Data description

Limitations of the data

How the data was downloaded, and processed

Reference List

Data on air traffic at LAS and SEA

Prepared by Alexandra Goh

March 24, 2023

Data description

The data is in file `flights.rda` in the `data` directory. It represents monthly data reported by U.S. certified air carriers that account for at least one percent of domestic scheduled passenger revenues, and contains these variables:

- **YEAR:** year of each flight (in this case, data was filtered to 2022)
- **MONTH:** month of each flight (in this case, data was filtered to include only September)
- **DAY_OF_MONTH:** day of the month that each flight operated on (for instance, 1st of September, etc.)
- **DAY_OF_WEEK:** day of the week that each flight operated on (for instance, 1 represents Monday)
- **OP_UNIQUE_CARRIER:** each reporting airline's unique carrier code/identifier (e.g. 9E = Endeavor Air)
- **TAIL_NUM:** unique identifier assigned to each aircraft
- **ORIGIN:** the origin airport that a flight departed from
- **DEST:** the destination airport that a flight flew to
- **CRS_DEP_TIME:** scheduled departure time for a flight
- **DEP_TIME:** actual departure time of a flight
- **DEP_DELAY:** difference in minutes between scheduled and actual departure time (negative numbers represent early departures)
- **CRS_ARR_TIME:** scheduled arrival time for a flight
- **ARR_TIME:** actual arrival time of a flight
- **ARR_DELAY:** difference in minutes between scheduled and actual arrival times (with early arrivals showing negative numbers)
- **CANCELLED:** binary variable indicating whether a flight was cancelled or not (1 = Cancelled)
- **DIVERTED:** binary variable indicating whether a flight was diverted or not (1 = Diverted)
- **DISTANCE:** distance between airports (miles)

The population of this dataset would be all domestic flights within the United States operated and reported by U.S. certified air carriers. In this scenario, the sample is a subset of flights with origin or destination airports in either LAS (Las Vegas) or SEA (Seattle-Tacoma International Airport). This is considered representative of the population as the QANTAS analyst team is interested in investigating which of the two airports (LAS or SEA) would be the most suitable in terms of providing connecting flights to other parts of the U.S, hence they can investigate how many flights are arriving to and leaving from these two specific airports.

For analyzing aviation data, September 2022 was chosen as it would be more helpful for the QANTAS analyst team to investigate recent trends in aviation patterns (instead of pre-COVID data, which would be better if they wanted to investigate long-term aviation trends). Furthermore, the international travel ban was lifted in the United States from early November 2021 onwards, hence it can be safe to assume there would be more frequent flights and higher numbers of travelers by September 2022 (Josephs, 2021). To reduce/avoid skewness in the data in certain months (e.g. due to peak holiday seasons, extreme weather conditions), we choose a "shoulder" month such as September. Shoulder months are the period of time after peak season but before off season; which are May, September and October in the U.S. (Sloss, 2019). Although there is lower demand for travel in September compared to holiday periods, this subset still consists of enough air traffic to provide a representative sample of the population.

The variables for scheduled and actual departure/arrival times as well as delays are useful in identifying how efficient airlines and airports are in meeting schedules and in their operations. For instance, if SEA was found to have a higher number of delays, this may be an unattractive option for QANTAS. Additionally, analyzing cancelled and diverted data can be helpful in understanding the frequency of flight disruptions in certain areas (e.g. flights may be cancelled/diverted due to weather conditions, airport congestion or other location-related factors). Finally, knowing the distance between airports could be an indication on whether the length/complexity of flight routes affect travel demand and/or flight delays.

Although this is considered open data, it should be noted there is no obvious license. However, this does fall under the US Open Data Policy, whereby datasets are made to be publicly available and accessible on their website (U.S. Department of Transportation, n.d.).

To get started in reading the data:

```
library(tidyverse)
library(dplyr)
library(here)
```

```
flights <- read_csv(here::here("data/T_ONTIME_REPORTING.csv")) %>%
  select(YEAR:DISTANCE)
save(flights, file=here::here("data/flights.rda"))
```

```
load(here::here("data/flights.rda"))
flights_subset <- flights %>%
  filter(ORIGIN %in% c("LAS", "SEA") | DEST %in% c("LAS", "SEA"))
```

Limitations of the data

There are certain limitations to be aware of when using this dataset. Firstly, the arrival and departure times in the aviation on-time performance data are reported in the local time zones of each airport. This may cause complexities when comparing aviation data across different time zones or when analyzing the delay minutes between different regions. To account for this, local time data should be converted to a standard time zone when comparing between airports in different time zones.

Besides that, there may be potential sampling bias involved. As this dataset covers a sample of flights/airlines flying to or from LAS or SEA, this shows that the sample involved is already biased towards these two large airports. This may not be representative of the broader population, as the on-time performance of smaller airports may be overlooked.

Last but not least, there are also external factors beyond the airlines'/airports' control which may affect aviation on-time performance. This includes weather, air traffic congestion, security/mechanical issues, crew availability and more. However, these factors are not recorded as variables in the Bureau of Transportation Statistics database. There may also be COVID-related biases as September 2022 is still considered a post-COVID era, with U.S. borders opening not that long ago; the COVID pandemic has impacted travel demand and airport/aviation operations, which should be kept in mind when analysing the dataset.

How the data was downloaded, and processed

The dataset was downloaded from the "Bureau of Transportation Statistics aviation ontime performance database" (https://www.transtats.bts.gov/), with aviation selected from the left box (i.e. "By Mode") followed by "Airline On-Time Performance Data". Under the "Reporting Carrier On-Time Performance (1987-present)" section in the resulting table, "Download" was selected to access the aviation on-time performance database. After filtering the dataset to only include data from September 2022, the variables aforementioned were then selected and the dataset downloaded as a .csv file.

Using the 'here' package, the dataset was read in as "flights" and later saved as a .rda file in the 'data' directory. The 'filter' function from the 'dplyr' package was then used to create a subset (known as "flights_subset") consisting of flights operating to (destination) and from (origin) LAS or SEA, as these were the airports of interest to QANTAS.

Reference List

Bureau of Transportation Statistics. (n.d.). On-Time: Reporting Carrier On-Time Performance (1987-present). Retrieved from https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr (https://www.transtats.bts.gov/DL_SelectFields.aspx?gnoyr_VQ=FGJ&QO_fu146_anzr=b0-gvzr)

Josephs, L. (2021). The U.S. is about to lift a nearly 20-month international travel ban. Here's what you need to know. Retrieved from https://www.cnbc.com/2021/11/07/us-to-lift-international-travel-entry-ban-here-are-the-rules.html (https://www.cnbc.com/2021/11/07/us-to-lift-international-travel-entry-ban-here-are-the-rules.html)

Sloss, L. (2019). How to Take Advantage of Shoulder Season. The New York Times. Retrieved from https://www.nytimes.com/2019/09/17/travel/travel-deals-shoulder-season.html# (https://www.nytimes.com/2019/09/17/travel/travel-deals-shoulder-season.html#):~:text=%E2%80%9CShoulder%20season%20is%20a%20travel,prices%20really%20start%20to%20decline.%E2%80%9D

U.S. Department of Transportation. (n.d.). Data Inventory. Retrieved from https://www.transportation.gov/data (https://www.transportation.gov/data)