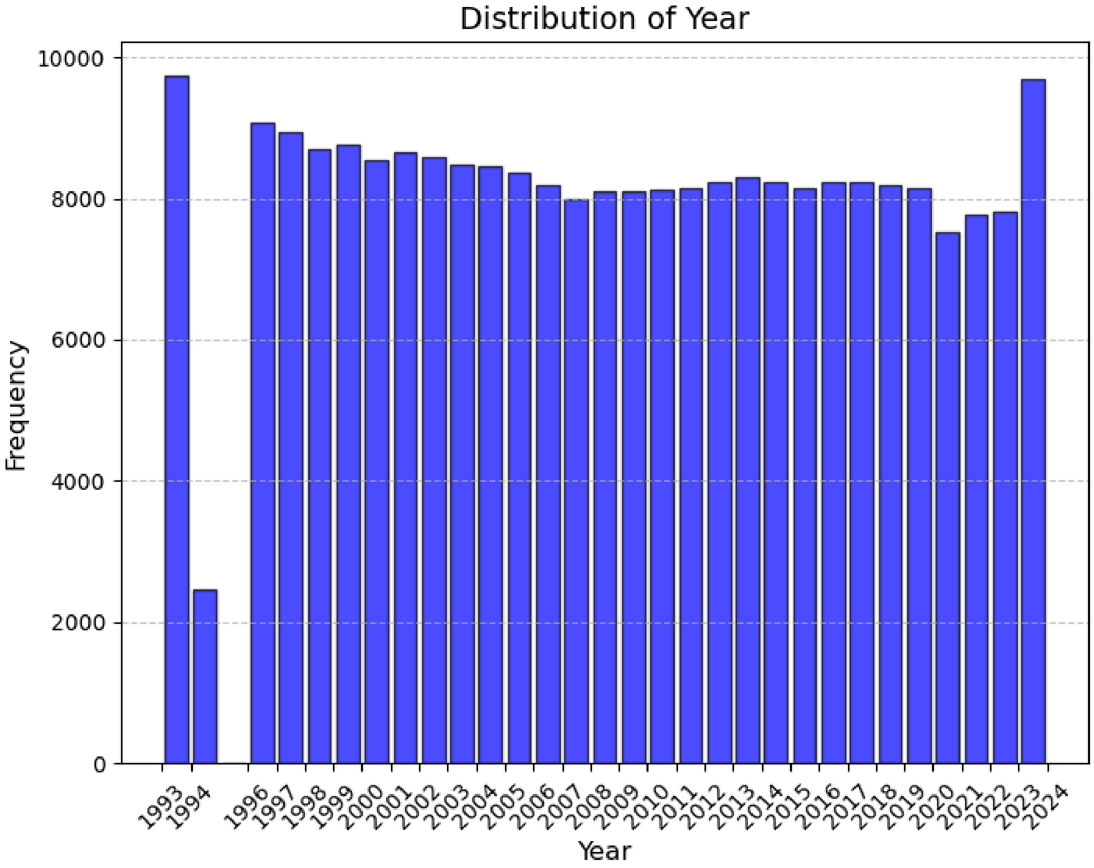
17-634: Dataset Feature Review & Evaluation

Xuchao Zhou

Feature 1 - Year

- This feature represents the year of the air travel data collected.
- It is a numerical feature.
- I decided that this feature would be an important predictor of my target variable “fare” because from common business knowledge we know the air travel price increases every year as a result of inflation, and there are also other factors that cause the increase in ticket price such as the increase in cost of energy and labor. However, I should be careful when dealing with the data from the pandemic because the fare between 2020 and 2023 is somewhat atypical.
-



This graph shows that the observed range of Year in the dataset is 1993-2024. The distribution of frequency is quite equal among the years so there are no significant peaks or skewness. Mean and standard deviation does not apply to this variable because calculating these statistics on years does not make sense.

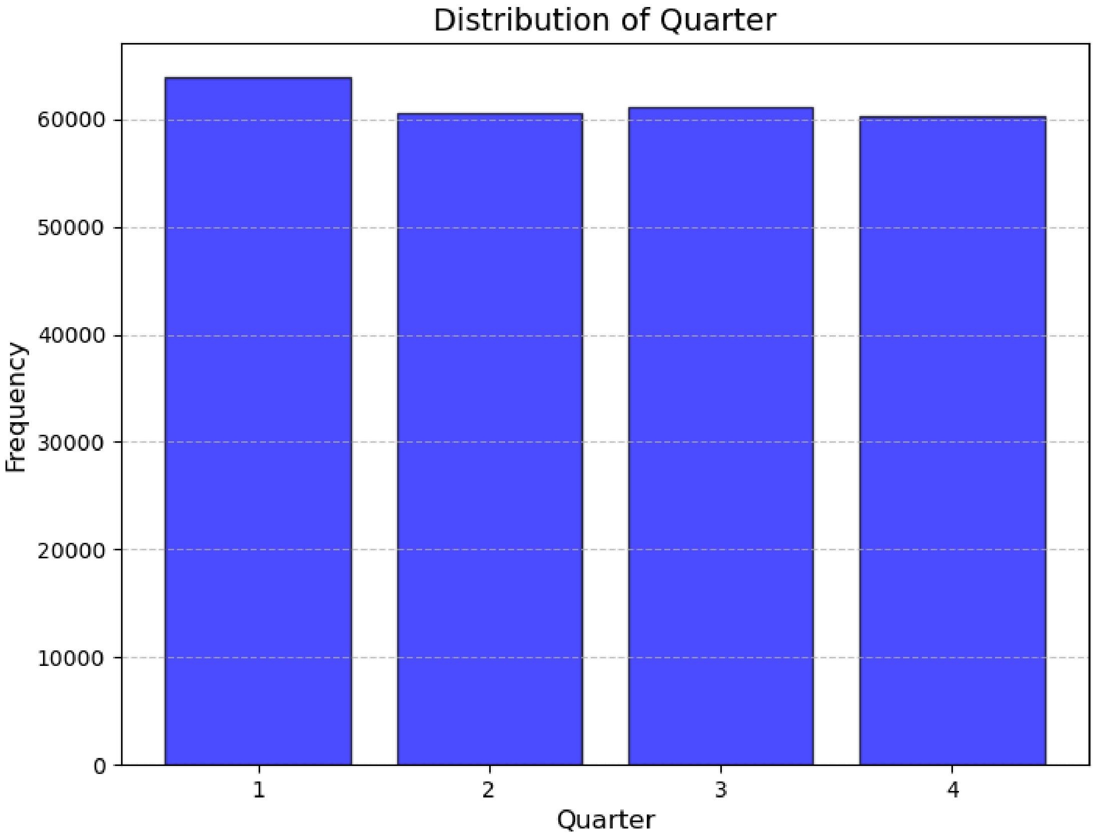
- We do not encoding for Year as this is a numerical feature. Log transformation also does not apply.

17-634: Dataset Feature Review & Evaluation

Xuchao Zhou

Feature 2 - quarter

- This feature represents the quarter of the year when the air travel data is collected.
- It is a categorical feature.
- From a business perspective, this should be a predictor of the air ticket price because the market has seasonal patterns (i.e. there are peaks of demand for travelling during summer and winter), which can shift the price of service. However, an ANOVA test shows that there is no statistically significant relationship between this feature and the fare according to the data in the dataset. The last part in the previous homework also shows that including this feature does not contribute much to the model accuracy. I decided that this feature is not an important predictor.
- The distribution of this feature is shown below:



- If we need to include this feature in a predictive model, we need to encode it because it is a categorical feature. I suggest one-hot encoding because there is no ranking among the four quarters.

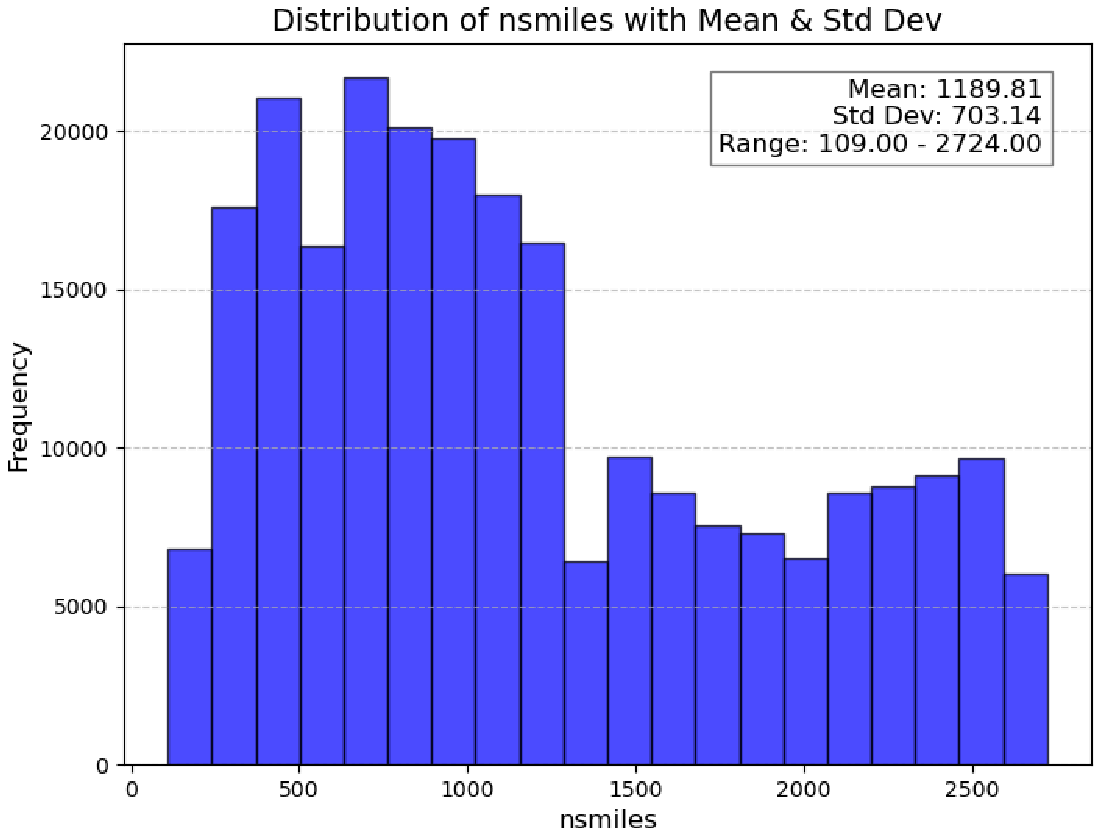
Feature 3 - route

- This feature represents the route of the trip.
- It is a categorical feature.
- The route is an important predictor of air ticket price because by common sense different routes have different fares and airlines always consider the demand at a route when making the price. Even with similar travel distance, popular routes such as NYC-LAX usually have significantly higher prices than low-demand ones such as ALB-EUG.
- This feature is generated by the combination of the features “airport_1” and “airport_2” and it shows the departure and arrival airport in the form of “XXX-YYY.” There are a gigantic number of categories in this feature, but the first and last three digits in the string should be within the range of the code of all US airports.
- We should encode this feature as it is categorical. Ideally, we should use one-hot encoding because there is no ranking for the routes, and there is unequal distribution of data so one-hot encoding can help avoid bias. However, since there are too many categories, one-hot encoding may make the data too large and significantly slow down the training process. In this case we may consider ordinal encoding as an alternative.

Feature 4 - nsmiles

- This feature represents the travel distance of the trip in miles.
- It is a numerical feature.
- This is an important predictor of the fare because the travel distance has a strong correlation with the cost of operation for the airline, and this directly decides the price on the market. Although the influence of travel distance and route popularity may cancel out, this feature should not be ignored when predicting fares.

- The distribution and related statistics of this feature is shown below:



- The distribution of this feature is a bit skewed to the right, so a log transformation may be applied to reduce the impact of the large values. The condition for a log transformation is met as there are no negative values.

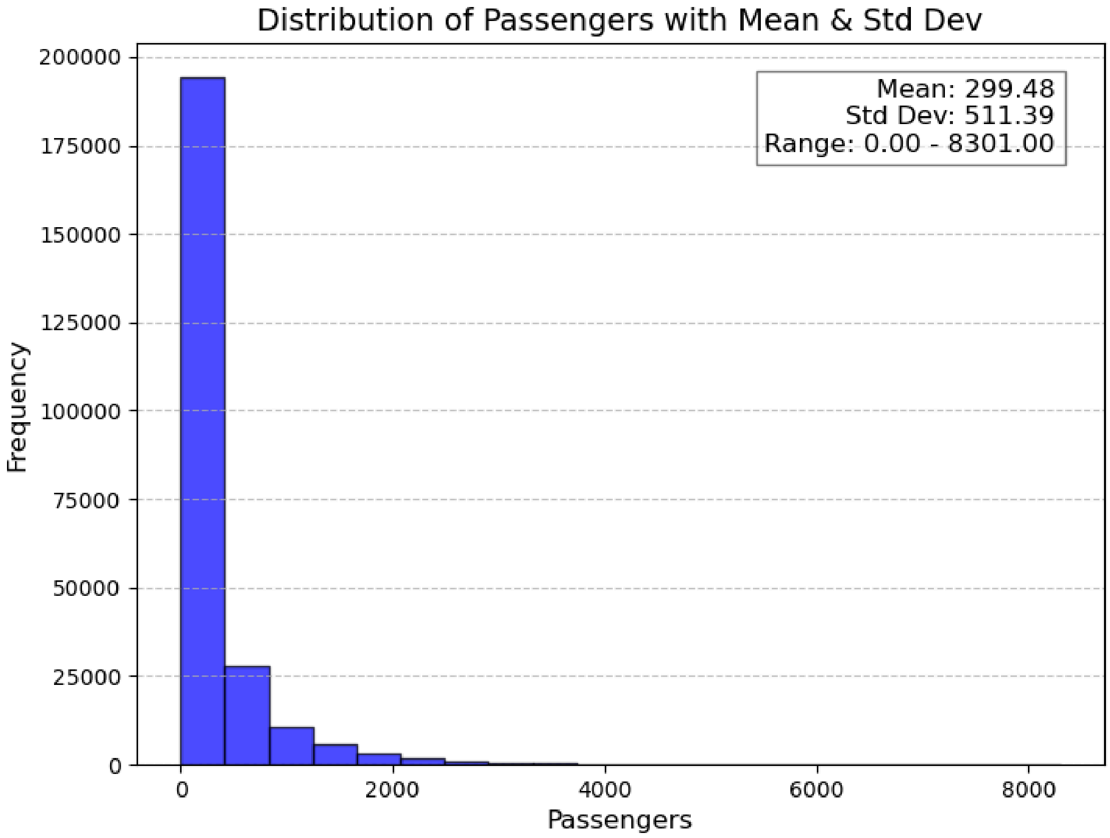
Feature 5 - passengers

- This feature represents the number of passengers that took the trip.
- It is a numerical feature.
- This feature indicates the popularity of the trip and shows the demand on the market, so it is closely related to the price. Airlines usually raise the price if there are more passengers taking the plane. Therefore, it is reasonable to say that this is an important predictor for the fare.

17-634: Dataset Feature Review & Evaluation

Xuchao Zhou

- The distribution and related statistics of this feature is shown below:



- The distribution of this feature is strongly right-skewed, so we should apply a log transformation or Box-Cox transformation to lower the influence of very large values. Because these transformations require the values to be all positive, we can drop the 0s in this feature.

Features - citymarketid_1, citymarketid_2, city1, city2, airportid_1, airportid_2, Geocoded_City1, Geocoded_City2, large_ms, lf_ms

- These features are redundant to "route" therefore should not be included in the model.

Features - tbl, carrier_lg, carrier_low, tbl1apk

- These features are irrelevant therefore should not be included in the model.