Katie Mowry

DATA ENG 300 Homework 1 Responses

1a. CARRIER

Constant imputation with NA is possible.

Based on the table of carriers and carrier names, it appears that ‘NaN’ corresponds to North American Airlines and isn’t actually a missing value. I found that there are 10 other rows where North American Airlines is referred to with a carrier of ‘NA.’ Since there are 59 rows with the carrier name as ‘North American Airlines,’ 49 rows where the carrier is missing, and 10 rows with ‘NA’ as the carrier corresponding to ‘North American Airlines,’ I can assume that the missing values can be replaced with NA through simple imputation with a constant of ‘NA.’ After replacing this value, I realized that the carrier was actually listed with a space after NA and realized that this was not standardized across the carrier values. I then removed the spaces from the ends of all the carrier names and confirmed that there were 59 rows with NA, matching the 59 rows with North American Airlines as the carrier name.

1b. CARRIER_NAME

Constant imputation with Comair Inc. for when the Carrier was OH. Constant imputation with ‘Lynx Aviation d/b/a Frontier Airlines’ for when the Carrier was L4.

It seemed like all of the rows with null Carrier Names had listed Carrier values and the only missing Carrier Names were when the Carrier was OH or L4. Upon inspection of the data with OH as the Carrier, it was clear that all of the missing values were from the year 2013. After inspecting the data for OH Carrier around the year 2013, it appeared that the years prior to and including 2011 were labeled as Comair Inc., and the years after and including 2015 were labeled as PSA Airlines Inc. Upon further inspection, the model was the same from 2011 through 2013 ("CRJ100-Passanger") and switched to "CRJ701" once the carrier name changed in 2015. Therefore, it seems most likely that the carrier name during 2013 was still Comair Inc.. Simple Imputation with a constant was used to replace the missing values with Comair Inc.

The Carrier L4 only mapped to one Carrier Name: ‘Lynx Aviation d/b/a Frontier Airlines.’ All remaining NaN values mapped to L4, so all values were imputed with ‘Lynx Aviation d/b/a Frontier Airlines.’

1c. MANUFACTURE_YEAR

Constant imputation with the mode year was used.

There were only three instances where the manufacture year was missing, and two of those instances were the same. To determine the appropriate manufacture year, other rows with the same manufacturer and model were printed, which showed that those values also had the same number of seats, carrier, and capacity. Therefore, the mode (2003) was used for constant imputation since that is the value that is most likely the correct manufacture year.

For the last missing manufacture year, there were 283 rows with the same model, aircraft type, and carrier. The mode was used again for imputation since it represents the most common and therefore most likely manufacture year for this aircraft.

1d. NUMBER_OF_SEATS

Constant imputation with 0.0.

There are seven instances where the number of seats were missing with the same manufacturer and carrier name, with very similar aircraft types. When filtering by manufacturer, aircraft type, and carrier name, all other similar aircrafts had a seat number of 0.0. This makes sense because the aircrafts have CARGO at the end of the name, which indicates that there probably wouldn't be passenger seats. Therefore, all missing values for the number of seats was replaced with 0.0.

1e. CAPACITY_IN_POUNDS

Constant imputation with the median value given the model, manufacturer, and aircraft type.

After inspecting the rows with missing capacity values, I found that there were 6 different unique instances when grouped by manufacturer, aircraft type, and model, which are likely the most indicative of the capacity. When looking at the fist aircraft (AIRCRAFT_TYPE = 6251), there was one outlier of 0.0, while the other capacity values were very similar. To avoid the outlier skewing the data too much, I used the median capacity because it best reflects the expected capacity of the aircraft given the specific model, manufacturer, and aircraft type. I looked at the next aircraft and found that the values given the same grouping were the same, so I used the median value again. The next aircraft type had a range of values, so the median was used again. This was repeated for the remaining 3 missing capacities.

1f. AIRLINE_ID

Constant imputation with 20417.0 for when the Carrier was OH. Constant imputation with 21217.0 for when the Carrier was L4.

It appeared that the rows that are missing the Airline ID are all in 2013, indicating that they are the same values that were missing for the carrier name. The years prior to and including 2011 are labeled as 20417.0, and the years after and including 2015 are labeled as 20397.0 for Airline_ID. From the previous inspection of the data, since the model stayed the same from 2011 through 2013 and changed in 2015 when the carrier name/airline ID changed, it seems likely that the airline ID was the same as before 2011 (20417.0). Therefore, simple imputation with a constant is used to replace the missing value for when the carries is OH. L4 only maps to one Airline_ID, so simple imputation was used to replace all the missing values with L4 as the carrier with 21217.0.

2a. MANUFACTURER

I standardized the values by converting everything to uppercase and removed any whitespace to ensure consistency.

First, I normalized the casing and got rid of any whitespaces, which helped combine some of the repeating manufacturers. From there, there were already visible duplicated (ex. BOEING, BOEINGCO, BOEINGCOMPANY). I placed all manufacturers in alphabetical order and combined any manufacturers that I assumed were referring to the same thing. I also fixed a few that looked like typos to further standardize the data. This shortened the number of manufacturers from 183 to 94.

2b. MODEL

I standardized the values by converting everything to uppercase and removed any whitespace to ensure consistency.

However, when inspecting models of similar structure, it doesn't seem like the values can be grouped just by the first string (ex. A319) because the different models have different capacities and aircraft types, indicating that they are unique. This was checked with two sets of similar model names, and both had differences between number of seats, capacity, aircraft type, and tail number for the slightly different model names. Therefore, since the values seem unique and I can’t confidently determine if there are typos, I’ve decided to leave the model values unchanged.

2c. AIRCRAFT_STATUS

I standardized the values by converting everything to uppercase and removed any whitespace to ensure consistency.

From the documentation, there are only aircraft statuses of A, B, and O, so any differences in casing were standardized. However, there were still 122 rows where the AIRCRAFT_STATUS was listed as ‘L.’ After inspecting the data, there is no clear difference between aircraft statuses listed as A, B, O, or L. There seems to be no pattern, so this is likely a mistake from the airline. Since it isn’t clear what L is referring to, I am going to leave the values as is, but will note this for any future analysis (ex. L was not included in #5 because it isn’t clear what the meaning is).

2d. OPERATING_STATUS

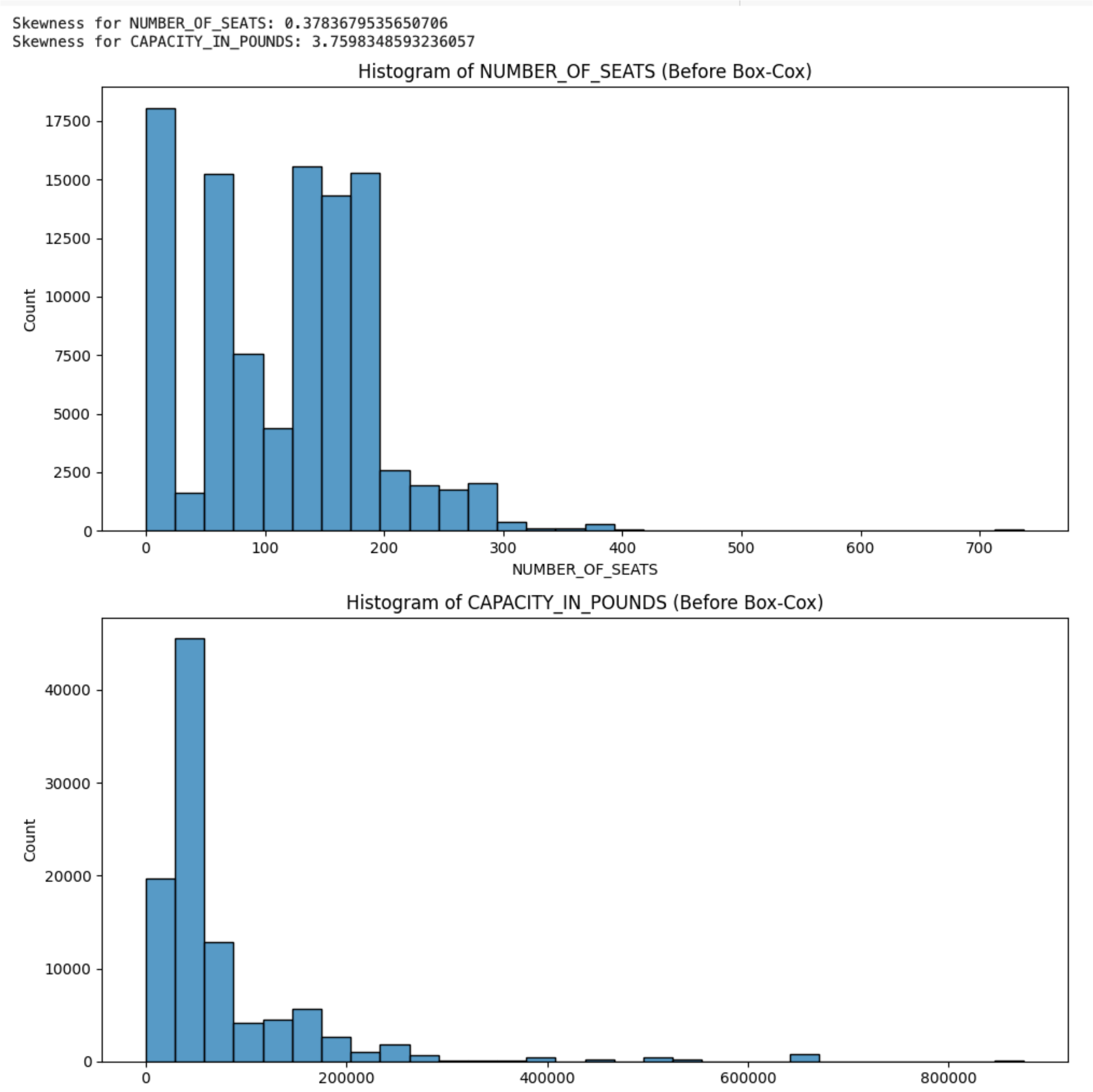
I standardized the values by converting everything to uppercase and removed any whitespace to ensure consistency.

Based on the documentation of the dataset, only Y and N are approved values for the operating status. Therefore, the casing was standardized so that any places where there was a ‘y’ was converted to a ‘Y.’ There was one instance where the operating status was missing and similar aircrafts with the same manufacturer and aircraft type were inspected to determine if that value could be replaced. Since 3029 out of the 3033 similar aircrafts have an operating status of Y and those with N have a different aircraft status and capacity than the one with the missing operating status, I am replacing the blank operating status with a Y.

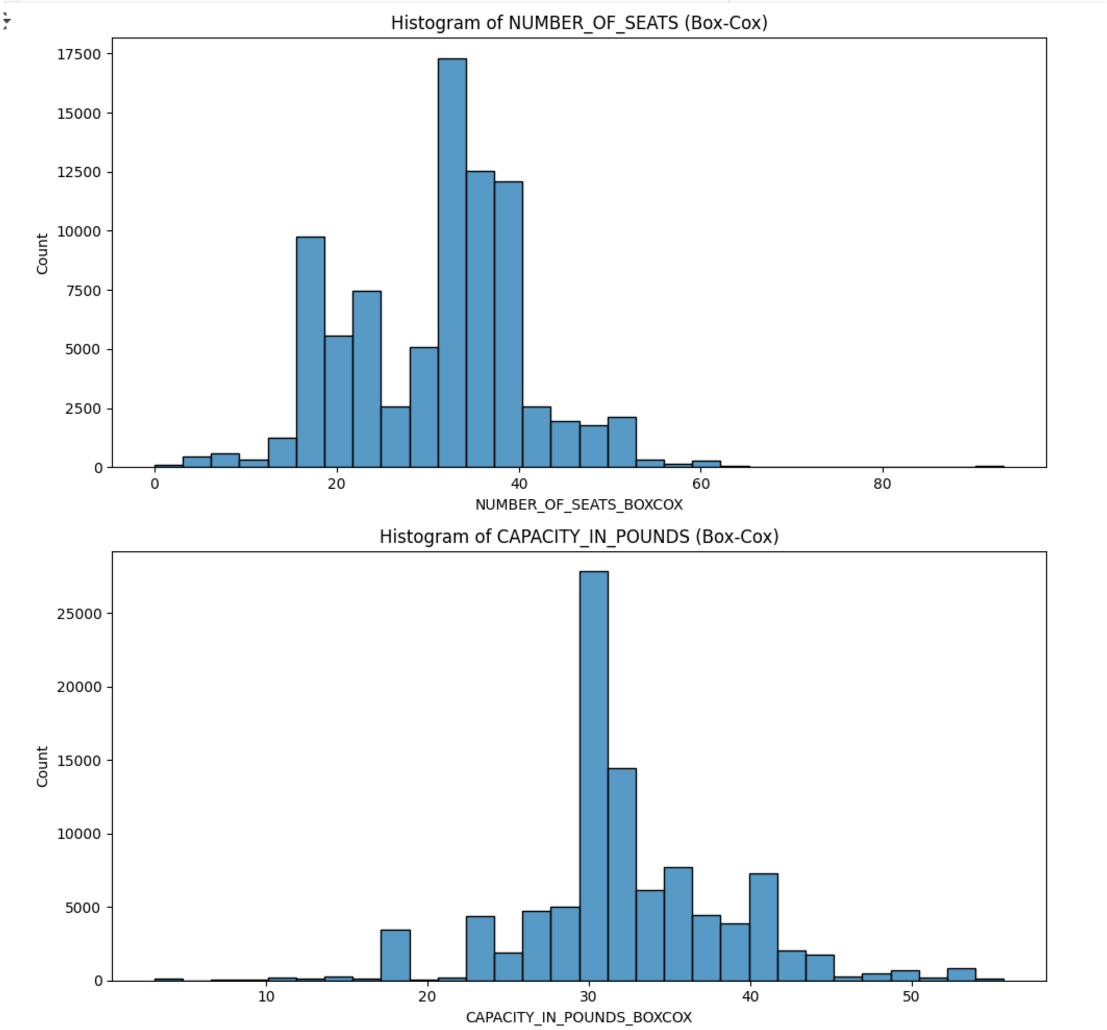
3. There are 101,276 rows after dropping the remaining missing values.

4. Skewness for NUMBER_OF_SEATS: 0.3783679535650706

Skewness for CAPACITY_IN_POUNDS: 3.7598348593236057



```
Skewness for NUMBER_OF_SEATS: 0.3783679535650706
Skewness for CAPACITY_IN_POUNDS: 3.7598348593236057
```



Before the Box-Cox transformation, the data was very skewed, especially for the CAPACITY_IN_POUNDS plot, which was also depicted by the skewness value. Both plots were skewed to the right before transformation. After the transformation, the distributions look much closer to normal and are centered around similar values.

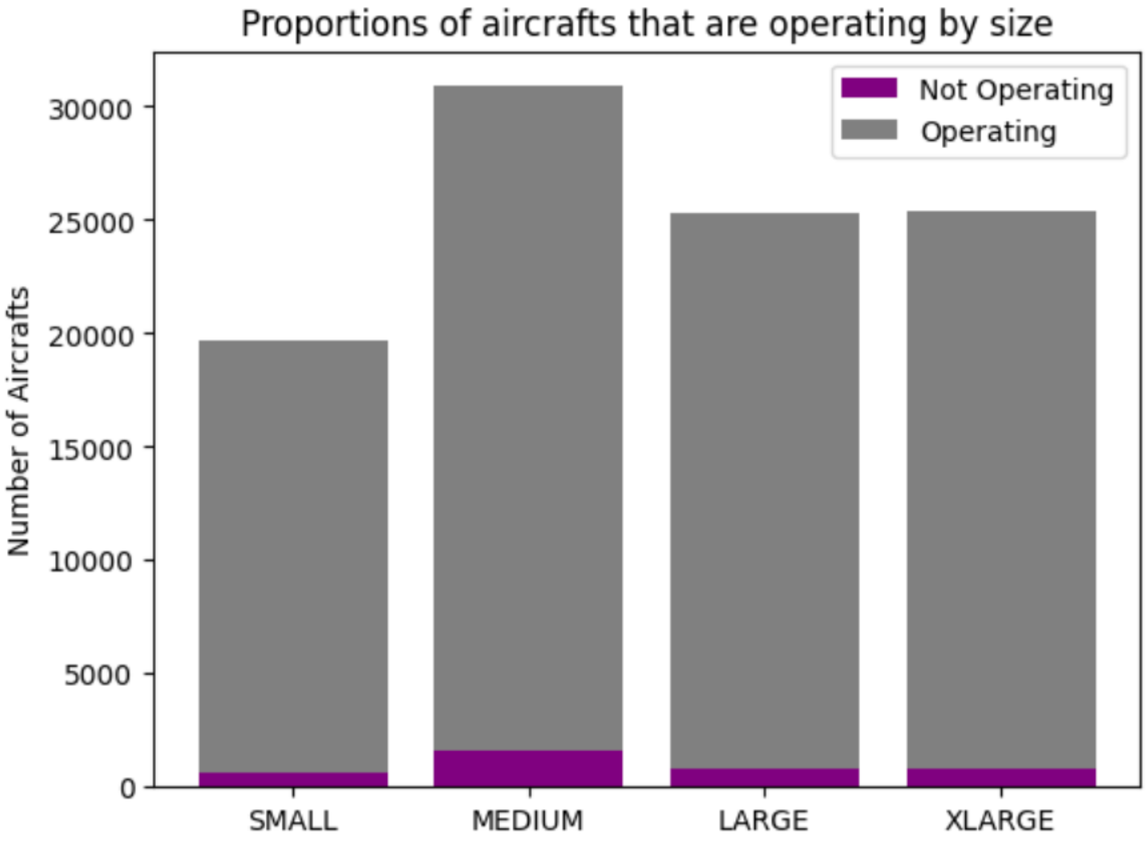
5a. Proportion of aircrafts that are operating:

SMALL: 0.9691650129751183

MEDIUM: 0.9476290661919404

LARGE: 0.966714048803601

XLARGE: 0.9694118573340682



This shows that almost all of the aircrafts are operating across all sizes because all sizes have proportions of over 0.94. Medium-sized aircrafts have a slightly lower operating rate, but the small, large, and xlarge sizes are all very similar. There is the greatest number of medium sized aircraft, and slightly less small aircrafts total than the other sizes. Overall, the proportions of aircrafts is relatively consistent across different sizes.

5b. Proportions of aircrafts belonging to each aircraft status group:

SMALL:

A: 0.037704167302701874

B: 0.23563832493766854

O: 0.7266575077596296

MEDIUM:

A: 0.066232461683030‌36

B: 0.513787628398302

O: 0.41997990991866757

LARGE:

A: 0.11529942251404161

B: 0.20144767027925006

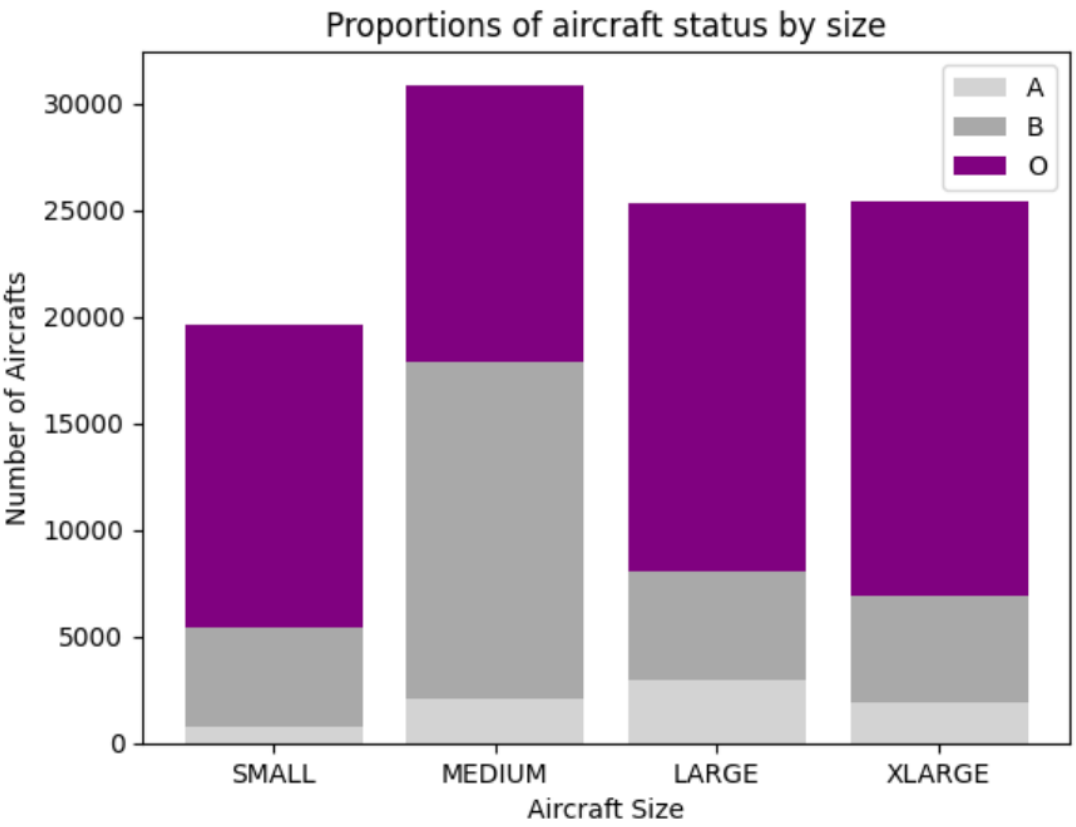
O: 0.6832529072067083

XLARGE:

A: 0.07638615032731289

B: 0.19413991639719222

O: 0.7294739332754949



This shows that there is the smallest proportion of status A among all sizes. Aircraft status of O is the most common across all sizes and is relatively consistent, with the exception of the medium aircrafts, where it is slightly lower. The medium aircrafts have the largest proportion of status B of 0.51, while the other sizes are relatively similar with a proportion of around 0.20. The stacked bar chart helps visualize these differences effectively.