King Saud University
College of Computer and Information Sciences
Department of Information Technology

IT362 – First semester 1446

Bias Awareness

Section	NAME	ID
56594	Najd Alsabi	443200578
	Leena Almusharraf	443200563
	Hissah Alhano	443200617
	Joud Almutairi	443200544
	Lujain Alharbi	443200811
	Asma Alshilash	443200439

Data Collection Considerations

1. **Airport Selection**
   Major Hubs: The dataset primarily focuses on major travel hubs, including Chicago O'Hare (ORD), Boston Logan (BOS), Los Angeles (LAX), San Francisco (SFO), and Las Vegas (LAS). While this concentration provides insights into high-traffic routes, it may limit the applicability of findings to smaller airports and regional routes.
   Diverse Geography: By selecting airports from the East Coast, West Coast, and Central U.S., the dataset does offer a geographically diverse view of travel patterns. However, the overemphasis on major hubs still limits the overall generalizability of the results.
2. **Flight Frequency and Availability**
   High Flight Frequency: Prioritizing frequently traveled routes enables the observation of pricing trends over time, as these routes often feature multiple price points and consumer behaviors. Nevertheless, this approach may introduce **flight frequency bias**, where the conclusions are more reflective of high-traffic routes and less applicable to low-frequency or regional flights.

   Data Availability: The dataset is enriched by comprehensive data from major airports, but this availability may skew results toward more popular, better-documented routes, which might not represent less common routes.
3. **Timeframe of Data Collection**
   Specific Search Dates: Data was collected only for specific months (July, September, and December 2024), focusing on periods with potentially high travel demand, such as holidays and peak travel seasons. This introduces seasonal bias, as price fluctuations during off-peak months (e.g., spring or fall) are not reflected in the dataset.
   Departure Date Consistency: Collecting data for the same departure date across various search dates facilitates controlled analysis of how search timing affects pricing.
4. **Sampling Methodology**
   Random Sampling: The dataset may utilize random sampling across various search dates to ensure a diverse representation of consumer behavior.
   Targeted Sampling: Alternatively, focusing on specific high-activity time frames could introduce bias if certain periods are overrepresented.
5. **Data Completeness and Quality**
   Quality Control: Implementing measures to ensure data accuracy—such as filtering out incomplete records or standardizing formats—is crucial for maintaining data integrity.
   Inclusion of Key Variables: The dataset includes specific variables (e.g., number of stops, airline, cabin class) that reflect an understanding of their relevance to ticket pricing and consumer choice.
6. **External Factors**
   Market Trends: The dataset may be influenced by current market conditions, including the rise of budget airlines or shifts in consumer behavior due to economic factors, impacting pricing strategies.

Potential Biases and Limitations

1. **Airport Bias:**
   The dataset primarily focuses on a few major airports (ORD, BOS, LAX, SFO, LAS), which may lead to **airport bias**, as insights drawn from these hubs may not be generalizable to smaller airports and regional routes.
2. **Class/Cabin Bias:**
   Although the dataset distinguishes between cabin classes (e.g., Economy, Basic Economy), there is a notable overrepresentation of economy options, with 33% of the data representing Economy and 27% representing Basic Economy. This bias could lead to skewed conclusions about pricing trends, service quality, and delays, possibly neglecting the experiences of premium travelers who represent a smaller portion of the dataset.
3. **Limited Date Range**:
   The data was collected only for July, September, and December 2024, which restricts the analysis of year-round pricing trends. As the data does not cover other months, it is difficult to assess how ticket prices fluctuate across all seasons, particularly during non-peak times.
4. **Temporal Information and Route Specificity:**
   The dataset includes both time and date in the departure and arrival columns, limiting flexibility for detailed analysis. Separating these fields would allow for a more granular evaluation of how specific flight times (e.g., overnight vs. daytime) and layover durations impact pricing. Additionally, this separation would enhance the ability to analyze how time of day influences demand and price trends, providing a clearer understanding of factors affecting ticket pricing.
5. **Flight Duration and Timing:**
   While the dataset indicates whether a flight lands the next day, without precise timing, analyzing the specific pricing impact of overnight versus daytime flights is challenging.
6. **Demand Fluctuations:**
   Although the dataset captures search dates as a proxy for ticket demand, it does not account for seasonal demand fluctuations (e.g., holidays) that significantly influence pricing.
7. **Population and Behavioral Biases:**
   The dataset may primarily reflect specific search behaviors and not encompass all travel patterns. For instance, it may disproportionately capture leisure travelers over business travelers, or reflect the booking habits of certain demographics, limiting the representation of broader travel trends.

Conclusion

Understanding the methodology behind data collection and acknowledging potential biases is critical for accurately interpreting the dataset. The limitations, such as the overrepresentation of major airports, class bias, and the restricted date range, influence the reliability of conclusions drawn regarding the correlation between search dates and ticket prices. A comprehensive analysis that considers these various factors, including consumer behaviors and market dynamics, is essential to derive accurate insights from the dataset.

Dataset in Kaggle: Flight Prices by date