BUDT758X - Data Processing and Analysis in Python

Spring 2018

Group 29 - Karthik Veeramalai

Project Proposal

Problem Statement

In an age where air travel is more common and accessible than ever, there appear to be a myriad of options for the consumer to choose from when he/she wants to book a ticket. But, an element that is often overlooked is the fuel efficiency of the trip. The goal of this project would analyze just that - what is the most fuel-efficient way to get from a source to a destination with the various options that different airlines offer? Although, airlines are a relatively small industry, they contribute disproportionately to the cause of climate change. U.S. airlines alone burned about 16.2 billion gallons of fuel during the twelve months between October 2013 and September 2014[1]. The fuel burn of an aircraft does not vary linearly with range but is instead a parabolic relationship that worsens with range because of the additional fuel payload. Each aircraft type has a specific range of distances at which flying it is most efficient. Airlines, more often than not use inefficient aircraft across their routes.

Developing a parameter for fuel economy would be great tool for educating consumers about healthier flying.

Approach

I'd be using the using the openflights[2] datasets in conjunction with other scraped data to perform this analysis. The openflights project contains three datasets - airlines, airports and routes. I'd be using all three. The airlines dataset contains trivial information like call-signs and the country of origin of the airline. The airports dataset has the list of all airports with their IATA codes and geographic locations (latitude and longitude). The routes dataset has a list of all the routes that these airlines operate to, the IATA and ICOA codes for the source and destination airports and the equipment that's being used to service the routes i.e. the aircraft. This dataset has conflicting entries that need to be cleaned. Data for calculating fuel economy would have to be scraped from a Wikipedia page [3].

Use of python

The datasets would be loaded into pandas dataframes for cleaning. The routes and airport datasets need to be cleaned extensively but can be done through regex. The fuel economy numbers would need to be scraped using beatifulsoup4. The fuel economy data that has been scraped needs to be further cleaned, again using regex. I would be using the geopy module for calculating the geodesic distance between routes. Calculation of fuel economy would be done greatly using numpy with matplotlib/vispy for visualization.

References

[1] Why airfare keeps rising despite lower oil prices, by Scott Mayerowitz, Assoc. Press Airlines Writer. Houston Chron., 17 November 2014.

[2] https://openflights.org/data.html

[3] https://en.wikipedia.org/wiki/Fuel_economy_in_aircraft