Name: Refiloe Kekana

October 13, 2023

Student number: u12087752

Date: 13 October 2023

TABLE OF CONTENTS

MIT 805 ASSIGNMENT 2: Visualising Flight Price Dataset

This academic paper delves into the visualising aspects of the flight price dataset using the PowerBI visualisation tool. PowerBi is a suitable tool to visualise this dataset because it is able to handle large data sets easily. It focuses on visualisation as opposed to programming which is the main aim of this assignment. It allows for easy dirriling down, which means we can focus on the most important aspects of the data without having to code to remove much sections from the data – keeping the integrity of the original dataset. We can focus more on visualisations and analysis, and PowerBI is a low-code environment. PowerBi also has a range of visualisation options which we will be making use of in this assignment. Since the data is a mix of alpha-numeric entries we can determine how an individual data point behaves or affects the rest of the data (Zhang et al, 2020).

0.1. Data preparation - Data Extraction and Transformation in PowerBI:

0.0.1. Details about the dataset as per assignment 1.

- Source: The dataset, "itineraries.csv," was collected through web scraping methodologies from prominent travel booking websites like Expedia (Wong, 2022).
- Size: The dataset is substantial, comprising approximately 6 GB of data with 27 columns and almost 6 million entries (Wong, 2022).
- Structure: The data is structured in a CSV file format, organized in a tabular structure. It includes 16 string columns, 3 integer columns, and 3 Boolean columns, capturing diverse aspects of flight details (Wong, 2022).
- Significance of the Dataset: The dataset is highly relevant in a business context, particularly for commercial companies in the aviation industry. Its focus on flight details from major US airports provides valuable insights into aviation industry dynamics, allowing for potential optimizations in flight pricing strategies (Wong, 2022).

0.0.2. Approach for Data Extraction and Transformation:

- Data Extraction: the data set was downloaded from the following URL: `https://www.kaggle.com/datasets/dilwong/flightprices` (Wong, 2022). It was then imported into PowerBI as a csv file.

0.0.1. Data Import Method: we imported the itineraries.csv file directly into PowerBI.

0.0.1. Power Query Steps:

Loading the itineraries.csv onto Power Query involved a meticulous examination of column relevance to align with the principles of data integrity and analytical focus [2]. The LegID column, serving as a unique identifier, was identified as crucial for overall data integrity but deemed less relevant for specific analyses, such as the influence of segmentsDepartureAirportCode on totalfare. Thus, for the analytical scope of this assignment, LegID was considered less significant. Additionally, redundancy in temporal information representation, exemplified by segmentsDepartureTimeEpochSeconds and segmentsDepartureTimeRaw, was addressed to optimize efficiency and eliminate unnecessary complexity in the dataset [2]. Decisions regarding column relevance and redundancy drew from both practical considerations and theoretical underpinnings in data processing.

0.0.1. Data cleaning:
Removing Duplicate Rows: there are no duplicate rows in the data sets. Each LegID is unique giving us unique data entries. All the columns of the data set are filtered to remove empty spaces and nulls.

- We used the "Remove Duplicate" function in Power Query on the following:
- Basefare and Totalfare Columns:
  - Empty entries were removed from these columns.
  - The columns were split into two by the comma delimiter.
  - Changed the data type to fixed decimal number.
  - Decimals were lost from these columns during the transformation.

- TotalTravelDistance Column:
  - Empty entries were removed from this column.

- SegmentsDepartureTimeEpochSeconds and SegmentsArrivalTimeEpochSeconds Columns:
  - These columns were removed entirely.
  - They could not be changed to the appropriate data type (duration) after being split, resulting in two columns with many null values.

- SegmentsDepartureTimeRaw and SegmentsArrivalTimeRaw Columns:
  - These columns were removed as they did not contribute significantly to the analysis.
  - The information was too complex for the aim of the assignment.

- SegmentsArrivalAirportCode and SegmentsDepartureAirportCode Columns:
  - These columns were split into three, and the "copy" columns were deleted.
  - The columns were then filtered to remove empty cells.

- SegmentsAirlineName and SegmentsAirlineCode Columns:

  - These columns showed that airline names and codes were duplicated in one column, respectively.
  - The columns were split into three, and the "copy" columns were deleted.

- SegmentsEquipmentDescription Column:
  - Empty cells were removed from this column, and the column was split.

- SegmentsDurationInSeconds and SegmentsDistance Columns:
  - These columns were split into three, and duplicated columns were removed.
  - Some information was lost as one of the copy columns had different information but also contained many null values.

- SegmentsCabinCode Column:
  - This column was split, and the copy columns were removed.
  - These were duplicated information columns.

0.0.1. Data Transformation:

Data transformation began with targeted cleaning, removing empty entries from columns like basefare, totalfare, and totaltraveldistance [2].

To align numeric columns, delimiter changes and data type conversions were applied, while redundant columns like segmentsDepartureTimeEpochSeconds and segmentsArrivalTimeEpochSeconds were removed [2]. Challenges in segmentsDepartureTimeRaw and segmentsArrivalTimeRaw led to their exclusion for streamlined analysis [2].

SegmentsArrivalAirportCode and SegmentsDepartureAirportCode underwent segmentation, enhancing clarity by deleting "copy" columns [2]. Columns with duplicated data, segmentsAirlineName and segmentsAirlineCode, were split and redundant "copy" columns removed [2]. The segmentsEquipmentDescription column was refined by removing empty cells and splitting, mirroring the treatment of segmentsDurationInSeconds and segmentsDistance [2]. The transformation included converting data from seconds to hours for accuracy [2]. SegmentsCabinCode underwent a similar splitting process with "copy" column removal [2].

0.0.1. Data Reshaping:

- Following targeted data cleaning and transformation, the reshaping process aimed at optimizing the dataset for analysis [2]. Columns, such as segmentsDepartureTimeRaw and segmentsArrivalTimeRaw, deemed non-contributory, were removed, streamlining the dataset [2]. The segmentation of segmentsArrivalAirportCode and SegmentsDepartureAirportCode, alongside the removal of "copy" columns, enhanced dataset clarity [2]. Additionally, refining segmentsAirlineName and segmentsAirlineCode by splitting and removing redundant columns contributed to a more focused dataset [2]. These steps align with best practices in data processing, preparing the dataset for meaningful analysis [2].

- Reasons for Specific Transformations:

  - Considering the removal of LegID:
    - Rationale: LegID, while essential for data integrity, was considered less relevant for specific analyses, aligning with the need for a focused dataset [2].
    - Business Objectives: Enhances data focus, supporting more precise analyses [2].
    - Insights Seeking: Streamlines data for improved analysis on factors like segmentsDepartureAirportCode's impact on totalfare [2].

- Redundant Temporal Columns Removal:
- Rationale: Eliminates redundancy in temporal information, optimizing dataset efficiency [2].
- Business Objectives: Ensures streamlined and efficient data processing, aligning with best practices [2].
- Insights Seeking: Enhances clarity and efficiency in handling temporal data [2].

- Handling Empty Entries:
  - Rationale: Cleanses the dataset by removing null values, ensuring accuracy in subsequent analyses [2].
  - Business Objectives: Improves data quality, crucial for reliable analysis and decision-making [2].
  - Insights Seeking: Prevents potential biases or inaccuracies in analyses due to null values [2].

- Delimiter Change and Data Type Conversion:
  - Rationale: Aligns numeric columns by changing delimiters and converting data types [2].
  - Business Objectives: Ensures uniformity in data representation, facilitating accurate numerical analyses [2].
  - Insights Seeking: Enables more straightforward numeric operations and computations [2].

- Columns Segmentation and "Copy" Column Deletion:
  - Rationale: Enhances dataset clarity by removing redundant "copy" columns and segmenting relevant ones [2].
  - Business Objectives: Improves dataset organization and readability, aiding in subsequent analyses [2].
  - Insights Seeking: Reduces complexity, allowing for more straightforward insights extraction [2].

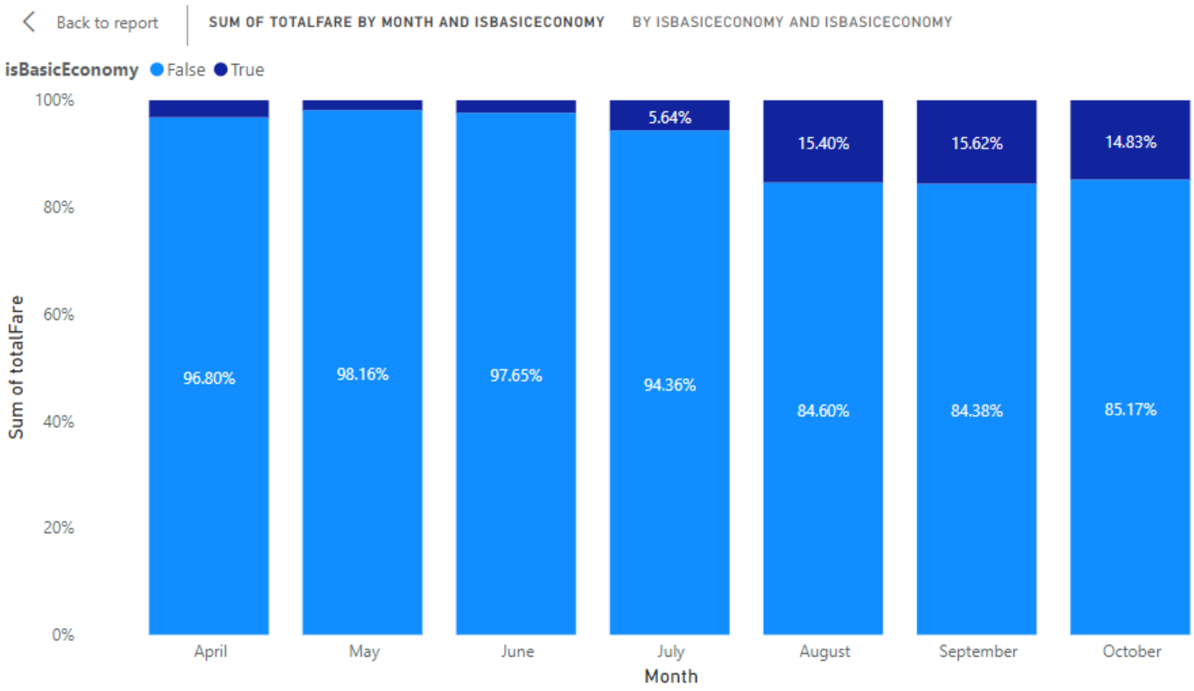
- Conversion of Time Units:
  - Rationale: Converts time data from seconds to hours for accuracy and consistency [2].
  - Business Objectives: Ensures data accuracy and aligns time units for coherent analyses [2].
  - Insights Seeking: Facilitates more accurate temporal analyses [2].

- Overall Business Objectives:
  - Dataset Optimization: Ensure the dataset is focused, streamlined, and free from redundancies, enhancing its suitability for in-depth analyses [2].
  - Data Accuracy: Implement transformations to remove empty entries and convert data types, ensuring accuracy for reliable business insights [2].
  - Efficient Data Processing: Remove redundant columns and streamline the dataset to optimize efficiency in subsequent analyses [2].

- Insights Seeking:
- Focused Analyses: By removing less relevant columns, the dataset is primed for specific analyses, such as the impact of departure airports on total fare [2].
- Improved Efficiency: Streamlining and cleaning contribute to more efficient data processing, allowing for a more straightforward extraction of meaningful insights [2].
- Data Quality: Ensuring data accuracy and consistency through transformations enhances the overall quality of insights derived from the dataset [2].

- Benefits of Reshaping:
  - The reshaping facilitates better analysis by breaking down the fare information into more manageable parts.
  - It enabled more straightforward calculations, comparisons, fare-related data.

0.1. Visuals



Figu
1: Stacked Bar Graph of Totalfare vs IsBasicEconomy

Figure 1 presents a stacked bar graph illustrating the distribution of total fares concerning the classification of flights into the basic economy and other fare categories over the observed time frame. The vertical axis represents the total fare values, while the horizontal axis signifies the progression of time, capturing shifts across different seasons though months.

The graph discerns a predominant inclination toward non-economy seats, denoting a preference for more expensive seating options throughout the year. Notably, during the latter part of the year, coinciding with United States holidays (October–December), there is a discernible increase in the percentage of searches for basic economy seats. This observation prompts a contemplation that the data might exhibit a surge in the number of basic economy seats if extended until December 2022. The graph offers a nuanced understanding of the seasonal dynamics and preferences in seat selection, potentially revealing evolving travel patterns and considerations during specific periods.

- Seasonal Variation: The graph suggests a potential correlation between seasons and seat preferences, indicating the need for a more detailed examination of seasonal influences on travel behavior.
- Holiday Impact: The notable rise in basic economy seat searches during the holiday season prompts an exploration into the specific factors driving this surge and its implications for travel planning.

- Projection for Year-End: The visual implies that extending the data collection until December 2022 may unveil a more comprehensive picture, especially regarding the surge in basic economy seat searches observed during the holiday period.

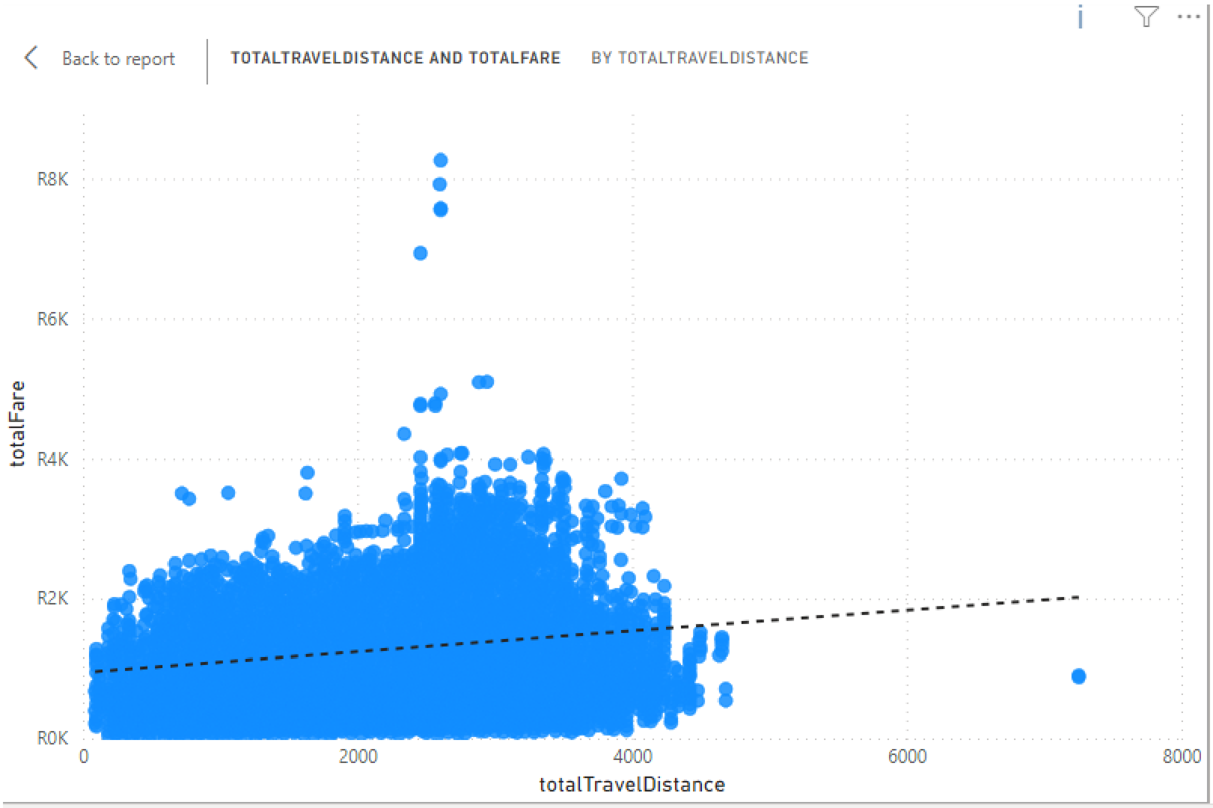


Figure 2: Scatter Plot of Total Fare vs Total Distance

Figure 2 depicts a scatter plot that explores the relationship between total fare and total distance traveled in flight itineraries. The horizontal axis represents the total distance covered, while the vertical axis signifies the corresponding total fare values. Each data point on the scatter plot represents an individual flight itinerary, providing insights into potential correlations between distance and fare.

The scatter plot reveals an intriguing observation—there is no substantial correlation between the total distance traveled and the total fare. The trend line, though present, indicates a weak positive correlation. This outcome is unexpected, as conventional assumptions would anticipate a stronger positive correlation, implying that longer distances result in higher fares. However, the data suggests a scenario where fares remain relatively standard regardless of the distance covered.

- Unexpected Relationship: The weak positive correlation challenges conventional expectations, prompting a deeper investigation into the factors influencing fare determination beyond distance.

- Standardized Fare: The data hints at a potential pattern where fares maintain a consistent level irrespective of the distance traveled, suggesting other influential factors in pricing strategies.
- Consideration of Variables: Further analysis should consider additional variables influencing fare determination, such as specific flight routes, airlines, or time of booking.

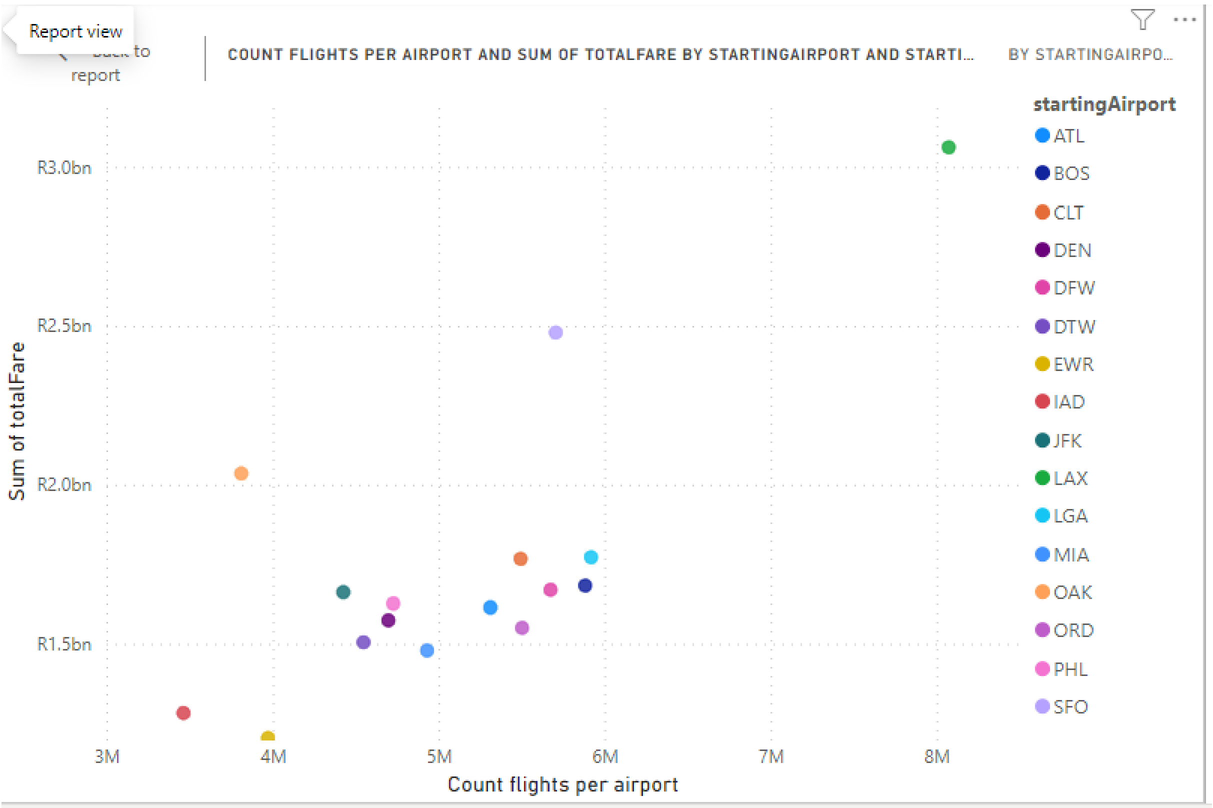


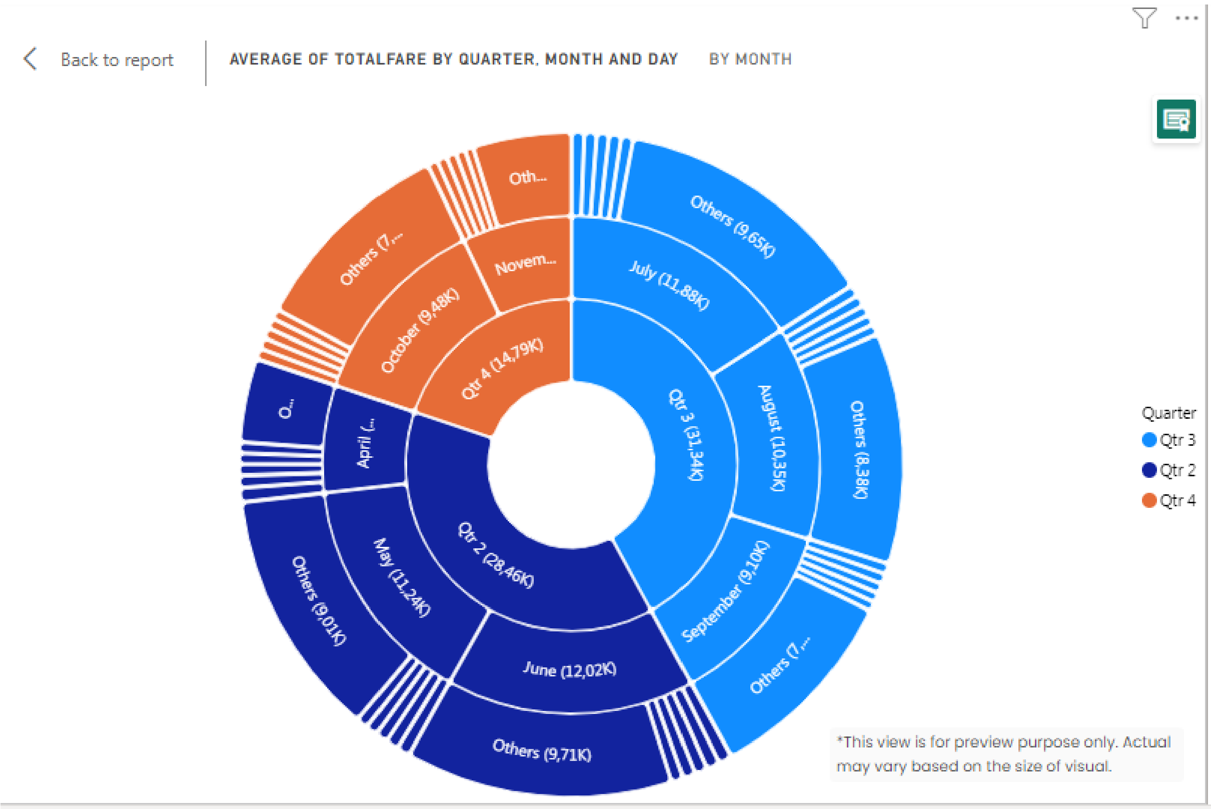
Figure 3: Scatter Plot - Flights per Airport vs. Total Fare per Airport

Figure 3 presents a scatter plot illustrating the relationship between the number of flights booked per airport and the corresponding total fare. Each point on the plot represents a distinct airport, showcasing insights into how booking frequency correlates with total fare values. The x-axis represents the number of flights per airport, and the y-axis signifies the total fare per airport.

The scatter plot demonstrates a positive correlation between the volume of flights booked at an airport and the total fare generated. Notably, LAX appears as a potential outlier, characterized by the highest number of flights and the largest total fare. This suggests a potential influence of airport size on both booking frequency and total fare.

- Positive Correlation: The general trend indicates that as the number of flights increases for an airport, the total fare also tends to rise, implying a positive correlation.

- Outlier Identification: LAX's exceptional position in terms of both flight volume and total fare prompts a closer examination to understand its unique factors.
- Airport Disparities: Insights from Mickeviciute and Mickeviciute (2023) [3] highlight differences in flight volumes between airports, providing context for the observed variations in total fare.



*This view is for preview purpose only. Actual may vary based on the size of visual.

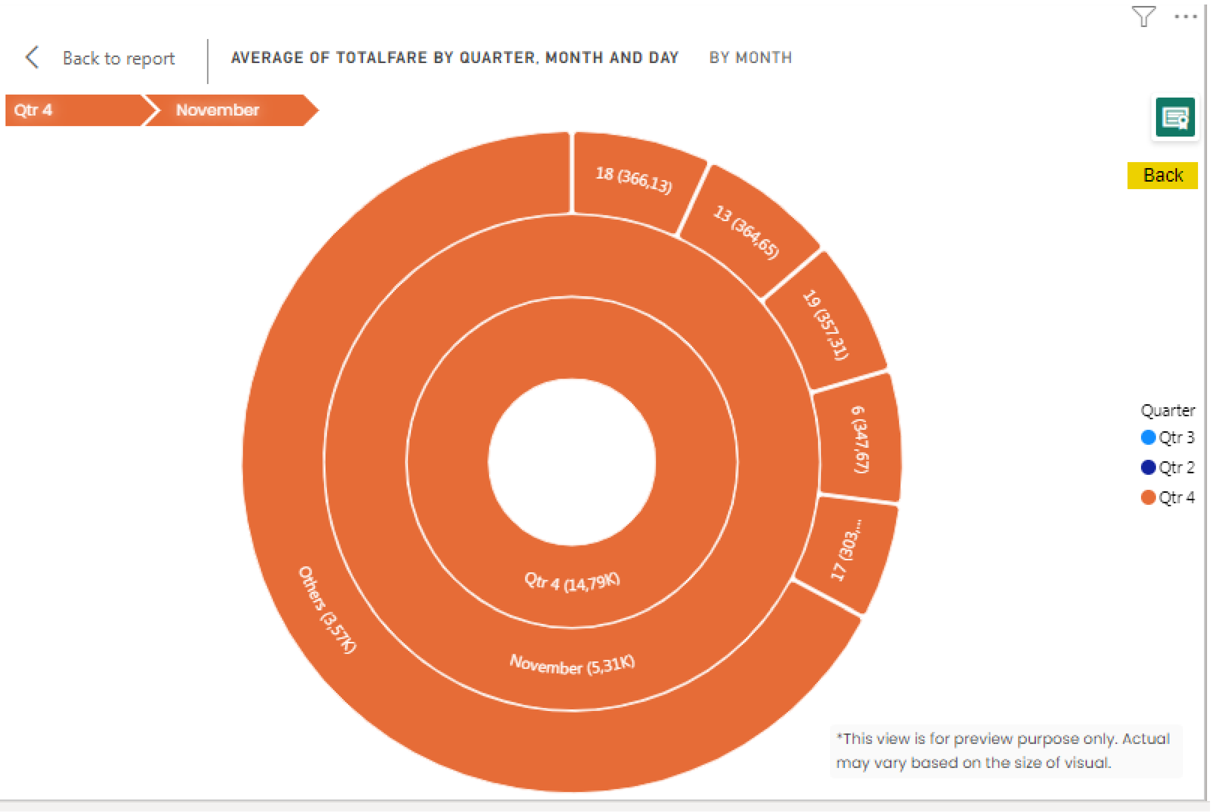
Figure 4: Sunburst Diagram - Average Fare Distribution Over 3 Quarters

Figure 4 presents a sunburst diagram depicting the distribution of average fare expenditure across three quarters. The visualization showcases the proportional spending on flights during different time periods, providing insights into seasonal variations.

The sunburst diagram reveals that Quarter 3 possesses the largest dataset, with notable concentrations in June, July, and May (Summer). This observation aligns with the typical surge in travel during the summer season, reflecting increased expenditure on flights.

The concentration of spending during the summer months is indicative of heightened travel activities, possibly influenced by the holidays. For instance, in 2022, Thanksgiving on November 24th may have contributed to the increased travel demand during the week before Thanksgiving.

- Seasonal Peaks: The diagram substantiates the notion that travel expenses peak during the summer months, emphasizing the impact of seasonal trends on flight expenditures.
- Holiday Travel:∗∗ The observed surge in spending a week before Thanksgiving corresponds with holiday-related travel patterns, aligning with common travel practices during significant holidays.



*This view is for preview purpose only. Actual may vary based on the size of visual.

Figure 5: Sunburst Diagram - Detailed Analysis of November 2022 Expenditure

Figure 5 provides an in-depth exploration of flight expenditure in November 2022 through a sunburst diagram. The visualization highlights specific dates during the month, offering a focused view of spending patterns.

The sunburst diagram illustrates that a significant portion of the expenditure on flights in November 2022 is concentrated on specific dates. Notably, the highest spending occurred on the 18th, 13th, 19th, 6th, and 17th of November.

These highlighted dates align with the week preceding Thanksgiving in 2022. The visual representation indicates heightened flight spending during this period, suggesting increased travel activity in anticipation of the Thanksgiving holiday.

- Pre-Thanksgiving Surge: The concentration of spending on the specified dates corresponds with the week leading up to Thanksgiving, reflecting a surge in travel-related expenses.
- Strategic Planning: Travelers appear to strategically plan their journeys, with notable peaks in spending occurring just before the holiday.

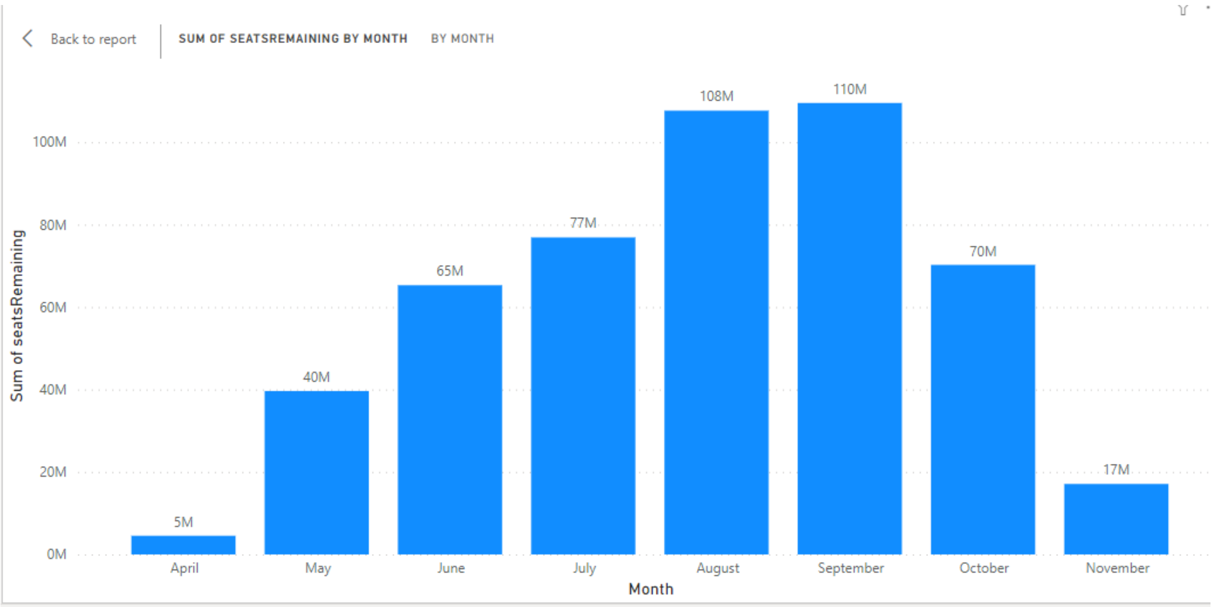
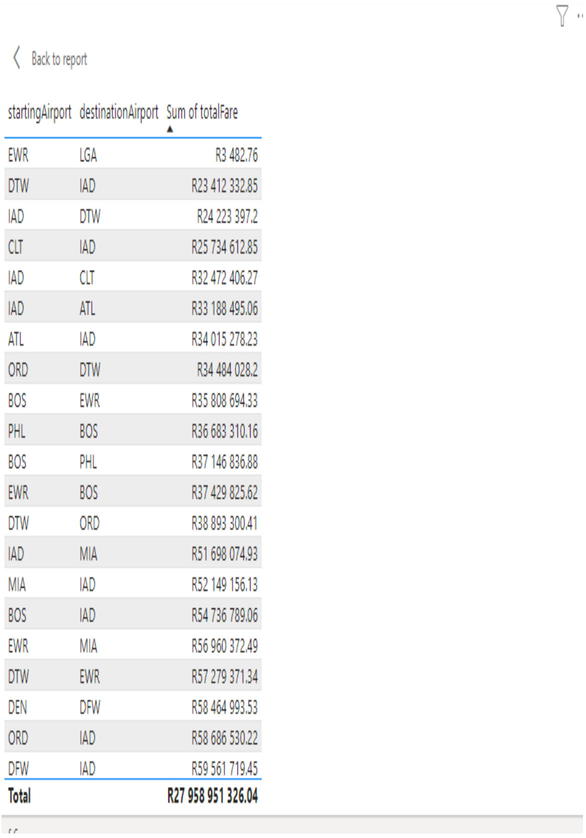


Figure 6: Bar Graph - Unoccupied Seats Analysis

Figure 6 presents a bar graph depicting the monthly variations in the number of unoccupied seats. The visual offers insights into the seasonal trends of seat occupancy throughout the year.

The graph highlights a noticeable increase in the number of unoccupied seats during the summer months, specifically in June, July, and August. Conversely, a sharp decline in unoccupied seats is observed during the holiday months.

- Summer Increase: The visual representation aligns with a surge in unoccupied seats during the summer season, indicating a potential decrease in travel demand during these months.
- Holiday Contraction: The significant decrease in unoccupied seats during holiday months suggests heightened travel activity, possibly driven by increased demand for flights during holiday periods.



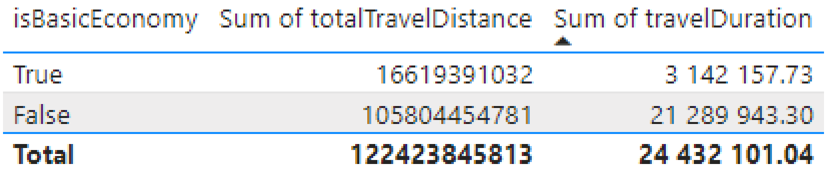
startingAirport	destinationAirport	Sum of totalFare
EWR	LGA	R3 482.76
DTW	IAD	R23 412 332.85
IAD	DTW	R24 223 397.2
CLT	IAD	R25 734 612.85
IAD	CLT	R32 472 406.27
IAD	ATL	R33 188 495.06
ATL	IAD	R34 015 278.23
ORD	DTW	R34 484 028.2
BOS	EWR	R35 808 694.33
PHL	BOS	R36 683 310.16
BOS	PHL	R37 146 836.88
EWR	BOS	R37 429 825.62
DTW	ORD	R38 893 300.41
IAD	MIA	R51 698 074.93
MIA	IAD	R52 149 156.13
BOS	IAD	R54 736 789.06
EWR	MIA	R56 960 372.49
DTW	EWR	R57 279 371.34
DEN	DFW	R58 464 993.53
ORD	IAD	R58 686 530.22
DFW	IAD	R59 561 719.45
Total		**R27 958 951 326.04**

Table 1: Flight Path Preference Analysis

Table 1 provides an insightful overview of flight path preferences based on key parameters such as distance between airports and total fare. The table aims to highlight popular flight routes and potential correlations between distance and fare.

The table prominently showcases the preference for flights from DFW to IDA, indicating a significant traveler choice for this route. Notably, these airports are relatively close, with a distance of 14.69 km. Additionally, the data reveals EWR to LGA as another frequently chosen route, covering a distance of 26.67 km and featuring the smallest total fare.

- DFW to IDA Preference: The popularity of the DFW to IDA route suggests that travelers may prioritize shorter distances, as evidenced by the relatively close proximity of these airports.
- EWR to LGA Route: The smaller total fare associated with the EWR to LGA route could be attributed to either higher demand or strategic pricing due to the longer distance, implying potential cost-conscious traveler behavior.



isBasicEconomy	Sum of totalTravelDistance	Sum of travelDuration
True	16619391032	3 142 157.73
False	105804454781	21 289 943.30
Total	**122423845813**	**24 432 101.04**

Table 2: Comparative Analysis of Economy and Non-Economy Flights

Table 2 presents a comprehensive comparison between economy and non-economy flights, considering both total distance traveled and travel duration. The purpose of this table is to shed light on potential differences in travel patterns and durations between these two categories.

The data in Table 2 clearly illustrates that non-economy seats have covered a greater distance in comparison to economy seats. Simultaneously, the travel duration for economy seats is notably shorter. This observation suggests a correlation between travel class and the extent of travel undertaken.

- Non-Economy Travel: Passengers in non-economy seats tend to cover more distance in their flights.
- Economy Travel: Economy-class travelers, on the other hand, exhibit shorter travel durations, implying relatively shorter distances covered.

These findings may signify varying travel purposes or preferences between passengers opting for different classes. Further analysis could explore the specific routes and destinations associated with these travel patterns.

1. Conclusion:

∗∗Conclusion:∗∗

In summary, this paper employed PowerBI to visualize a substantial flight price dataset, emphasizing visual exploration over programming. The data preparation process, from extraction to transformation, prioritized data integrity and focused analysis.

Visuals, including stacked bar graphs, scatter plots, and sunburst diagrams, uncovered nuanced patterns. Key findings encompassed seasonal seat preferences, unexpected fare-distance relationships, airport-wise dynamics, and detailed expenditure analysis around holidays.

Data cleaning and transformation decisions were grounded in both practical and theoretical considerations. Reshaping the data offered benefits of improved focus, processing efficiency, and enhanced data quality.

In essence, this assignment successfully translated raw data into meaningful insights, utilizing PowerBI for impactful visualizations. The visuals serve as gateways to further exploration of flight pricing dynamics and traveler behaviors.

LINK TO VIDE: `https://drive.google.com/file/d/135gKjQ9dfL-hWuh2D7db-eDorDWA0pNT/view?usp=sharing`

0.1 Bibliography

[1] Wong Dillon. “Flight Prices". In: (2022).

[2] Zhang, N., Wang, M., Duan, Z. and Tian, C., 2020. Verifying properties of mapreduce-based big data processing. IEEE Transactions on Reliability, 71(1), pp.321-338.

[3] Mickeviciute, R. and Mickeviciute, R. (2023) ’Top 10 busiest airports in the US during 2022: aviation hubs,’ AeroTime [Preprint].
https://www.aerotime.aero/articles/top-10-busiest-airports-in-the-us-during-2022.