



Pacific Airfields Data Analysis

***Developing Airfield Functional Categories
Using K-Means Clustering
Segment 2 Deliverable
24 Sep 2022***

Christopher Rice

Topic Selection

I consult with aerospace companies, investors, and other interested parties who need to understand business opportunities for new types of aircraft in Pacific Ocean regions.

This project will automate collection of data on airfields, build an airfield database, and employ an unsupervised learning model to classify airfields based on the length/width of their runway(s), surface material of their runway(s), number of radio frequencies and navigation aids, and number of airlines providing scheduled service.

Compiling this data will create a useful reference dataset that can be uploaded and browsed in Google Earth. Automating creation of this dataset will save large amounts of work required for manual data collection and analysis. The clustering analysis will provide a classification schema for use in detailed transportation studies that inform decisions involving, aircraft basing, aircraft routing, and investments in upgrading existing airfields and/or developing new facilities.

Data Sources

The data required for this study comes from two principal sources: Airportdatabase.net and Wikipedia. Google Earth can also be used to visually inspect for missing variables.

Airportdatabase.net source allows us to identify all of the airfields in a selected country or region and to collect airfield data including:

- Airfield name, GPS and IATA codes
- Latitude, longitude
- Elevation
- Runway length, width
- Number and frequencies of radio stations and navigation aids
- Names of airlines providing scheduled service
- URL for the airfields Wikipedia page

Wikipedia allows us to collect more detailed information including:

- Runway surface type

Google Earth allows us to visually assess runway surface and mensurate runway dimensions when no data is available from other sources.

Data Exploration Phase

An earlier, less complete version of this dataset compiled from Wikipedia and Google Earth is available for islands in the Northern Pacific Ocean: This dataset is stored in a .csv file and was used to test the preliminary machine learning model.

I originally planned to automate web scraping the HTML tables from airportdatabase.net, but eventually settled on manually collecting the URLs for each airfield and storing them in a csv.

The runway surface data was collected manually from Wikipedia and Google Earth and entered in the same csv with the URLs for airportdatabase.net.



Data Exploration Phase

The airport table is simply reference data; it has no use in the machine learning model.

The runway length/width are integer variables. For airfields with more than one runway, the length/width of their runways are summed.

Over a dozen runway surface types are reported in Wikipedia. These are aggregated into four categories and assigned a numeric value as follows:

- 3 – Paved (asphalt, concrete)
- 2 – Smooth hard surface (macadam, coral, chipseal)
- 1 – Rough hard surface (gravel)
- 0 – Unhardened surface (grass, turf, dirt)

Public and commercial airfields generally have one or more radio stations and navigation beacons; a count of the number of radio frequencies and navigation aids is used to create an integer variable.

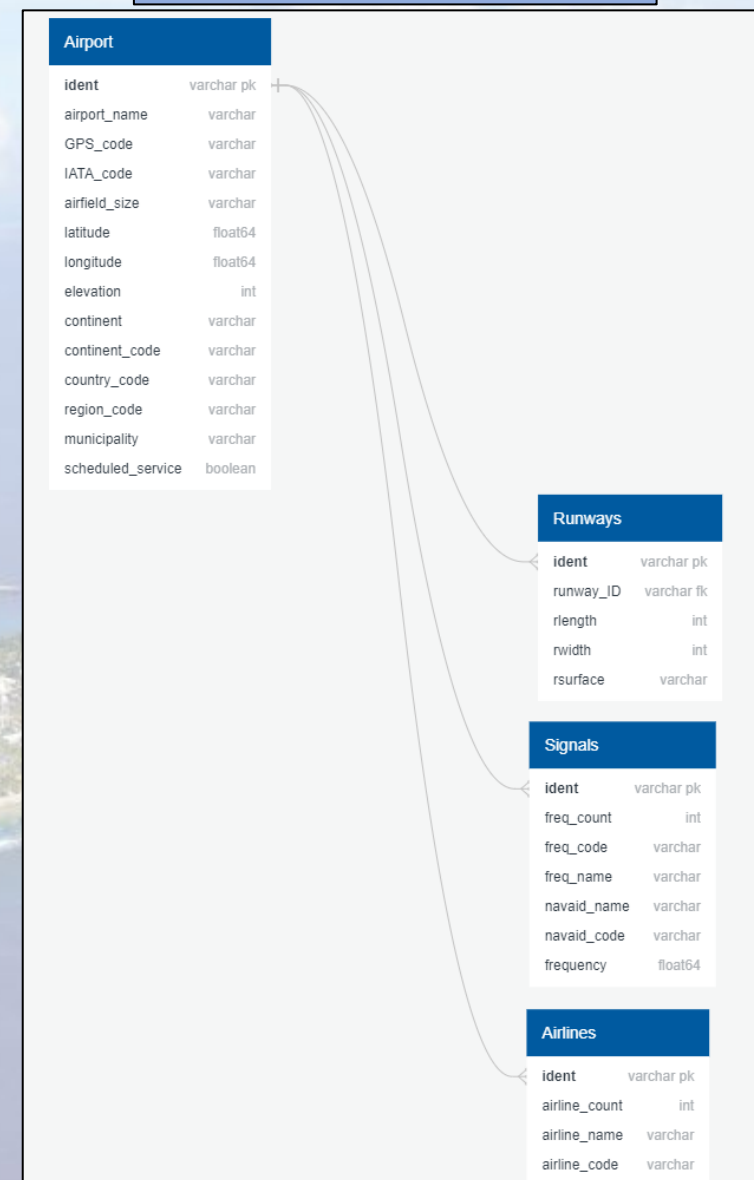
The name and nationality of each commercial airline providing scheduled service to an airfield are reported on airportdatabase.net. A count of the number of airlines is used to create an integer variable.

Raw Data -> SQL Database

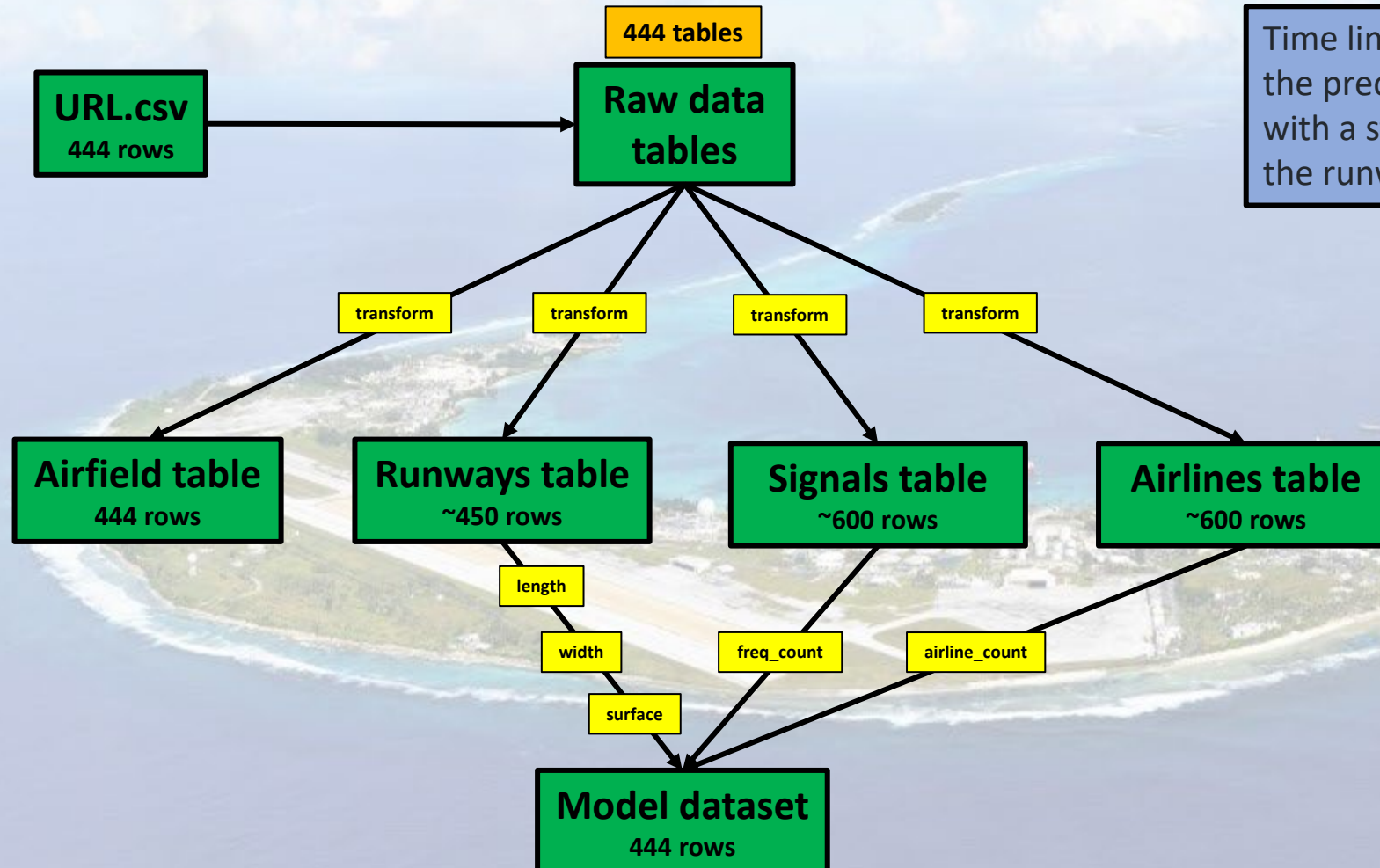
HTML table from Airportdatabase.net

Unnamed: 0		Unnamed: 1
0	Honiara International Airport details and info...	Honiara International Airport details and info...
1	ident:	AGGH
2	type:	medium airport
3	NaN	NaN
4	latitude:	-9.42800045013428
5	longitude:	160.05499267578125
6	elevation:	28 ft.
7	Airport data	continent: Oceania (OC)
8	iso country:	Solomon Islands (SB)
9	ISO Region:	Capital Territory (Honiara) (SB-CT)
10	Municipality:	Honiara
11	Scheduled Service:	yes
12	GPS Code:	AGGH
13	IATA Code:	HIR
14	wikipedia link:	Honiara International Airport in Wikipedia
15	Runway data	Honiara International Airport runways
16	06/24	7218x148 ft.
17	Honiara International Airport frequencies	Honiara International Airport frequencies
18	AFIS (INFO)	118.1 Mhz
19	Signals data	INFO (INFO) 342.5 Mhz
20	Honiara International Airport navaids	Honiara International Airport navaids
21	Honiara (HN)	112600 Mhz
22	Honiara (HN)	348 Mhz
23	Airline data	Airlines flying from/to Honiara International ...
24	Pacific Blue (DJ) , Polynesian Blue (DJ) , Vir...	Pacific Blue (DJ) , Polynesian Blue (DJ) , Vir...

ERD Diagram for SQL database

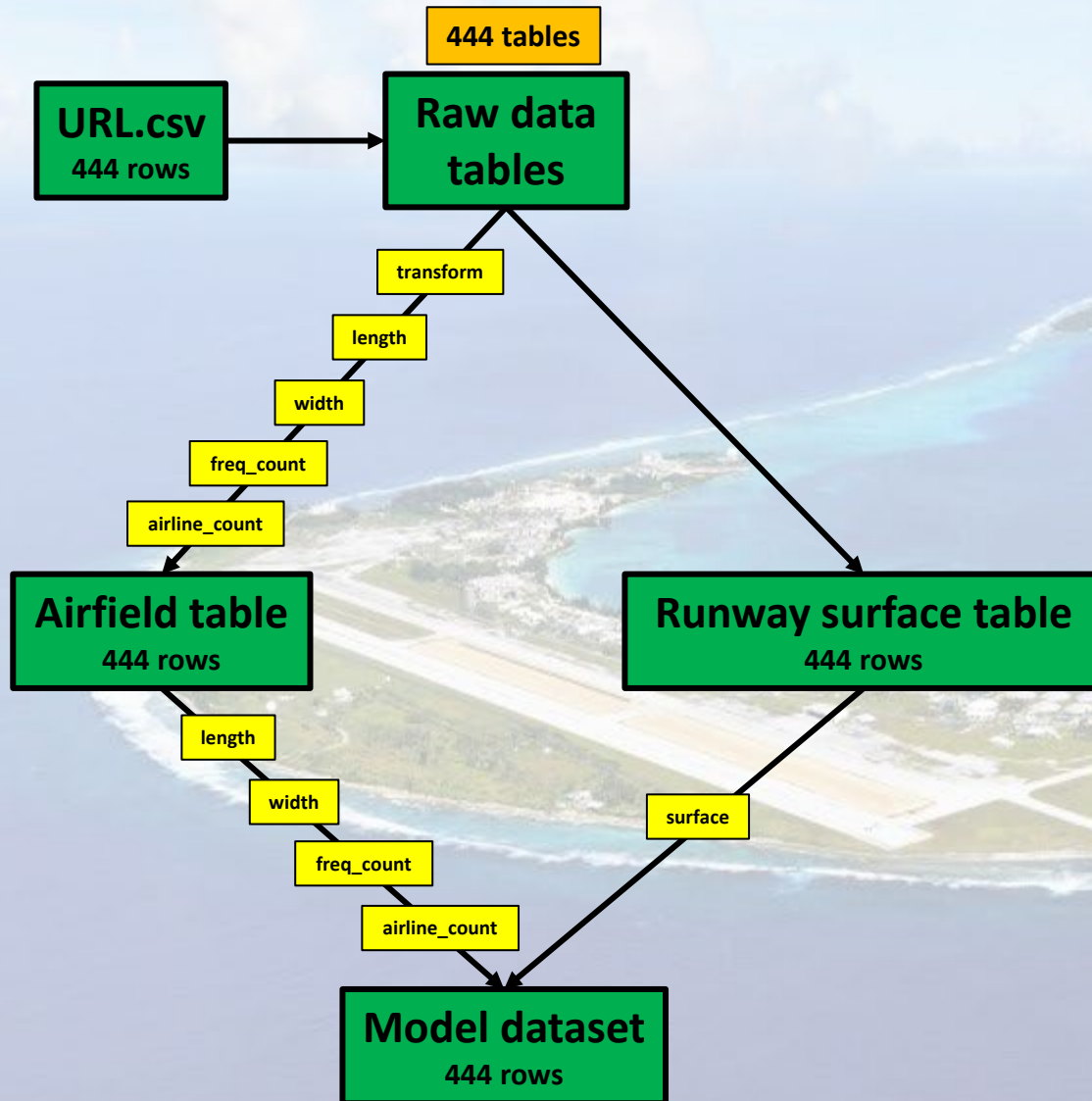


ETL Process



Time limits prevented me from implementing the preceding table structure; I had to go with a streamlined ETL process that ignored the runways, signals, and airlines tables.

Data Processing



Instead, I created four of the model variables (runway width, runway length, signals count, and airline count) during the transform process and stored them in the airfield table.

The runway surface variable was stored in a separate SQL table, then joined with the other model variables to create the input dataset for the model.

SQL Database

SQL connection information

PostgreSQL 12

General Connection SSL SSH Tunnel Advanced

Host name/addresses: localhost

Port: 5432

Maintenance database: postgres

Username: postgres

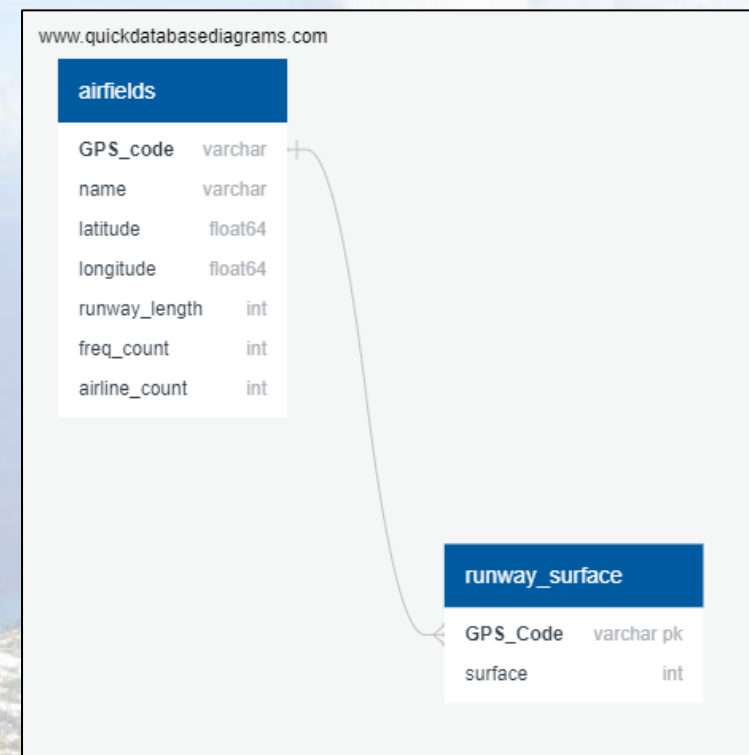
Kerberos authentication? ☐

Role:

Service:

Close Reset Save

ERD



The “Airfields” SQL database contains two tables:

“airfields” contains the data scraped from airportdatabase.net

“runway_surface” contains the data collected manually from Wikipedia and Google Earth.

Required database functionality is documented in the *.sql files and *.ipynb model files.

Feature Engineering

The available data from airportdatabase.net breaks airfields into three categories; small, medium, and large. These categories are not sufficient to distinguish between the functional categories of airfields that are important for operational and strategic analysis.

To support this analysis requires segregating airfields into at least six groups based on their operational attributes. To achieve this, an unsupervised machine learning model - K-means clustering - will be used to group the airfields into their functional categories.

Runway length/width are physical measurements that directly correlate with the operational capacity of an airfield.

Runway surface type is also correlated with operational capacity but is much less precisely measured. I chose to aggregate the numerous surface types into a numeric categorical variable.

The number of radio frequencies and navigation aids used by the airfield is also a direct measure of its capacity to handle air traffic.

Ideally, the final variable(s) would measure the airfield's refueling, servicing, repair, passenger and cargo handling capacities. This data is not freely available for most airfields and is not consistent when it is collected. Therefore, the number of airlines serving each airfield is used as a proxy variable for ground facilities.

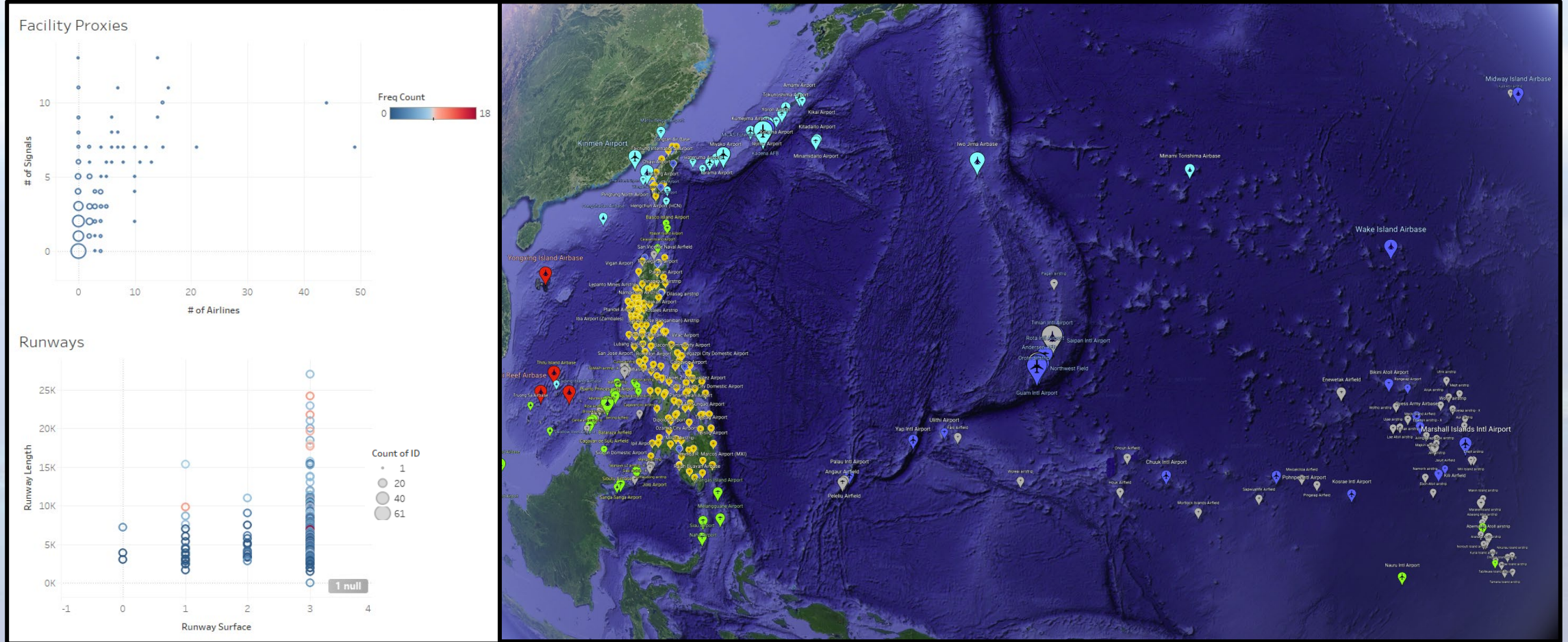
Model Selection

Initial modeling experiments explored using the provisional dataset as training data for a neural network, then for a multinomial logistic regression. After further analysis -- particularly the implications of finding additional data fields for radio stations, navigation aids, and airline service in airplannedatabase.net -- both of these approaches were discarded. It was decided that the best use of this data was to group airfields into functional categories with an unsupervised learning model.

The available data from airportdatabase.net breaks airfields into three categories; small, medium, and large. These categories are not sufficient to distinguish between the functional categories of airfields that are important for operational and strategic analysis. To support this analysis requires segregating airfields into at least six groups based on their operational attributes. To achieve this, an unsupervised machine learning model - K-means clustering - will be used to group the airfields into their functional categories.

This clustering approach accomplishes exactly what is required; creating more finely-grained categories of airfields that can be mapped and analyzed as part of market research and mobility studies. The principal limitation is that these categories are not direct measurements of operational capacity and cannot be linked to any explanatory variables in a quantified fashion.

Tableau Dashboard



The interactive elements for this dashboard will include:

- 1) The map will allow filtering the displayed airfields based on any of their model attributes (runway length, runway surface, # of signals, # of airlines, airfield class) as well as countries.
- 2) The analytic graphs will allow filtering by countries and airfield class.

The dashboard for this project will consist of two main sections:

- 1) A geographic view that displays airfields by location. This view will display airfield icons with their sizes determined by the output of the machine learning model.
- 2) An analytic view comprised of 2-4 graphs that show key empirical relationships. For example, the “Runways” graph shows runway surfaces vs runway sizes, while the “Facility Proxies” graph shows the # of signal stations vs # of airlines at each airfield (the size of the bubbles represents the number of airlines at each ordered pair).