DE300 Written Answers

1. **Reasoning for imputed missing values**

   CARRIER: None of the carrier values are actually missing, but are only being registered as missing because their value is “NA,” meaning the carrier is North America airlines. This was taken care of by ensuring the string type of these values.

   CARRIER_NAME: These values could be associated with the carrier, serial number, and tail number. I cross-checked those values with other existing, non-missing data, to confidently impute the missing data.

   MANUFACTURE_YEAR: Associated with serial and tail number.

   NUMBER_OF_SEATS: Associated with aircraft type and model.

   CAPACITY_IN_POUNDS: Associated with aircraft type and model.

   AIRLINE_ID: Associated with serial and tail number.

2. **Reasoning for standardization or transformation of data.**

   MANUFACTURER: Needs a lot of standardization.

   - Many of the values meant the same manufacturer but were inputted differently

   MODEL: Some standardization could be done, but generally didn’t want to mess with it too much because the models were mostly unique from looking at the data sheet

   AIRCRAFT_STATUS & OPERATING_STATUS: These values had given options based on the dataset attributes, so the standardization was straightforward and the invalid values were removed.

3. **Remaining Data**

   Rows of Data: 120866

4. **Description of Before and After**

   - the distributions for both number of seats and capacity in pounds after box cox transformation look more uniform, although there is still definitely some minor skew in both.
   - particularly, number of seats has a high frequency of lower bin values. this is explained by the 0 values that exist for many of the non-passenger aircrafts in the database
   - capacity in pounds has a much more normal distribution

5. **Written Summary of Findings**

   Operating Status:

Based on the plot of proportions for operating status, it's clear that the large majority of aircrafts in the data set are currently operating. For each size quartile, there is less than 15% of the aircrafts that are NOT operating, with aircrafts of medium size having the most aircrafts not operating.

Aircraft Status:
Based on the plot of proportions for aircraft status, the majority of aircrafts in the data set are owned (labeled "O"), with the exception of medium sized aircrafts having a larger proportion of operating leased (labeled "B"). All of the size buckets have a small (less than 17.5%) proportion of aircrafts under a capital lease.