**DS04 Activity 4: Remove Duplicates and Replace Missing Values**

Before exploring data, clean it by removing duplicate rows and replacing missing values

Task 1: Remove duplicate rows (retaining the first instance of each row). Rows are considered duplicates in this dataset if they have matching values for all the following fields:

* Year
* Month
* DayofMonth
* Carrier
* Origin Airport ID
* Dest Airport ID
* CRS Dep Time
* CRS Arr Time

**A screenshot of a computer

Description automatically generated**

Task 2: After removing the duplicate rows, replace missing values in the DepDelay and ArrDelay columns with the value 0 (zero). Use the built-in Azure Machine Learning Remove Duplicate Rows module,

**A computer screen shot of a computer

Description automatically generated**

Task 3: Answer the following questions, after you have removed duplicate rows and replaced missing values,

* How many rows remain in the dataset?
* What is the mean value of the ArrDelay column?

Include screen shots of each task of Activity 4 in the Project Report.

A computer screen shot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

Following the removal of duplicate rows and replacement of missing values, we observe the following:

The dataset maintains its original structure with 14 columns.

The number of rows now totals 2,719,397.

The calculated mean value for "ArrDelay" is 6.5669.

This demonstrates the effectiveness of the duplicate removal and data cleaning processes in maintaining the dataset's integrity and statistical properties.

After completing the Data Cleanup, perform activities and tasks related to Data Exploration.

At Activity 5