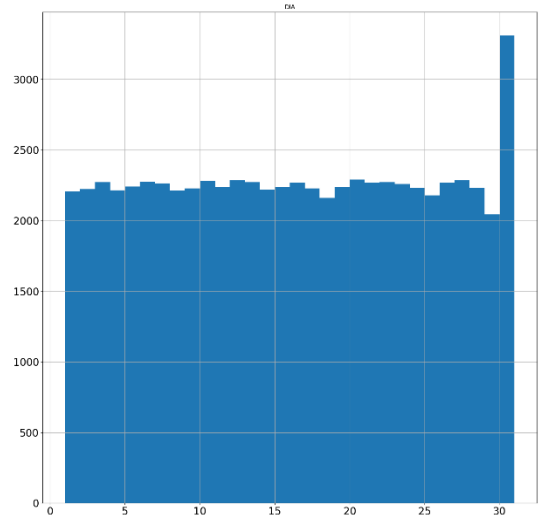
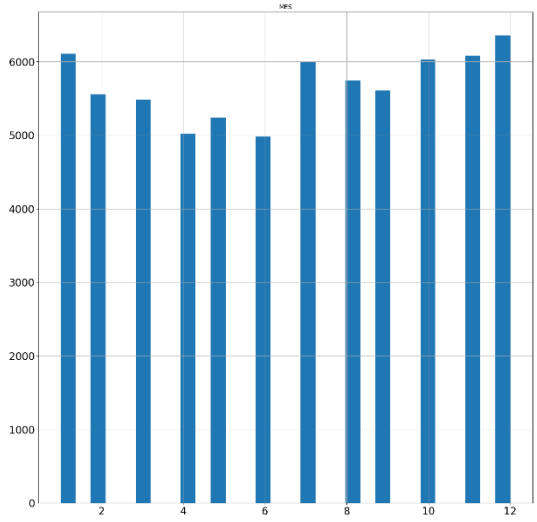
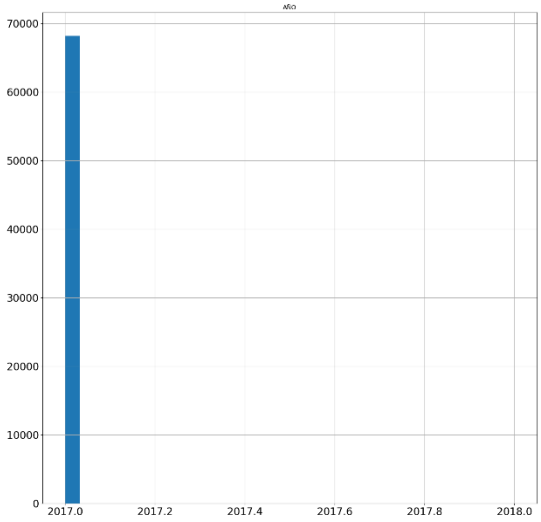
LATAM Project

1. **How is the data distributed? Did you find any noteworthy insight to share? What can you conclude about this?**







According to the graphs shown in the image, the distribution of the numeric variables is of the uniform type. It can be seen that for the case of the variable "DÍA", there is a tendency to carry out flights on the last day of the month.

2. **What is the behavior of the delay rate across destination, airline, month of the year, day of the week, season, type of flight?What variables would you expect to have the most influence in predicting delays?**

The variables that could have the greatest influence could be the day, flight time, and destination. Also, the airline could have an influence on the delay or not of the flight.

To verify this, a model analysis will be carried out and the importance of the variables in it will be observed. First, the necessary features for the training and testing of these will be created.
Afterwards, different models will be trained to perform a binary classification of the variable "delay_15".

3. Train one or several models (using the algorithm(s) of your choice) to estimate the likelihood of a flight delay. Feel free to generate additional variables and/or supplement with external variables.

   Four different models were used. Logistic Regression, Decision Trees, Random Forest, and Gradient Boosting.
   As the variables are categorical and are to be used for different models, what was done was a conversion of the variables to one-hot encoding. This way, the models can be compared and trained without difficulty.

4. Evaluate model performance in the predictive task across each model that you trained. Define and justify what metrics you used to assess model performance. Pick the best trained model and evaluate the following: What variables were the most influential in the prediction task? How could you improve the Performance?

   First, a Logistic Regression model was trained. With the following results:

	precision	recall	f1-score	support
0	0.82	0.99	0.90	13898
1	0.50	0.05	0.09	3154
accuracy			0.82	17052
macro avg	0.66	0.52	0.50	17052
weighted avg	0.76	0.82	0.75	17052

   Then a Decision Tree model was trained, and the following results were obtained:

	precision	recall	f1-score	support
0	0.84	0.88	0.86	13898
1	0.35	0.28	0.31	3154
accuracy			0.77	17052
macro avg	0.60	0.58	0.59	17052
weighted avg	0.75	0.77	0.76	17052

In third place, a Decision Tree model was trained and the following metrics were obtained:

	precision	recall	f1-score	support
0	0.84	0.93	0.88	13898
1	0.40	0.21	0.28	3154
accuracy			0.80	17052
macro avg	0.62	0.57	0.58	17052
weighted avg	0.76	0.80	0.77	17052

Finally, a Gradient Boosting model was trained:

	precision	recall	f1-score	support
0	0.82	1.00	0.90	13898
1	0.66	0.03	0.07	3154
accuracy			0.82	17052
macro avg	0.74	0.52	0.48	17052
weighted avg	0.79	0.82	0.75	17052

To evaluate the performance of the models, F-1 Score is the best indicator, because the classes are imbalanced and the delayed flights are not the majority. Furthermore, as it is a binary classification problem, it is recommended to use the F-1 Score indicator.

It can be observed that there are models that have good performance for flights without delay. However, the model that has good performance for all flights is the Decision Tree model. Therefore, variable tuning will be carried out to verify the performance it could achieve.

To improve the selected model, the first step was to balance the classes and train the model in this way. Afterwards, the model was evaluated to see how it performed. Finally, the importance of the variables was verified.

By training the model with balanced classes and evaluating it again on the original (imbalanced) dataset, an improvement in performance was obtained.

Accuracy: 0.8479357260145437

	precision	recall	f1-score	support
0	0.93	0.89	0.91	15181
1	0.36	0.48	0.41	1871
accuracy			0.85	17052
macro avg	0.64	0.69	0.66	17052
weighted avg	0.87	0.85	0.86	17052

Finally, the most important variables for that model were verified.

```
{'Des-I': 0.3563400123611429,
 'DIA': 0.29221273091922395,
 'MES': 0.07755802154313274,
 'DIANOM': 0.03995971845038059,
 'TIPOVUELO': 0.03438164544791378,
 'OPERA': 0.10484899308067572,
 'high_season': 0.002144641835051591,
 'period_day': 0.09255423636247871}
```

It can be seen that the most important variables are Des-I, DIA, and OPERA.

To improve the current model, we can reduce the number of variables and only leave the most important ones. We can also adjust the parameters used and even try with the previous models using a similar procedure to the one used with this model.