Kumar (KJ) Gandhi
NetID: KJG2

Flight Data Analysis

Methods

The files airports.dat, airlines.dat, and routes.dat are not SAS datafiles (needs MySQL to be imported) which means that those files will need delimiting (,) and converting the '\N' values from a CHAR value to 'NULL', so that it makes it easier for further analysis. I utilized proc freq and proc print pretty early on to validate the data by looking for missing values in important columns such as Airline Name or Airport Country, etc. and removing those from the dataset so I could prepare the tables for a merge for later.

To clean the '\N' values I used a loop with an array in the data step for all three datasets which just turned it into missing values. For data preparation, I used outer joins for all airline and route data by merging the airline IATA and route airline variables to create a table with all the routes that used a specific airline and cleaned this by removing all rows where there was no airline id.

The next step for data preparation was to create two separate tables for airport information to make the complete table more legible. I created a table for SourceAirports and DestinationAirport and used an outer right join with the newly created airline+routes table by merging the source/destination airport variable, originally from routes, with the source/destination airport IATA. This gives us a complete dataset with Routes, Airlines, and Source/Destination airports, which means we can do some analysis on it now.

The tasks were addressed by then using the complete table generated and breaking that table down into smaller tables with specific guidelines for example, if we wanted to know where flights flying into Chicago originate from, we create a table by using a where statement with DestinationAirportCity = 'Chicago'.

Results

So, the first thing I wanted to do was see what the most common airline was worldwide. This resulted in RyanAir as first, American Airlines in second, and United Airlines in third, which all makes sense because they are pretty popular airlines, especially in the United States (except for Ryanair).

The next few things I wanted to check out was to see which Airports were most popular in terms of flights leaving and flights flying in. One flaw in this is that the data is somewhat skewed because of the termination of rows that included missing values. I also ended up doing a count of # of airports by country.

# of Flights by Airline

Name	NameCount
Ryanair	2,483
American Airlines	2,347
United Airlines	2,176
Delta Air Lines	1,979
US Airways	1,954
China Southern Airlines	1,437
Air China	1,249
China Eastern Airlines	1,239
Southwest Airlines	1,143
United Feeder Service	1,130

Source Airport Top 10

SourceAirport	SourceAirportCount
Hartsfield Jackson Atlanta International Airport	938
Frankfurt am Main Airport	671
Charles de Gaulle International Airport	587
Chicago O'Hare International Airport	568
London Heathrow Airport	549
Beijing Capital International Airport	545
Singapore Changi Airport	529
London Gatwick Airport	523
Barcelona International Airport	515
Munich Airport	515

De

DestinationAirport	DestinationAirportCount
Hartsfield Jackson Atlanta International Airport	934
Frankfurt am Main Airport	664
Charles de Gaulle International Airport	578
Chicago O'Hare International Airport	559
London Heathrow Airport	546
Beijing Capital International Airport	544
Singapore Changi Airport	535
Los Angeles International Airport	516
Barcelona International Airport	515
London Gatwick Airport	515

# of Airports by Country

SourceAirportCountry	AirportCount
United States	1,503
Canada	428
Australia	328
Brazil	260
Russia	251
Germany	248
France	217
China	207
United Kingdom	165
India	145

I also did the same for the Chicago data, showing which airports are most popular, popular cities to fly to and # of airports by Olson Time zone in America, to show the division of airports by east to west splits. Most of the tables make sense, like flying into Atlanta makes sense for Chicago flights on a layover, as well as all of the other popular destinations to fly into. I also was not shocked by the # of flights by airline in Chicago, as United Airlines is also headquartered in the City and has a massive presence.

In addition to that, in a battle between O'Hare and Midway in terms of popularity, O'Hare takes the crown by 400 flights flying in/flying out more than Midway.

Popular Destinations to fly from Chicago

DestinationCity	DestinationCityCount
Atlanta	28
New Orleans	16
London	12
New York	12
Paris	11
Duesseldorf	9
Minneapolis	9
Los Angeles	8
Portland	7
Rome	7

# of Flights by Airline in Chicago

Name	NameCount
United Airlines	161
American Airlines	123
US Airways	118
Southwest Airlines	61
AirTran Airways	61
Delta Air Lines	20
Spirit Airlines	15
Air Canada	12
British Airways	9
Iberia Airlines	9

Chicago Destination Airport Count

DestinationAirport	DestinationAirportCount
Chicago O'Hare International Airport	559
Chicago Midway International Airport	131

Chicago Source Airport Count

SourceAirport	SourceAirportCount
Chicago O'Hare International Airport	568
Chicago Midway International Airport	138

# of Airports by USA Olson Timezone

SourceOlson_Timezone	AirportCount
America/New_York	444
America/Chicago	341
	201
America/Anchorage	174
America/Los_Angeles	163
America/Denver	117
America/Phoenix	36
Pacific/Honolulu	19
America/Adak	3
America/La_Paz	1

```
/* Airline Analysis */
proc format;
    invalue null
    '\N' = .A;
    value null
    .A = '.';
run;

data airlines;
    length Airline_ID 8.
           Name $50
           Alias $20
           IATA $2
           ICAO $3
           Callsign $40
           Country $30
           Active $3;
    infile '/home/kjg20/airlines.dat' dsd MISSOVER;
    input Airline_ID
          Name $
          Alias $
          IATA $
          ICAO $
          Callsign $
          Country $
          Active $;
    drop Airline_ID;
    format Name Alias IATA
           ICAO Callsign Country
           Active $null.;
    proc sort;
        by IATA;
run;

data airports;
    length Airport_ID 8.
           airportName $50
           City $20
           airportCountry $20
           airportIATA $3
           airportICAO $4
           Latitude 8
           Longitude 8
           Altitude 8
           Timezone $4
           DST $3
           Olson_Timezone $20
           Type $15
           Source $20;
    infile '/home/kjg20/airports.dat' dsd MISSOVER;
    input  Airport_ID
           airportName $
           City $

```

```
            airportCountry $
            airportIATA $
            airportICAO $
            Latitude
            Longitude
            Altitude
            Timezone
            DST $
            Olson_Timezone $
            Type $
            Source $;
run;

data routes;
    length Airline $3
            Airline_ID
            Source_airport $3
            Source_airport_ID
            Destination_airport $3
            Destination_airport_ID
            Codeshare
            Stops
            Equipment $50;
    infile '/home/kjg20/routes.dat' dsd MISSOVER;

    input Airline $
          Airline_ID
          Source_airport $
          Source_airport_ID
          Destination_airport $
          Destination_airport_ID
          Codeshare $
          Stops
          Equipment $;
          format
          Airline
          Source_airport
          Destination_airport
          Codeshare
          Equipment $null.
          Airline_ID
          Source_airport_ID
          Destination_airport_ID
          Stops null.;
run;


proc sql;
    create table airline_routes as
    select*
    from airlines
    full join routes
    on airlines.IATA = routes.Airline;
quit;
```

```
108
109 proc sql;
110     create table airline_routes_airports as
111     select*
112     from airports
113     right join airline_routes
114     on airports.airportIATA = airline_routes.Source_airport
115     where City = 'Chicago';
116 quit;
117
118
119
120
121
122
123
124
125
```