

Case Study – Airlines Data

Download Airlines dataset for the year 2008 from stat-computing.org

<http://stat-computing.org/dataexpo/2009/the-data.html>

Extract csv file from the zip folder and save it into data folder

note: Refer to <http://www.airportcodes.org/> for a list of airport codes

0.0.1 Import CSV file into your R session and name the dataframe as myDF?

1 Extraction of Flights data, Arrival & Departure time analysis:

1.0.1 Write down How many observations and variables our myDF dataframe has?

1.0.2 Extract the Head and Tail of a Data Set?

1.0.3 Which were the origin cities for the first six flights in the airline data file for 2008?

1.0.4 What are the destination cities for the first six flights in the airline data file for 2008?

1.0.5 Which were the origin cities for the last six flights in the airline data file for 2008?

1.0.6 What are the destination cities for the last six flights in the airline data file for 2008?

1.0.7 How many number of flights departure from Indianapolis (IND) in 2008?

1.0.8 similarly how many flights that destination city are Indianapolis (IND) in 2008?

1.0.9 What are the first 6 flights that had IND as the origin?

1.0.10 In 2008, how many flights departed from O'Hare Airport (ORD)?

1.0.11 In 2008, how many flights arrived at O'Hare Airport (ORD)?

1.0.12 How many flights in the data set departed from the Indianapolis Airport (IND) and arrived at the O'Hare Airport (ORD)?

1.0.13 Create a dataframe with Subset of *only* the flights in which IND as the origin city, and create another data frame with the flights for which IND is the destination city then Cross check the number of observation & variables for each dataframe have below!

42750 29

42732 29

1.0.14 Get the first 6 flights of each of those new data frames created from above statements?

- 1.0.15 How many indianapolis flights departed during each month?
- 1.0.16 Use the same question above to represent in a graphical/plotting
- 1.0.17 write down your own observations or interpretations from the above 2 questions of numerical and graphical representation of indianapolis flights of each month ?
- 1.0.18 Apply the same interpretations for the flight for which indianapolis was the destination?
- 1.0.19 How many flights departed from TUP in 2008?
- 1.0.20 What is the average departure delay of the flights that depart from TUP in 2008?
- 1.0.21 Find out the number of flights in the early morning before 6 AM departed out of IND origin?
- 1.0.22 Similarly, flights that departed before 12 noon, 6 PM, 12 midnight
- 1.0.23 Now checkout how many NA values in Departed time column contains?
- 1.0.24 Cross check whether the total number of flights departed from IND is the sum of the flights departed before or equal to midnight and number of NA's
- 1.0.25 How many flights arrived at LAX in 2008?
- 1.0.26 How many flights departed from ATL and landed at LAX in 2008?
- 1.0.27 Among the flights from ATL to LAX in 2008, how many departed before 12 noon?

2 Popular Flights Paths & Arrival Delay Analysis:

- 2.0.1 Build a table that shows how many cities are the origin for the flights throughout 2008?
- 2.0.2 Sort the above results in a decreasing order?
- 2.0.3 which airport has highest flights departed in 2008, what is the code of that airport and how many number of flights departed?
- 2.0.4 what are the top 6 airports which has most flights departed in 2008?
- 2.0.5 plot all the flights departed within each hour range during the course of the day
- 2.0.6 Make a table of all the origin-to-destination pairs
- 2.0.7 what are the first 6 values from the above table?
- 2.0.8 what are the last 6 values from the above table?

- 2.0.9 what are the 20 most such popular pairs?
- 2.0.10 How many origin-to-destination paths were only flown one time (each) in 2008?
- 2.0.11 Find the average departure delay at each airport
- 2.0.12 Which airports has shortest and longest average departure delay
- 2.0.13 Which airports has shortest and longest average arrival delay
- 2.0.14 When considering all flights to an airport, take an average of the distances (in miles) of the flights to that airport. Suppose that we do such an analysis of all airports. Which airport has the longest average distance of flights (in miles) arriving to that airport?
- 2.0.15 Which day of the week should we fly, if we want to minimize the expected arrival delay of the flight?
- 2.0.16 what is the average arrival day of the only IND arrivals flights for each day of the week?
- 2.0.17 How many flights occur per month?
- 2.0.18 Which airline has the worst average departure delay?
- 2.0.19 Which airlines have the best average arrival delays
- 2.0.20 Which airlines have the worst average arrival delays
- 2.0.21 sort the expected ArrDelay for each day of the year 2008 ?
- 2.0.22 What is the expected ArrDelay for the flights arriving to IND for each date?
- 2.0.23 On which day of the year were the average departure delays the worst?
- 2.0.24 On which day of the year were the average departure times the worst for flights departing from O'Hare (ORD)?
- 2.0.25 Categorize the Expected arrival delay for the flights from ORD to IND according to the date?
- 2.0.26 Consider flights that departed from ATL and landed at LAX in 2008. For how many days of the year were the average departure delays more than 90 minutes?

3 Final note: *Create Your Own Data Frame*

Build a data frame of your own, containing the flights that depart from an airport of interest to you (New York City's JFK airport, for instance). try to answer some questions about your own data frame. How do the questions and answers compare/contrast for

bigger or smaller airports? Another possibility is to try to analyze other years of the airline flight data, using the rest of the data provided by the ASA Data Expo 2009. The flight data from this website goes back to 1987.