Impact of COVID-19 on U.S. Industries

Group 11: Ravali Anupoju (015359604), Wan Chen (015293278), Sharvaree Kshirsagar (015384044), Sruthi Mallarapu (015355405), Nidhi Parihar (015298244)

Abstract — The COVID-19 pandemic has drastically changed our lifestyles and has also affected various industries across the globe. We analyzed the publicly available dataset of COVID-19 impact on various U.S. Industries such as Consumer Spending, Employment, Unemployment claims, and small business revenue. There is a sharp decline in Consumer Spending in April and this is the time when the lockdown was imposed and gradually the consumer spending has been restored to its normalcy over the period. The entertainment and recreation industry is the most affected one in terms of consumer spend and the leisure and hospitality industry employment rates have declined. The grocery and food store industry has gained profits during this pandemic. Small businesses haven't recovered yet even though spending patterns of the consumers have increased which has affected the low-wage worker's employment rate. We used MongoDB, Mysql, Microsoft SQL Server to store the data, Talend tool to perform ETL, Postgres as a Data Warehouse, Tableau and Excel as visualization tools

Keywords—Database, Data Analytics, Data Visualization, Data Warehouse

1 Introduction

A COVID-19 pandemic is a one-in-a-lifetime event for everyone. The whole world is in this together with no exceptions. It has changed the way all of us live. With the increasing number of confirmed cases and death cases, People have started to wear masks every day and keep a 6.ft
of distance to avoid the spread of the virus. The majority of the people begin to work from home. Education is done virtually. People are staying home a lot more than before which changes the way of their spending behaviors. Therefore, the consumer spending pattern is shifting significantly which leads to a decline in overall small businesses revenue. As a result, the employment rate also drops to a historic low and the unemployment claim is rising sky-high.

In this project, we used the publicly available dataset of impacts of COVID-19 on various industries that include affinity(consumer spending data), unemployment claims, employment counts, womply (small business revenue) to get meaningful insights on behavior changes during this time.

2 Design Flow

In this section, we documented all the detailed procedures that we've performed to analyze the data using the knowledge that we have learned from the lecture, homework, and other tutorials

during research. There are a lot of tools available in the data-analytic area and choosing which tools to use and explore gives us a great opportunity to dive deeper into the subject and be familiar with the industry.

2.1 Overview of the Data

In this Analysis, We used United States data of 6 different industries such as Employment, Small Business Revenue, Unemployment Claims, Consumer spending, and Covid data. All the industries data except Unemployment Claims are at state, daily level ranging from Jan 2020 to April 2021. GeoID sheet is the reference table for all the other tables that have details of the states such as population, name of the state, etc.

Metadata Information

Table name	Record Count
COVID	22083/20
Affinity	21713/18
Womply	23046/24
UI Claims	3213/18
Employment	19839/12
Geoid	51/4

2.2 Cleaning the Data

- Performed data cleaning using python.
    - Replaced the '.' values as "0" for easier data transportation.
    - Dropped the columns that are not necessary for analysis
    - Modified the column data types appropriately
    - The delimiters should be the same across each column. We have used the comma "," separate delimiter.
    - Each Line is terminated by '\n'.
- Created Entity Relationship diagram to form well-structured relations.
- Removed the transitive and partial dependencies from the data and converted the data into 3NF.

2.3 Importing the data

We imported our data into the Relational Database - MYSQL and NoSQL database - MongoDB to familiarize with different types of databases.

- **MYSQL**
    - Created python script to import .csv file data into MYSQL database using **mysql-connector.**
- **MongoDB**
    - Created a cluster on MongoDB cloud.
    - Developed the python script to create a database, collections, and import .csv file data into collections using pymongo.
    - Using the **MongoDB BI** connector, connect MongoDB to Tableau desktop.
- **Microsoft SQL Server 2008 R2**
    - Used create a query to create database project in Microsoft SQL
    - Server Management studio
    - Under Task, use the import data wizard to import the cleaned CSV as a flat-file source into the project. Make necessary changes for data types.
    - Once the data is imported, the tables are ready to be used for query and analysis.
    - These data are later used in Microsoft Excel for visualization.

2.4 Data Analysis using SQL

We have analyzed the data by executing SQL queries. We have some insightful information.

- The entertainment and recreation industry are the most affected industries where the average consumer spending has dropped by 41.7 %. However, the spending has increased by 12.4 % in the grocery and food stores industry when compared to Jan 2020.

100% 1:44 3 errors found

Result Grid Filter Rows: Search Export:

Industry	percentage_change	affect_rank
arts, entertainment, and recreation	-41.77	1
transportation and warehousing	-38.75	2
accomodation and food service	-25.17	3
health care and social assistance	-11.69	4
generalstores and apparel and accessories	-2.91	5
grocery and food store	12.44	6

Fig. 1. Top affected industries

- Low wage workers employment_rate income has affected much when compared to high wage employment_rate

Result Grid Filter Rows:

inc_low	inc_high	y_m
0.2	0.03	2020-1
-0.01	0.95	2020-2
-5.11	-0.97	2020-3
-29.32	-8.86	2020-4
-27.31	-4.05	2020-5
-18.77	-0.43	2020-6
-15.57	-0.34	2020-7
-14.77	0.3	2020-8
-13.97	1.68	2020-9
-13.01	2.29	2020-10
-12.85	2.91	2020-11
-14.59	3.24	2020-12
-16.01	2.63	2021-1
-14.47	3.51	2021-2

Fig. 2. Demonstration of Microsoft Excel Power Pivot layout

- California has almost double the claims than New York (rank #2).

statename	TotalClaims
California	4493019
New York	2320849
Texas	1229798
Pennsylvania	1148808
Illinois	902177
Ohio	796385
Massachusetts	756313
Michigan	706128
New Jersey	700410
Indiana	418580

Fig. 3. Total 10 states with highest Unemployment claims counts

- The employment level of all the workers in specific industries

Result Grid Filter Rows: Export: Wrap

Worker_Industry	Employment_level_for_workers
leisure and hospitality	-16.3
trade,transportation and utilities	-5.68
education and health services	-4.32
professional and business services	-3.31

Fig. 4. Top affected industries

- Top affected states on employment rate during the year 2020 and 2021

Result Grid Filter Rows: Export: Wrap

STATENAME	YEAR	total_employment_rate
Hawaii	2020	-15.05
Vermont	2020	-13.08
New York	2020	-10.86
West Virginia	2020	-10.84
Rhode Island	2020	-10.58
New Jersey	2020	-8.82

Fig. 5. Top affected states on the employment rate in 2020

Result Grid Filter Rows: Expo

STATENAME	YEAR	total_employment_rate
District of Columbia	2021	-12.44
Rhode Island	2021	-12.28
Vermont	2021	-12.07
New York	2021	-11.37
West Virginia	2021	-10.85
Connecticut	2021	-10.75

Fig. 6. Top affected states on the employment rate in 2021

- Top affected states on spend rate during the year 2020 and 2021

Result Grid Filter Rows: Sea

statename	year	spend_rate
District of Columbia	2021	-15.56
Alaska	2021	-7.74
Missouri	2021	-5.91
Nevada	2021	-5.29
Colorado	2021	-3.26

Result Grid Filter Rows:

statename	year	spend_rate
District of Columbia	2020	-27.77
South Dakota	2020	-12.73
Alaska	2020	-12.65
California	2020	-12.11
New Mexico	2020	-11.96

Fig. 7. Top affected states on the employment rate in 2021

Data Analysis Using MongoDB:

Monthwise new case count

```
{ "_id" : { "year" : 2020, "month" : 1 }, "TotalCases" : 3 }
{ "_id" : { "year" : 2020, "month" : 2 }, "TotalCases" : 70 }
{ "_id" : { "year" : 2020, "month" : 3 }, "TotalCases" : 122194 }
{ "_id" : { "year" : 2020, "month" : 4 }, "TotalCases" : 856780 }
{ "_id" : { "year" : 2020, "month" : 5 }, "TotalCases" : 727942 }
{ "_id" : { "year" : 2020, "month" : 6 }, "TotalCases" : 772255 }
{ "_id" : { "year" : 2020, "month" : 7 }, "TotalCases" : 1836303 }
{ "_id" : { "year" : 2020, "month" : 8 }, "TotalCases" : 1525409 }
{ "_id" : { "year" : 2020, "month" : 9 }, "TotalCases" : 1186066 }
{ "_id" : { "year" : 2020, "month" : 10 }, "TotalCases" : 1749108 }
{ "_id" : { "year" : 2020, "month" : 11 }, "TotalCases" : 4228296 }
{ "_id" : { "year" : 2020, "month" : 12 }, "TotalCases" : 6174880 }
{ "_id" : { "year" : 2021, "month" : 1 }, "TotalCases" : 6275299 }
{ "_id" : { "year" : 2021, "month" : 2 }, "TotalCases" : 2616982 }
{ "_id" : { "year" : 2021, "month" : 3 }, "TotalCases" : 1805195 }
{ "_id" : { "year" : 2021, "month" : 4 }, "TotalCases" : 1324794 }
```

Fig. 8. Monthwise case count

2.5 Performing ETL

- Talend is an open-source tool and widely used for ETL (Extract, Transform and Load)
- This tool can easily manage all the steps involved in the ETL process from the design, format, cleaning, transform to the execution of ETL data load from different systems.
- In this project, we have used this tool to load the data from MySQL Database to Postgres Database. Below are the steps we have followed to complete the data load.
    - Create source (MySQL) and target connection (Postgres) in the metadata node.
    - Retrieve the required objects from the schema of a source connection.
    - Create a new job under the Job Designs node in the Repository tree view.
        - ❖ An empty design workspace opens up once a job is created.
        - ❖ Drag source(Table) component from the metadata and add into the workspace and configure the component and in the same way configure the target component.
- Using the tMap component, map the source and target columns in the mapping editor
- Once all the components are added, run the jobs and extract the data from the Mysql table into the Postgres target table.

2.6 Creation of fact and dimension tables

We used Postgres as a DataWarehouse and have developed code in PostgreSQL to create one fact table and one dimension table.

Fact table: It is created by aggregating all the tables at a monthly level and keeping only the desired columns for our analysis.

Dimension Tables: Enriched the time dimension with columns such as Quarter, Month Name, and year_month from the date column that are helpful for our analysis.

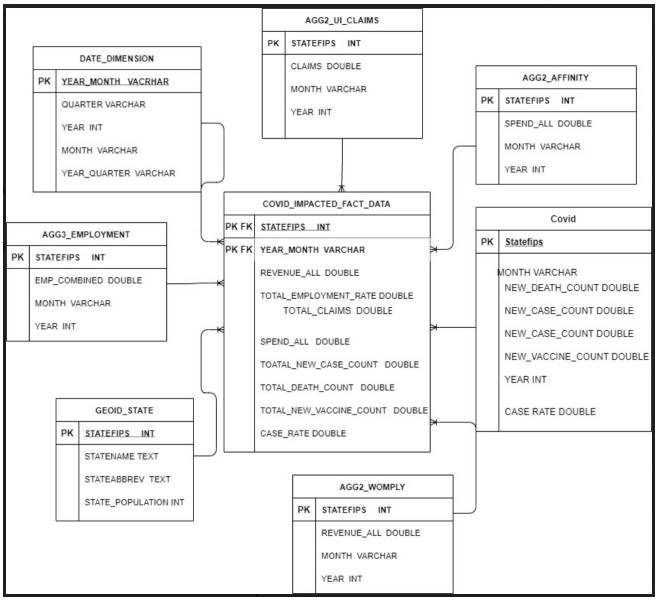

Fig. 9. Star Schema for fact and dimension tables

3 Analyzing the data and data visualization

We explored multiple visualizations tools. Since Microsoft Excel is still widely used in the industry and Tableau is a rising star, we have decided to use these two softwares. Microsoft Excel has

been in use for a very long time and most people had already been very familiar with it. Therefore, We decided to use the Power Pivot function in Microsoft Excel for a basic summary table and chart and use Tableau as our main tool for dashboards.

3.1 Microsoft Excel Power Pivot

Power Pivot is an add-in tool that is available for recent release excels (2010 and after). The biggest advantage of power pivot over regular pivot is that Power Pivot enables users to use data from different tables once the relationship is created for the data model. Vlookup will no longer be necessary when trying to get insights from multiple tables. Power Pivot also allows users to create measures and hierarchies of the table which is very helpful for data analysis.

- Connect the Microsoft SQL database as an external source and import the tables from the “project” database.
- Created the table relationship using the “statefips” common field for all tables imported.
- After setting up the data model, the pivot table and chart is ready to be generated.
- In Fig.4, Table for Top 10 states had the most new case count for this pandemic period (Mar 2020 - Apr. 2021) is created.
- California has the most counts of all states followed by Texas and Florida. This is also corresponding to the population ranking in the US.

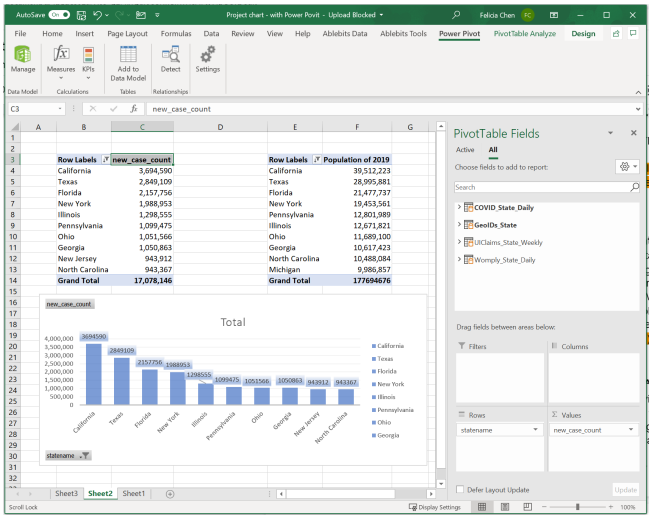

Fig. 10. Demonstration of Microsoft Excel Power Pivot layout

3.2 Tableau Dashboards

Tableau is a frontend visualization tool and can connect to hundreds of data sources with a tableau desktop. We used Tableau to connect with MongoDB and PostgreSQL. Since we experience some performance issues when directly loading all the data into tableau and running the queries there. We perform data enrichment and import only the fact and dimension table into the dashboard. We also divided our dashboard into multiple dashboards with fewer functionalities to avoid having complex queries which cause longer query elapsed time.

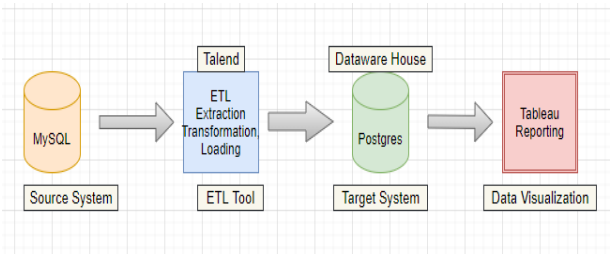

Fig. 11. Process flow Diagram

Covid impact on the overall United States:

Case rate: The covid-19 pandemic began in the year 2020 in the United States. The below Dashboard showcases the rates at which case increases every month and for every state.

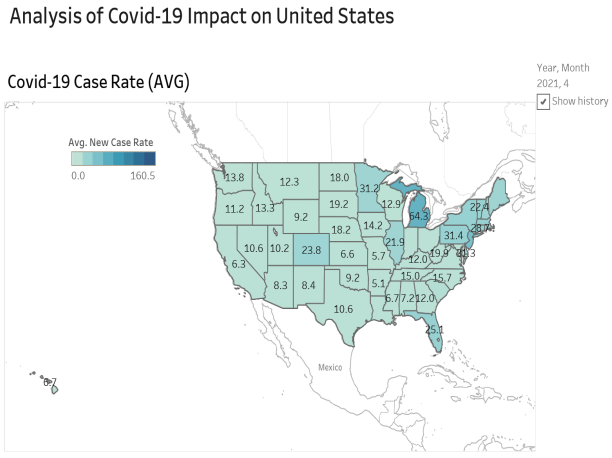

Fig.12. Dashboard for Covid-19 Case Rate

Case count: Since the covid-19 pandemic began in March 2020 in the United States, we took monthly new cases count for the whole year 2020 which shows the month has the maximum case count.

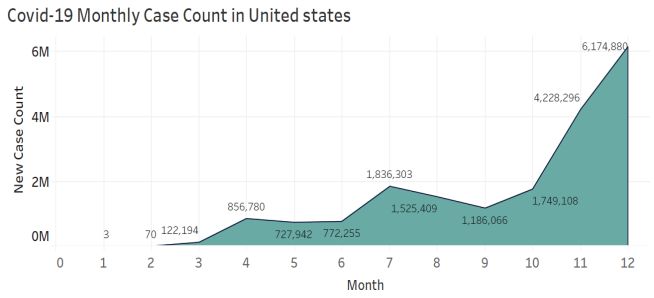

Fig. 13. Dashboard of Covid-19 monthly case rise

Top 10 affected states: In this graph(Fig. 14), we can list the top 10 states having the highest number of death counts till April 2021. California has the highest number of death counts followed by New York and Texas.

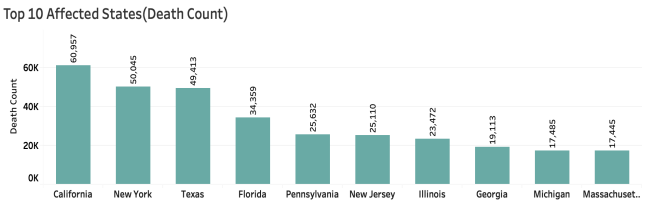

Fig. 14. Demonstration of top10 affected states in United States

Case Count VS Vaccine Count: Since vaccination began in the United States in Feb 2021, vaccination count started increasing. We can clearly see that the new cases count started decreasing due to vaccination, social distancing, and lockdown. In the month of March vaccination count was the highest.

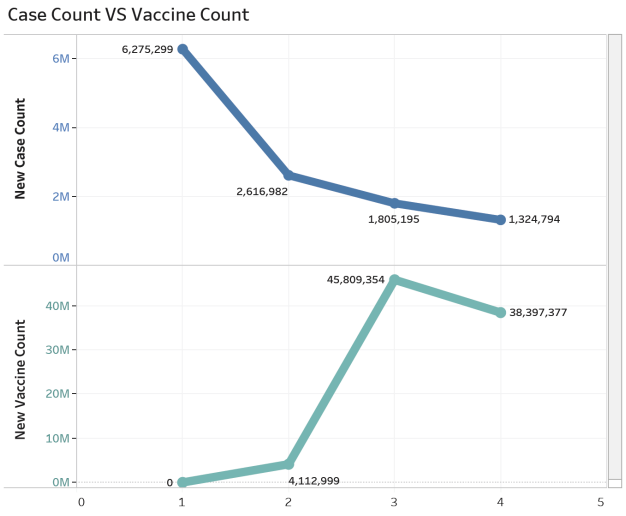

Fig. 15. Demonstration of case count VS vaccine count

Employment Rate and UI Claims Insights

UI Claims: Unemployment Insurance Claims Count on monthly basis.

Employment Rate: Employee level of all the workers compared to Jan 2020.

Since the unemployment claims data is on a weekly basis, we use last Saturday as the last day of the month to calculate the count of unemployment insurance claims. We combine the unemployment claim with the pandemic unemployment assistance and the pandemic unemployment emergency compensation as the aggregates UI Claims.

For the Employment rate, we take the average of the data for the month to calculate the percentage of each month.

We created these two graphs in the dashboard together and we can see the curve of bar charts are correlating. The UI claims increase for a couple of months when the unemployment rate drops dramatically in April 2020. This is due to EDD requiring more time to respond to the sudden elevation of incoming claims.

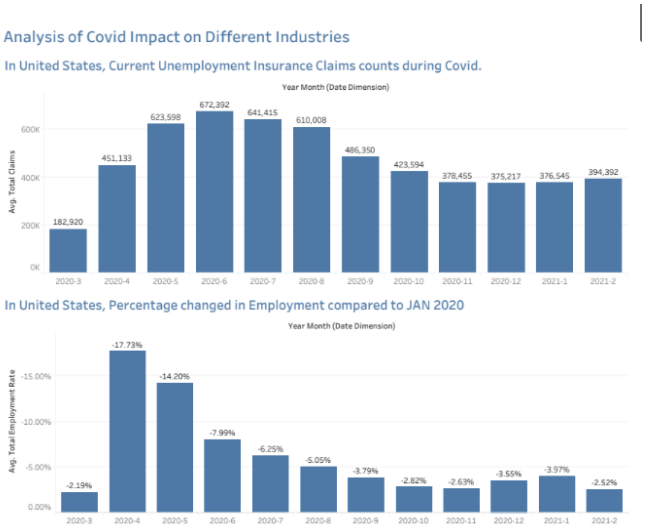

Fig. 16. Dashboard of Job situation impact

Customer Spending and Small Business Revenue Insights

Consumer Spending: Total spending by all consumers compared to Jan 2020. Consumer Spending has decreased dramatically in the

month of April and gradually picking up the pace over time.

Small Business Revenue: Net revenue earned from transportation, professional, business services, education, health services, and hospitality has decreased in the month of April combined for all the states.

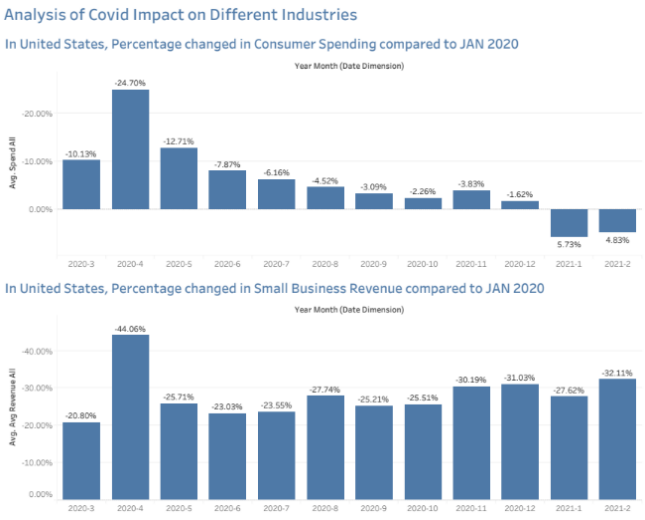

Fig. 17. Dashboard of consumer spending and small business revenue impact

4 Lesson Learned

- Pair programming - Since we have different areas to analyze for this project, We consistently do pair programming to ensure our results are corresponding with the related area. Example: When creating the fact table, Sruthi acted as the organizer and met with the corresponding team members who are familiar with each area one by one to make sure the aggregates that are included in the fact table are accurate.

- Teamwork - Our team meets regularly to review everyone's progress and voice any concerns or ideas about the project. We will discuss the objective and purpose of the work and divide them up. We also set up a due date for deliverables. Since everyone has different backgrounds of specialties, when anyone has questions or issues, other team members are very helpful and are willing to provide additional help.

- Technical difficulties - There are many tools to choose from for this project. Some tools are for subscription and have a short trial period. Choosing which tools to use and to allocate when and who to sign up for which tool needs planning. We try to use as many tools that are relevant as possible, however, we are only able to use the basic features for most of them. This project introduces us to many useful tools as data analysts and we will continue to learn and explore more advanced features of the tools going forward.

5 Conclusion

Even though Consumer Spend is increasing, small business revenue is still affected which in turn is affecting the low-wage workers employment. Entertainment and recreation industry is the most affected industry as the government has imposed strict lockdown during the pandemic. Food and grocery Industry has gained profits as people are habituated to eat at home. The federal government should issue funds to help these small businesses and unemployed workers to sustain their daily lives.

Appendices

1. Presentation Skills - Elevated by professor
2. Significance to the real world - Introduction
3. Code Walkthrough - In presentation
4. Report - Elevated by professor
5. Version Control - https://github.com/sruthimallarapu/Data225-Group11
6. Discussion/Q&A
7. Lessons Learned - section 4 lesson learned
8. Innovation - Abstract and section 2.6
9. Teamwork - section 4 lesson learned
10. Technical Difficulties - section 4 lesson learned
11. Practiced pair programming? - section 4 lesson learned
12. Practice Agile/Scrum - see DATA 225 Project | MINUTES - group 11 document.
13. Used Grammarly/other tools for language - Uploaded screenshot
14. Elevator pitch video (see youtube link)
15. Slides - Project powerpoint
16. Demo - In presentation
17. Used Unique tools - Microsoft Excel Power Pivot, TALEND
18. Performed substantial analysis using database techniques - section 2.4 Data Analysis using SQL
19. Used a new database or data warehouse tool not covered in the HW or class - Microsoft SQL Server, MongoDB
20. Used appropriate data models (figure 3 for star schema)
21. Used ETL tools - Talend
22. Data Cleaning - section 2.2
23. Demonstrated how analytics support business decisions - Conclusion
24. Used NOSQL database - MongoDB
25. Used RDBMS - MySQL
26. Used Data Warehouse - PostgreSQL
27. Includes DB Connectivity/API calls: pymongo for mongodb, mysql-connector-python for mysql (see python code)

Acknowledgement

We would like to sincerely thank Vishnu S. Pendyala, PhD for the excellent materials on DATA-225 Sec 21 - Db Systems for Analytics Lectures.

References

[1] Chetty, Friedman, Hendren, Stepner, and the OI Team (2020) Economic Tracker [Online] Available at: https://tracktherecovery.org. [Accessed 3rd Mar 2021]

[2] "The Economic Impacts of COVID-19: Evidence from a New Public Database Built Using Private Sector Data", by Raj Chetty, John Friedman, Nathaniel Hendren, Michael Stepner, and the Opportunity Insights Team. November 2020. Available at: https://opportunityinsights.org/wp-content/uploads/2020/05/tracker_paper.pdf

[3] Introduction to MongoDB. [Online] Available at https://docs.mongodb.com/manual/tutorial/getting-started/ [Accessed 3rd Apr 2021]

[4] Power Pivot: Powerful data analysis and data modeling in Excel. [Online] Available at https://support.microsoft.com/en-us/office/power-pivot-powerful-data-analysis-and-data-modeling-in-excel-a9c2c6e2-cc49-4976-a7d7-40896795d045 [Accessed 6th May 2021]

[5] Github: https://github.com/OpportunityInsights/EconomicTracker[Accessed 8th Apr 2021]