Introduction

According to CRAN, the ‘nycflights13’ dataset consists of “[a]irline on-time data for all flights departing New York City in 2013. Also includes useful 'metadata' on airlines, airports, weather, and planes.”

The purpose of this exploratory data analysis is to evaluate the data, extract information on it, and perform transformations in order to draw useful and valuable conclusions from it. This report documents “all findings, steps, conclusions, lessons learned, insights, etc.” from the EDA. In particular, this analysis evaluated factors such as the busiest airports, delay times, and how busy each airport was by month or by year. This information will aid the individuals and companies who possess it in increasing efficiency and customer satisfaction with the airline they choose to fly with. The conclusions drawn from this EDA also have several other implications. This analysis can help airline companies identify their weak areas and compete more effectively with other carriers. By increasing efficiency through the reduction of factors such as flight and delay times, carriers can lessen the toll that air travel takes on the environment.

Lastly, the data in this analysis is represented clearly and visually aesthetically so that individuals with both technical and non-technical backgrounds can comprehend and benefit from the information.

Data Description

As stated in the Introduction, the 'nycflights13' dataset contains information about all flights departed from New York City in 2013. The dataset was authored by and is maintained by Hadley Wickham. Below is a summary of the variables (features) of the dataset:

year, month, day Date of departure.
dep_time, arr_time Actual departure and arrival times (format HHMM or HMM), local tz.
sched_dep_time, sched_arr_time Scheduled departure and arrival times (format HHMM or HMM), local tz.
dep_delay, arr_delay Departure and arrival delays, in minutes. Negative times represent early departures/arrivals. carrier Two letter carrier abbreviation. See airlines to get name.
flight Flight number.
tailnum Plane tail number. See planes for additional metadata.
origin, dest Origin and destination. See airports for additional metadata. 4 planes
air_time Amount of time spent in the air, in minutes.
Distance Distance between airports, in miles.
hour, minute Time of scheduled departure broken into hour and minutes.
time_hour Scheduled date and hour of the flight as a POSIXct date. Along with origin, can be used to join flights data to weather data.

Data Cleaning

Extensive data cleaning was not necessary in this EDA. This could be due to the fact that the ‘nycflights13’ dataset is maintained by its author.

Analysis

To perform the exploratory data analysis on this dataset, several methods and techniques were used. Detailed below are some of the most relevant/most common functions and lines of code that were used to perform specific operations on the data:

1. **summarise()** - used to calculate summary statistics from the data (e.g. mean, minimum, maximum). Useful to find information about longest/shortest variables, or least/largest variables
2. **summarise(count = (n))%%** - used to calculate (count) occurrences within the data (e.g. number of flights departing from an airport, number of flights on a specific day/date, etc).
3. **filter()** - used to select specific rows of data. Useful to calculate information involving data for a specific date or range of numbers
4. **group_by()** - used to narrow down the data when necessary (e.g. group_by(carrier) was used to split that data so that the operations that followed would only be performed on the category/subset of airline carrier)

Listed below are the libraries used to perform functions and operations on the data:

1. **library(nycflights13)** - loads the package that contains the 'nycflights13' dataset
2. **library(dplyr)** - loads the 'dplyr' library, allowing access to functions such as 'filter()' and 'summarise()' (listed in the previous section)
3. **library(ggplot2)** - loads the 'ggplot2' package, which is used to visualize the data (see section below)

The following functions and operations were used to visually display the data and its insights:

1. **ggplot()** - this function and the operations within it were used to assign data to the aesthetic properties of the plot (e.g. axes, color, shape, and size)
2. **geom_*()** - In this particular EDA, the geom functions were used to create bar charts, line graphs, and violin graphs
3. **labs()** - used to title the plot and label it (x and y axes)
4. **theme_minimal()** - used to make the plot more aesthetically pleasing by simplifying its appearance

In order to access all of the lines of code used for this EDA, as well as their output, please refer to the attached .Rmd file titled, 'Joseph_Homework3_report.pdf'.

Key Findings

Here are several insights from the exploratory analysis of the nycflights13 dataset (from 'Joseph_Homework3_report.pdf'):

- There are three primary airports in NYC from which flights depart.
- The busiest airport in terms of departures is EWR.
- A total of 187 flights departed from LGA on July 4, 2013.
- The busiest day of the year was 11/27/2013 with 1014 flights.
- The most popular flight destination was ORD.
- Flight durations varied considerably, with the longest flight lasting 695 minutes and the shortest just 20 minutes.
- The airline with the largest number of flights was UA.

Future Work

As stated in the introduction, insights such as these can help determine which airports are the busiest and when. In the case of an airport such as EWR, if this analysis were to be taken further, the correlation between the number of EWR departures and how busy the airport generally is could be studied. Are the airports with fewer departures generally less busy? Or are they equally likely to be as busy when arrivals are taken into account as well?

Carrier efficiency can also be studied in terms of flight duration. Based on the EDA, it is observed that the shortest flight is 20 minutes. Which flight destination did the 20-minute flight go to? Based on that information, further analysis can observe how much time the flights for other carriers take to reach that destination. Is it about the same for every carrier (20 minutes)? Outliers can then be determined and studied further. If the average flight time to a specific destination is 20 minutes, why would a flight take significantly more time than that? These are just a few examples of further analysis that can be performed on this dataset and the conclusions that could be drawn.

References

‘Nycflights13’ - CRAN
https://cran.r-project.org/web/packages/nycflights13/nycflights13.pdf

Data Analysis Report: NYC Flights 2013 - Abigail Joseph
file:///C:/Users/ajose/OneDrive/Documents/R/Joseph_Homework3_report.pdf