Cara Nugent 04/18/2025

Data Eng 300

Homework 01

NOTE: All code is included in the google collab notebook with comments explaining.

Question 01:

a. For 'CARRIER' I imputed 'NA ' into all of the missing values because 'North American Airlines' was the only 'CARRIER_NAME' missing the 'CARRIER' field and the rest of the rows with 'North American Airlines' had 'NA ' in the 'CARRIER' column.
b. For 'CARRIER_NAME' I imputed the mode for other rows with the corresponding 'CARRIER.' I did this because there were only two distinct carriers with missing carrier names and the first one only had one unique carrier value while the second only had two unique carrier values.
c. For 'MANUFACTURE_YEAR' and 'NUMBER_OF_SEATS' I did not impute any data.
d. For 'CAPACITY_IN_POUNDS' I imputed the median capacity in pounds of the given model. I did this because the models had pretty consistent capacities and using the median accounts for a limited range like those observed in this case.
e. For 'AIRLINE_ID' I imputed the mode for other rows with the corresponding 'CARRIER.' I did this because there were only two distinct carriers with missing airline IDs and the first one only had one unique ID while the second had two unique IDs, but the mode appeared twice as often as the other value.

Question 02:

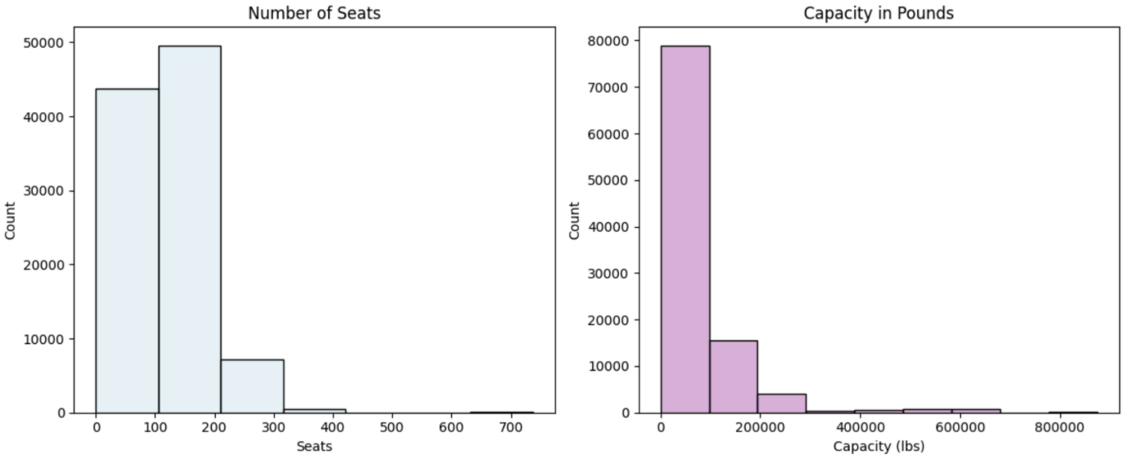
a. For 'MANUFACTURER' there were many identical names that either had trailing whitespaces or different uppercase/lowercase values so I stripped the names and made them all uppercase. I also made a mapping based on the popular manufacturer names to convert all variations of a manufacturer to the most basic version since they were all pointing to the same name.
b. For 'MODEL' I converted all of the strings to uppercase and removed any separators {-/ () } from their strings since many models had the same strings just with random separators. I also removed a bunch of trailing descriptions to convert the models to their base form since these descriptions had various spellings etc. Finally, for the most popular number models I converted models with the first chunk of letters to their base form to consolidate them as well.
c. For 'AIRCRAFT_STATUS' I made the statuses all uppercase since they had identical uppercase and lowercase versions.
d. For 'OPERATING_STATUS' I made the statuses all uppercase since they had identical uppercase and lowercase versions and also manually removed the row with " " as the value since it was not detected as null to be dropped later.

Question 03:

a. After removing the remaining rows with missing values, there are 101167 rows of data.

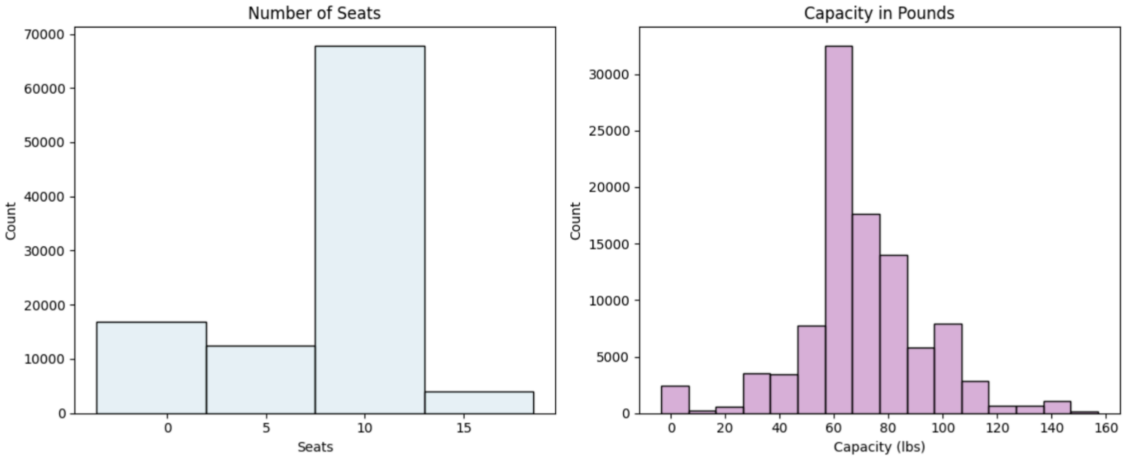
Question 04:

a. Before Transformation:



Both data frames are skewed to the left. The ranges of the data and bins are also very large before transformation.

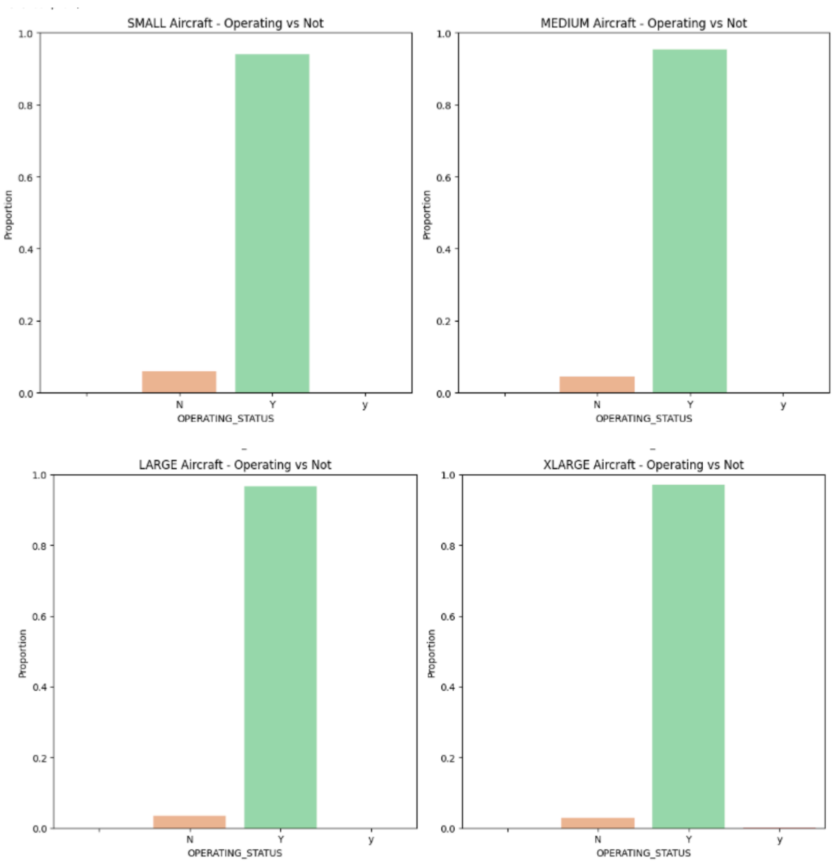
b. After Transformation:



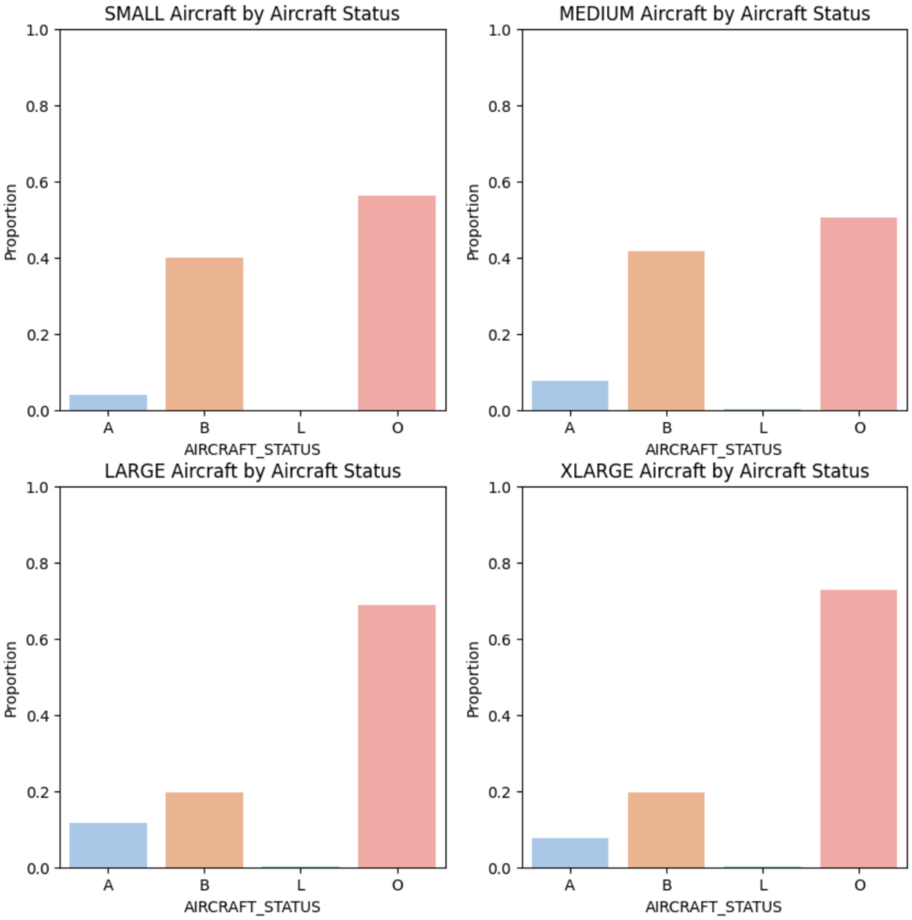
Both datasets are normally distributed now. Due to the small range for the number of seats, the data looks slightly less normal, but this is because there are only 4 bins.

Question 05

a. Based on the barcharts below, it is apparent that most of the aircraft are operating. The highest percentage of not-operating aircrafts are the small planes with 4.6%. The lowest percentage of not-operating aircrafts are the x-large planes with 3.1%. Overall, there does not appear to be any correlation between aircraft size and operating percentages.



b. Looking at the charts below, all four sizes of aircraft have the most of status ‘O’ then ‘B’ then ‘A’ and finally almost no aircrafts of status ‘L.’ Both medium and small aircrafts have significantly larger proportions of ‘B’ and smaller proportions of ‘O’ than large and x-large aircrafts. Large aircrafts also have the highest percentage of ‘A’ aircrafts while small aircrafts have the lowest percentage of this type.


SMALL Aircraft by Aircraft Status
MEDIUM Aircraft by Aircraft Status
LARGE Aircraft by Aircraft Status
XLARGE Aircraft by Aircraft Status
Proportion
A
B
L
O
AIRCRAFT_STATUS