IMT 563
Group 9 | Meghana Parab, Yiliang Wang, Janice Chun-Ping Yang, Ken Yu-Ru Yang
March 16, 2021

COVID-19 Travel Navigator

Problem Statement

COVID-19 has greatly impacted air travel on a global scale. As many countries continue to proceed towards reopening, there will be a growing need for people to make travel decisions, both internationally and domestically. We wanted to find a relational database approach to help people make informed travel decisions in light of uncertainty surrounding the pandemic. Our users are those with air travel needs such as travel agencies and business/independent travelers. We envision that our model be integrated into systems of online travel shopping companies such as Expedia or metasearch engines such as Skyscanner. The system should provide accurate information, store data safely, and provide the latest information possible.

Domain and Scope

Our team established a relational database system with an interactive front end that engages with users. The system provides present-day information on the state of air travels, COVID-19 case and death tolls, and COVID-19 vaccination progresses across 15 countries worldwide and 50 states domestically. The following lists in-scope and out-of-scope components.

- In-scope
    - Front end
        - Compare flight rates with COVID-19 cases by country. This will help us understand if rise/fall in COVID-19 cases is impacting air travel across the international/national boundaries
        - Users can filter air travel and vaccination information by country. This helps us answer questions such as whether an increased vaccination rate of a country has enhanced air travel to that country
    - Back end
        - Store data on a cloud platform (AWS) with data sources being updated daily (every 12 hours).
- Out-of-scope
    - Out project does not provide future predictions on COVID-19 or air travel trends as that will require additional machine learning or statistical modeling techniques
    - Data sources are not updated in real time

Data Source

We used datasets from the following sources. A comprehensive reference list is provided in the final section of the report.

- COVID-19 case/death: COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Format is CSV.

- COVID-19 vaccination progress: Our World in Data, a scientific publication based at the University of Oxford. Format is CSV.
- Air travel:
    - 15 countries: OAG (Official Airline Guide), a global travel data provider with headquarters in the UK. Format is CSV.
    - 50 domestic states: OpenSky Network, a non-profit receiver network based in Switzerland that collects unfiltered raw data of air traffic surveillance. Format is CSV.

Database Implementation

Back end

To provide an integrated and timely information of vaccination, COVID-19 cases and flights, our database should have good scalability to accommodate data from multiple sources and make updates on a daily basis. Thus, we used AWS to meet all the needs in a serverless manner.

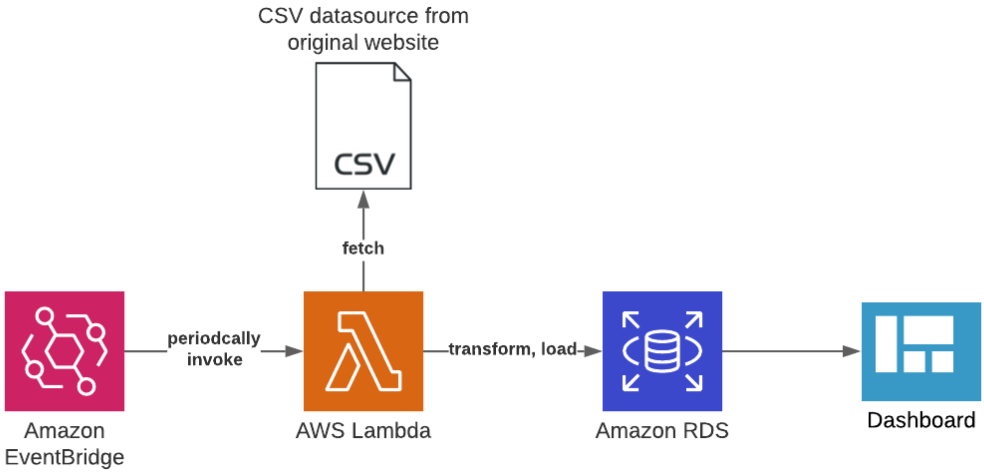


Figure 1: Database implementation process flow diagram

Tools and techniques:

- Data extraction and transformation:
    - Because all of our data is in the form of CSV files, we do not need to extract the data from webpages using scraper. We simply used pandas to create a dataframe from the web source file directly and transform the data into a normalized form in order to better load it into our database on the cloud.
- Data update:
    - To achieve an automated update of the dataset periodically, we used AWS Lambda to deploy the ETL process. Our COVID-19 and vaccination data was updated daily, and flight data was updated monthly. So we configured Amazon EventBridge to invoke our Lambda function twice a day to listen to the update of data sources. And once the data is updated, the Lambda function conducts the ETL process and writes the updated part to our database.

- Data storage:
    - To better integrate our ETL and update jobs with data storage, we created a MySQL database instance on Amazon RDS as our cloud data storage. Because RDS can communicate with Lambda functions in the RDS proxy network which provides a better scalability and secure environment for querying and updating.

Database architecture:

- Entities relation & data sources integration
    - We created a schema that includes both COVID-19 related datasets and air travel dataset in 12 tables. We have two layers of granularities for the location: country and state. All COVID-19 data connects with either country or state table in a many-to-one relation, and we designed two sources of flight in different ways: for the OpenSky Network dataset, each flight was provided by departure and arrival airport, so we created the airport intermediate table to map each individual flight into different granularities of locations; for the OAG dataset, the data is provided in an aggregated way, therefore the design is similar to the COVID-19 data.
    - The state table is designed in such a way that it separates from the country table, because by far, only US dataset is provided in the granularity of state. Considering that no query from the dashboard will need the relation of state and country, we leave this as a future plan to improve the scalability to accommodate more countries with statewide data.

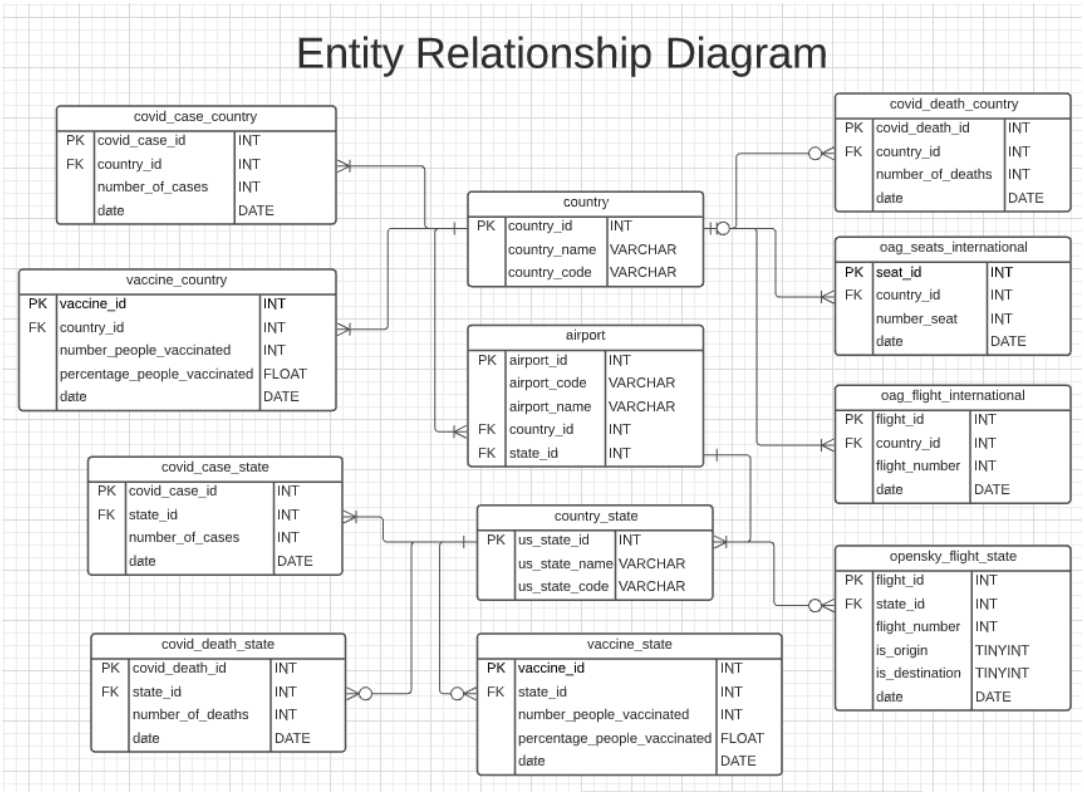


Figure 2: Entity relationship diagram

Front end

After we set up our databases and loaded the data, joined tables by writing SQL queries, connected Tableau with the joined tables, and started creating our interactive dashboards. We aimed to achieve five features as follows.

1. Provide dashboards on global scope and country scope (United States)
2. Visualize information on air travels, and users can filter information by fields such as country
3. Visualize information on COVID-19 cases, deaths, and vaccination progress where users can sort information by different metrics
4. Compare number of flights before and after COVID-19
5. Compare number of flights with total COVID-19 cases

The dashboard is targeted to serve two types of users—travel agencies and individual travelers.

User Type	User Story	Success Criteria
Travel Agency	As a user that advises and sells travel itineraries to groups of travellers, I would like to view data for multiple countries at once and know which are the most popular destinations to travel during the COVID-19 pandemic	Users can derive necessary insights from visualization that shows: 1. Number of COVID-19 cases, deaths, and vaccination progress by country 2. Flight information including how many flights were booked to a particular country and flight capacity
Individual Travelers	As a user traveling for business to different countries, I want to assess flight reliability by understanding the impact of COVID-19 on flights to a particular destination	A visualization that allows user to- 1. Track COVID-19 cases by country 2. Filter data and focus on specific locations 3. Check vaccination rates by country 4. Correlate covid cases and number of flights to see the impact of COVID-19 on the travel industry

Table 1: User types

Finally, we built our dashboards, and each page contains different information to achieve our user requirements. We include screenshots of the dashboard in appendices.

- Page 1: Contains data from the 15 selected countries
    - Global Total Cases & Vaccination Progress: A geomap that presents total cases and percentage of people vaccinated in each country

    - Global COVID statistics: A table that lists down important COVID-19 metrics. Users are able to sort the data with every column
    - Global Flights & Seats: A chart that shows the total number of flights and seats in each month
    - Flights of Each Country: A bar chart that shows the total number of flights of each quarter. Users can select the country they want to focus on or click on all to view all countries at once

- Page 2: Contains data from the 50 states in the United States
    - United States Total Cases & Vaccination Progress: A geomap that presents total cases and percentage of people vaccinated in each state
    - United States COVID statistics: A table that lists down important COVID-19 metrics of each state. Users are able to sort the data with every column
    - Total Number of Flights: A bar chart that shows the total number of local flights in each month. Users are able to select the year that they want to include in the chart
    - Total Number of Flights: A bar chart that shows the total number of flights of each state. Users can select the state they want to focus on or click on all to view all countries at once

- Page 3: Combines total COVID-19 cases with number of flights of each country

Limitations

- Fight data from all countries was not available in our data sources
    - Solution: We conducted our analysis for 15 available countries
- State-specific flight data available only for US, not for other countries
    - Solution: We created separate state-based chart for US and country-based charts for the remaining countries
- Flight cancellation data was not available in our dataset
    - Solution: We used overall flight count as a metric to visualize the number of flights operating in a given region in comparison to COVID-19 cases for that region

Challenges

- Updating data sources on a periodic basis
    - Solution: We leveraged AWS Lambda features to update data sources daily
- Different data sources were stored at different levels of granularity
    - Solution: We wrote custom SQL query in Tableau to join data sources

References

COVID-19 Data Repository by the Center for Systems Science and Engineering (CSSE) at Johns Hopkins University. Retrieved March 13, 2021 from https://github.com/CSSEGISandData/COVID-19.

OAG Aviation Worldwide Limited. (2021). Coronavirus airline schedules data. Retrieved March 13, 2021 from https://www.oag.com/coronavirus-airline-schedules-data

Olive, X. (2019, July). Traffic, a toolbox for processing and analysing air traffic data. Journal of Open Source Software 4(39).

Ritchie, H., Ortiz-Ospina, E., Beltekian, D., Mathieu, E., Hasell, J., Macdonald, B., Giattino, C., Appel, C., & Roser, M. (2021). Coronavirus (COVID-19) vaccinations. Our World in Data. Retrieved March 13, 2021 from https://ourworldindata.org/covid-vaccinations

Schäfer, M., Strohmeier, M., Lenders, V., Martinovic I., & Wilhelm, M. (2014, April). Bringing up OpenSky: a large-scale ADS-B sensor network for research. In Proceedings of the 13th IEEE/ACM International Symposium on Information Processing in Sensor Networks (IPSN), pages 83-94.

Appendices

Project github repository: https://github.com/LouisYLWang/corornavigator

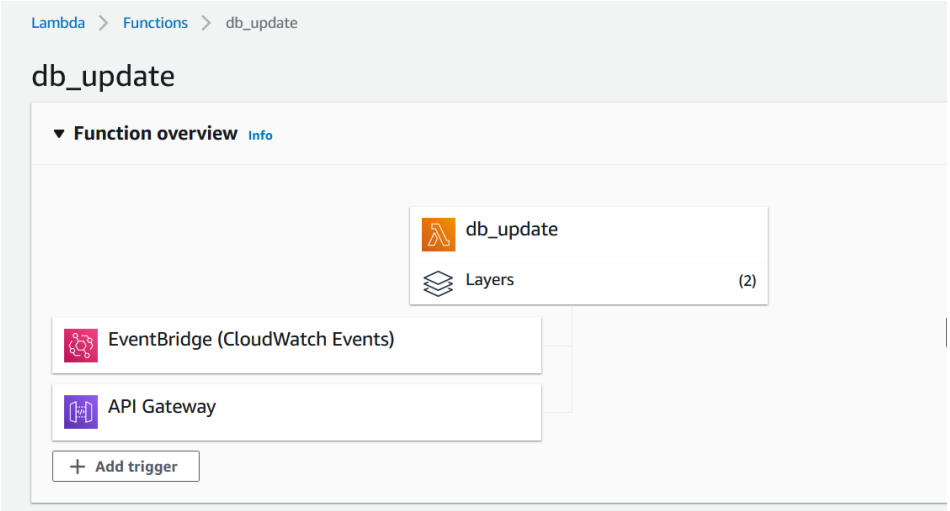


Figure 3: AWS lambda schema

Recent invocations

#	Timestamp	RequestID	LogStream
1	2021-03-16T19:49:01.831Z	01948149-82ed-4cf1-8325-5dae51508bd7	2021/03/16/[$LATEST]040c13ee985245f7a9d21383ef687f58
2	2021-03-16T07:49:01.529Z	e105faf9-5d6b-4d8e-b251-cbdc3a5a0786	2021/03/16/[$LATEST]7b8b4acd78634501926b439515876343
3	2021-03-15T19:49:02.057Z	4df0408f-644b-41dc-bd47-faf4fca44eaa	2021/03/15/[$LATEST]622c5de61b8147e4bac538352b13cd36
4	2021-03-15T07:49:01.622Z	ca2298b6-e0a5-43a4-a672-e9bc7149026c	2021/03/15/[$LATEST]3af852fdb88e45ffbea3c532b0bb9e6c
5	2021-03-14T19:49:01.165Z	974997c9-6c5a-49f9-a5ce-b65f3a1ec78b	2021/03/14/[$LATEST]a3034e8adb034c0cbb71ac4295a695be
6	2021-03-14T07:49:01.496Z	debc07a5-ac75-4322-af23-f4cdccbfb1eb	2021/03/14/[$LATEST]cb4fe6e9c3af472ab35422b03e514545
7	2021-03-13T19:49:02.313Z	f1a4a7fb-ef62-4447-95f4-1ab7ab07d068	2021/03/13/[$LATEST]ccf1231c0d11459c983592b2da6dc8b1
8	2021-03-13T07:49:01.341Z	b341b6f5-af3c-4977-bbd5-0de620c3570e	2021/03/13/[$LATEST]78a7236bb006482a979bdfdfe12a041a
9	2021-03-12T19:49:01.694Z	d784a1a1-3739-4239-99c4-2474b12c55b5	2021/03/12/[$LATEST]e64ddd46c9f842a0ae30be21db2615ce

Figure 4: Lambda function invocations log 1

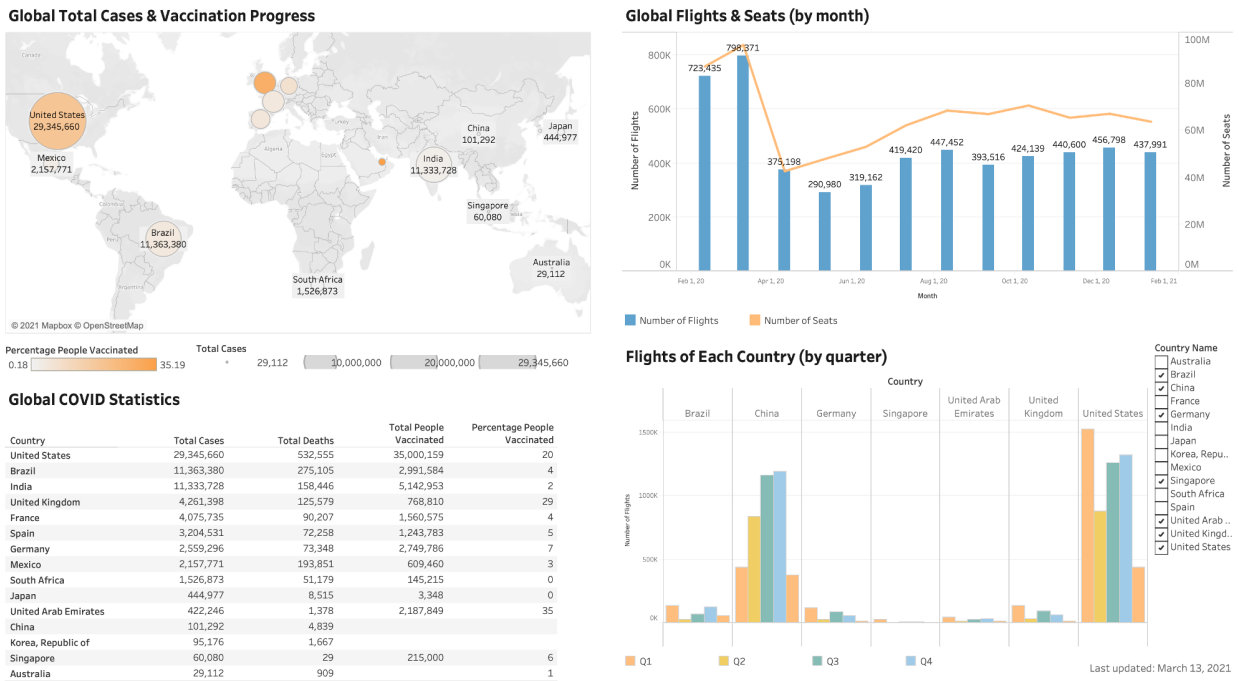
Log events
You can use the filter bar below to search for and match terms, phrases, or values in your log events. Learn more about filter patterns

View as text | Acti

Filter events | Clear | 1m | 30m | 1h

Timestamp	Message
	No older events at this moment. Retry
2021-03-16T00:48:56.280-07:00	START RequestId: e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 Version: $LATEST
2021-03-16T00:49:00.959-07:00	OpenBLAS WARNING - could not determine the L2 cache size on this system, assuming 256k
2021-03-16T00:49:00.959-07:00	[INFO] 2021-03-16T07:49:00.957Z SUCCESS: Connection to RDS MySQL instance succeeded
2021-03-16T00:49:01.151-07:00	[INFO] 2021-03-16T07:49:01.151Z e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 SUCCESS: No new data on schema vaccine_state.
2021-03-16T00:49:01.211-07:00	[INFO] 2021-03-16T07:49:01.210Z e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 SUCCESS: No new data on schema vaccine_country.
2021-03-16T00:49:01.288-07:00	/var/task/lambda_function.py:70: SettingWithCopyWarning:
2021-03-16T00:49:01.288-07:00	A value is trying to be set on a copy of a slice from a DataFrame.
2021-03-16T00:49:01.288-07:00	Try using .loc[row_indexer,col_indexer] = value instead
2021-03-16T00:49:01.288-07:00	See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
2021-03-16T00:49:01.288-07:00	dataset['date'] = dataset['date'].astype(str)
2021-03-16T00:49:01.328-07:00	[INFO] 2021-03-16T07:49:01.328Z e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 SUCCESS: Update 51 queries on schema covid_death_state, updated to 2021-03-15.
2021-03-16T00:49:01.394-07:00	[INFO] 2021-03-16T07:49:01.393Z e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 SUCCESS: Update 15 queries on schema covid_death_country, updated to 2021-03-15.
2021-03-16T00:49:01.474-07:00	[INFO] 2021-03-16T07:49:01.474Z e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 SUCCESS: Update 51 queries on schema covid_case_state, updated to 2021-03-15.
2021-03-16T00:49:01.512-07:00	[INFO] 2021-03-16T07:49:01.512Z e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 SUCCESS: Update 15 queries on schema covid_case_country, updated to 2021-03-15.
2021-03-16T00:49:01.529-07:00	END RequestId: e105faf9-5d6b-4d8e-b251-cbdc3a5a0786
2021-03-16T00:49:01.529-07:00	REPORT RequestId: e105faf9-5d6b-4d8e-b251-cbdc3a5a0786 Duration: 564.03 ms Billed Duration: 565 ms Memory Size: 128 MB Max Memory Used: 128 MB Init Duration: 4849.96 ms

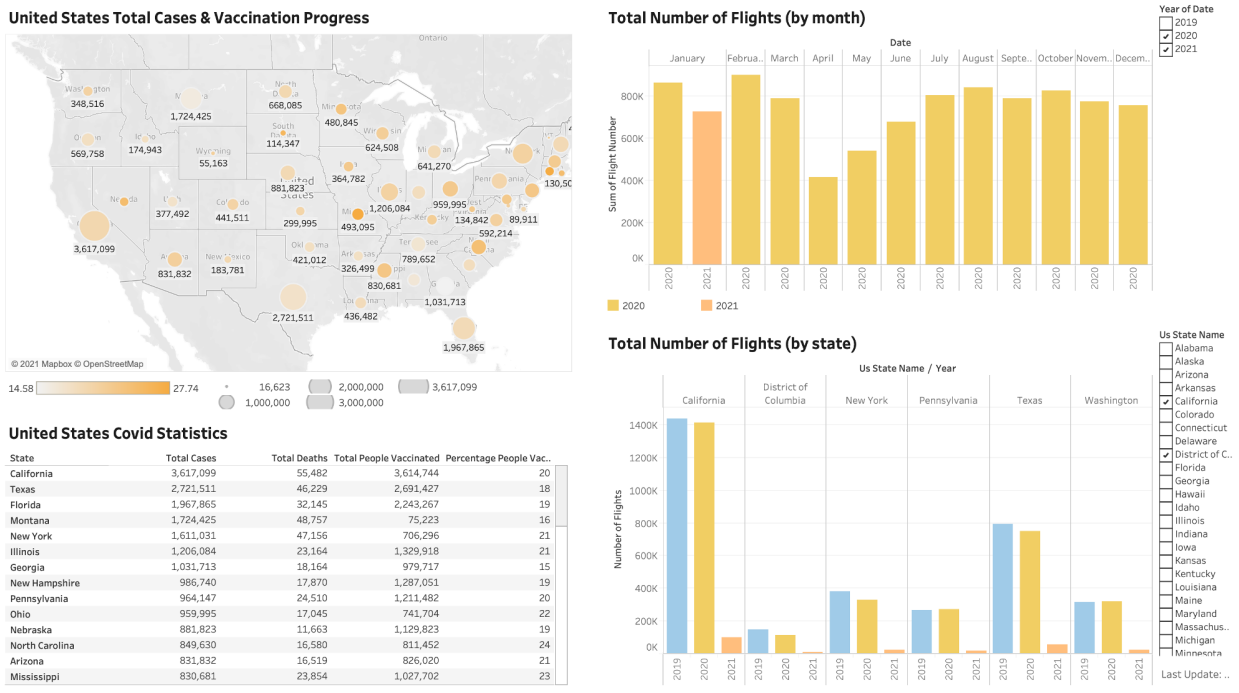
Figure 5: Lambda function invocations log 2



Global COVID Statistics

Country	Total Cases	Total Deaths	Total People Vaccinated	Percentage People Vaccinated
United States	29,345,660	532,555	35,000,159	20
Brazil	11,363,380	275,105	2,991,584	4
India	11,333,728	158,446	5,142,953	2
United Kingdom	4,261,398	125,579	768,810	29
France	4,075,735	90,207	1,560,575	4
Spain	3,204,531	72,258	1,243,783	5
Germany	2,559,296	73,348	2,749,786	7
Mexico	2,157,771	193,851	609,460	3
South Africa	1,526,873	51,179	145,215	0
Japan	444,977	8,515	3,348	0
United Arab Emirates	422,246	1,378	2,187,849	35
China	101,292	4,839		
Korea, Republic of	95,176	1,667		
Singapore	60,080	29	215,000	6
Australia	29,112	909		1

Figure 6: Tableau dashboard 1



United States Covid Statistics

State	Total Cases	Total Deaths	Total People Vaccinated	Percentage People Vac..
California	3,617,099	55,482	3,614,744	20
Texas	2,721,511	46,229	2,691,427	18
Florida	1,967,865	32,145	2,243,267	19
Montana	1,724,425	48,757	75,223	16
New York	1,611,031	47,156	706,296	21
Illinois	1,206,084	23,164	1,329,918	21
Georgia	1,031,713	18,164	979,717	15
New Hampshire	986,740	17,870	1,287,051	19
Pennsylvania	964,147	24,510	1,211,482	20
Ohio	959,995	17,045	741,704	22
Nebraska	881,823	11,663	1,129,823	19
North Carolina	849,630	16,580	811,452	24
Arizona	831,832	16,519	826,020	21
Mississippi	830,681	23,854	1,027,702	23

Figure 7: Tableau dashboard 2

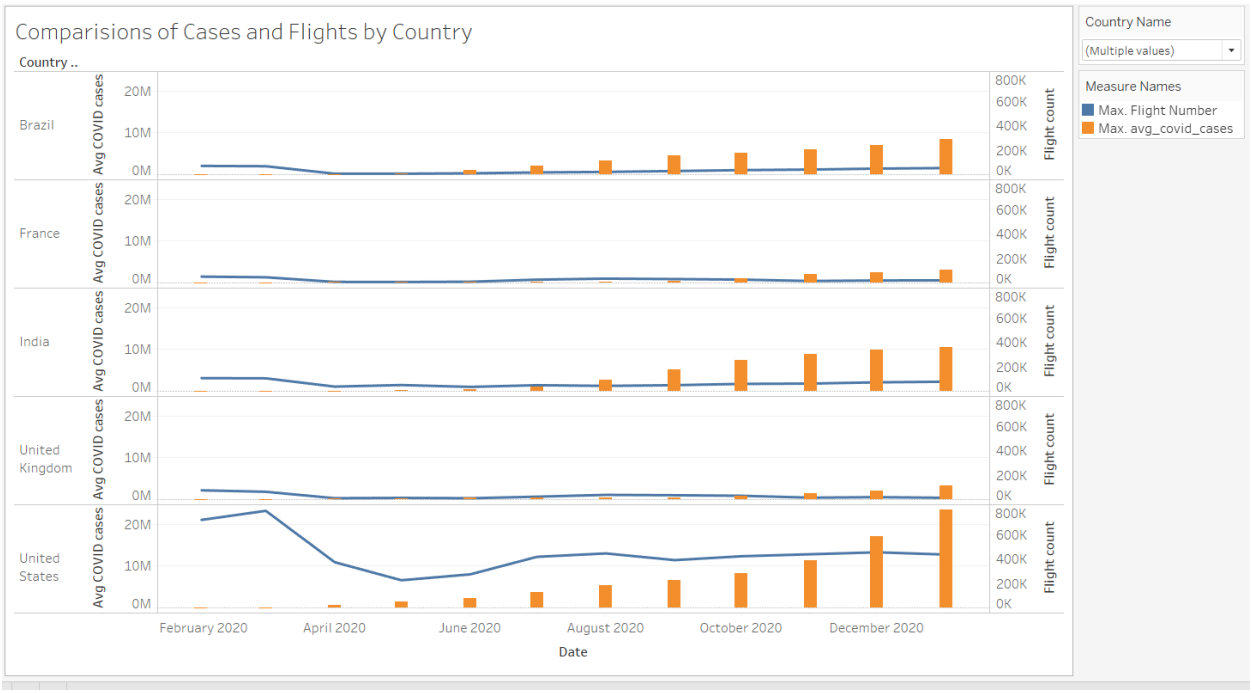


Figure 8: Tableau dashboard 3