[www.edureka.co/r-for-analytics](http://www.edureka.co/r-for-analytics)

|  |
| --- |
| Project Guide |
| Airline Dataset |

|  |
| --- |
|  |



© 2014 Brain4ce Education Solutions Pvt. Ltd.

Project Guide

Airline Dataset

Table of Contents

[Real World Problem Statement & Data 2](#_Toc399249331)

[(i) Problem Statement 2](#_Toc399249332)

[(ii) About the Data 2](#_Toc399249333)

[How to get the data? 2](#_Toc399249334)

[Variable Descriptions in the Data 2](#_Toc399249335)

# Real World Problem Statement & Data

## (i) Problem Statement

1. Check the skewness of Distance travelled by airlines.
2. Calculate the mean, median and quantiles of the distance travelled by US Airlines (US).
3. Check the standard deviation of distance travelled by American Airlines (AA).
4. Draw a boxplot of UniqueCarrier with Distance.
5. Draw the direction of relationship between ArrDelay and DepDelay by drawing a scatterplot.

## (ii) Problem Statement

1. What is the probability that a flight which is landing/taking off is “WN” Airlines **(marginal probability)**
2. What is the probability that a flight which is landing/taking off is either “WN” or “AA” Airlines **(disjoint events)**
3. What is the joint probability that a flight is both “WN” and travels less than 600 miles **(joint probability)**
4. What is the conditional probability that the flight travels less than 2500 miles given that the flight is “AA” Airlines **(conditional probability)**
5. What is the joint probability of a flight getting cancelled and is supposed to travel less than 2500 miles given that the flight is “AA” Airlines **(joint + conditional probability)**

## (iii) Problem Statement

1. Suppose arrival delays of flights belonging to “AA” are normally distributed with mean 15 minutes and standard deviation 3 minutes. If the “AA” plans to announce a scheme where it will give 50% cash back if their flights are delayed by 20 minutes, how much percentage of the trips “AA” is supposed to loose this money. (Hint: pnorm)
2. Assume that 65% of flights are diverted due to bad weather through the Weather System. What is the probability that in a random sample of 10 flights, 6 are diverted through the Weather System. (Hint: dbinorm)
3. Do linear regression between the Arrival Delay and Departure Delay of the flights.
4. Find out the confidence interval of the fitted linear regression line.
5. Perform a multiple linear regression between the Arrival Delay along with the Departure Delay and Distance travelled by flights.

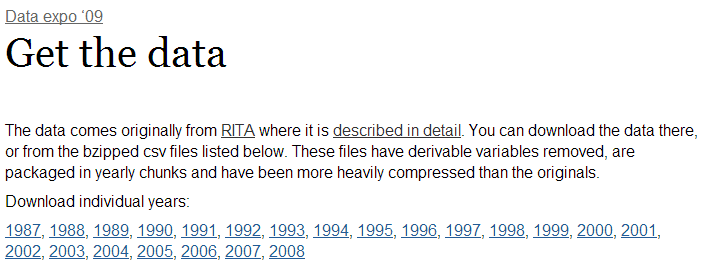
## (ii) About the Data

The data used in this project is real and is based on collection of over 20 years. The total number of record in this dataset is roughly around 120 million rows and the size of the data is approximately 12GB. The data consists of flight arrival and departure details for all commercial flights within the USA, from October 1987 to April 2008. This is a large dataset. There are around 29 attributes.

### How to get the data?

The data originally comes from [http://stat-computing.org/dataexpo/2009/the-data.html](http://stat-computing.org/dataexpo/2009/the-data.html%20)

In the site, you will come across something like this:



You can download the data for each year by clicking the appropriate link in the above website (Remember the size is going to be more than 12GB).

### Variable Descriptions in the Data

In order to understand the data, one has to follow the following variable descriptions.

|  |  |  |
| --- | --- | --- |
| **S. No** | **Variable** | **Description** |
| 1 | Year | 1987-2008 |
| 2 | Month | 1-12 |
| 3 | DayofMonth | 1-31 |
| 4 | DayOfWeek | 1 (Monday) - 7 (Sunday) |
| 5 | DepTime | actual departure time (local, hhmm) |
| 6 | CRSDepTime | scheduled departure time (local, hhmm) |
| 7 | ArrTime | actual arrival time (local, hhmm) |
| 8 | CRSArrTime | scheduled arrival time (local, hhmm) |
| 9 | UniqueCarrier | [unique carrier code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 10 | FlightNum | flight number |
| 11 | TailNum | plane tail number |
| 12 | ActualElapsedTime | in minutes |
| 13 | CRSElapsedTime | in minutes |
| 14 | AirTime | in minutes |
| 15 | ArrDelay | arrival delay, in minutes |
| 16 | DepDelay | departure delay, in minutes |
| 17 | Origin | origin [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 18 | Dest | destination [IATA airport code](http://stat-computing.org/dataexpo/2009/supplemental-data.html) |
| 19 | Distance | in miles |
| 20 | TaxiIn | taxi in time, in minutes |
| 21 | TaxiOut | taxi out time in minutes |
| 22 | Cancelled | was the flight cancelled? |
| 23 | CancellationCode | reason for cancellation (A = carrier, B = weather, C = NAS, D = security) |
| 24 | Diverted | 1 = yes, 0 = no |
| 25 | CarrierDelay | in minutes |
| 26 | WeatherDelay | in minutes |
| 27 | NASDelay | in minutes |
| 28 | SecurityDelay | in minutes |
| 29 | LateAircraftDelay | in minutes |

You can take data for any year and try to solve the following problems.

A screenshot containing the 25 first lines may look like this:

