Question 1:

Carrier Name and Airline ID: After verifying that all rows missing a carrier name were also missing an airline ID, I decided they could be handled in conjunction. I first found the carriers with missing carrier names, which were either OH or L4. L4 was easy to impute because it corresponded to one carrier name - Lynx Aviation d/b/a Frontier Airlines. The corresponding airline ID was also added. OH was either Comair or PSA, but only Comair enters their models with "-Passanger" which was observed in the missing rows. Thus, Comair and its airline ID was used to impute those with carrier OH.

Carrier: All carriers missing a carrier ID had carrier name North American Airlines. In 2007, the carrier was entered as NaN. In 2008, the carrier was entered as NA. In 2009, the carrier was once again entered as NaN. It's likely that NA was the intended carrier ID and was confused for NaN. All missing values were imputed to NA.

Manufacture Year: Two serial numbers missing manufacture years could be directly imputed based on other rows with the same numbers - both were made in 2004. The third serial number, 26259, was the only record with the serial number. I assumed that the carrier, Atlas Air, had some kind of sequence when assigning serial numbers, so I printed a list of all records from Atlas Air with the same model and serial number within 100 of 26259. I found that nearly all B747-400s with a similar serial number were made in 1992, so that value was used.

Number of Seats: All rows with missing number of seats were cargo planes. Other similar cargo planes had the number of seats assigned to 0 instead of NaN, so the same was done for those rows.

Capacity in Pounds: Capacity in pounds was difficult to deterministically impute because many records were missing models or did not exist elsewhere in the dataframe. Since it's a numerical value I decided to use KNN to mathematically impute the missing values. I chose to use the number of seats, manufacturer year, and capacity in pounds as the features for the KNN imputer.

Question 2:

Manufacturers: Several things had to be done to clean the data, including capitalizing all letters, removing any punctuation and empty spaces, and removing suffixes like CO, INC, and INDUSTRIES using regex. Pictured below are the top 5 manufacturers before and after cleaning.

MANUFACTURER	
BOEING	15922
Embraer	11508
THEBOEINGCO	9223
Bombardier	8871
Boeing	8392

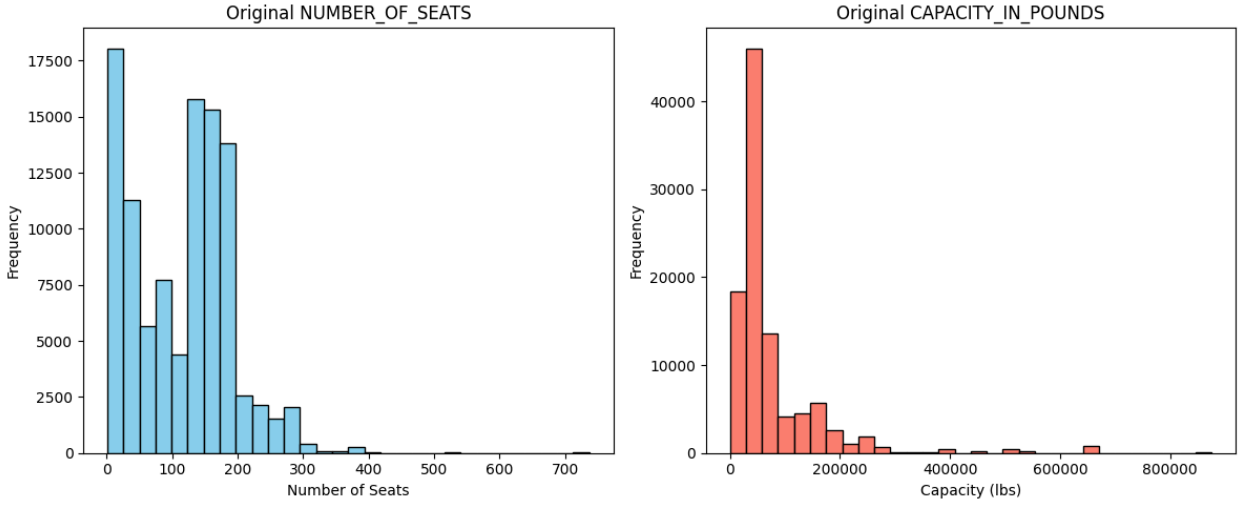
MANUFACTURER	
BOEING	55519
AIRBUS	23513
EMBRAER	15554
BOMBARDIER	11834
MCDONNELLDOUGLAS	8465

Model: The model column had fewer glaring issues, although there were several methods that different models indicated that they were passenger planes. I used regex to replace many common instances with “-PSGR”. I also made all letters upper case.

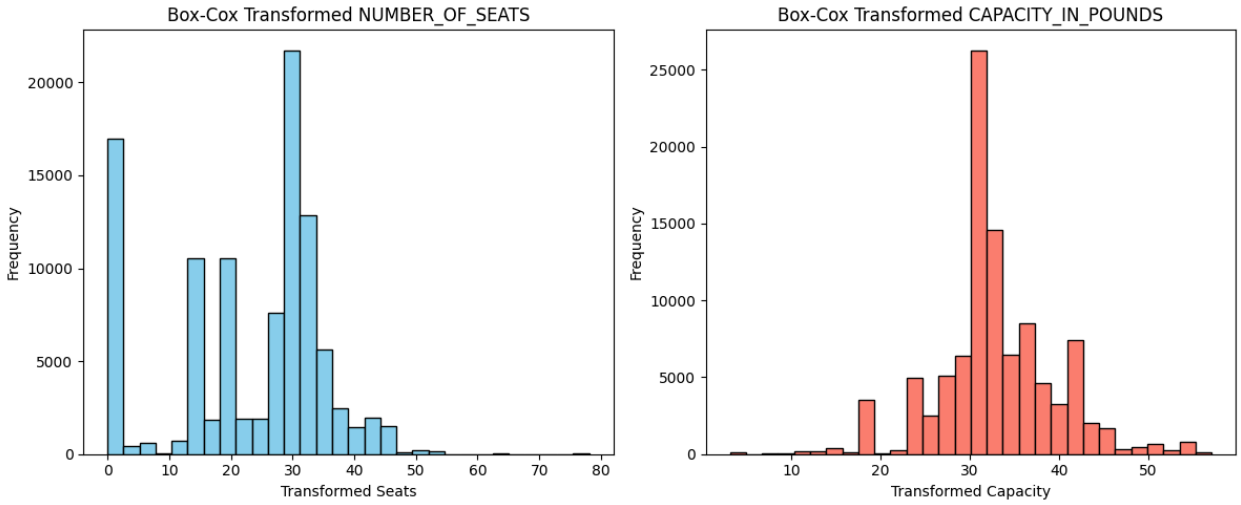
Aircraft and Operation Status: In both cases there was no need to have both upper and lower case status letters, so all were set to upper case.

Problem 3:
After removing all remaining rows with missing values, I had 101276 rows left. This is 76.5427% of the original data (with imputations).

Problem 4:
The original histograms are pictured below. Both are considerably right-skewed.



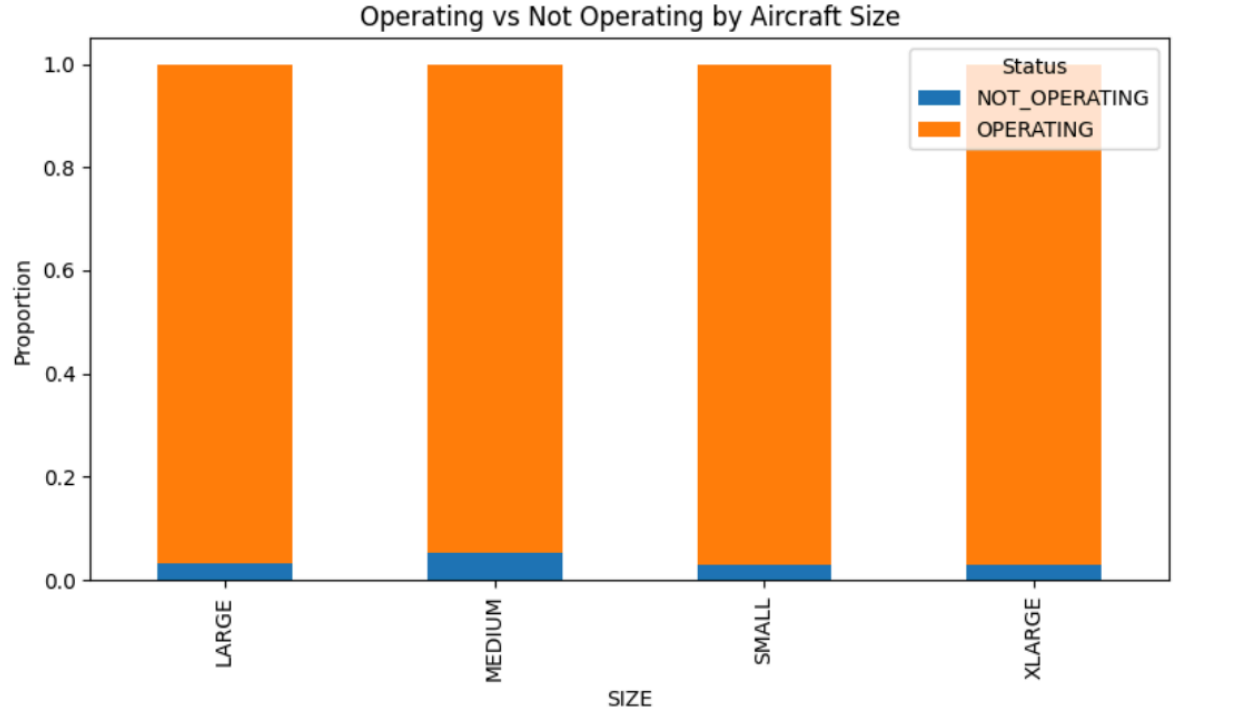
After applying the boxcox transformation, the new columns are much closer to a normal distribution (especially capacity in pounds). However, there is still a prominent peak around zero with number of seats.



Problem 5:

After adding the SIZE column which classifies each row by quartiles, I got the following information from sorting by SIZE and operating status:

```
Aircraft counts by SIZE and operating status:
        NOT_OPERATING  OPERATING
SIZE
LARGE             843       24483
MEDIUM           1619       29276
SMALL             606       19047
XLARGE            777       24625
```



From sorting by SIZE and aircraft status, I got the following information:

```
Aircraft counts by SIZE and aircraft status:
AIRCRAFT_STATUS     A      B   L      O
SIZE
LARGE            2915   5093  44  17274
MEDIUM           2044  15856  34  12961
SMALL             741   4631   0  14281
XLARGE           1937   4923  44  18498
```