Small Components' Analysis and Flight Delay Prediction

Team 22

Introduction

In a more and more connected world, flight routes gave us a great framework in order to detect countries/cities that are socially isolated from the rest of the world. We will use Open Flight database which contains more than 67000 routes and includes data about airports and airlines. Since we have already explored much of the network related with this data set during the first assignments we decided to focus on the components that were not seen.

Table Of Content

Table Of Content

Table Of Content

Table Of Content

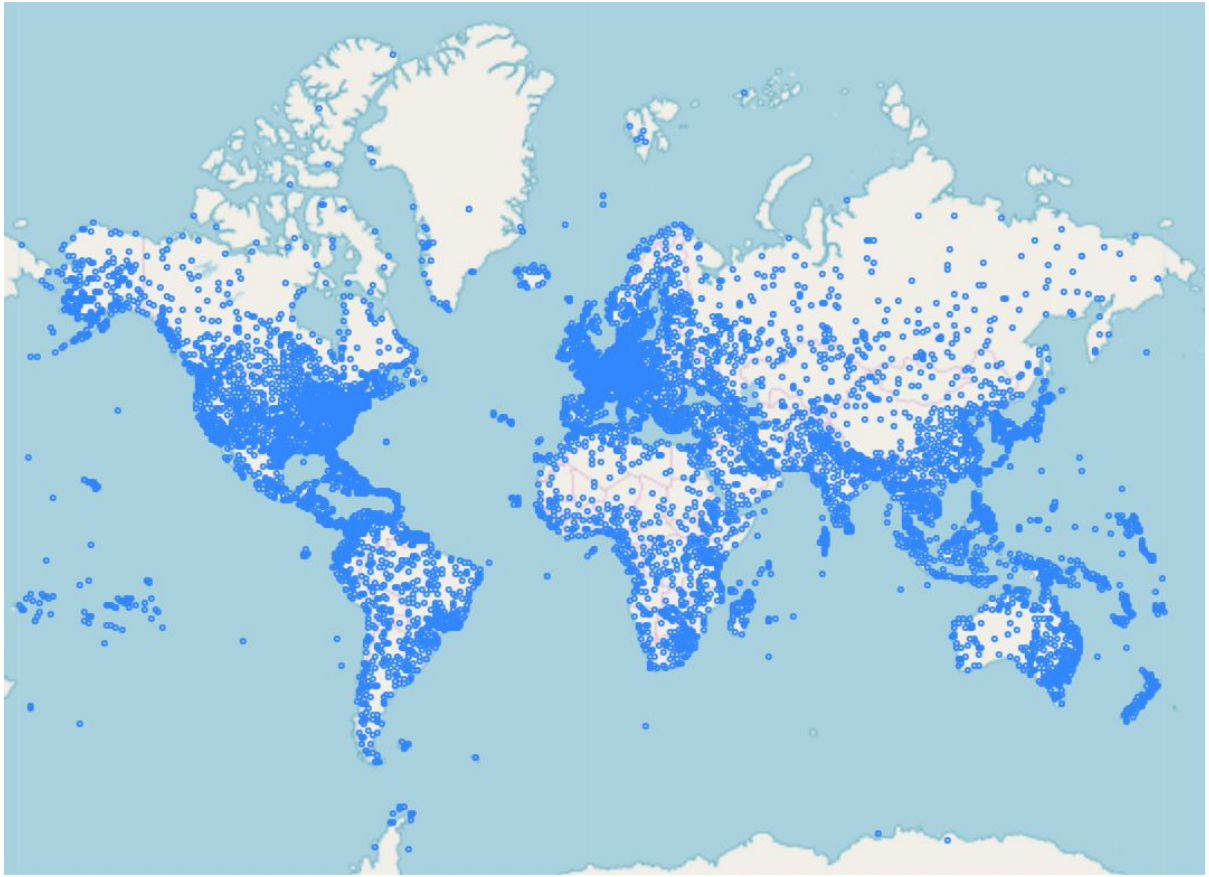
Table Of Content

Data Analysis

Airport locations

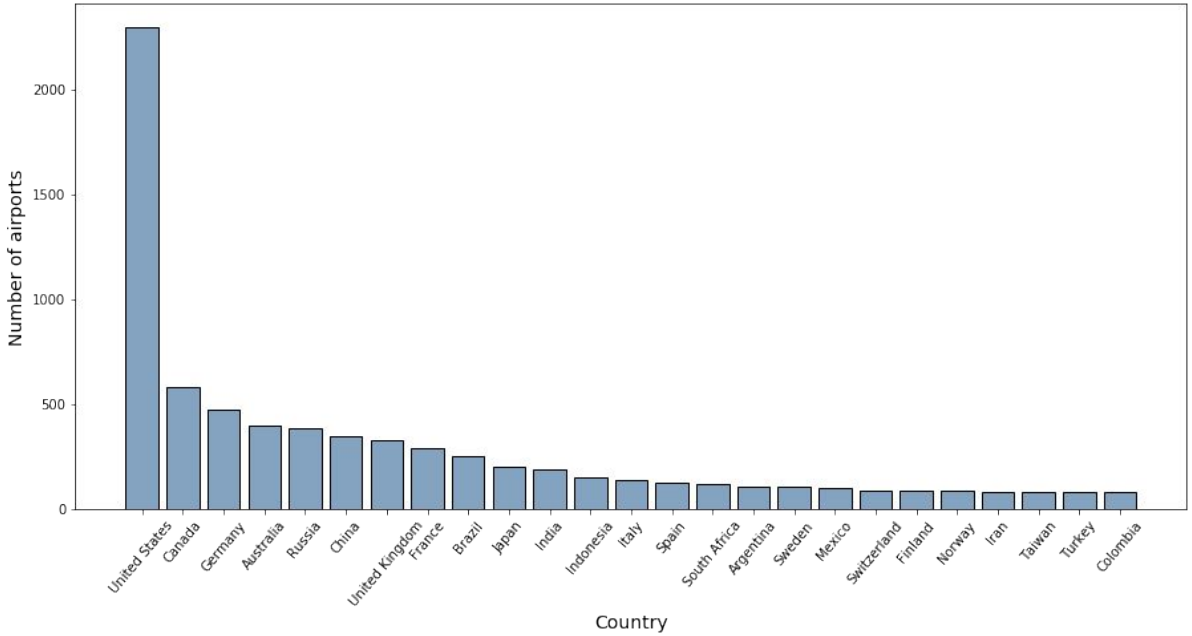
We start off by visualizing all the airports across the countries.

We can see that first world countries are the ones who have the biggest number of airports. While bigger countries like Russia and Algeria have a lower number per area due to their large surface.



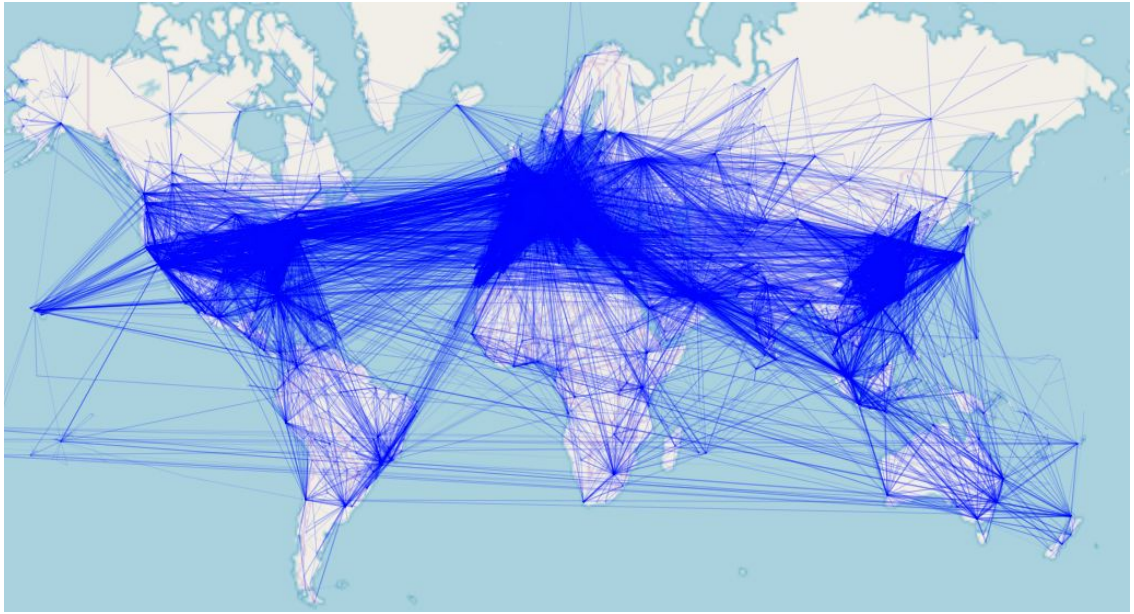
Countries with the most airports

After displaying each node and edge, a question one might ask is "Which are the countries that have the most airports?". As one might have guessed, we can see that the United States are leading the whole world in terms of number of airports.



Airport Connections

Next, we take a look at how the airports are connected. Here we see all the edges of our graph on a map. We can notice that countries close to each other geographically have the most connections and that powerful countries are the most connected ones to the rest of the world.



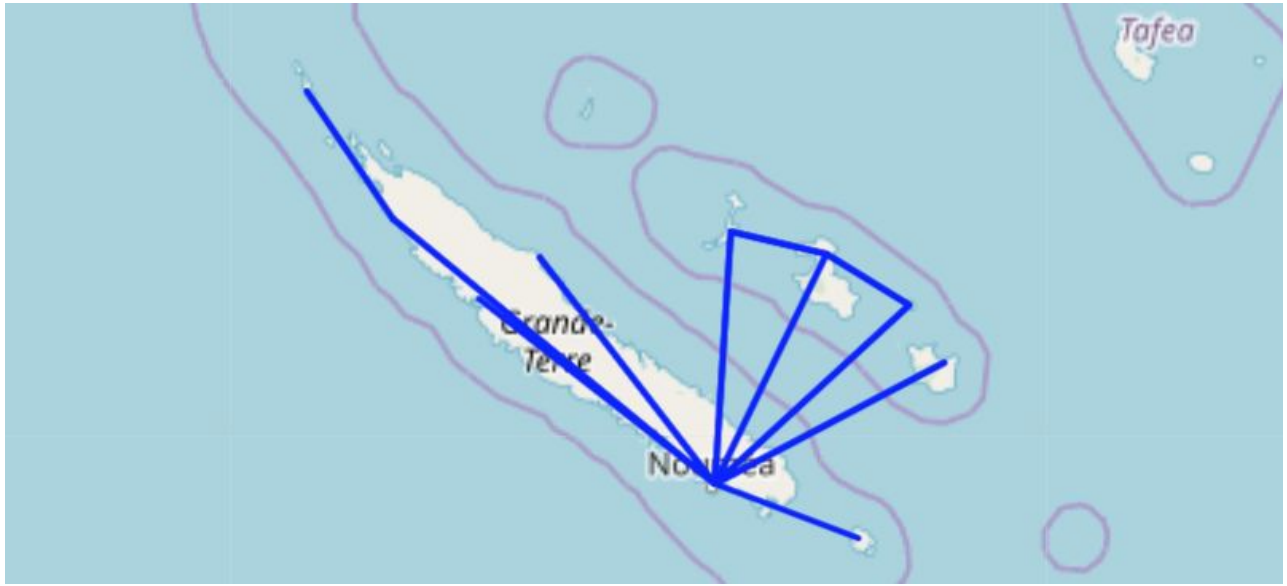
Smaller components analysis

Some Examples Of Isolated Flights Routes

British Virgin Islands:

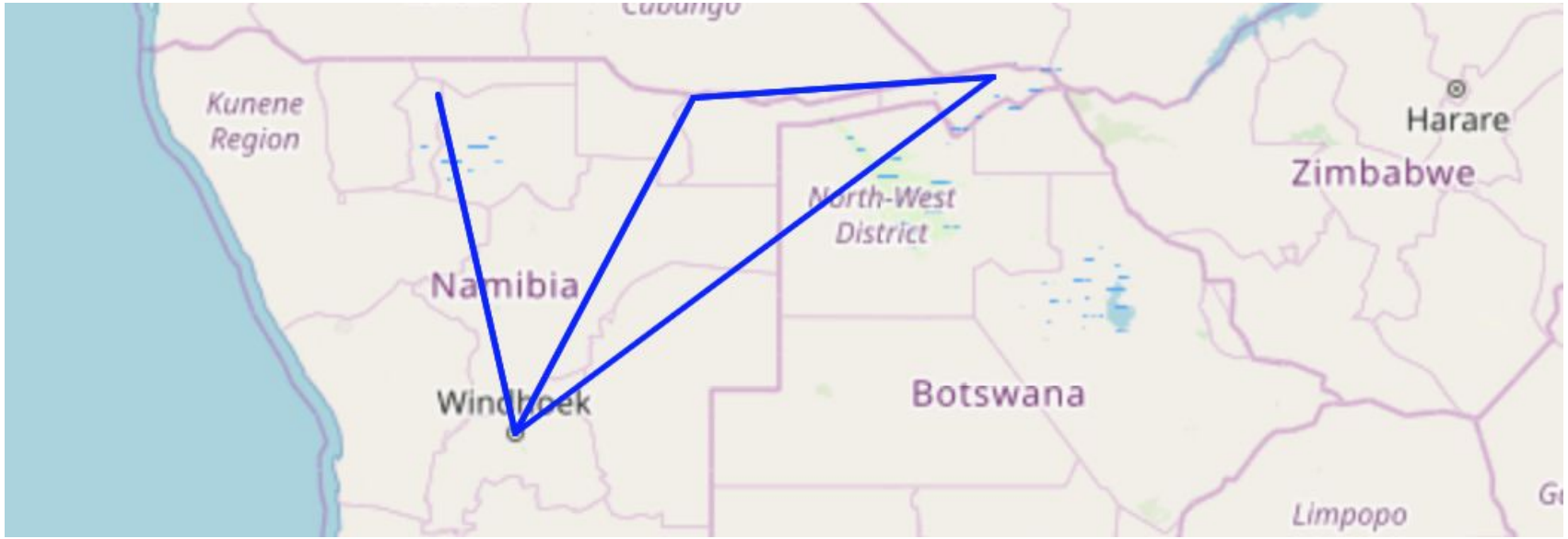
Some Examples Of Isolated Flights Routes

New Caledonia:



Some Examples Of Isolated Flights Routes

Namibia:



Some Examples Of Isolated Flights Routes

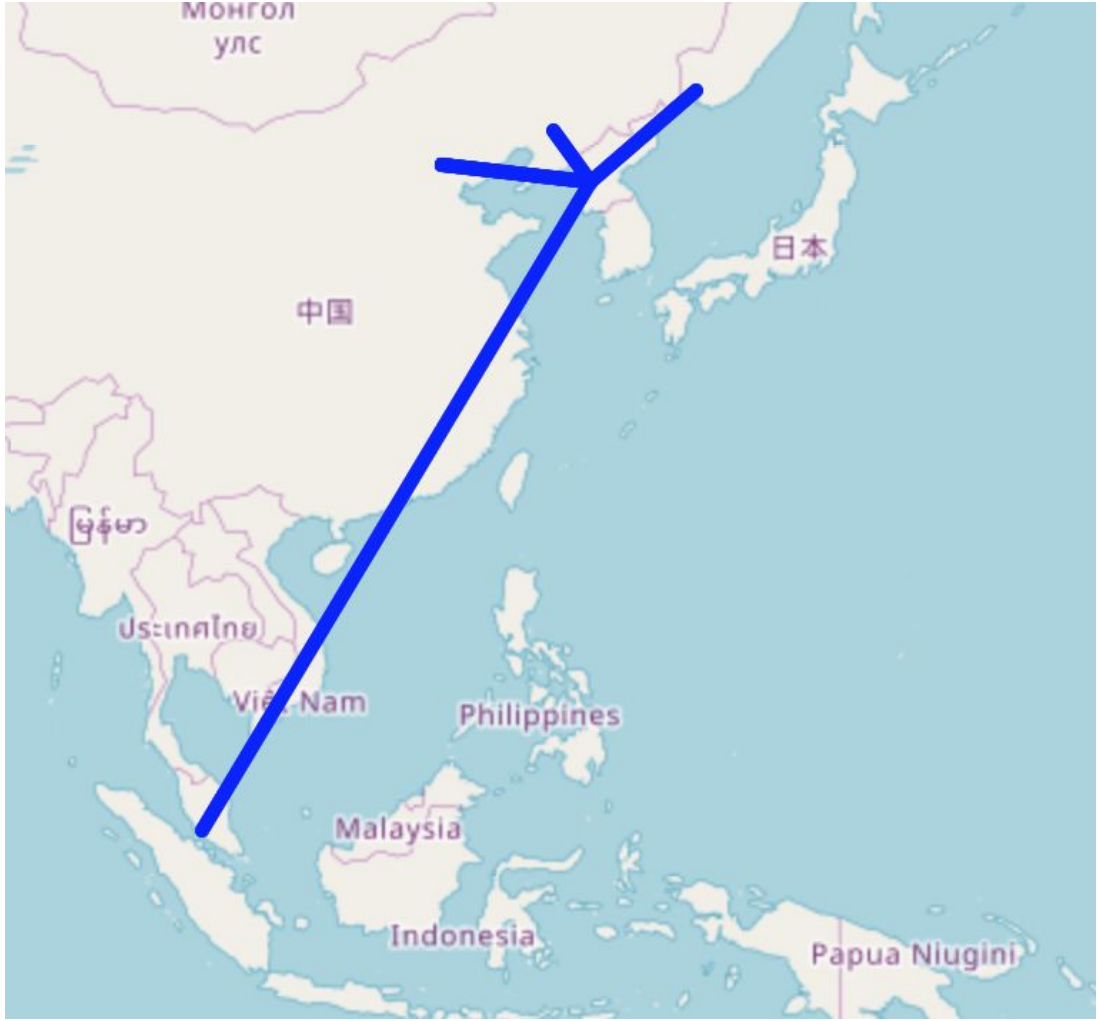
Nevada - Arizona :

An Example Of an Anti-Social Country

North Korea

1st Level Connections

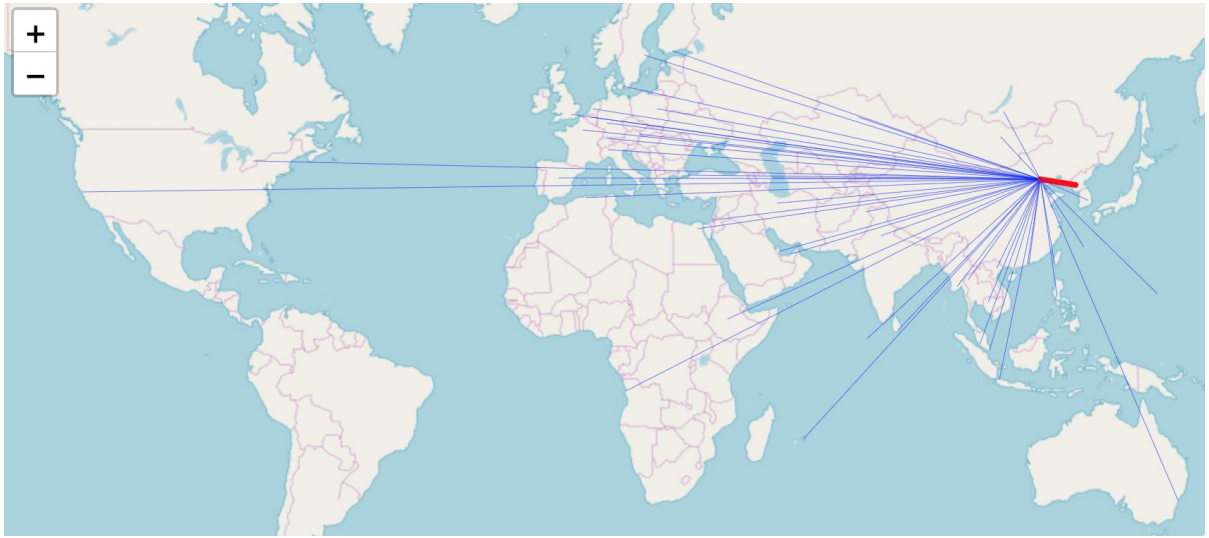
We can see from the figure that it is only connected to 3 countries :'Malaysia', 'China', 'Russia' through only one airport in each country except China where the number is 2.



Aircrafts going in and out of north Korea

2nd Level Connections

Here we can see that we can reach North Korea with at least one layover going through Pekin airport which is the only one that offers commercial flights.



Predicting flight delays

Data exploration :

The Features :

1. AIRLINE
2. ORIGIN AIRPORT
3. DESTINATION AIRPORT
4. SCHEDULED DEPARTURE
5. DEPARTURE TIME
6. DEPARTURE DELAY
7. SCHEDULED ARRIVAL

The Target :

1. DELAY LEVEL

ML Algorithms :

1. Gradient Boosted Trees :

   Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees.

2. Random Forest :

   Random forests creates decision trees on randomly selected data samples, gets prediction from each tree and selects the best solution by means of voting.

3. K Nearest Neighbors :

   In KNN, an object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors.

4. Logistic Regression

   Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis.

Prediction Results :

Algorithm	Hyper-Parameter value	Accuracy
Gradient Boosted Trees	17	87.021%
Random Forest Classifier	100	86.02%
K Nearest Neighbors	29	86.97%
Logistic Regression	0	87.02%

Conclusion

THANK YOU FOR YOUR ATTENTION