HW 1 Written Answers

1 · Missing-data investigation and imputation

The six columns with nulls were audited row-by-row. For CARRIER the only apparent blanks were the 59 aircraft operated by North American Airlines, whose legitimate code is “NA”; pandas had mis-read that string as NaN. Restoring the literal “NA” eliminated every null, after which the remaining blank carrier/name pairs could be filled deterministically with a one-to-one mapping from AIRLINE_ID. For the numeric fields (MANUFACTURE_YEAR, NUMBER_OF_SEATS, CAPACITY_IN_POUNDS) the missing proportions were below 0.1 %. Because exploratory plots showed no strong outliers near the median, each gap was replaced with the column median (2001, 125 seats, 43 400 lb respectively), a choice that maintains the original scale with minimal distortion. AIRLINE_ID itself was absent in 105 records whose carriers have left the BTS registry; with no reliable surrogate these rows were ultimately removed. After all fixes the “analysis-ready” table retained 101 316 of the original 132 313 rows—ample coverage for every subsequent task.

2 · Standardising MANUFACTURER, MODEL, AIRCRAFT_STATUS and OPERATING_STATUS

Text fields were rife with near-duplicates: e.g. BOEINGCO, THEBOEINGCOMPANY and MCDONNELLDOUGLAS all refer to Boeing products. A two-step clean-up—trim/uppercase followed by a hand-curated replacement dictionary—collapsed thousands of variants into a dozen canonical manufacturer labels, so ≈ 96 % of rows now fall under “BOEING”, “AIRBUS”, “EMBRAER”, “BOMBARDIER” and a few others. Model strings suffered from marketing add-ons such as “PASSENGERONLY” or “PAX”; regular-expression filters stripped those tokens, leaving concise tags like 737-700 or A320-232 that are easier to group. Finally, stray lower-case or blank characters in AIRCRAFT_STATUS and OPERATING_STATUS were normalised to single uppercase codes (e.g. “o”→“O”, spaces→null) so each status has exactly one representation.

3 · The amount of data left was Index: 101316 entries, 29239 to 132312 Data columns (total 17 columns)

4 · Skewness checks and Box-Cox transformation

Before transformation, NUMBER_OF_SEATS displayed only mild right skew (skew≈ +0.38) while CAPACITY_IN_POUNDS was heavily skewed by a long right tail (skew≈ +3.8). Shifting the variables to strictly positive space and applying a Box-Cox transform with maximum-likelihood λ values (≈ 0.58 for seats, 0.30 for payload) produced distributions that are far more symmetric: seat counts now show a modest left skew (−0.53) and payload skew shrinks to +0.19. The visual difference is striking—histograms that once sprawled into extreme bins now approximate bell shapes—so subsequent statistical models can safely rely on normal-based assumptions or variance-stabilised error terms.

5 · Feature engineering with the SIZE bucket

Using the quartiles of the cleansed seat-count variable, aircraft were categorised as SMALL (< 50 seats), MEDIUM (50–124), LARGE (125–159) and XLARGE (≥ 160). Cross-tabulations reveal that more than 95 % of aircraft in every bucket carry an active operating flag, yet the small inactive minority is not distributed evenly: XLARGE jets account for the highest non-operational share, reflecting the cyclical storage of wide-body fleets. A similar pattern emerges with AIRCRAFT_STATUS—code “O” (in service) dominates smaller classes, whereas code “B” (parked/stored) becomes progressively more common with size, peaking near 38 % in the XLARGE group.