ILLINOIS TECH

Illinois Institute of Technology

ONLINE SOCIAL NETWORK ANALYSIS

Project 2

Lucia Colin Cosano A20552447

CS 579

Contents

1 PROJECT PROPOSAL

The project proposal is Analyzing Airline Network Dynamics and Passenger Flow. For it the an existing air traffic dataset contianing the following features is going to be used:

- **PASSENGERS:** Number of passengers on a specific flight departure.
- **DISTANCE:** Distance of the flight route (potentially missing).
- **UNIQUE_CARRIER:** Unique identifier for the airline operating the flight.
- **UNIQUE_CARRIER_NAME:** Name of the airline operating the flight.
- **ORIGIN_AIRPORT_ID:** Unique identifier for the origin airport.
- **ORIGIN:** Name of the origin airport.
- **DEST_AIRPORT_ID:** Unique identifier for the destination airport.
- **DEST:** Name of the destination airport.
- **YEAR:** Year of the flight departure.
- **QUARTER:** Quarter of the year (e.g., Q1, Q2) for the flight departure.
- **MONTH:** Month of the flight departure.

The data source is the Bureau of Transportation Statistics (BTS) of the US Department of Transportation: Link here.

Network Modeling Approach: two network models can be developed to explore airline network dynamics and passenger flow.

- **Airport Network:** This model represents the air transportation system as a directed and weighted network. Nodes represent individual airports, identified by their unique identifiers or names. Edges connect airports with flight departures occurring between them. The weight of each edge signifies the distance covered by the route.
- **Airline Network:** This model depicts the airline landscape as a directed and weighted network. Nodes represent airlines, distinguished by their unique carrier codes or names. Edges connect airlines that operate flights between the same origin-destination pairs. Edge weights can represent the total passenger volume for a specific airline on a particular route.

By constructing these networks, we aim to gain insights into airport connectivity, airline presence on specific routes, and the overall passenger flow within the air transportation network.

Insights and Analysis:

- **Identify Major Routes and Hubs:** Analyze the airport network to identify airports with high incoming and outgoing passenger volume, potentially representing major hubs within the network.

- **Seasonal Variations:** Analyze passenger numbers by month and quarter to understand seasonal travel patterns on specific routes.
- **Carrier Performance:** Analyze passenger numbers and network connectivity for different airlines to identify their strengths and potential areas for improvement.

2 DATA COLLECTION AND ANALYSIS

The Bureau of Transportation Statistics (BTS), a part of the U.S. Department of Transportation, serves as a pivotal resource for comprehensive transportation data and statistics. This project aims to leverage BTS datasets to enhance transportation insights, inform decision-making, and drive innovation in the transportation sector.

The primary objective of this project is to utilize public datasets provided by the BTS to analyze and extract valuable insights into various aspects of the U.S. transportation system.

The features used for the development of the project, as previously explained, include information such as the number of passengers, distance of the flight route, airline details, airport identifiers, and timing of the flight. These features have been categorized into three types based on their data characteristics:

Numeric: 5 features (PASSENGERS, DISTANCE, YEAR, QUARTER, MONTH) Categorical: 6 features (UNIQUE_CARRIER, ORIGIN_AIRPORT_ID, DEST_AIRPORT_ID, ORIGIN, DEST, UNIQUE_CARRIER_NAME)

The dataset contains a total of 11 variables and 69,784 observations. There are no missing cells in the dataset. Additionally, there are 386 duplicate rows in the dataset, accounting for 0.6% of the total observations. This information provides an overview of the dataset's structure and quality, which is essential for conducting meaningful analysis and deriving accurate insights.

3 DATA CLEANING AND TRANSFORMATION

3.1 HANDLING MISSING VALUES, UNIQUE VALUES AND DUPLICATE ROWS

The only feature in the dataset with null values is 'PASSENGERS'. With approximately 19.92% of its values as 0, indicating null entries, it has been decided to remove rows where 'PASSENGERS' is null. This approach allows us to maintain a substantial amount of data for analysis while ensuring data integrity.

After analyzing the unique values in each feature of the dataset, it was found that the 'YEAR' feature has only one unique value. Since this feature does not provide any variability or useful information for our analysis, we have decided to remove it from the dataset. This decision ensures that we focus on features that contribute meaningful insights to our analysis and modeling process.

Regarding duplicate rows, they have been successfully identified and removed from the dataset. Initially, the dataset contained 386 duplicate rows. However, after performing the necessary data cleaning operations, only 2 duplicate rows remained, which were subsequently removed.

This process ensures that our dataset is free from duplicate entries, which could have otherwise skewed our analysis.

3.2 OUTLIERS DETECTION

In this graphical representation, two box plots are considered for numerical features outliers detection. The first plot illustrates the distribution of passenger counts, while the second plot shows the distribution of distances traveled.

Passenger Distribution:

The box plot for passenger counts reveals a wide range of values. The box (interquartile range) spans from the 25th percentile to the 75th percentile, indicating where most data points lie. Outliers (represented by individual points beyond the whiskers) suggest extreme cases—perhaps unusually high or low passenger counts. The median (middle line within the box) provides the central tendency, showing the typical passenger count.

Distance Distribution:

The box plot for distances traveled follows a similar pattern. The interquartile range captures the majority of data points. Outliers in the distance plot may indicate exceptionally long or short trips. The median distance provides insight into the typical journey length.

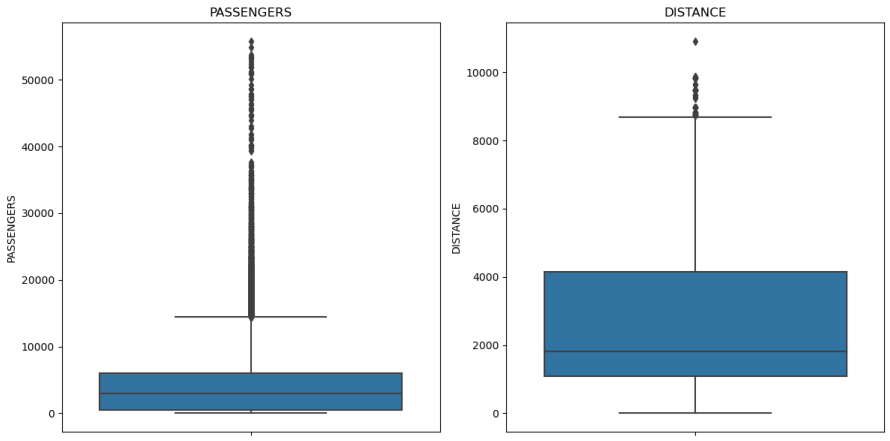

Figure 3.1: Outliers detection.

Related to categorical features, the most relevants have been considered: Quarter and Month. Firstly, the Quarter Distribution chart displays the distribution of data across quarters, each representing a three-month period in a year. The bars are nearly equal in height, indicating a uniform distribution across the quarters. This uniformity suggests a consistent pattern in the underlying data throughout the year, without significant seasonal variations.

Secondly, the month Distribution chart represents the distribution of data across months. The heights of the bars do not vary significantly, reaching the same conclusion as in the quarter one.

The uniform distribution implies stable business activity throughout the year, without pronounced seasonal fluctuations. This could indicate consistent performance or operations across different quarters. Within each quarter, the is a slight vary in the heights of the bars in the month distribution which suggest that higher counts or values may correspond to peak seasons

or specific events, while lower months may indicate off-peak periods.

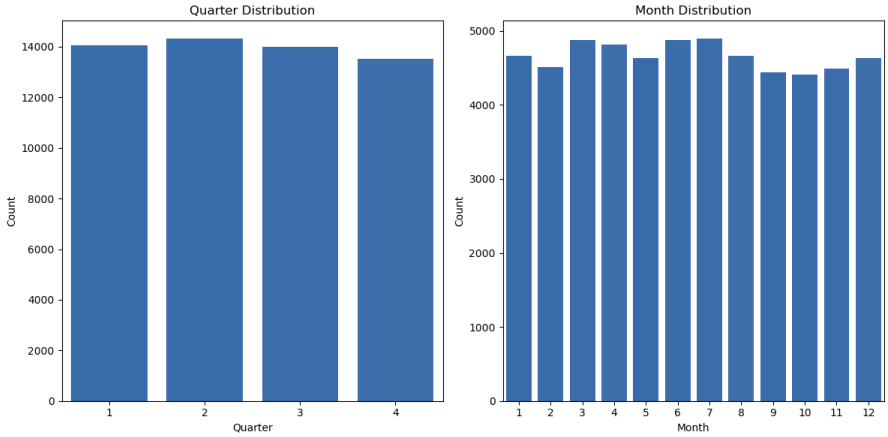

Figure 3.2: Outliers detection.

3.3 FEATURE ENGINEERING

The first feature that has been created is *'AVG_DISTANCE_PER_PASSENGER'*. For it, the average distance traveled per passenger for each flight has been calculated. This feature provides valuable insight into the efficiency of passenger travel, indicating how far each passenger typically travels on a given flight.

A higher value for *AVG_DISTANCE_PER_PASSENGER* suggests that passengers are traveling shorter distances on average, which could indicate more short-haul flights or a lower density of passengers on each flight. Conversely, a lower value may indicate longer flights or a higher passenger load per flight.

References

[1] "Bureau of Transportation Statistics," March 20, 2024. https://www.bts.gov/.