The Impact of Covid-19 on Air Traffic: Spatiotemporal Forecasting and Benchmarking

Adelle Driker, Bo Yan, Yuan Hu

Advisor: Professor Rose Yu

Key Findings Through EDA

1. Data Quality Issues

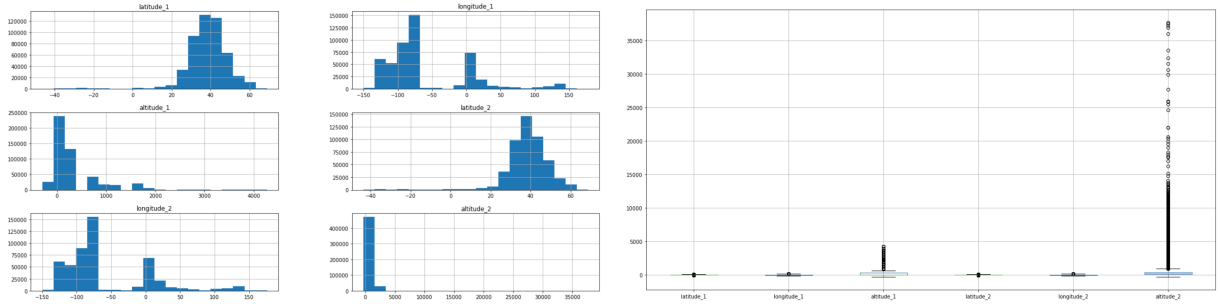
Through exploring the data, there are several data quality issues below that need to be cleaned.

- Missing values. Drop fields like number, registration, typecode, origin, and so on.
- Duplicated rows. Drop some duplicated rows in our datasets.
- Unused columns. Drop unused columns such as registration, and typecode.
- Data type. Some data types don't match our needs, so we convert them to the data types we need, such as parsing datetime.

2. Data Distribution

Through exploring the data distribution, we found some fields are highly skewed while others are moderately skewed or approximately symmetric. Also, there are some outliers for some fields. Please refer to the screenshots below.

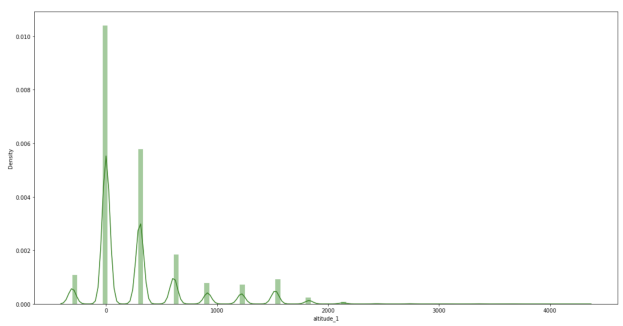
- Highly skewed distribution: latitude_1, longitude_1, altitude_1, latitude_2, longitude_2, and altitude_2.
- Positive skewness values such as longitude_1, altitude_1, longitude_2, and altitude_2 indicate asymmetry in the distribution and the tail is larger towards the right hand side of the distribution.
- Negative skewness values such as latitude_1 and latitude_2 indicate asymmetry in the distribution and the tail is larger towards the left-hand side of the distribution.
- Both altitude_1 and altitude_2 attributes have lots of outliers.

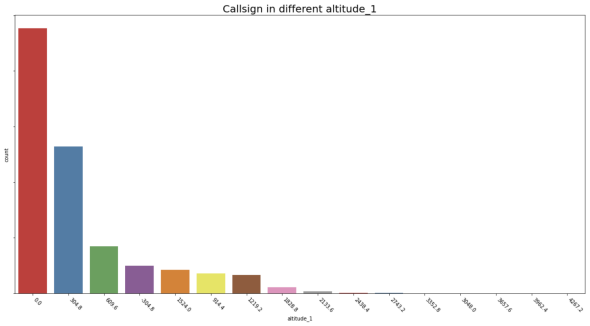


3. Common Trends

Through exploring the common trends, we can find the following results. Please refer to the screenshots below.

- Most of the callsigns are at an altitude of 0.0.
- Most of latitude_1 and latitude_2 belong to 30 to 50, meaning most airports are in the northern hemisphere.
- Most of the longitude_1 is in the range -175 to -80, meaning most airports are in the western hemisphere.

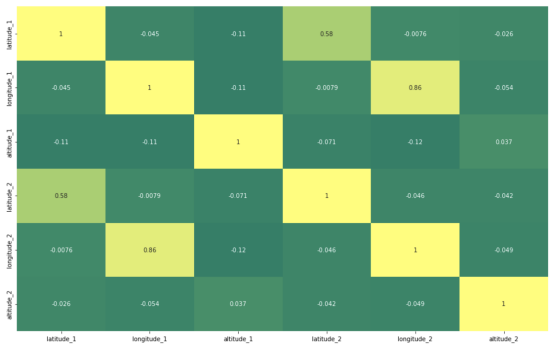


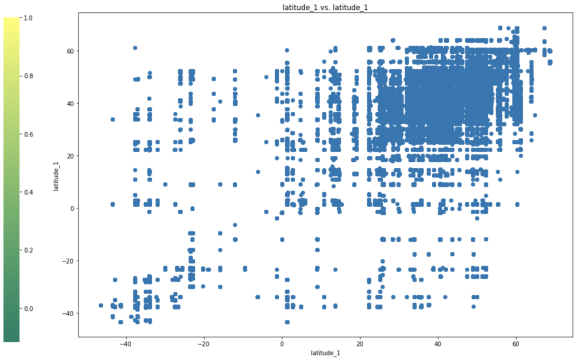


4. Variable Relationships

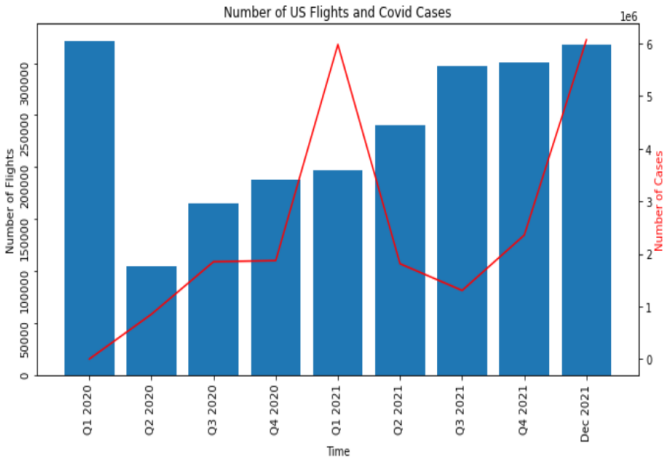
Some features are highly correlated with each other and dependent on each other. The highest correlation between features is 0.86. Please refer to the screenshots below.

- Latitude_1 and latitude_2, longtitude_1 and longtitude_2, and altitude_1 and altitude_2 are positively correlated.
- The relationship between longtitude_1 and latitude_1, altitude_1, longtitude_2, altitude_2 are slightly negatively correlated.
- The relationship between latitude_1 and longtitude_1, altitude_1, longtitude_2, altitude_2 are slightly negatively correlated.



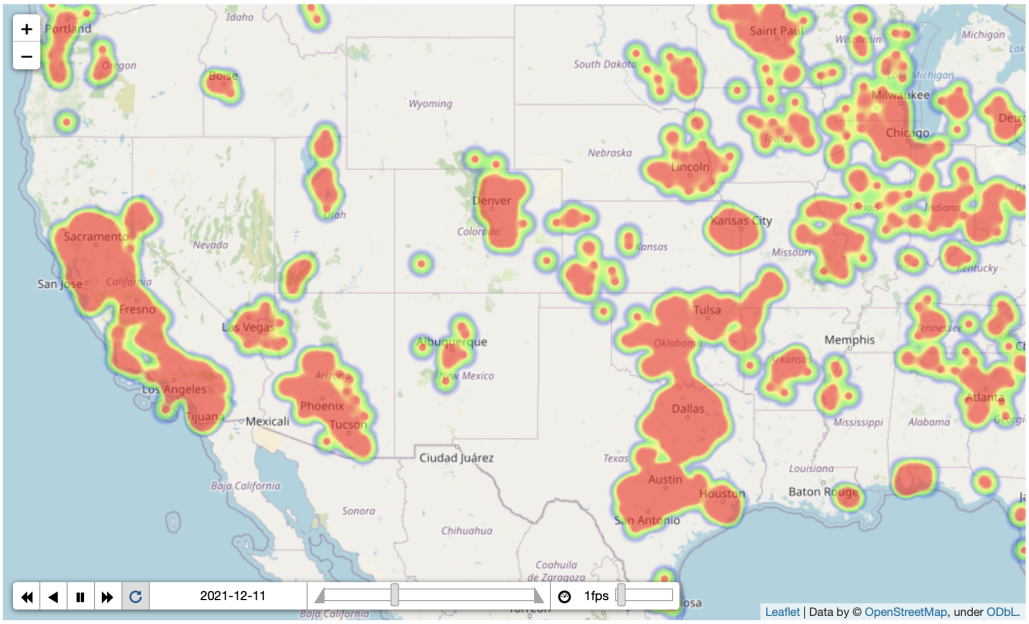


5. An analysis of the time series data from both the flight and Covid perspective in the US shown in the figure to the left shows an expected dip in the beginning of 2020, just as Covid began. A period of uncertainty persisted into late 2020 as cases continued to climb and airline companies tentatively began to allow more flights to resume. Around Q1 2020, a large spike is apparent due to a major Covid surge after the holiday season and a dip ensues as vaccines begin to become available to the public. Another wave begins between Q3 and Q4 2021 and grows into the Delta and eventually the Omicron waves. Although the Covid data depicts significant peaks and dips, the airline data portrays a single dip at the beginning of the pandemic and a steady recovery afterwards. As a result, the data shows that the airline data



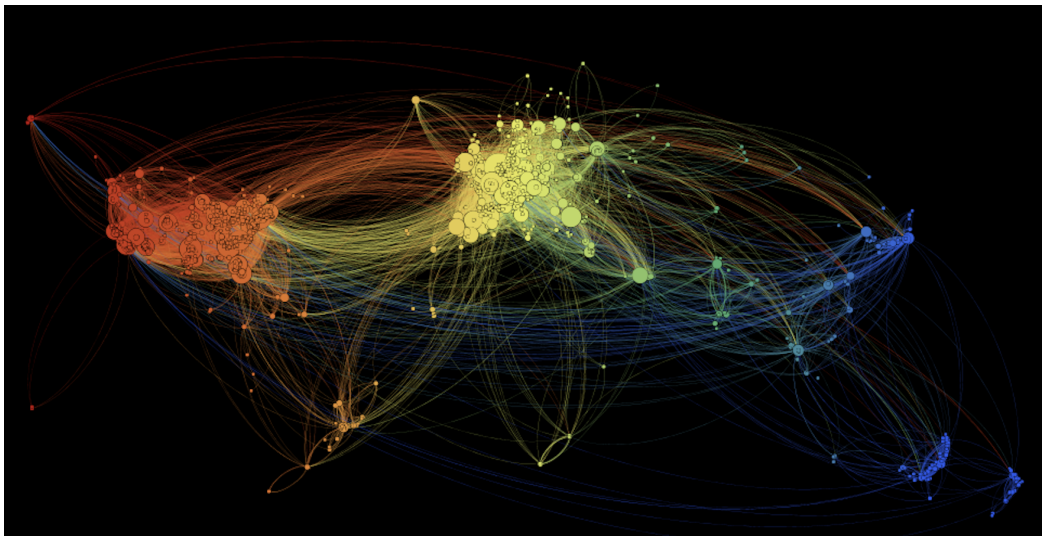
is mostly independent of the Covid data as additional factors such as vaccines, the holiday season, and increased testing availability in parallel with federal/global guidelines affected the trend.

6. Flight trend heat map representation with time slider animation



Graph above represents the number of departing flights from the majority of US airports. As of now, the data that is shown only covers Dec 12, 2021, and the time sliding interval has been set to days. Once all the data has been processed, the animation will cover the span of the entire pandemic with a monthly sliding window which will help monitor a whole picture of flight trends after the outbreak of COVID-19.

7. Flight trajectory network representation using Gephi



This directed node-link network exhibits worldwide flight trajectories on Dec 26, 2021. Nodes represent source and destination airports, edges represent a flight trajectory from source airport to destination. Node size indicates its degree of centrality, which evaluates the number of flights

from the same source airport. A gradient color palette was chosen to help distinguish the geo-location of different continental and countries. We can see that most flights are concentrated within North America (specifically the US) and Europe, and in between the two continents.

Data Exploration, Cleaning, Wrangling, and Engineering

Data Exploration Summary

As we have had a chance to utilize more of the files in our datasets,we have found that files from different data sources mostly share the same data qualities. For example, after loading in several more files of flight data, we found that callsigns, latitudes and longitudes of origin airports consistently hold non-NaN values, and latitude and longitudes of destination airports have fewer than 100 empty values. This means that the origin and destination airport names will be the main focus of data imputing. The Covid dataset is much cleaner since it is consistently maintained.

Data Preprocessing

Open Sky Network Flight Data

The Data Preprocessing method for the flight data was broken into two main steps: merging the Flight and Covid data, and imputing as many missing values as possible. We have used a single file of flight data to create and test the data preprocessing method and expect this to apply to the other files with little to no additional adjustments. Using the Bansard Airline to Country Mapping data, we added two more columns - Airline Name and Country - to the Flight dataset with the help of the ICAO Designator.

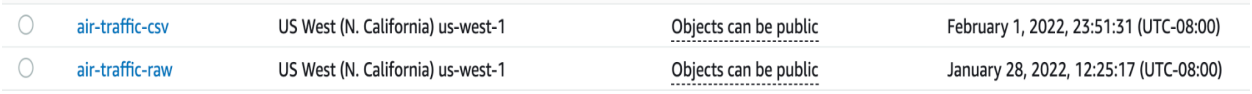
Once completed, we created a list of distinct airport names and their GPS coordinates, which initially had very slightly different readings due to the nature of Open Sky Network's data collection process. To resolve this, we looked at one of two methods - order the entries in alphabetical order by airport name and take the first entry for each airport, or take the average of each coordinate reading for a single airport. The first idea, while easier to perform, may end up including dirty data and the second, while more computationally expensive, would allow for better accuracy when creating visualizations and models.

With the cleaned list of airports and their coordinates, we filled in more than half of the missing origin and destination airports. Since the records in the flight data are not dependent on each other when filling in the data, we plan to employ parallel processing to make the preprocessing more efficient.

Johns Hopkins Covid-19 Data

The Johns Hopkins Covid dataset contains a transposed format of data, with countries/regions in the first column and dates beginning on 01/22/2020 as columns holding cumulative counts of confirmed cases, recoveries, and deaths. In order to be stored in the InfluxDB database for easier querying, the time series data will need to be transposed once more so that the date columns will become rows with NULLs replaced with zeros. For EDA purposes however, the file was loaded into a pandas dataframe with NaNs replaced with zeros. We expect any preprocessing of the Covid data to be smooth due to the consistency and regular maintenance by Johns Hopkins.

Storing Processed And/Or Integrated Data



- As shown above, two buckets on S3 were created to store data, 'air-traffic-raw' stores raw zipped files, 'air-traffic-csv' stores automated unzipped csv files.
- S3 lambda function to unzip files is not complete yet
- The Influxdb database has been set up on an AWS EC2 instance, to ensure data transfer speed 'gp3' was chosen and 'IOPS' was set to '3000'. This instance will serve as the main data entry point when the modeling stage starts.
- Majority of the processed data for EDA were kept locally

Original Dataset	**Processed Dataset Description**
OpenSky_ Flight_Data	1. Airline Name and Country Columns added 2. Origin and Destination airport names imputed 3. Timestamped airports with respective counted number of flights each day along with lat/long in Dec 2021 4. Nodes and edges dataframe with number of flights from the same source airports as weights on Dec 26, 2021

Processed Dataset Name	**Input Datasets**	**Link to Processing Scripts and Notebooks**	**Provisional Data Size**
OpenSky_ Flight_Data_Final	OpenSky_ Flight_Data	Notebook/Script: Flight_Preproc.ipynb (new) Folium_timesliding_heatmap.ipynb (new) network_process_gephi.ipynb(new) nov_21_top_airline.ipynb	10GB

JH_Global_ Confirmed_Final JH_Global_ Recovered_Final JH_Global_ Deaths_Final	JH_Global_ Confirmed, JH_Global_ Recovered, JH_Global_ Deaths	Notebook/Script: confirmed_covid_us.ipynb recovered_us.ipynb death_us.ipynb	2MB Each
Airline_Code_to_ Country_Final	Airline_Code_to_ Country	N/A	1MB

Feature Engineering and Data Modeling

- Folium heatmap with sliding time window

A list of airports was extracted and reference data was used for the inner join to map exact airports' latitude and longitude. Then 'groupby' was used to count the number of departing flights from each airport per day. Also, Timestamps were parsed as dataframe indexes. Finally, dataframe was processed to points and indices for rendering folium map

- Node-link directed Worldwide network

First, reference data was used to map exact airports' latitude and longitude for all the airports worldwide. Flight on Dec 26, 2021 was extracted. Airports nodes were generated. Edges were generated from source airports to destinations. The number of flights from the same origin to the same destination were counted as edges' weight.

Link to Input Datasets	**Link to Feature Engineering Scripts and Notebooks**	**Provisional Data Size**
OpenSky_Flight_Data	Notebook/Script: Flight_Preproc.ipynb(new) Folium_timesliding_heatmap.ipynb (new) network_process_gephi.ipynb(new) nov_21_top_airline.ipynb	10GB
JH_Global_Confirmed JH_Global_Recovered JH_Global_Deaths	Notebook/Script: confirmed_covid_us.ipynb recovered_us.ipynb death_us.ipynb	2MB Each
Airline_Code_to_ Country	N/A	1MB

Data Access Design

Initial design of data querying interface: data is processed and loaded into influxDB for access and querying. To query the data, the python API(influxdb-client-python) will be used to connect influxDB and query data via influxQL or Flux language.
We programmatically access the data. Data is stored in both Amazon S3 and InfluxDB. For accessing the data, we use InfluxDB to query and graph in dashboards. InfluxDB allows us to quickly see the data that we have stored via the Data Explorer UI. We can also use templates or Flux (InfluxData's functional data scripting language designed for querying and analyzing), which can rapidly build dashboards with real-time visualizations and alerting capabilities across measurements.

Team Member Contributions

Bo

- Maintain our Capstone Project Timeline and Planning.
- Wrote notebooks 'flight_8.ipynb" and "covid19_confirmed_global.ipynb" to do the EDA.
- Practice on deep learning models such as NN using pytorch.
- Contributed to Key Findings Through EDA (1-4), Data Access Design, Next Steps/Plan of Action.

Yuan

- Created folium map and Gephi node-link network visualization for key findings 6, 7
- Wrote "Folium_timesliding_heatmap.ipynb","network_process_gephi.ipynb"
- Set up influxdb on AWS EC2
- Contributed to "Storing processed and/or integrated data", "Feature engineering and data modeling"

Adelle

- Coordinated/Scheduled Team meetings
- Wrote "Flight_Preproc.ipynb" and "US_Flight_Covid_EDA.ipynb" files for flight data preprocessing and EDA
- Contributed to Key Findings Through EDA (5), Data Exploration Summary, Data Preprocessing, Updates to Steps 1 and 2, and Next Steps/Plan of Action sections

Updates to Steps 1 and 2

Since our first two reports, we have finalized our S3 buckets and stood up our influxDB database and plan to run our full datasets through the preprocessing steps to store the cleaned data in a single location. A more concrete data preprocessing methodology has been developed

and we now have a better understanding of our datasets. Below are several steps we plan to take in order to begin creating our minimum viable model product.

Next Steps/Plan of Action

Hypothesis Testing

Initial statistical inference will be conducted with hypothesis testing, such as comparing the impact of Covid on major airline companies, major airports and different countries.

Preliminary Models

- A spatio-temporal forecasting problem will be defined
- We plan to discover 4 deep learning methods (MLPs, CNNs, LSTMs, and Hybrids of these methods) and 3 baseline methods (ARIMA, SARIMA, and exponential smoothing forecasting methods).

Measures of Success

We need to consider which measures and benchmarks to rely on to conclude that our model is accurate, robust, and efficient.

Performance Metrics

For model evaluation metrics, we plan to use Confusion Matrix and F1 Score to evaluate our models. For system evaluation metrics, we plan to use the User Satisfaction / Apdex Scores to evaluate our application performance.