**Assignment 4:** Your task in this week’s assignment is to identify an interesting set of network data that is available on the web (either through web scraping or web APIs) that could be used for analyzing and comparing centrality measures across nodes. As an additional constraint, there should be at least one categorical variable available for each node (such as “Male” or “Female”; “Republican”, “Democrat,” or “Undecided”, etc.). In addition to identifying your data source, you should create a high level plan that describes how you would load the data for analysis, and describe a hypothetical outcome that could be predicted from comparing degree centrality across categorical groups.

For week 4 assignment, I will be using the OpenFlights/Airline Route Mapper Route Database contains **67663** routes between **3321** airports on **548** airlines spanning the globe, as shown in the map above. Each entry contains the following information:

|  |  |
| --- | --- |
| Airline | 2-letter (IATA) or 3-letter (ICAO) code of the airline. |
| Airline ID | Unique OpenFlights identifier for airline (see [Airline](https://openflights.org/data.html#airline)). |
| Source airport | 3-letter (IATA) or 4-letter (ICAO) code of the source airport. |
| Source airport ID | Unique OpenFlights identifier for source airport (see [Airport](https://openflights.org/data.html#airport)) |
| Destination airport | 3-letter (IATA) or 4-letter (ICAO) code of the destination airport. |
| Destination airport ID | Unique OpenFlights identifier for destination airport (see [Airport](https://openflights.org/data.html#airport)) |
| Codeshare | "Y" if this flight is a codeshare (that is, not operated by *Airline*, but another carrier), empty otherwise. |
| Stops | Number of stops on this flight ("0" for direct) |
| Equipment | 3-letter codes for plane type(s) generally used on this flight, separated by spaces |

Node type **Stops** contains 0 for direct flights or the number of stops. We will dummy code this variable, 1 for direct flights and 0 for non-direct flights. **This is our categorical variable**.

We will calculate degree centrality, eigenvector centrality, between-ness centrality, and possible closeness centrality.

Degree centrality counts the number of edges incident on a node (both in and out bound). It is a measure of the most popular flight types. Eigenvector centrality is a measure of how a node’s importance increases based on how it is connected to other important nodes.

Between-ness centrality is a count of the number of times a node acts as a bridge the shortest path between two nodes. It is a way to find a nodes importance based on its connectivity.