Stage1-Final

February 14, 2022

1 Project: Stage I (Problem and Tasks)

2 COVID-19 Data Analysis :

- GitHub Path: https://github.com/BalaMallampatiGIT/IAC621_DataScience
- Bala Mallampati
- Neethu Sreerangam
- Samip Thapa

2.1 Description

COVID-19 is a continuing worldwide pandemic which has affected a lot of people, including you. Our goal with the major project in this class is to develop an analytical framework to study the data coming from the United States, to understand patterns of COVID-19 effects and spread.

In order to achieve that, the project is separated into four stages:

- Stage I: Data and Project Understanding
- Stage II: Data Modeling and Hypothesis Testing
- Stage III: Basic Machine Learning
- Stage IV: Dashboard

2.2 Project Stage I: Data and Project Understanding

2.3 COVID-19 Dataset

We'll utilize data from usfacts.org. The dataset contains daily county-level trackers of COVID-19 cases. This makes it easy to follow COVID-19 cases on a granular level, as does the ability to break down infections per 100,000 people (with population data). The underlying data is available for download below the US county map, and has helped government agencies like the Centers for Disease Control and Prevention in its nationwide efforts.

- USA Facts: US COVID-19 cases and deaths by state
- USA Facts: Number of Cases (.csv file)
- USA Facts: Number of Deaths (.csv file)
- USA Facts: Population by County (.csv file)

2.4 Task 1

The entire team looks at the COVID-19 dataset and understands the type of variables present in each of the data.

2.4.1 Team

- Section in the report describing the COVID-19 dataset and datatype: variable dictionary.
- Preliminary intuitions from the data.
- Each member of the team takes on an enrichment dataset. They read the data descriptions and understand the variables present in the data.

2.4.2 Individual

- Section in the report describing the enrichment data and datatype: variable dictionary.
- How can you merge the data with the primary COVID-19 dataset? Identify the individual variable which maps between the datasets.
- Describe how your enrichment data can help in the analysis of COVID-19 spread. Pose initial hypothesis questions.

2.5 Introduction

As COVID-19 pandemic continues to spread, many people around the world have been affected in many different ways. We can understand patterns of COVID-19 effect and the spread by designing and developing an analytical framework with real life data. The data used for this report is COVID-19 data and relevant impact data, which reflects aspects of demographics, socioeconomics, employment, politics, and hospital beds as enrichment datasets. These datasets will provide additional insight into the American municipalities. This project provides us with opportunities to not only demonstrate our technical skills and knowledge in data science, but also to turn uncovered scientific outcomes into insightful knowledge and information that will be highly beneficial to public health and quality of living.

We have imported all required libraries for Team task, Covid-19 data analysis

[148]:
```
USConfirmCasesDF = pd.read_csv("..\\..\\Data\\covid_confirmed_usafacts.csv")
USDeathsDF = pd.read_csv("..\\..\\Data\\covid_deaths_usafacts.csv")
USCountyPopulationDF = pd.read_csv("..\\..
↪\\Data\\covid_county_population_usafacts.csv")
```

We have downloaded latest data from provided website *USA Facts: US COVID-19 cases and deaths by state* for below datasets. US County level data available until Feb 10th 2022 for all states

- Covid19 Confirmed Cases
- Covid19 Death Cases
- County Population

We see that US Covid19 Confirmed Cases, Death Dataset contains 3193 observations and 755 columns(753 Integers & 2 Object datatypes)

We observe that US County Population Dataset contains 3195 observations and 4 columns(2 Integers & 2 Object datatypes). All of these data columns does not contain any NaN, NA values.

We also verified US Covid19 Confirm Case, Death data has any null columns and didnt find any, above results shows all Null counts are Zero(0)

As per unique naming standards and easy understand code , we have renamed all columns with Init CAPS. Ex: County Name to CountyName, countyFIPS to CountyFIPS

We have observe that US Covid19 Confirmed Cases is wide dataset, which needs to make long dataset to combine & use with other datasets.

- Taken all common columns for Indexing into variable VarIndex. i.e. State, StateFIPS, CountyFIPS, CountyName
- And then taken rest of all columns(755-4=751) into another variable VarValue
- We have used Pandas Melt function to convert Pivot to unpivot table by passing VarIndex & Varvalue
- Named New columns as "ReportDate", "ConfirmCaseCount" & DeathCount
- And also rearranged columns as per standards and easy understand

We have verified after converision of long format datatypes, observations & columns

- Observations increased from 3193 to 2,397,943 rows for each report date, it shows Confirm Case Count by State, StateFIPS, CountyFIPS, CountyName
- Total 6 columns with ReportDate, State, CountyName are Object, StateFIPS, CountyFIPS, ConfirmCaseCount are Integers

We have merged US Covid19 Confirmed Cases and Death datasets by using below steps.

- Taken all common columns between data sets. i.e. State, StateFIPS, CountyFIPS, CountyName
- Even though we can just join based on StateFIPS, CountyFIPS, we have taken Names as well since they all come from same soure i.e. usafacts.org
- We have tested with and without names, both gives same results.
- Also to avoid missing data due to joins, we have used Outer join.
- So merged based on above columns and verified data, they are all looks good

We observe that combined data for US Confirm and Deaths data "USConfirmCasesDeathsDF" shows same amount of rows 2,397,943 with 7 columns

- Total 7 columns with ReportDate, State, CountyName are Object, StateFIPS, CountyFIPS, ConfirmCaseCount, DeathCount are Integers
- Verified data and no nulls exists in all columns.

Before join population data, we have verified US Population data for all counties.

We have merged US Covid19 Confirmed Death combined and Population datasets by using below steps.

- Taken all common columns between data sets. i.e. State, CountyFIPS, CountyName

- Even though we can just join based on CountyFIPS, we have taken Names as well since they all come from same soure i.e. usafacts.org
- We have tested with and without names, both gives same results.
- Also to avoid missing data due to joins, we have used left join on County Population to ConfirmCase&Deaths.
- So merged based on above columns and verified data, they are all looks good except Population column as float64

Since most of datasets(Enrichment Data) contains only State Names instead of State&StateFIPS, which is missing in final combined COVID19 dataset.

- We have taken US States Codes & Names mapping in a list

Covid19 combined dataset improvements and make useful & easy next steps for analysis

- Added StateName column based on States Mapping list
- Rearranged columns for easy readability
- Converted Population column datatype from Float to Int
- Converted ReportDate datatype from Object to Date datatype
- Verified data and now Population&StateName added to final dataset

Verified final Covid19 Dataset has any null values, observe data Population column has 43,558 blanks(due to left join) outof 2,397,943

[168]:
```
USCovid19DF.to_csv('..\\..\\Data\\US_Covid_19_Dataset.csv', index=False)
USCovid19ReadDF = pd.read_csv('..\\..\\Data\\US_Covid_19_Dataset.csv')
USCovid19ReadDF.head()
```

[168]:

	ReportDate	StateFIPS	State	StateName	CountyFIPS	CountyName	\
0	2020-01-22	1	AL	Alabama	0	Statewide Unallocated	
1	2020-01-22	1	AL	Alabama	1001	Autauga County	
2	2020-01-22	1	AL	Alabama	1003	Baldwin County	
3	2020-01-22	1	AL	Alabama	1005	Barbour County	
4	2020-01-22	1	AL	Alabama	1007	Bibb County	

	ConfirmCasesCount	DeathCount	Population
0	0	0	0.0
1	0	0	55869.0
2	0	0	223234.0
3	0	0	24686.0
4	0	0	22394.0

Exported Python Dataframe data into csv file and stored under Data directory named: US_Covid_19_Dataset.csv

- And also read again from final dataset data into dataframe and verified all working fine for next Enrichment analysis
- We have observed that again Population datatype changed from Int to Float, which we handled in Enrichment data integration

Converted Population Datatype from Float to Int, since its changed during reimport final dataset from CSV

2.5.1 Preliminary Observations (intuition from the data)

Both confirmed and death cases each county seems to exponentially grow over time given the time period. They are highly likely to be correlated to each other In looking at the data, the data presented is from January 22, 2020 to February 10, 2022. It shows the increase in the number of confirmed cases and the number of deaths in the various counties across America. The population ranges from 0 to over 10,000,000 people. Underlying data sets have been leveraged to facilitate analysis of the independent variables that have an influence on the progress of the COVID-19 pandemic. In tandem with the variance of population density and socioeconomic status throughout the country, newly discovered trends, such as the political majority by state gives a picture of how the human element is steering the direction of the COVID-19 outbreak.