DATA_ENG 300
Homework 1 Written Answers
Nayeon Kim

Question 1: Investigate missing data

CARRIER: I found that all the rows with a missing entry under the ‘CARRIER’ column are under North American Airlines. Because NA is the respective carrier for North American Airlines, we can manually impute this into the data. After imputing this data, we are left with zero null entries under the ‘CARRIER’ column.

CARRIER_NAME: I found that L4 and OH are the only carriers that show up for the entries with a missing carrier name. Because Lynx Aviation d/b/a Frontier Airline is the only carrier name that shows up for the carrier L4, we can manually impute any missing entries under ‘CARRIER’ that have a Lynx Aviation d/b/a Frontier Airline carrier name. After analyzing the data, we see that the carrier OH switches from Comair Inc to PSA Airlines after 2011. We can manually impute the respective carrier name depending on the year. After imputing this data, we are left with zero null entries under the ’CARRIER_NAME’ column.

MANUFACTURE_YEAR: For ‘MANUFACTURE_YEAR’, it is not possible to accurately impute any missing values because there is no one-to-one correspondence with any specific columns nor set of columns.

NUMBER_OF_SEATS: All the missing data found under the ‘NUMBER_OF_SEATS’ column are cargo planes. We can fill in these missing entries as "0" because cargo planes do not carry seats for passengers.

CAPACITY_IN_POUNDS: It is not possible to impute the exact missing values for ‘CAPACITY_IN_POUNDS’. However, one possible method is taking the mean weight of the specific model and imputing it. After imputing this data, we are left with zero null entries under the ‘CAPACITY_IN_POUNDS’ column.

AIRLINE_ID: I found that Lynx Aviation d/b/a Frontier Airlines and PSA Airlines Inc. are the only carrier names that show up for the entries with a missing airline ID. Seeing that Lynx Aviation d/b/a Frontier Airlines has a unique ID of 21217 and that PSA Airlines Inc. has a unique ID of 20397, we can manually input the respective missing IDs into the data. After imputing this data, we are left with zero null entries under the ‘AIRLINE_ID’ column.

Question 2: Cleaning the data through standardization and transformation

MANUFACTURER: First, I standardized all the data to get rid of any extra whitespace and make it all uppercase. I also removed periods and hyphens. This made it easier for the following transformations that need to be made and helped reduce some of the repeated manufacturers. To address the different variations of the same manufacturer, I transformed these variations into

a standardized version of that manufacturer to minimize overcounting the number of actual unique manufacturers. For example, anything containing “BOE” was substituted with “BOEING.” After these transformations, the number of unique manufacturers was reduced from 183 to 98.

MANUFACTURER	count
BOEING	55640
EMBRAER	15554
AIRBUS	13440
BOMBARDIER	12483
DOUGLAS	10812
...	...
SAABSCANIA	1
ROLLSROYCE	1
CSSNACITATIONX	1
AMA/EXPR	1
ISRAELAIRCRAFTINDUSTRIES	1

98 rows × 1 columns

MODEL: First, I standardized all the data to get rid of any extra whitespace and make it all uppercase. I also removed periods, hyphens, and parentheses. This made it easier for the following transformations that need to be made and helped reduce some of the repeated models. To address the different variations of the same model, I transformed these variations into a standardized version of that model to minimize overcounting the number of actual unique models. I also removed extra words from the model that I didn't think were necessary for classification (e.g. passenger, only). After this standardization and transformations, the number of unique models was reduced from 1340 to 939.

MODEL	count
CRJ200	3344
B737823	3326
EMB145	3064
B7377H4	2474
A320232	2463
...	...
C340/335	1
757200SIES	1
767281	1
7572S7	1
75727B	1

939 rows × 1 columns

AIRCRAFT_STATUS: Currently, ‘AIRCRAFT_STATUS’ contains uppercase and lowercase variations of the letters O, B, A, and L. To standardize the data, we can make all the entries capitalized. This is the only transformation needed because there are no other types of variations between the letters.

AIRCRAFT_STATUS	count
O	79506
B	43551
A	9134
L	122

OPERATING_STATUS: Currently, ‘OPERATING_STATUS’ contains uppercase and lowercase variations of Y and N along with a single blank entry. To standardize the data, we can make all the entries capitalized. This is the only transformation that needs to be made because there are no other variations within the data. We do not know what the operating status is for the empty row so we can leave it.

OPERATING_STATUS	count
Y	126648
N	5664
	1

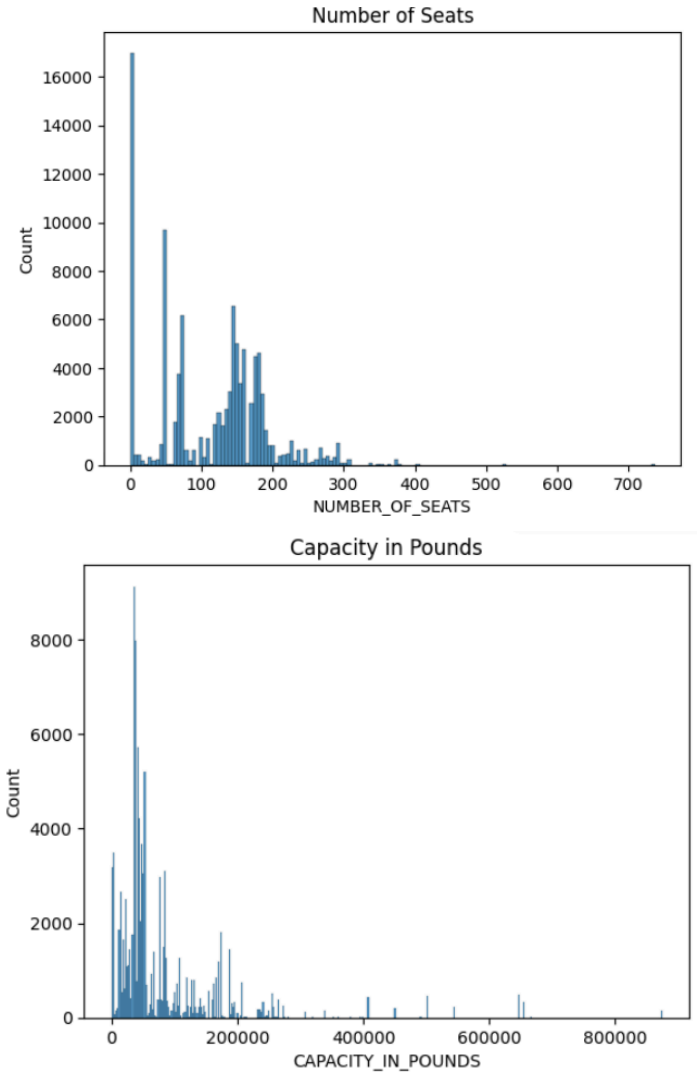
Question 3: Remove missing data

After removing the missing entries from the cleaned dataset, we are left with 101274 non-null entries. We initially started with 132313 entries from the cleaned dataset.

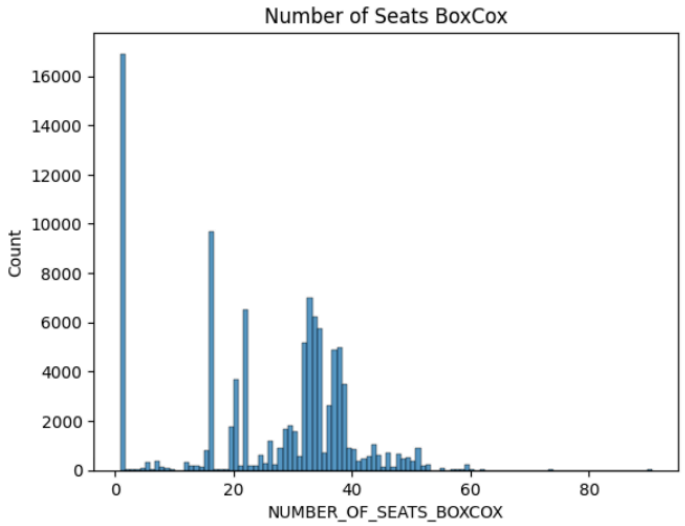
```
Data columns (total 17 columns):
 #   Column               Non-Null Count   Dtype
---  ------               --------------   -----
 0   YEAR                 101274 non-null  int64
 1   CARRIER              101274 non-null  object
 2   CARRIER_NAME         101274 non-null  object
 3   MANUFACTURE_YEAR     101274 non-null  float64
 4   UNIQUE_CARRIER_NAME  101274 non-null  object
 5   SERIAL_NUMBER        101274 non-null  object
 6   TAIL_NUMBER          101274 non-null  object
 7   AIRCRAFT_STATUS      101274 non-null  object
 8   OPERATING_STATUS     101274 non-null  object
 9   NUMBER_OF_SEATS      101274 non-null  float64
 10  MANUFACTURER         101274 non-null  object
 11  AIRCRAFT_TYPE        101274 non-null  object
 12  MODEL                101274 non-null  object
 13  CAPACITY_IN_POUNDS   101274 non-null  float64
 14  ACQUISITION_DATE     101274 non-null  object
 15  AIRLINE_ID           101274 non-null  float64
 16  UNIQUE_CARRIER       101274 non-null  object
dtypes: float64(4), int64(1), object(12)
memory usage: 13.9+ MB
```

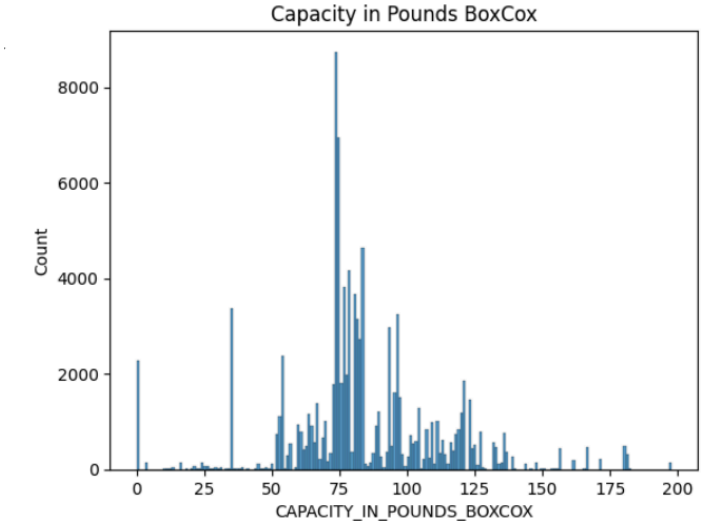
Question 4: Transformation and derivative variables

Before the transformation, ‘NUMBER_OF_SEATS’ had a skewness of around 0.3783 while ‘CAPACITY_IN_POUNDS’ had a skewness of around 3.7610. The histogram for these two columns before the transformation are shown below.



After the Box-Cox transformation, ‘NUMBER_OF_SEATS’ had a skewness of around -0.4531 while ‘CAPACITY_IN_POUNDS’ had a skewness of around 0.1901 The histogram for these two columns after the transformation are shown below.





Before the transformation of NUMBER_OF_SEATS, the histogram depicted a positively skewed distribution. Specifically, there was a large concentration of aircrafts with 0 seats, depicted by cargo planes. There was also a long tail extending towards the aircrafts with higher seat counts. After the Box-Cox transformation, there still is a large concentration of aircrafts with 1 seat (all entries were shifted by 1 to account for entries with 0 when performing the transformation). However, the distribution is more symmetric and bell-shaped compared to the original distribution. This can be seen from how the skew decreased from 0.3783 to -0.4531.

The original histogram of CAPACITY_IN_POUNDS also depicts a positive skew with most of the aircrafts having lower capacity and fewer aircrafts with greater capacity. After the transformation, the distribution is more symmetric and bell-shaped compared to the original distribution. There appears to be less of a skewed tail with a distribution that looks closer to normal. This can be seen from how the skew decreased from 3.7610 to 0.1901.

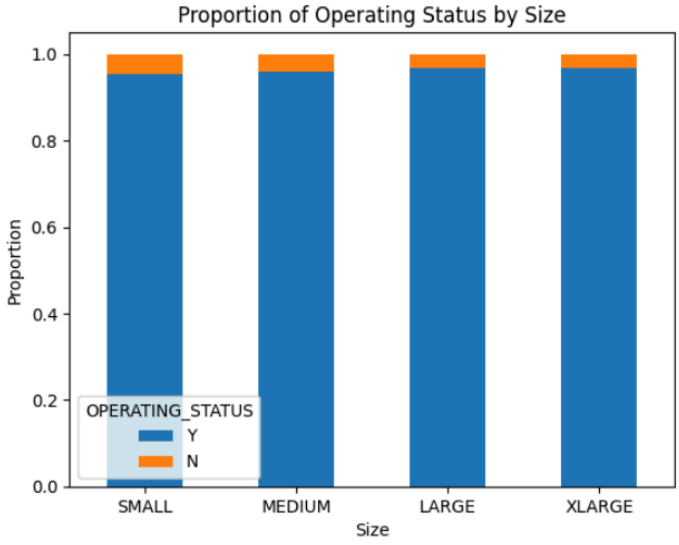
Question 5: Feature engineering

After dividing up the entries in 'NUMBER_OF_SEATS', we obtained the following counts for each aircraft size:

SIZE	count
SMALL	29317
LARGE	25687
XLARGE	24909
MEDIUM	21361

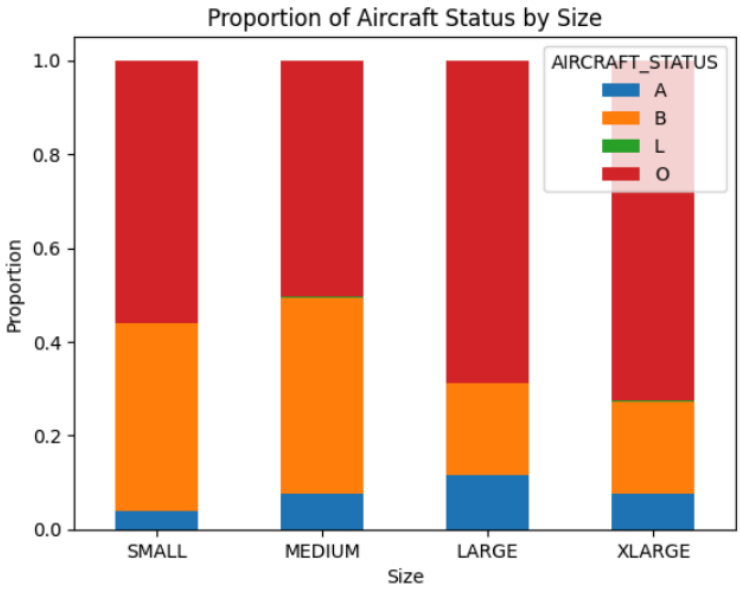
Results for the proportion of operating status by size:

OPERATING_STATUS	Y	N
SIZE		
SMALL	0.953304	0.046696
MEDIUM	0.959599	0.040401
LARGE	0.967493	0.032507
XLARGE	0.968806	0.031194



Results for the proportion of aircraft status by size:

AIRCRAFT_STATUS	A	B	L	O
SIZE				
SMALL	0.039329	0.398847	0.000000	0.561824
MEDIUM	0.076401	0.417115	0.001592	0.504892
LARGE	0.115545	0.195040	0.001869	0.687546
XLARGE	0.075635	0.196234	0.001606	0.726525



After dividing up the data into the four size groups using quartile-based binning, I found that the Small category had the largest number of aircrafts (29,317). This was then followed by Large (25,687), X-Large (24,909), and Medium (21,361).

When analyzing the operating status across each size group, I found that a high proportion of aircrafts were operational (OPERATING_STATUS = 'Y') across all four groups. Specifically, 95.3% of Small aircrafts were operational, 95.9% for Medium, 96.7% for Large, and 96.8% for X-Large aircrafts.

The aircraft status across the four size groups had more variation. Status A, was more common in larger aircrafts, with 3.9% in Small to 11.6% in Large. Status B was most prevalent in Small and Medium aircraft, at approximately 39.9% and 41.7% respectively, but dropped significantly in Large and X-Large categories to around 19.5%. Status L was rare in all size categories, appearing in less than 0.2% of aircraft. Status O was highest among the largest aircraft, reaching 72.7% in the X-Large group and 68.7% in the Large group.