DATA_ENG 300
Mo Babgi

Homework 1: Aircraft Inventory Analysis

Question 1: Missing Data Investigation

For CARRIER and CARRIER_NAME, I choose to impute entries, a CARRIER value but no CARRIER_NAME, since those missing values seem to be of the Missing Completely at Random type. I used the most frequent carrier name for each carrier number to impute.

MANUFACTURE_YEAR also seems to be MCAR, but imputing would require a complex calculation. Considering that there are only 3 missing values, the missing values would likely have a negligible effect, so I choose to not impute and instead focus on the complete entries.

For NUMBER_OF_SEATS and CAPACITY_IN_POUNDS, I choose to impute with the median, using the reported MODEL of the aircraft. Due to the large number of models of over 1000, I only used models with over 100 entries to impute (i.e., the top 250 models).

For AIRLINE_ID, I choose to impute with the mode, using the reported UNIQUE_CARRIER of the aircraft since those two values seem to be reflective of each other. I used the most frequent ID number for each unique carrier number to impute.

Question 2: Standardization and Transformation of Categorical Fields

The MANUFACTURER column has a lot of the same manufacturer names being capitalized differently, or being written out in full/short (e.g., Airbus vs. AIRBUS vs. AirbusIndustries). Standardization is needed to make such that each manufacturer represented is unique. I choose to standardize the data by changing all values to be in uppercase and removing any whitespace, and then standardizing the names of each major aircraft to be represented unifromly.

After inspecting the column MODEL, I don't think standardization is necessary in the same way as with MANUFACTURER, since each value represents the model of an aircraft. I choose to simply uppercasing the values and remove whitespace to ensure that all entries of aircraft of the same model can be grouped together effectively.

The column AIRCRAFT_STATUS seems to feature 4 different possible statuses for aircraft: A, B, and L, O. Similar to MODEL, I standardize by simply uppercasing the values to ensure consistency.

OPERATING_STATUS values are either Yes (Y, y) or No (N). I choose to standardize the values first by changing the values to uppercasing. Then, I also transform the column to make it boolean for ease of analysis.

Question 3 — Dropping Remaining Missing Data

After dropping all the remaining missing values, considering the columns imputed in Question 1, the dataset retained with complete data has 132144 entries, down 169 entries from the original dataset.

Question 4 — Skewness & Box-Cox Transformation

The skewness is positive for both columns, at 0.43 and 4.06 for NUMBER_OF_SEATS and CAPACITY_IN_POUNDS, respectively. This indicates the data is skewed right, with CAPACITY_IN_POUNDS being more skewed right. This is confirmed by the histograms of the two columns, which suggest that smaller airplanes appear more frequently in the dataset.

After transforming the data using the Box-Cox algorithm, its normality increased significantly. The skewness of NUMBER_OF_SEATS became negative, at -0.51, while the skewness of CAPACITY_IN_POUNDS remained positive but decreased significantly, at 0.23. This is confirmed by the histograms, showing more normal-like distributions for both variables, with NUMBER_OF_SEATS showing a slight skewness to the left.

Question 5 — Feature Engineering and Group-wise Analysis

Inspecting the plot of OPERATING_STATUS by size shows that the largest number of aircraft in the dataset is of medium size, followed by the xlarge size, with small aircraft appearing least frequently. Medium-sized aircraft have the highest proportion of non-operating aircraft at 6.3%, but all 4 sizes have comparable proportions of non-operating aircraft, with the lowest proportion being for xlarge-sized aircraft at 2.7%

Inspecting the plot of AIRCRAFT_STATUS by size shows that all sizes of aircraft have have Status O as the most common aircraft status, except the medium size, which has more Status B aircraft. That said, for medium-sized aircraft, Status O still appears at a similar proportion to Status B. Status L appears very rarely for all sizes of aircraft, while Status A appears infrequently although its proportions are still considerable.