**Applying Machine Learning**

The goal for us is to predict the number of empty seats for a given airline which can be inferred by predicting the seat utilization.

As the input data has all the required variables for the computation and also has the output variable (Seat Utilization), We classified this problem as a Supervised regression problem as the output variable (Seat Utilization) is not a category as well.

**Features:**

In this case, Seat Utilization would be the dependent variable.

The independent feature variables are:

* Month and Year
* Date
* Departures
* Monthly Revenues

**Linear Regression:**

We are using the Linear Regression machine learning technique as this is best suited for the problem and the data at hand. Multiple linear regression allows us to use all of our variables to improve our predictive ability.

We will measure the performance of the model by using the R-squared value where a R-squared value close to 1 indicates a good model.

**Approach:**

We use the recipes package to build up the data frame that can be used in any modelling technique.

We created dummy variable columns for each airline (151 airlines), so that each of these columns can be treated as independent variables.

To be able to predict airlines that are not in the data set, we introduce an “OTHER” airline:



We now use the lm function to perform linear regression analysis.

First, we use the lm function with seat utilization against all the dummy variable columns except the seats, date and Monthly revenue columns.



Which gets us an R squared value of 0.8694



lm function with all the columns including revenue



R squared value is 0.8726

