

Qure.ai Data Engineer assignment

Cleaning and storing the data:

I started by observing complete data in each individual file. Then thought about how each file is related to the other. I found out a few things by following data,

i)Countyleveldata have covid death data from 22/01/20 to 28/07/20, but in state-level data most of the states have data from 13/03 to 18/07, So by using the county-level data I found out the deaths in each state on every day and added it into state-level data. Now state-level data consists of deaths from 22/01 to 28/07. (In End we only need the deaths from march, April, May, and June, I thought adding the data for Jan and Feb maybe become redundant but I have added it because more data helps in visualizing the change better.)

ii)latitude and longitude value, in county-level data for unallocated deaths statewide deaths there are no latitude and longitude values. Filling the values into this column didn't make sense to me. Because if we add 0 it resembles some location on earth and if we add the mean of the state longitude and latitude the deaths will be allocated to the middle of the state which is not ideal. So, I left those columns empty.

iii)Mobility data, this data confused me. Because there are negative values, I have not understood the hypothesis behind the negative mobility cases. Thought of relation between previous and present-day but didn't make sense. Then I converted all negative values into positive values and filled empty values with the mean of each column.

iv)Outlier detection, the phenomena which we do for almost every dataset. But in this case, outlier detection is not needed because there is a particular rule that the values should be in this range only and values can differ largely from the mean. But if the Machine learning team requires this, we can remove the values which differ largely from the mean by removing the tails of the distribution

Data Cleaning code can be found in the file – “Qure_data_cleaning_Ramsai.ipynb”

I have chosen the MYSQL database to store the data, I have created a database and loaded 3 files as 3 tables into the data.

Creating Database loading tables into database code can be found in –
“Qure_data_storing_Ramsai.ipynb”

Displaying the data:

I have chosen grafana visualization tool to visualize the data, I have created a data source in grafana using the Database created in storing data step.

I have added filters to visualize the data with convenience.

Graphs added:

i) Total number of States in the given Data

ii) Total number of Counties in the Selected State (select using State Fips Number)

iii) Average Deaths per a day in the Selected Month (select using Month)

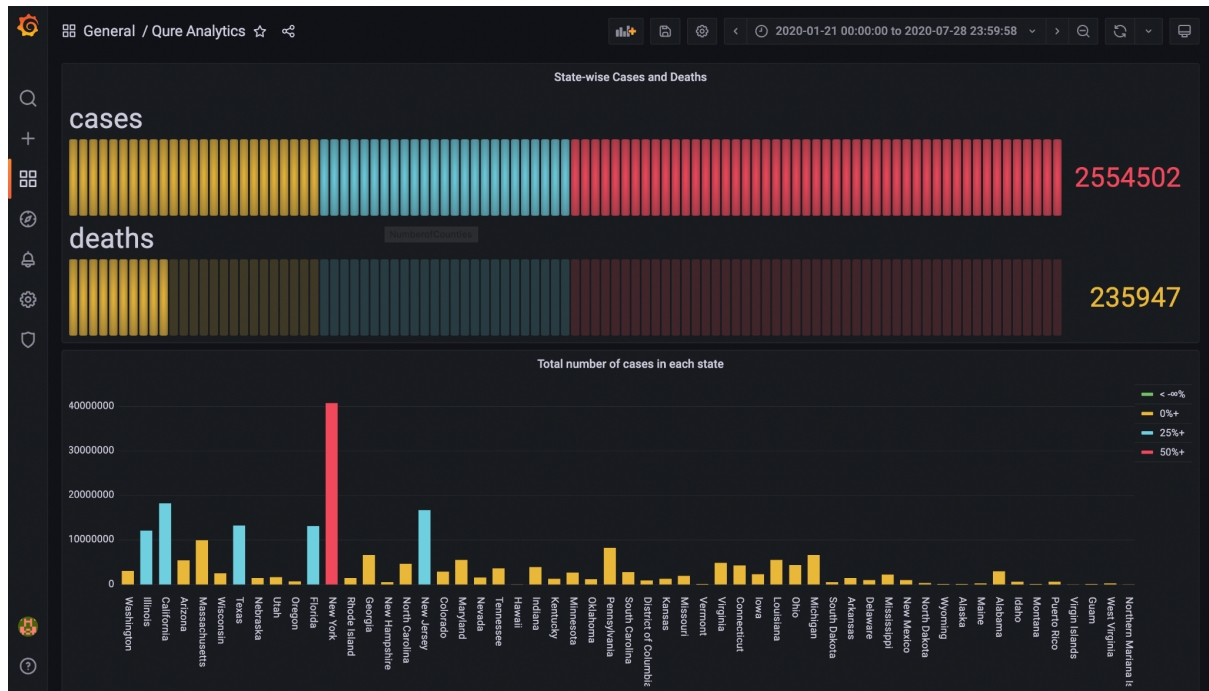
iv) Bar graph showing the total County wise Deaths and average deaths in the selected state and county (if you select State Fips Number from the dropdown, County Fips Number updates itself with the counties present in the selected state, there you can choose the required county)

v) Average covid cases, Total covid cases, Average covid deaths and Total covid deaths in the selcted month (select using month)

vi) Total covid cases and deaths in the selected state (select state name using State)

vii) Total Number of cases in each state





Sending the data:

Created Slack webhook by building an in slack app builder and channel in slack to receive the data.

Written queries and code to get the required data and posted to slack using ost request. I have encountered LIMIT,IN not supported in the current mysql version issue so followed different approach after finding no luck with multiple subqueries and joins.

You can find the code written in - 'Qure_slack_update_Ramsai.py' file.

Slack Updates:

ramsal

Browse Slack

Channels

api

data-updates

general

random

Add channels

Direct messages

ramsai.chamakura you

Add teammates

Apps

pure-data

Add apps

data-updates

Search ramsai

data-updates

1:29 Top 3 counties with highest number of covid deaths from top3 states(by covid deaths) for month of April

Month-April

New York--425198

Kings County--86067

Queens County--84238

Bronx County--58916

New Jersey--102708

Essex County--17920

Bergen County--17886

Hudson County--10778

Michigan--59519

Wayne County--27930

Oakland County--11879

Macomb County--9481

Top 3 counties with highest number of covid deaths from top3 states(by covid deaths) for month of May

Month-May

New York--854088

Kings County--86067

Queens County--84238

Bronx County--58916

New Jersey--308935

Essex County--17920








Bergen County--17886

Hudson County--10778





Massachusetts--170827

Middlesex County--9516

Suffolk County--6061

B I       

Send a message to #data-updates

ramsal

Browse Slack

Channels

api

data-updates

general

random

Add channels

Direct messages

ramsai.chamakura you

Add teammates

Apps

pure-data

Add apps

data-updates

Search ramsai

data-updates

ramsai

Today

1:29 Top 3 counties with highest number of covid deaths from top3 states(by covid deaths) for month of March

Month-March

New York--7943

Kings County--86067

Queens County--84238

Bronx County--58916

Washington--2377

King County--9398

Snohomish County--2379

Yakima County--841

New Jersey--1145

Essex County--17920

Bergen County--17886

Hudson County--10778

Top 3 counties with highest number of covid deaths from top3 states(by covid deaths) for month of April

Month-April

New York--425198

Kings County--86067

Queens County--84238

Bronx County--58916








New Jersey--102708

Essex County--17920





Bergen County--17886

Hudson County--10778

Michigan--59519

B I       

Send a message to #data-updates

ramsai

Browse Slack

Channels

api

data-updates

general

random

Add channels

Direct messages

ramsai.chamakura you

Add teammates

Apps

pure-data

Add apps

data-updates

1

Top 3 counties with highest number of covid deaths from top3 states(by co Today for month of May

Month-May

New York--854088

Kings County--86067

Queens County--84238

Bronx County--58916

New Jersey--308935

Essex County--17920

Bergen County--17886

Hudson County--10778

Massachusetts--170827

Middlesex County--9516

Suffolk County--6061

Norfolk County--5959

Top 3 counties with highest number of covid deaths from top3 states(by covid deaths) for month of June

Month-June

New York--918476

Kings County--86067

Queens County--84238

Bronx County--58916

New Jersey--388821

Essex County--17920

Bergen County--17886

Hudson County--10778

Massachusetts--228975

Middlesex County--9516

Suffolk County--6061

B I [link] [code] [list] [table] [quote] [code] [link]

Send a message to #data-updates

+ [emojis] [mentions] [text]

ramsai

Browse Slack

Channels

api

data-updates

general

random

Add channels

Direct messages

ramsai.chamakura you

Add teammates

Apps

pure-data

Add apps

data-updates

1

+ Add a bookmark

Kings County--86067

Queens County--84238

Bronx County--58916

New Jersey--308935

Essex County--17920

Bergen County--17886

Hudson County--10778

Massachusetts--170827

Middlesex County--9516

Suffolk County--6061

Norfolk County--5959

1:29 Top 3 counties with highest number of covid deaths from top3 states(by covid deaths) for month of June

Month-June

New York--918476

Kings County--86067

Queens County--84238

Bronx County--58916

New Jersey--388821

Essex County--17920

Bergen County--17886

Hudson County--10778

Massachusetts--228975

Middlesex County--9516

Suffolk County--6061

Norfolk County--5959

B I [link] [code] [list] [table] [quote] [code] [link]

Send a message to #data-updates

+ [emojis] [mentions] [text]

It is my first time working with Grafana and Practical implementation of MYSQL. I learned and enjoyed doing the assignment. Thank you for the wonderful opportunity.

RAMSAI REDDY CHAMAKURA