Problem 1: Data Imputation

- Null carrier values were all actually just NA for North American, these were imputed to have carrier and unique carrier values of “NA”
- We can see CARRIER_NAME was null for carrier “OH” and “L4”. OH is the carrier for Comair, which became PSA Airlines Inc. in 2012. Because the year for each row was 2013, these were imputed to be PSA Airlines Inc. L4 rows were imputed to be Lynx Aviation/Frontier Airlines.
- Only three rows had missing manufacturing years. Based on the model of the plane, which has a very tight distribution of manufacturing years, we imputed the year to be the median of the manufacturing years for that model.
- Rows missing data on the number of seats were all cargo planes with zero seats. These were imputed to have “0” instead of null.
- Missing weight capacities were imputed by randomly sampling from the IQR of weight capacities for the given model. Each model had very tight ranges of capacities, so I think this method should be fairly accurate. It should be noted that I did this homework before we learned modern imputation techniques, but I think this should hold up fine.

Problem 2: Cleaning

- I cleaned up and consolidated all the various string formats of the major manufacturers names, accounting for 98.5% of datapoints.
- I standardized model names using regex to a general format of A320 or B747 etc.
- OPERATING_STATUS and AIRCRAFT_STATUS were all capitalized

Problem 3: Removing MIssing

- Initial rows: 132313
- Final rows: 101398
- Percentage of data retained: 76.63%
- I was unable to impute aircraft type to reduce the number of missing rows. 🙁
    - The vast majority of missing/removed rows were from this.

Problem 4: Transformations and Operationality

Boxcox worked to make the data resemble a more normal profile. However, the transformed variables still aren't very normally distributed.The data is highly skewed because, in both cases, there are large groupings of common types/models of planes that lead to concentrations of both capacities and seat counts at zero and a select number of values that correspond to popular models. Not only are seat counts inherently discrete (must be a whole number), they are centered at a few specific numbers, making the variable not well-described by a Gaussian distribution.

Most aircraft are operational. Operating status did not substantially differ depending on the plane size, and ~96% of all planes in this database are operational.

According to the BTS website, the codes are as follows: A - Capital Lease, B - Operating Lease, O - Owned. The small number of "L" entries are likely mistaken as short for "Lease" but can't be mapped to A or O. Overall, planes were most often owned (~62%). Operating leases were also common (~30%), and capital leases were least common. Small planes were least frequently capital leased aircraft and were almost entirely either owned or on operating leases. Extra large planes were most frequently owned relative to other plane sizes, followed by large planes.