Data_Eng 300 Homework 1
Jonathan Wu
4/18/2025

Data_Eng 300 Homework 1 Written Answers

1. For the columns: CARRIER, CARRIER_NAME, MANUFACTURE_YEAR, NUMBER_OF_SEATS, CAPACITY_IN_POUNDS, and AIRLINE_ID:
    a. Carrier: the carriers with the tag “NA” have a carrier name of “North American Airlines.” It reasons that the NA abbreviation was misinterpreted. We can change the tag to “NA “ and it will fill the other North American Airlines carriers.
    b. Carrier Name: There are two missing carrier names with the carrier tags, L4 and OH. The “L4” tag is simple. The L4 tags that are filled in carrier name columns are dubbed Lynx Aviation. We can fill that in. For OH tags, there are two possible tags, Comair and PSA Airlines. When we check the year, we see that there is a split in between 2013 and 2015. Subsetting on this, we see that the carrier before 2014 with the OH tag is Comair and after 2014 is PSA Airlines. There is no overlap, so we can change the names for both.
    c. Manufacture Year: We see there is an Atlas Air carrier plane and two Endeavor Air planes with no manufacture years. Based on the surrounding Atlas Air planes, we can’t make any conclusions for that plane. If we looks for the Endeavor Air planes, we can check the serial numbers. Unfortunately, when we connect the serial numbers, there are jumps where these planes should be, and they would fit into two different manufacturing years. For example, a missing serial number is 10134. This is in between serial numbers 10112 and 10153. The manufacturing years are 2003 and 2004. We can’t impute this.
    d. Number of Seats: The missing planes here are cargo planes. These planes don’t carry passengers, so we can impute 0.
    e. Capacity in Lbs: We can’t do anything here. When we split based on manufacturer and by carrier, there are not any common capacity. We’ll leave this untouched
    f. Airline ID: We see the missing airline ids are carriers OH and L4. Using the same strategy, we can replace L4 with 21217. We replaced OH carriers with 20417 for Comair and 20397 for PSA airlines.
    g. We notice that the unique carriers are only L4, OH, and NA. These are the carriers we had issues with. We can drop this column, it doesn’t really help us.
2. For the columns: MANUFACTURER, MODEL, AIRCRAFT_STATUS, and OPERATING_STATUS:
    a. Manufacturer: We need to use transformation and standardization to clean this column up. First, we can standardize everything to lowercase, which blends many different values together.

```
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].str.lower()
```

Afterwards, we can transform many of the common manufacturers such as Boeing, Airbus, Bombardier, Cessna, and Mcdonnell Douglas. This is done by looking for an instance of “boeing” and replacing the column’s value with “Boeing.”

```
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "boeing" if isinstance(x, str) and "boeing" in x.lower() else x
)
```

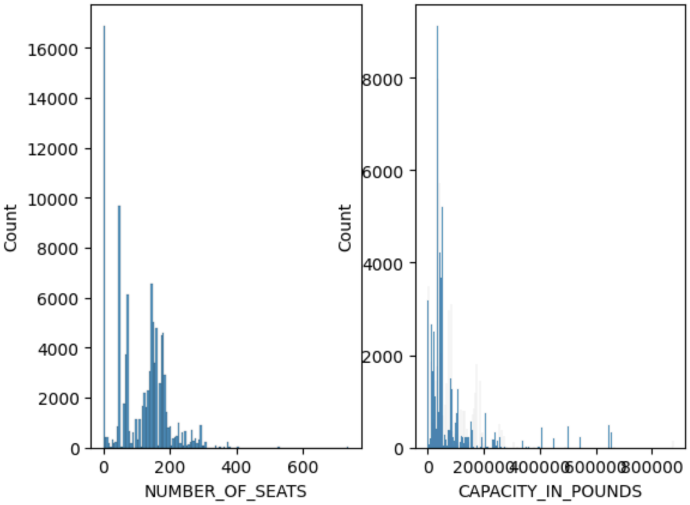
```
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "airbus" if isinstance(x, str) and "airbus" in x.lower() else x
)
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "mcdonnelldouglas" if isinstance(x, str) and "mcdonnel" in x.lower() else x
)
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "mcdonnelldouglas" if isinstance(x, str) and "dougla" in x.lower() else x
)
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "embraer" if isinstance(x, str) and "embrae" in x.lower() else x
)
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "bombardier" if isinstance(x, str) and "bombardier" in x.lower() else x
)
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "cessna" if isinstance(x, str) and "cessna" in x.lower() else x
)
inventory["MANUFACTURER"] = inventory["MANUFACTURER"].apply(
    lambda x: "canadair" if isinstance(x, str) and "canada" in x.lower() else x
)
```

b. Model: We can both standardize and transform the data. First we can set everything to uppercase. Afterwards, we can remove dashes and slashes (- and /) with nothing. This drops the number of unique values from 1342 to 1200. Given how much complexity there is with different plane types and configurations, I can't really do much better without having additional institutional knowledge (ie what's the difference between B737XXX and B737XXX)

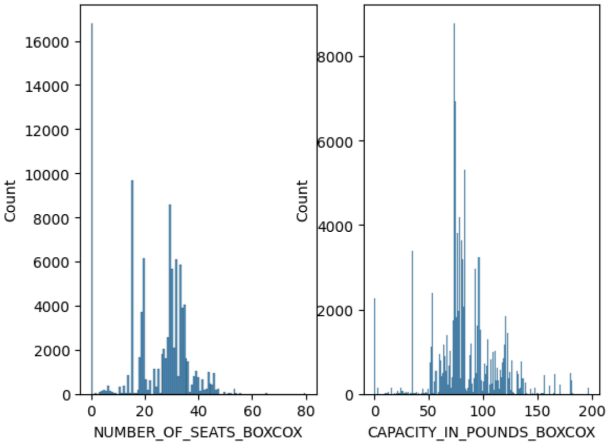
```
inventory["MODEL"]=inventory["MODEL"].str.replace(r"[-/]", "", regex = True)
inventory["MODEL"]=inventory["MODEL"].str.title()
inventory["MODEL"].value_counts().head(10)
```

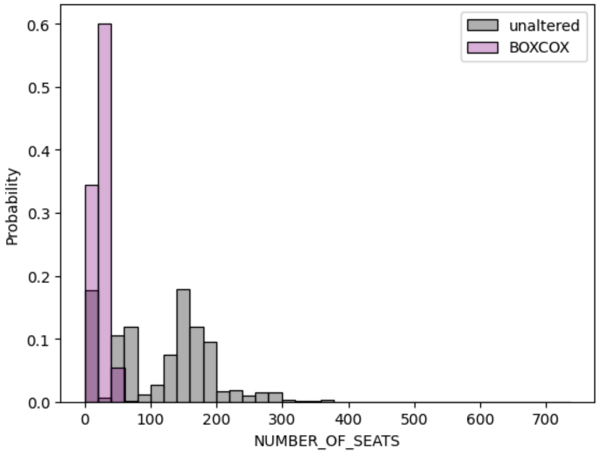
c. Aircraft and Operating Status: These can be mostly handled with standardization. We can change everything to capital letters. However, there is a status of 21217. On the BTS website, this is clearly not allowed, so we drop these rows.

3. After removing the data rows that still have missing values, we are left with 101,203 rows leftover. This means we dropped 31,013 rows.
4. For the columns, there is a skew of 3.766 in capacity in pounds and 0.3778 in number of seats by the Fisher-Pearson coefficient of skewness. This is done using the pandas .skew() method. The variables are distributed as shown below.

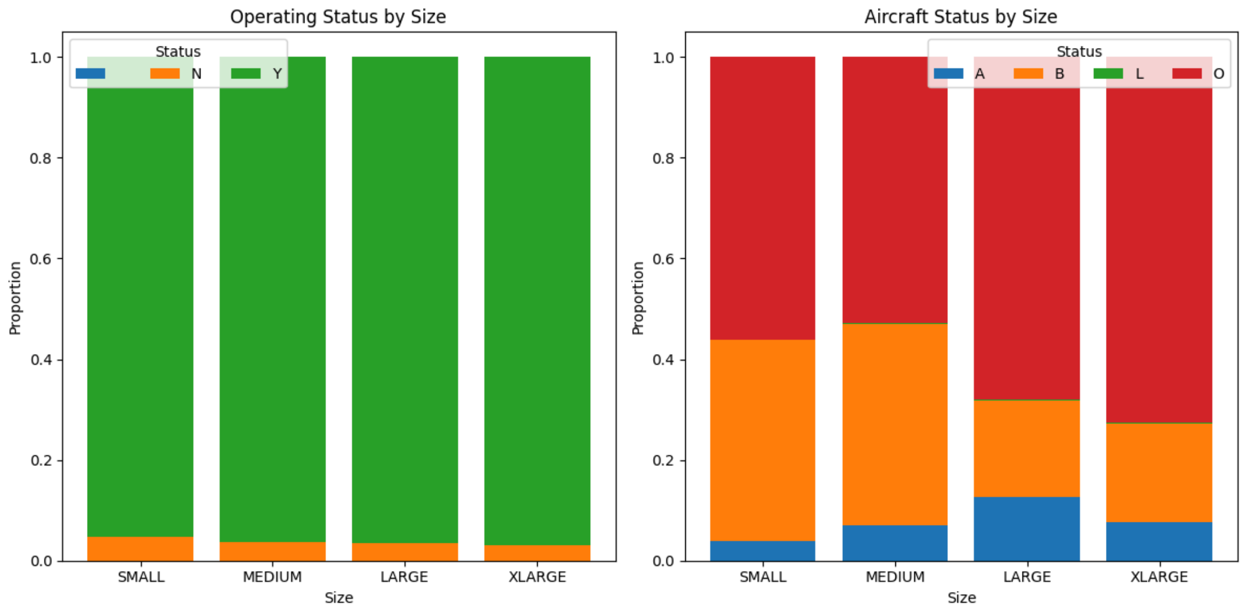


a. After the Box-Cox transformation, we see the x-axis has a much smaller range. It seems like the distribution appears more normal as outliers are limited much more. This can be seen in the number of seats histogram especially. It appears that in the capacity in pounds histograms, the largest spike appears to be close to the median.





5. It appears that the operating status of different plane sizes are not very different. There is a noticeable uptick of non operating status for small planes, but they all appear to be less than 5%. A larger plane is more likely to be operational.
There are many differences in aircraft status though. Primarily, the large and extra large planes are more likely to be owned than smaller planes. Medium planes are not owned as much as smaller planes, and there is a more than 10% jump of ownership from small planes to large and extra large planes. Medium planes have the highest percentage of operating leases, and large planes have the highest percentage of capital leases.



GAI disclosure:

I used chat gpt to help me debug code that errors and use it to build plots. This can be found in the supplementary GAI Disclosure pdf.