Task 1 Data Imputation
NA in CARRIER/UNIQUE_CARRIER for "North American Airlines" is not missing data. It's a string. When reading the CSV file, I count NA as a string but count "", N/A, and Null as missing data.

CARRIER doesn't have missing data. Other variables do have missing data. Checked by seeing the head of each missing column.

CARRIER_NAME is safe to impute because we have CARRIER and UNIQUE_CARRIER_NAME filled. This is probably due to MCAR. Data are imputed by mapping CARRIER and CARRIER_NAME.

MANUFACTURE_YEAR is safe to impute because it is likely due to MCAR. Since there are only 3 data points missing, median imputation is acceptable.

NUMBER_OF_SEATS is safe to impute. It is probably due to MAR because, upon inspection, only cargo planes miss the number of seats data. Constant imputation is applied with constant = 0.

CAPACITY_IN_POUNDS is safe to impute. Likely MAR - tied to aircraft specs. Multiple imputation can be applied.

AIRLINE_ID is missing as it's correlated with CARRIER_NAME. We can impute from CARRIER. This is due to MAR. The airline ID is based on CARRIER.

Task 2 Transformation or Standardization of Data
For MANUFACTURER, upon inspection, I found out there are typos, extra white spaces, and capitalization errors. The standardization I did was to capitalize all strings, strip extra white spaces. I also discovered that there are some inconsistencies for some major manufacturers. Therefore, I mapped inconsistent names to the standard manufacturer names.

For MODEL, I found capitalization errors and formatting issues, so I capitalized all strings and stripped extra white spaces. I don’t want to generalize the models, so I only removed the common manufacturer identifiers or config tags like “PAX” or “PASSENGER” at the end of the model string.

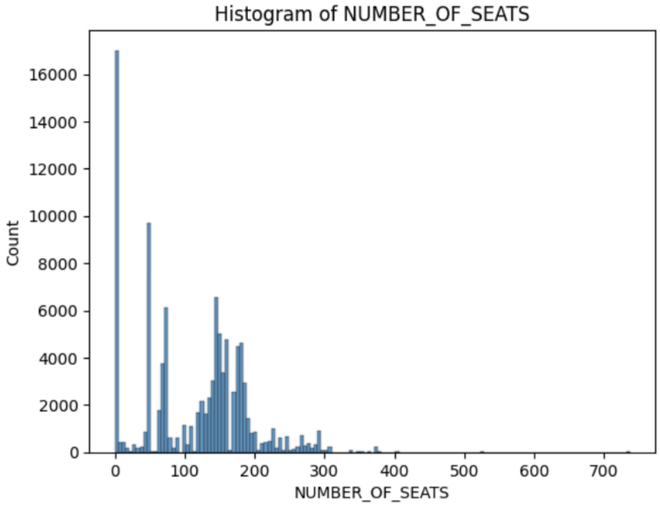
For AIRCRAFT_STATUS, I found there were only capitalization errors, so I capitalized all strings for this column.

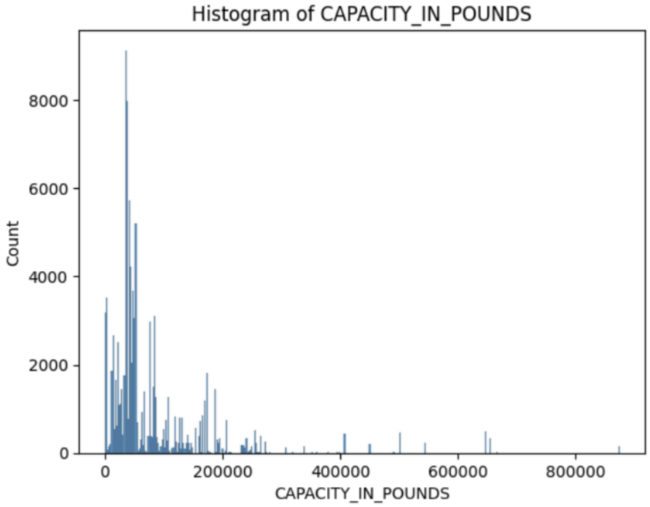
For OPERATING_STATUS, I found there were only capitalization errors, so I capitalized all strings for this column.

Task 3 Number of Remaining Rows for Cleaned Data
Remaining rows: 101316
Original number of rows: 132313

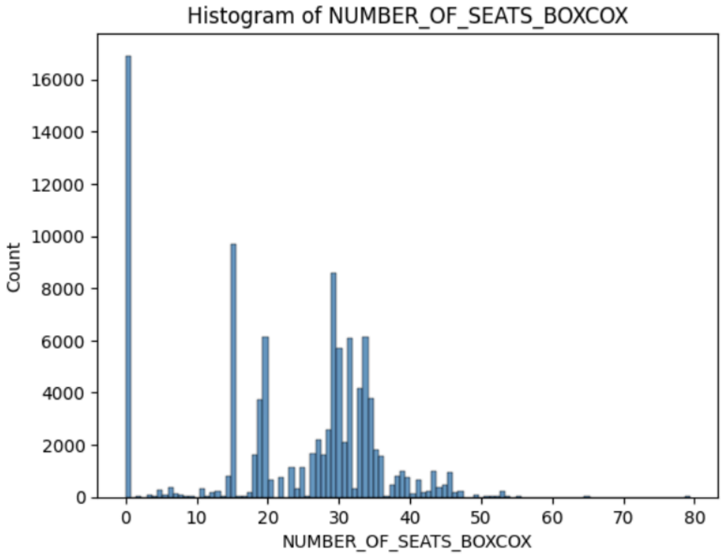
Task 4 Transformation and Derivative Variables

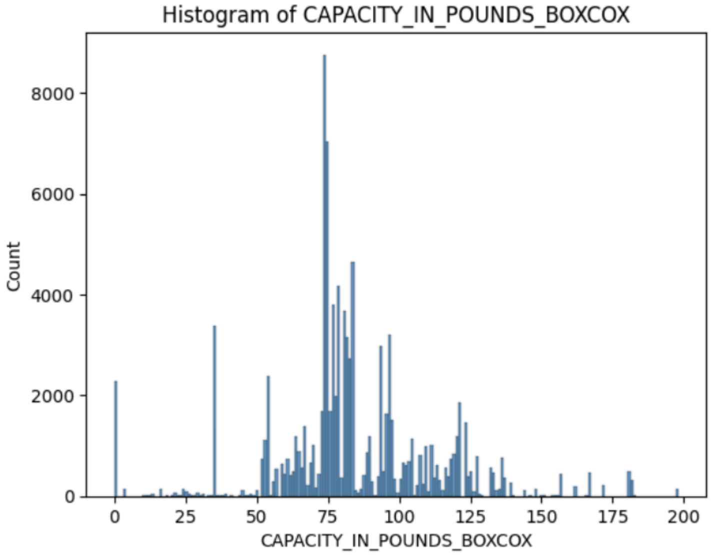
1. Skewness of NUMBER_OF_SEATS is: 0.3779649862249041
   Skewness of CAPACITY_IN_POUNDS is: 3.766052520334328





2. No output. Two columns are tasnformed by Box-Cox

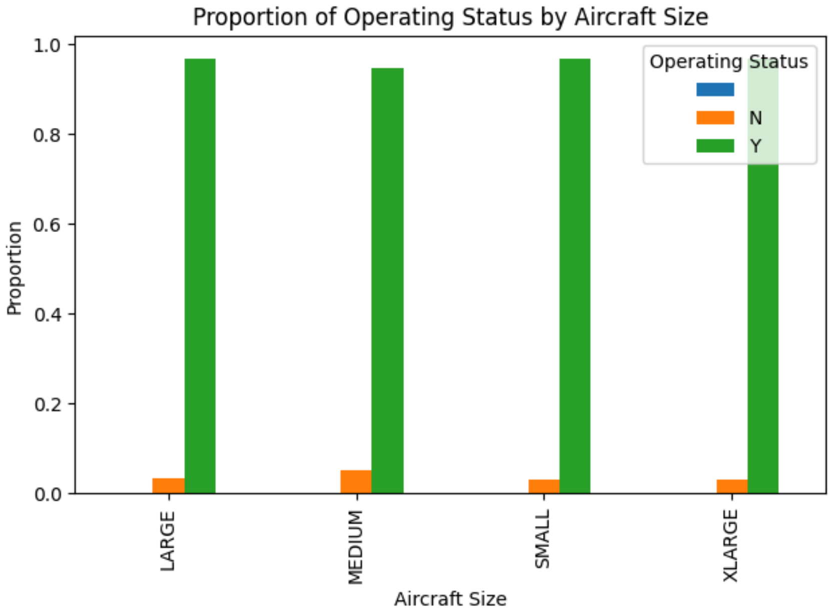




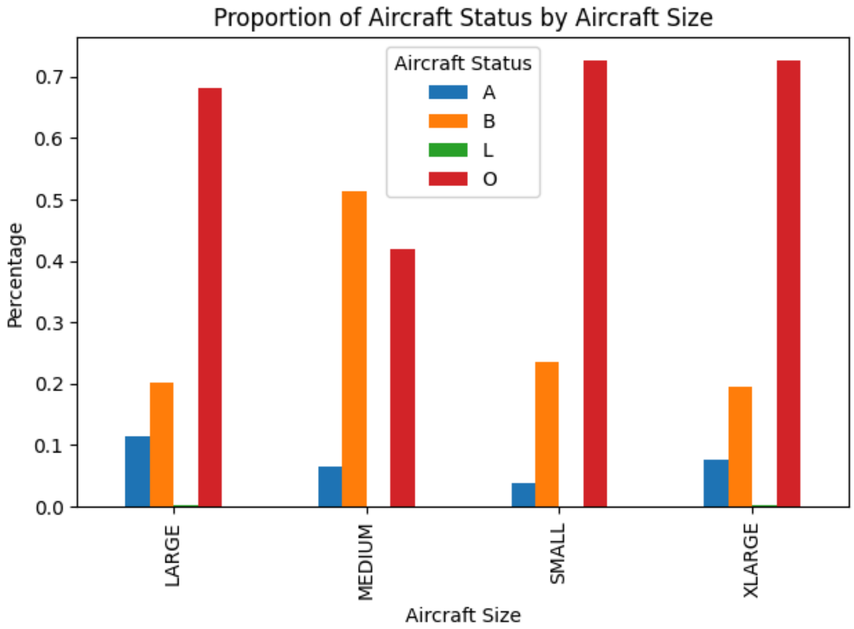
3.
4. Observation: Box-Cox transformation reduces skewness and normalizes the data. The two distributions look more symmetric and resemble a normal distribution. The transformation also adjusts the scale of the data. The x-axis numbers have changed. The transformation seems to make extreme values more compressed and alter the scale.

Task 5 Feature Engineering

1. No output, SIZE column is created based on quartiles of NUMBER_OF_SEATS



2.



3.

4. Summary of findings: most planes are in operation. Medium-sized planes have more non-operating planes in proportion. For the aircraft status, large, small, and xlarge airplanes have proportionally more "O" status. Medium-sized aircrafts have proportionally more "B" status. Status "L" is relatively less in all aircraft sizes.