CS777 – Term Project Proposal Submission Template

Student: First_name Last_name

1. Data set description: Provide a detailed description of the public data set you have selected, including its source, format, and any relevant details about the data.

Airport database

As of January 2017, the OpenFlights Airports Database contains **over 10,000** airports, train stations and ferry terminals spanning the globe, as shown in the map above. Each entry contains the following information:

Airport ID	Unique OpenFlights identifier for this airport.
Name	Name of airport. May or may not contain the **City** name.
City	Main city served by airport. May be spelled differently from **Name**.
Country	Country or territory where airport is located. See Countries to cross-reference to ISO 3166-1 codes.
IATA	3-letter IATA code. Null if not assigned/unknown.
ICAO	4-letter ICAO code. Null if not assigned.
Latitude	Decimal degrees, usually to six significant digits. Negative is South, positive is North.
Longitude	Decimal degrees, usually to six significant digits. Negative is West, positive is East.
Altitude	In feet.
Timezone	Hours offset from UTC. Fractional hours are expressed as decimals, eg. India is 5.5.
DST	Daylight savings time. One of E (Europe), A (US/Canada), S (South America), O (Australia), Z (New Zealand), N (None) or U (Unknown). *See also: Help: Time*
Tz database time zone	Timezone in "tz" (Olson) format, eg. "America/Los_Angeles".
Type	Type of the airport. Value "airport" for air terminals, "station" for train stations, "port" for ferry terminals and "unknown" if not known. *In airports.csv, only type=airport is included.*
Source	Source of this data. "OurAirports" for data sourced from OurAirports, "Legacy" for old data not matched to OurAirports (mostly DAFIF), "User" for unverified user contributions. *In airports.csv, only source=OurAirports is included.*

The data is UTF-8 encoded.

Note: Rules for daylight savings time change from year to year and from country to country. The current data is an approximation for 2009, built on a country level. Most airports in DST-less regions in countries that generally observe DST (eg. AL, HI in the USA, NT, QL in Australia, parts of Canada) are marked incorrectly.

Sample entries

```
507,"London Heathrow Airport","London","United Kingdom","LHR","EGLL",51.4706,-0.461941,83,0,"E","Europe/London","airport","OurAirports"
26,"Kugaaruk Airport","Pelly Bay","Canada","YBB","CYBB",68.534401,-89.808098,56,-7,"A","America/Edmonton","airport","OurAirports"
3127,"Pokhara Airport","Pokhara","Nepal","PKR","VNPK",28.200899124145508,83.98210144042969,2712,5.75,"N","Asia/Katmandu","airport","OurAirports"
8810,"Hamburg Hbf","Hamburg","Germany","ZMB",\N,53.552776,10.006683,30,1,"E","Europe/Berlin","station","User"
```

2. Research question: Clearly define your research question and explain why studying is important. What do you want to learn from the data?

Research question*: Can we cluster airports based on their time zones to understand global aviation patterns?*

why studying is important: *By conducting a thorough time zone analysis, stakeholders in the aviation industry can gain actionable insights that can be used for strategic planning and operational efficiency.*

What want to learn: *similar to research question, I want to gain some valuable insight into the aviation industry, such as understanding the operational hours and peak times for airports and geographical distribution of airports*

3. Machine Learning model: Specify the type of machine learning model you plan to use, such as classification or clustering, and explain why you have chosen this model.

Clustering model. I plan to use GMM with EM algorithm. For clustering model, GMM can offer a more flexible model and it can deal with overlapping data, like airports close to time zone boundaries.

4. Expected outcomes: What do you expect to achieve after implementing your learning model? What do you hope to learn or discover from your data analysis?

I can get the optimal number of clusters and probabilities that a given airport belongs to each cluster. Also some stats that allow me to gain insights into the characteristics that define each cluster.

From data analysis, I may find how airports cluster based on time zones can provide insights into the temporal dynamics affecting flight schedules and passenger flow or find the opportunities for better operational synchronization between airports in the same time zone guiding staffing better resource allocation. Moreover, I may hope to find some interconnection between certain clusters, which could be valuable information for route planning.

5. Evaluation plan: Explain how you plan to evaluate your project and assess the correctness of your model. What metrics or methods will you use to evaluate the effectiveness of your learning model? How well do you expect the model to work, and how will you measure its performance?

For model evaluating, I plan to use some stats to measure the model such as log likelihood, bic, aic and Silhouette Score .

For model expectation, I may expect the EM algorithm to converge, and specifically give me the prob of airports close to time zone boundaries to each cluster