 data for all provinces in USA, a list of queries has been created by using the intermediate provinces list. In the process of sending request for each province, the irrelevant province, which has no information or has no cities information has been deleted, the erroneous information, for example, many "Unassigned" cities, has been check and deleted, cities with empty geographical information, such as fips, lat or long, are deleted. Finally, the intermediate province list and the Covid-19 data of cities for each province are refined by each other. All data are cleaned, there is no duplicated records and no empty entries. A new province list and the final Covid-19 cities data in USA have been generated.

- The "provinces_of_US_collection" and the "reports_US_provinces_cities_collection" have been stored in "jialin_Mongo_database".
- "DBAP_CA2_provinces_US.csv" and "DBAP_CA2_reports_US_provinces_cities.csv" have been generated by the above collections.
- Then the contents of these CSV files have been copied into tables "new_provinces_data" and "reports_us_provinces_cities" respectively in PostgreSQL database "jialin_postgresql_database".

An ETL (Extract, Transform, Load) pipeline has been built by using Python and its libraries like pandas. Since the data process in this project is not too complex, library pandas provides data structure and analysis tools are enough to create the ETL pipeline. The original data is extracted from the "Covid-19 Statistics" API in Rapidapi website [14], after preprocessing, these data are stored into a MongoDB database, eventually they are loaded in a PostgreSQL database as structured data and ready for visualising.

E. Combination

The hardest hit area together with the other Categorical data from dataset 1 were chosen and imported as CSV file, stored in PostgresDB, specially for the purpose of combination with dataset 2, code script see Fig. 14. There are 7 distinct status at different levels, see Fig. 15.

```
CREATE TABLE final_provinces_data_hha AS
WITH STEP AS(
     SELECT p.state_code, p.confirmed, p.deaths, h.hha
     FROM new_provinces_with_state_code AS p
     INNER JOIN min_hha AS h ON state_code = st_abbr
     GROUP BY p.state_code, p.confirmed, p.deaths, h.hha
     ORDER BY p.confirmed DESC, p.deaths DESC
)
SELECT STEP.*
FROM STEP;
```

Fig. 14. Combination Code Script

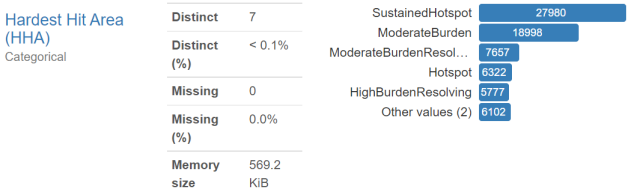


Fig. 15. HHA Categories

IV. Result and Evaluation

A. Dataset 1. Covid-19 Vulnerable Community Crosswalks

HHA decision tree model metrics has achieved 87.65% accuracy, which discovered that apart from Covid-19 impact, the status of hardest hit area can also be affected by rural area, low-income area, tribal community, poverty score and percentage of those areas.

B. Dataset 2. Covid-19 Statistics

In order to present the spread of Covid-19 in the entire world and USA, a dashboard with three tabs has been created by using library dash in Python. The dashboard is running on localhost, port 8050 (see Fig. 18, 19 and 20).

- Tab1: Covid-19 Global Data (see Fig. 18)
  In the first tab, the left side shows four line charts of the global Covid-19 data. The top left presents the daily new confirmed cases in 7 days, by tickling the point, the specific number of new confirmed cases appear. The top right display the daily new deaths cases in 7 days, by checking the point, although the number of new confirm is large, the corresponding deaths number is small. The bottom two line plots show more peaks and troughs, the

left one is the new confirmed cases changed in 30 days and the right one is the new deaths cases changed in 30 days. Compared to the number of new confirmed cases one exceeded one million, the new deaths cases less than 3000. These results indicates the fatality rate is very low. The right side exhibit three line charts of the number of new confirm cases, the number of new deaths cases and the fatality rate of the global Covid-19 data in 1000 days. The top chart display that since around 20/01/22, there was a peak number of new confirmed cases, the number of new confirmed cases has flattened. The middle chart shows the last peak number of new deaths cases was on 01/06/21, recently the number of deaths case are very small. The bottom plot displays the fatality rate change in the past 3 years, after a peak data on 09/06/2020, it has been declining, the latest fatality rate is stable at 0.0103. Changes in global data deserve continued attention, especially after some countries have adjusted their epidemic prevention policies.

- Tab2: Covid-19 US Data (see Fig. 19)
  The left side of the second tab presents three line charts of Covid-19 data in 365 days. The top one shows the change of the number of new confirmed cases in US, the middle one display the change of the number of new deaths cases in US, and the bottom one exhibit the fatality rate changed in US. By comparing these three plots, the last peak and trough the numbers of new confirmed and new deaths cases were appeared on 13/07/22 and 14/07/22, while the fatality rate also had a noticeable ups and downs in those two days, once up to 0.0084. Now the deaths rate is stable at 0.0079, less than the rate of entire world.
  The right side of the second tab, two pie charts have been used to demonstrate the number of the confirmed cases and the number of the deaths cases in each state in USA. Firstly, the quantity numbers of the number of the confirmed cases and the number of the deaths cases have a huge difference on size. Hence the right pie is very small in purpose. But the number of each province can be seen by touching a particular part of the pie chart. Secondly, the state code has been sorted by the number of confirmed cases and the number of the deaths cases. For example, the biggest part in the pie chart, with state code CA represent for California, has the worst pandemic, followed by TX (Texas), FL (Florida), NY (New York) and IL (Illinois).
- Tab3: Covid-19 US Map by State Code (see Fig. 20)
  The third tab contains two choropleth maps of USA. The left one presents the number of confirmed cases by state code. The right one displays the number of deaths cases by state code. For both of them, the light colour indicates the number is small, the dark colour shows the state has a large quantities. These two maps demonstrate the pandemic in USA in a geographical look.

C. Identification of the Hardest Hit Areas

Fig. 16 presents the distribution map of the hardest hit areas according to the degree of disaster. One SustainedHotspot area has been identified, whose state code is DE, represents Delaware in USA. One LowBurden area, several Hotspots areas have been identified as well. Most areas are Moderate-BurdenResolving areas.

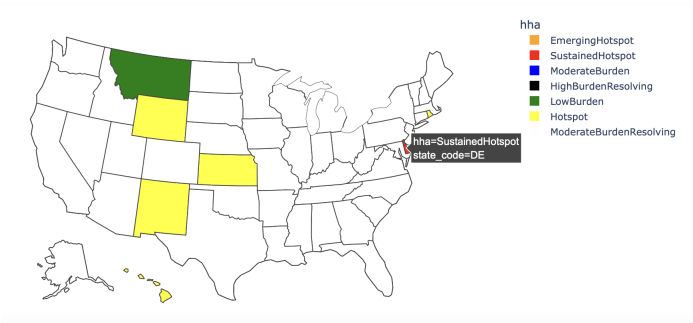


Fig. 16. Screenshot Identification of HHA in USA Map

V. Conclusions and Future Work

Covid -19 pandemic has put negative impact on our society. It has changed the way we used to live. The status of hardest hit area by the pandemic is highly associated with low-income, rural area, and tribal community, as well as poverty percentage of those areas.

While those community issues might have been there even before the pandemic, more research and effort would be urgently needed for local government to consider regarding how to improve the status of those communities, especially on how to increase income and reduce poverty for those communities.

The impact of Covid-19 on human society is irreversible, but it may not be the last virus. The epidemic has taught us lessons and experience, allowing us to better face unknown risks. In the future, with the continuous attention to Covid-19, more clear and intuitive display is required. The authors would like to develop a more detailed dashboard to present the pandemic situation and make contributions to support vulnerable communities.

Acknowledgment

We would like to express our grateful thanks to Dr. Rashid Mijumbi for his professional academic teaching for the module Database & Analytics Programming and his guidance on our project. And we would like to thank Jaswinder Singh, a member in teaching assistant group, for his patience and help.

References

[1] Johns Hopkins. Coronavirus Resource Centre [online]. Available at: https://coronavirus.jhu.edu/ (Accessed: Dec. 2022).

[2] T. J. Mueller, et al. “Impacts of the COVID-19 pandemic on rural America”, PNAS, 2020. [online]. Available at: https://www.pnas.org/doi/10.1073/pnas.2019378118 (Accessed: December 2022).

[3] Columbia Research. [online]. Available at: https://research.columbia.edu/covid/community (Accessed: November 2022).

[4] Q. Gao, "COVID-19's Immediate Impacts on Poverty, Hardship, and Well-Being among New Yorkers" Columbia Research [online]. Available at: https://research.columbia.edu/covid/community/immediateimpacts (Accessed at: December 2022).

[5] S.J. Islam, G. Malla, and R. W. Yeh, et al. "County-Level Social Vulnerability is Associated With In-Hospital Death and Major Adverse Cardiovascular Events in Patients Hospitalized With COVID-19: An Analysis of the American Heart Association COVID-19 Cardiovascular Disease Registry", 2022. [online]. Available at: https://www.ahajournals.org/doi/10.1161/CIRCOUTCOMES.121.008612 (Accessed: December 2022).

[6] World Health Organisation (WHO) [online] https://www.who.int (Accessed: December 2022).

[7] WHO Covid-19 Dashboard [online] https://covid19.who.int (Accessed: December 2022).

[8] Johns Hopkins University [online] https://www.jhu.edu (Accessed: December 2022).

[9] Johns Hopkins Center for Systems Science and Engineering (CSSE) [online] https://systems.jhu.edu (Accessed: December 2022).

[10] Johns Hopkins Center for Systems Science and Engineering (CSSE) Covid-19 Dashboard [online] https://coronavirus.jhu.edu/map.html (Accessed: December 2022).

[11] Ireland Government Covid-19 Hub [online] https://covid19ireland-geohive.hub.arcgis.com (Accessed: December 2022).

[12] P. Chapman, et all. "CRISP-DM 1.0 Step-by-step data mining guide ", 1999, NCI Moodle [online]. Available at: https://mymoodle.ncirl.ie/mod/assign/view.php?id=12344 (Accessed: December 2022).

[13] DATA.GOV. "COVID-19 Community Vulnerability Crosswalk - Rank Ordered by Score" [online]. Available at: https://catalog.data.gov/dataset/covid-19-community-vulnerability-crosswalk-rank-ordered-by-score (Accessed: November 2022).

[14] Rapidapi Covid-19 Statistics [online] https://rapidapi.com/axisbits-axisbits-default/api/covid-19-statistics/ (Accessed: December 2022).

[15] E. Andersson, and Z. Berggren, "A Comparison Between MongoDB and MySQL Document Store Considering Performance", 2017. [online] Available at: https://www.diva-portal.org/smash/get/diva2:1161166/FULLTEXT01.pdf (Accessed: December 2022).

[16] V. Kanade, "What Is a Decision Tree? Algorithms, Template, Examples, and Best Practices", 2022. [online]. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-decision-tree/ (Accessed: December 2022).

[17] Python https://www.python.org

[18] G. Van Rossum, and F. L. Drake Jr, "The python language reference", Python software foundation, Python, 2014. [online] Available at: https://dev.rbcafe.com/python/python-3.5.1-pdf/reference.pdf (Accessed: December 2022)

[19] Lutz M. BOOK "Learning Python" Chapter 1"A Python Q&A Session " Why Do People Use Python, p3 O'Reilly, 5th edition (12 July 2013) ISBN-13: 978-1449355739

[20] Jupyter Notebook https://jupyter.org

[21] R https://www.r-project.org

[22] AcaStat "Applied Statistics Handbook", 2015. [online]. Available at: https://www.acastat.com/statbook/chisqassoc.htm (Accessed: December 2022).

[23] V. Kanade, "Decision trees are tree-like visual models that illustrate every possible outcome of a decision", 2022. [online]. Available at: https://www.spiceworks.com/tech/artificial-intelligence/articles/what-is-decision-tree/ (Accessed: December 2022).

[24] H. Lau, V. Khosrawipour, P. Kocbach, A. Mikolajczyk, J. Schubert, J. Bania, and T. Khosrawipour, "The positive impact of lockdown in Wuhan on containing the COVID-19 outbreak in China", Journal of travel medicine, 2020. [online] Available at: https://covid-19.conacyt.mx/jspui/bitstream/1000/5688/1/1109910.pdf (Accessed: December 2022).

[25] FIPS Code of USA [online] https://nitaac.nih.gov/resources/frequently-asked-questions/what-fips-code-and-why-do-i-need-one (Accessed: December 2022).

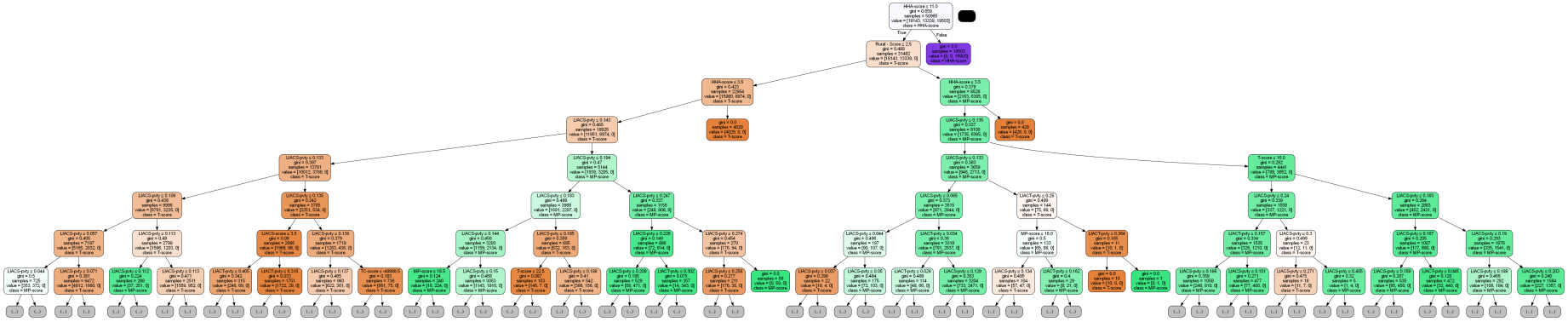


Fig. 17. The Complete HHA Tree (Dataset 1)

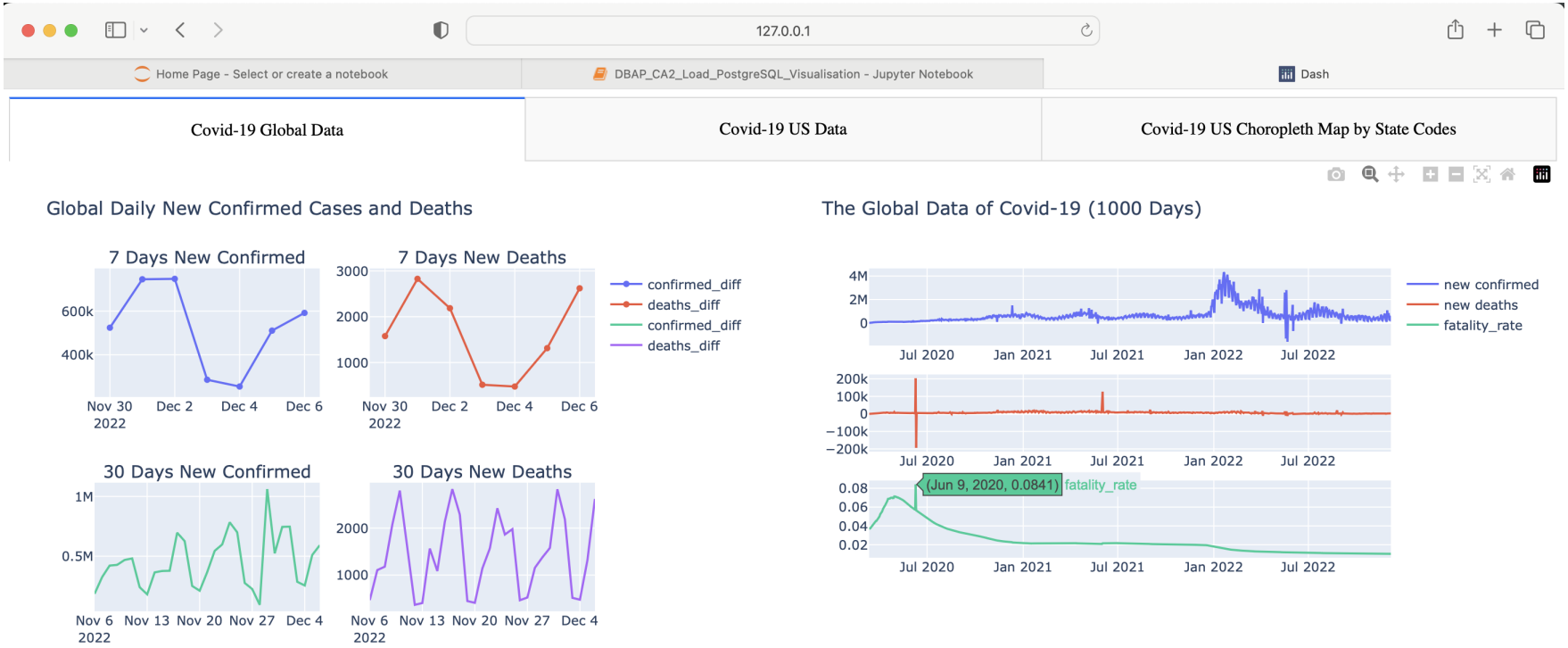


Fig. 18. Screenshot of the Dashboard Tab1 Global Data.

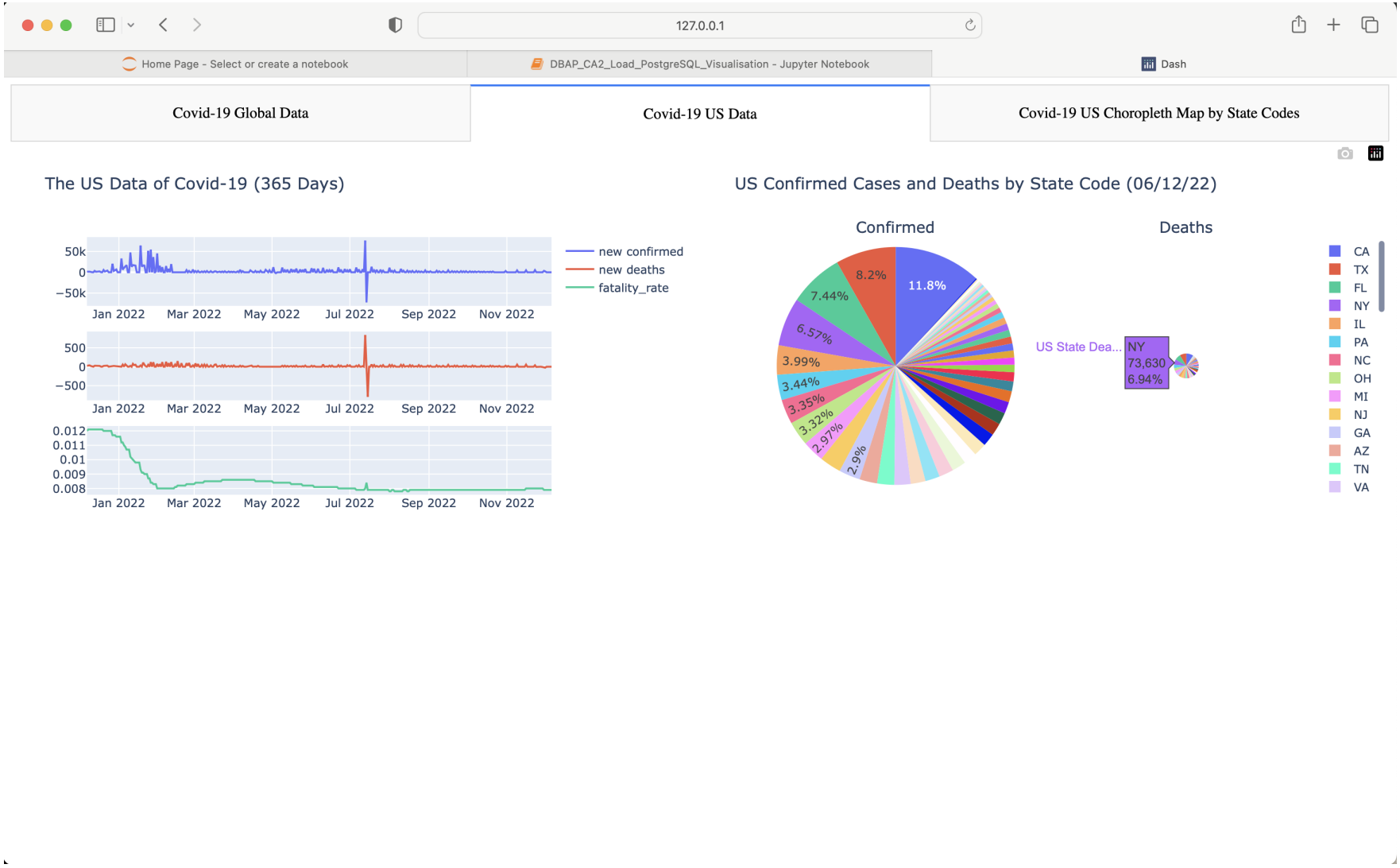


Fig. 19. Screenshot of the Dashboard Tab2 US Data

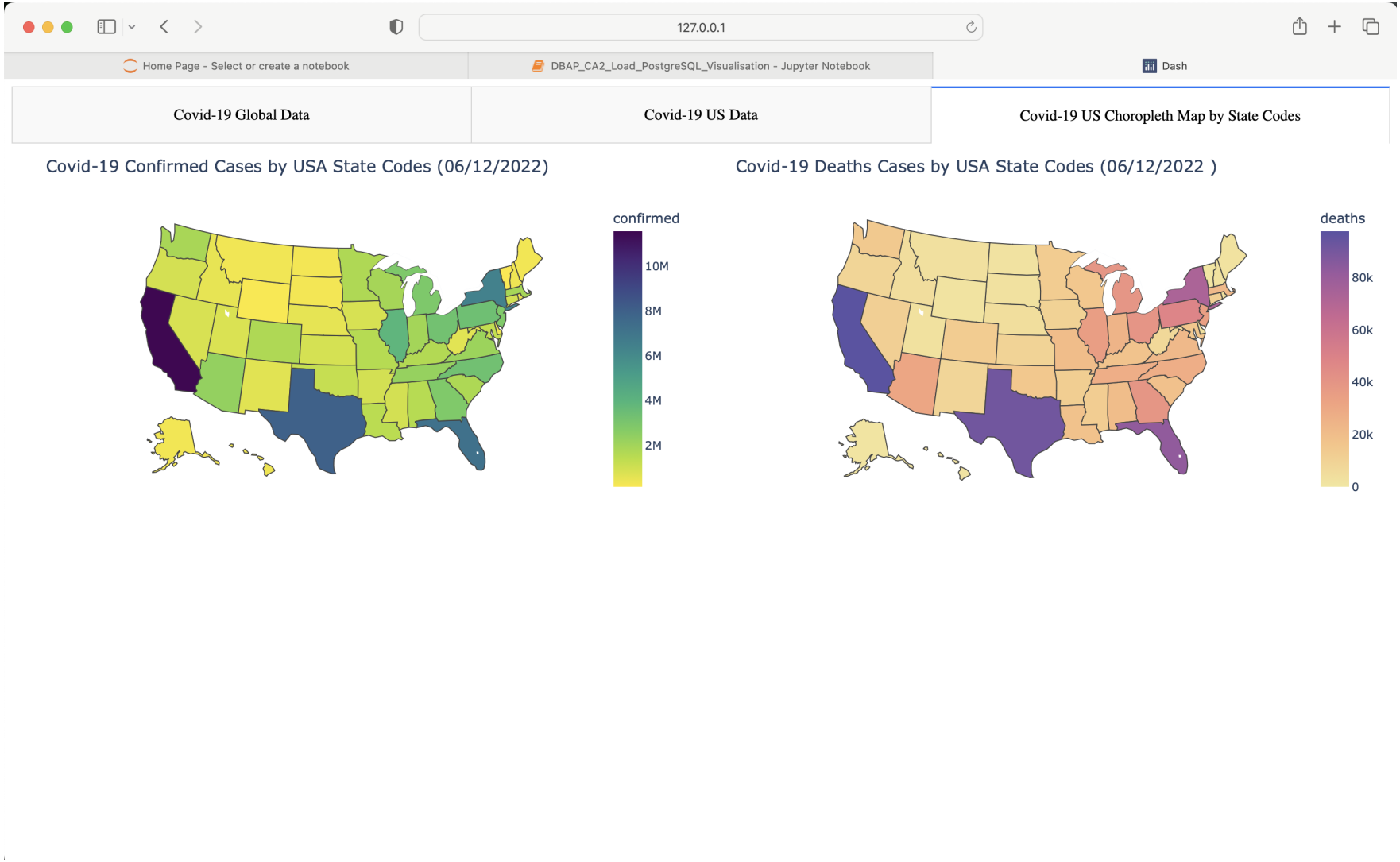


Fig. 20. Screenshot of the Dashboard Tab3 US Chropleth Map.