

# Introduction

This case comprises of 4 sections:

- Getting Familiar with data
- Preparing the data
- Exploring the data
- Forming hypotheses

## ➤ Getting Familiar with the data

Study the flights data along with the planes, and weather metadata. Generate a list of variables in each dataset along with their attributes. Use the data dictionary to make sense of the data and the variables in each of the dataset.

## ➤ Preparing the data

### *Dealing with Missing data*

Missing values are to be treated separately and are an important part of data preparation. If data is missing for key variables, then we might decide to delete the observation. If the variable is not important, we can also delete the variable. Missing values can also be imputed. In some cases, we replace missing values with aggregated numbers from the entire dataset, but in some cases these replacements have to be calculated particular to sector and used accordingly. While answering the below questions, try to understand the reason for taking different approaches while dealing with missing data.

Q 1).

- In the flights data:
  - Calculate the missing values present in each variable.
  - Delete all observations where a missing value in any of the following variables: tail number departure time and arrival time.
  - Replace the missing values for Air\_time with the average airtime on the specific route (origin -> destination) for the specific carrier.
- In the weather data:
  - Calculate the missing values present in each variable
  - Replace the missing data for weather conditions with average weather conditions at that airport on that day.
- In the planes data:
  - Calculate the missing values present in each variable
  - Remove redundant variables with more than 70% missing values.
  - Remove all the observations with any missing values.

### *Extracting information from the existing variables*

Q 2). Flights data contains information for all flights that departed New York City i.e. from John F. Kennedy International Airport (JFK), LaGuardia Airport (LGA) or from Newark Liberty International Airport (EWR) in 2013.

Create the following new variables:

- Year from date variable
- Month from date variable
- Day from date variable
- Hour from scheduled departure time variable
- Departure delay – this captures the difference between departure time and scheduled departure time.
- Arrival delay – this captures the difference between arrival time and scheduled arrival time.

### ➤ Exploring the data -: Data manipulation to extract relevant information

Q 3). Busiest routes

- i. Identify the busiest routes for the year 2013 ie which origin-dest had the maximum flights.
- ii. Calculate the number of flights for each of the carriers for the top five routes.
- iii. Compare the numbers calculated in (ii) with total number of flights for each carrier

Q 4). Busiest time of the day (maximum flights taking off)

- i. Identify the busiest time of the day for each carrier.
- ii. Identify the busiest time of the day for three airports, John F. Kennedy International Airport (JFK), LaGuardia Airport (LGA) and Newark Liberty International Airport (EWR).

Q 5). Origin and Destinations

- i. Out of all flights departing from JFK, what percentage of flights got delayed?
- ii. Which origin airport had the least number of total delays? (Since this is origin airport, please track delay basis departure delay)
- iii. Which destination(s) has the highest delays?

## ➤ Forming hypotheses -: Checking for relationships

Q 6). Understanding weather conditions related with delays

- i. Join the weather and flights data using the variables: date, hour and origin variables.
- ii. Calculate averages for the weather condition parameters provided and the departure delay, grouped by months.

Q 7). Years of operation and Fuel consumption cost

- i. Is there a relationship between manufacturing date of the plane and average annual fuel consumption cost of the plane i.e. do older planes use more fuel?
- ii. Also understand check the relationships between fuel consumption with other plane variables like number of seats, engine type, number of engines, type of plane.

Q 8). Variation of delays over the course of the day

On average, how do departure delays vary over the course of a day? Does it increase or decrease? (You might want to analyse average departure delays for each hour and check the trend)