**Data Acquisition:**

**Acquiring Chinese Data:**

The idea of the project is to acquire the Chinese travel data and analyse the destination countries for covid cases. We tried to get the data from different sources(websites) e.g. Flight aware [1]*,* Aviation Stack [2] but it turned out that Chinese flight data is not available on any of the sources. The reason behind this was that the Chinese Government doesn’t allow ADS-B method. The ADS-B method is the method which is widely use to track flight data by the websites. All the sources (including paid sources) we looked for the data were either taking too long to response or doesn’t have the Chinese flights data.

Due to all these complications, we decided to move on instead of spending more time on Chinese flights data, we started focusing on the second variant i.e. Omicron variant.

**Acquiring Data of Omicron Variant:**

According to WHO, the [SARS-CoV-2 Omicron variant](https://en.wikipedia.org/wiki/SARS-CoV-2_Omicron_variant) (B.1.1.529), first reported from South Africa in November 2021. As South Africa and United Kingdom were the first countries to witness the first omicron cases, we decided to look for the data for these two countries.

Unlike Chinese flights, the data for South Africa and United Kingdom was available in abundance. We were able to acquire data from Zenodo [3] which were providing data from the Open sky network [4].

**Data Preparation**

After acquiring the data, the data is available from the start of 2019 till march 2022 but as the first Omicron case was hit on 27th November 2021, we decided to filter our data from November 2021 till end of January 2022. The data is available in many CSV format files which were splitted periodically, so we transformed those files into one csv file, applied data cleaning techniques by removing null values and irrelevant fields. Then we again filtered this data into two different CSVs, one for United Kingdom and the other for South Africa.

From those Unique Destinations we have provided the file to Data Analysis Team to Identify countries for Analysis

**Identifying Countries:**

After preparing the first phase of our data we have selected destination countries having high, moderate and low incoming traffic from South Africa and United Kingdom each. These countries from South Africa are Qatar, Netherlands, Great Britain, Germany, France having high amount of traffic, Israel, Morocco, Spain, Algeria, Guatemala having moderate amount of traffic and Iceland, Barbados, Norway, Poland, Bulgaria having low amount of traffic. Similarly, the destination countries from United Kingdom are Germany, Spain, Ireland, France, Italy having high amount of traffic, Barbados, Malta, Latvia, Luxembourg, Singapore having moderate amount of traffic and Malaysia, Kazakhstan, Philippines, Vietnam, Kenya having low amount of traffic. So we filtered our data again with only the above mentioned destination countries.

**Population Density:**

After identifying the destination countries and filtering the data according to that, we need more attributes in the data which can help to achieve our goal. One important factor in the covid transmission is population density, so we include a column of population density of the destination countries. We downloaded another dataset I.e., of population density of the countries [5] and mapped the population density to destination countries in the flights data by countries names.

**Finding the Seating Capacity:**

After adding the population density, another important attribute is to know how many passengers are traveling in the flight. But the challenge is that, no source is providing the information about the number of passengers travelling in a flight. Also, the dataset does not have the information about the seating capacity of a flight. So we downloaded another dataset which consists of aircraft model and its seating capacity. We use this dataset and map the seating capacity of a plane to the flight’s dataset by aircraft model.

After mapping the the seating capacity of the flights, we still have the problem of number of passengers travelling on a flight, as this type of the data is not available. So, according to the travel restrictions and CNN travel article [6] we considered that about 60 percent people were travelling in the flights according to the flight capacity.

References:

1. [www.flightaware.com](http://www.flightaware.com)

2. [www.aviationstack.com](http://www.aviationstack.com)

3. <https://zenodo.org/record/6411336>

4. <https://opensky-network.org>

5. <https://data.worldbank.org/indicator/EN.POP.DNST>

6. <https://edition.cnn.com/travel/article/flight-capacity-united-states-coronavirus/index.html>