COVID 19 SPREAD INSIGHTS FOR THE US

IBM Applied Data Science Certification Capstone Project

ABSTRACT

This is the initial report of the COVID 19 Insights Capstone Project.

Rajesh Ramachandran

Coursera IBM Applied Data Science Capstone

s

CONTENTS

Introduction 2

Business Problem 2

Target Audience for this project 2

Data Sources 2

1. JHU CSSE COVID-19 Dataset 3

File naming convention 3

Field description 3

2. US Cities and Counties 3

File structure 3

Example 3

3. Foursquare Database 4

Structure of the API 4

INTRODUCTION

The world has never experienced such a widespread pandemic like COVID-19. The uncertainty over this epidemic has marred nations. Fortunately, we live in an era, where unimaginable data is being generated every second. To tackle COVID -19 and make intelligent decisions, organizations are heavily relying data that is being generated and assembled by various sources in the world. More and more insights driven by machine learning algorithms and fast data processing has enabled to derive strategies and action plans.

A platform to provide COVID 19 case spread at various levels (County, State, National) insights and visualizations help frontline working strategies to be streamlined. Targeted programmes like increase in testing centers, can be deployed to areas affected worse and with minimal healthcare facilities.

BUSINESS PROBLEM

The attempt of this Capstone Project is to find and analyze COVID 19 data available in the public forum for the United States. Using Data Science Methodologies the outcome of this analysis we will try to answer these business questions.

1. Which are the worst and least affected areas in the US?
2. Can we classify Counties of any US State into High, Medium, Low risk based on their health care infrastructure?

TARGET AUDIENCE FOR THIS PROJECT

Mainly the project is targeted at helping Health Care Organizations and teams tackling the COVID19 pandemic from a planning perspective. It includes, Hospitals, State and National Healthcare officials and any organizations involved in COVID 19 frontline work. It helps them to make data-driven decisions.

In addition to that this project is also aimed anyone who is directly or indirectly affected by COVID 19 who would like to get a perspective of the COVID 19 spread.

DATA SOURCES

The problem presented above would need the following data from available sources

1. Daily COVID 19 Cases reported data in the US. **The John Hopkins University Center for Systems Science and Engineering Data** repository will be used as source for this purpose. The GitHub page for JHU CSSE has multiple datasets in CSV format with active reported cases of COVID 19 throughout the world. The datasets are available here (https://github.com/CSSEGISandData/COVID-19).
2. For the classification problem, we will need data from 2 sources
    a. To get a list of cities in each county we will be using a publicly available GitHub US cities and Counties database in CSV format. The dataset is available here (https://github.com/grammakov/USA-cities-and-states/blob/master/us_cities_states_counties.csv)
    b. To get Hospital venue information for each city, we would be accessing the Foursquare Database using the Foursquare Developer API. Foursquare has one of largest database of 105+ million places and is used by over 125,000 users worldwide. More info on Foursquare APIs can be found here (https://developer.foursquare.com/docs/places-api/)

1. JHU CSSE COVID-19 DATASET

For our project we will be usign the CSSE COVID 19 Daily Report, which is uploaded daily in the CSSE GitHub page.

Daily reports (csse_covid_19_daily_reports))

This table contains an aggregation of each USA State level data.

FILE NAMING CONVENTION

MM-DD-YYYY.csv in UTC.

FIELD DESCRIPTION

1. **FIPS**: US only. Federal Information Processing Standards code that uniquely identifies counties within the USA.
2. **Admin2**: County name. US only.
3. **Province_State**: Province, state or dependency name.
4. **Country_Region**: Country, region or sovereignty name. The names of locations included on the Website correspond with the official designations used by the U.S. Department of State.
5. **Last Update**: MM/DD/YYYY HH:mm:ss (24 hour format, in UTC).
6. **Lat and Long** : Dot locations on the dashboard. All points (except for Australia) shown on the map are based on geographic centroids, and are not representative of a specific address, building or any location at a spatial scale finer than a province/state. Australian dots are located at the centroid of the largest city in each state.
7. **Confirmed**: Confirmed cases include presumptive positive cases and probable cases, in accordance with CDC guidelines as of April 14.
8. **Deaths**: Death totals in the US include confirmed and probable, in accordance with CDC guidelines as of April 14.
9. **Recovered**: Recovered cases outside China are estimates based on local media reports, and state and local reporting when available, and therefore may be substantially lower than the true number. US state-level recovered cases are from COVID Tracking Project.
10. **Active**: Active cases = total confirmed - total recovered - total deaths.
11. **Incidence_Rate**: confirmed cases per 100,000 persons.
12. **Case-Fatality Ratio (%)**: = number recorded deaths / number confirmed cases

2. US CITIES AND COUNTIES

This is a GitHib respository of US Cities, Counties and States. it contains a full list of US cities, states and counties (60k+ entries) formatted. The file is available in pipe delimited format here

FILE STRUCTURE

City|State short name|State full name|County|City Alias Mixed Case

EXAMPLE

Bronx|NY|New York|BRONX|University Heights

3. FOURSQUARE DATABASE

The Foursquare Database is used in the is projects to get the number of hospitals per county. Since County is a big geographical entity we will be searching the Foursquare API for hospital venues at city level and group them at county level later.

The Foursquare Database is accessed via the Foursquare Developer API which required credentials. The credentials have been previously setup and kept in secure python file, which is used by this project.

We will be calling the API iteratively till we get all hospital venues from the API within the given distance. The maximum venues the API can fetch is 100, so we will fetch all venues by iteratively calling this API and increasing the offset each time.

We will be using the Foursquare Venue Search API. Credentials required to access the Foursquare API are

- CLIENT_ID
- CLIENT_SECRET
- VERSION

Parameters Required for the Venue Search API

- **ll** = Latitude and longitude of the search location. This will be the Lat and Longitude of the city we wish to search for.
- **radius** = currently we will set this as 20 miles (~32000 meters). Assumption here is that each city is at the max 40 miles wide.
- **categoryId** = A comma separated list of categories to limit results to. In our case this will be list of categories for hospitals is
    - **Hospital** - 4bf58dd8d48988d196941735
    - **Hospital Ward** - 58daa1558bbb0b01f18ec1f7
    - **Medical Lab** - 4f4531b14b9074f6e4fb0103
    - **Urgent Care Center** - 56aa371be4b08b9a8d573526
    - **Doctor's Office** - 4bf58dd8d48988d177941735
    - **Emergency Room** - 4bf58dd8d48988d194941735
    - **Medical Center** - 4bf58dd8d48988d104941735

STRUCTURE OF THE API

The API request URL would look like this

https://api.foursquare.com/v2/venues/search?client_id=XXXX&client_secret=XXXX&v=XXXX&ll=XX,XX&radius=32000&categoryId=LIST_OF_CATEGORIES