Amrita Natarajan
DataEng300
Homework 1 PDF

Task 1: Investigate the missing data in this dataset.

Carrier: The missing values for the "Carrier" column only occur for North American Airlines. This likely resulted from a parsing issue when importing the CSV into pandas, where the abbreviation "NA" was misinterpreted as NaN rather than its intended meaning of North American Airlines. This was confirmed by inspecting the original CSV file, where “NA” is clearly used as an abbreviation rather than representing missing data. Therefore, the Carrier column values in these rows are not truly missing and can be safely corrected.

Carrier Name: For the "Carrier Name" column, there are two specific cases where the values are missing: Carrier “L4” in 2007 and Carrier “OH” in 2013. Since the missingness is associated with specific years, this suggests that the data is Missing At Random (MAR). This was verified by filtering the dataset by carrier and year. Based on other existing records in the dataset, it was determined that all missing “L4” entries correspond to "Lynx Aviation d/b/a Frontier Airlines" and all missing “OH” entries correspond to "Comair Inc." These imputations were made accordingly.

Manufacture Year: The "Manufacture Year" column has a missing value for an entry where the Carrier is “5Y” and the Carrier Name is “Atlas Air Inc.” Because all other values in the row are present and align with similar rows in the dataset, this missingness appears to be Missing Completely At Random (MCAR). The year was imputed using the median of the manufacture year distribution, which was 1998. The median was chosen over the mean (which was 1998.23) due to a slight right skew in the distribution. Two additional missing entries with Carrier “9E” and aircraft type 6311 were also considered MCAR and imputed with the year 2003, based on similar logic. Again, the median (2003) was used because the mean was slightly lower at 2002.9 and the distribution appeared right-skewed.

Number of Seats: The "Number of Seats" column contains missing values that all occur in 2019 for Carrier “M6.” This pattern suggests the data is MAR, and not randomly missing across the dataset. However, since the aircraft models in question (6252 and 6262) consistently have 0.0 seats in the rest of the dataset, it was appropriate to impute the missing values with 0.0 to maintain consistency.

Capacity in Pounds: For the "Capacity in Pounds" column, unit 78991 had missing data, but all other variables matched unit 71830, which has a known capacity of 237756.0. Therefore, the missing value was filled using this matched value. Two more missing values for units 79015 and 79016 were also matched to similar entries and filled with the same value. For the aircraft models DC-10-10, DC-10-30, and MD-11, which had multiple entries with missing capacities, a K-Nearest Neighbors (KNN) imputation approach was used. This method was selected due to the presence of multiple known capacity values within each model category. Using k=5 and numerical features such as number of seats and manufacture year, KNN imputed the missing values based on the average of similar entries. Post-imputation histograms showed a strong alignment between the distributions of original and imputed data, supporting that the approach preserved the structure of the dataset.

Airline ID: The "Airline ID" column had 97 missing entries associated with Carrier “OH” and 8 associated with Carrier “L4.” For L4, the matching airline ID across all other aircraft was 20107, and for OH, it was OH (1) across the models CRJ100-Passenger, CRJ700-Passenger, CRJ900-Passenger, and CRJ200-Passenger. These consistent patterns allowed for confident imputation of the missing airline IDs based on carrier and model type.

Task 2: Data Standardization

For the “Manufacturer” column, multiple representations of the same company were found due to inconsistent naming conventions. For example, “Boeing” appeared in various forms such as “BOEING,” “THEBOEINGCO,” and even model-specific references like “B747.” To address this, values were analyzed using case-insensitive string matching to identify and consolidate variations of the same manufacturer under a single standardized name. A similar process was applied to the “Model” column, though with more caution. Unlike manufacturers, airplane model names often include meaningful alphanumeric distinctions, making it more difficult to determine whether slight variations indicate genuine differences or simple formatting inconsistencies. Therefore, only clear and repetitive formatting differences (e.g., spacing or use of hyphens) were standardized on a per airline basis. For both “Aircraft Status” and “Operating Status”, all values were converted to uppercase to ensure consistency in categorical labeling and prevent mismatches due to case sensitivity.

Task 3: Amount of data obtained is 101316 rows from the original 132312 rows.

Task 4: Transformation and derivative variables

Before transformation, both the NUMBER_OF_SEATS and the CAPACITY_IN_POUNDS were skewed right, with NUMBER_OF_SEATS having a smaller skew (.378) and CAPACITY_IN_POUNDS being more heavily skewed (3.76). Post transformation, both histograms were a lot closer to normal visually, although NUMBER_OF_SEATS appears to have a large outlier of 0 while the rest are centered around 11. CAPACITY_IN_POUNDS is more evenly spread from 0 to 160 centered around 65. It is key to note that the units shifted post transformation (apparently no longer in seats, pounds respectively).

Task 5: Feature Engineering

Across all size categories, over 90% of aircraft are operational (under Operating Status), with many categories exceeding 95%, suggesting that planes of all sizes are generally in active use. Focusing on aircraft status within the SMALL size category, statuses O (Operational) and B (Being Worked On) are nearly evenly split — O accounts for about 50% and B just over 40%. Status A (Available) makes up less than 5%, while L (Left Fleet) is not represented. This makes sense because L is not included in the description of possible aircraft statuses, however seeing as there was representation in the data, it was included. There is no clear indication of what L represents. As aircraft size increases, the proportion of O steadily rises — from 50% in SMALL to over 70% in XLARGE — while B shows a noticeable decline, dropping from over 40% in SMALL to about 20% in XLARGE. The proportion of A remains relatively stable, with a modest increase from about 5% to just under 10%. Interestingly, L status appears only in the MEDIUM to XLARGE categories, though in very small proportions, around 1%.

General AI Usage: For areas where I knew I wanted to use REGEX for formatting and standardizing, like identifying spaces, inserting dashes and making many of the same variation of a word the same word, I asked ChatGPT to generate the regex pattern.