Knowledge Streams

Power BI Assignment 1A

In this assignment, we will focus on EDA and a bit of feature engineering on the flight dataset.

This assignment will not give you direct instructions or commands as given in the previous 2 notebooks. Instead, there will be hints and you have to use your **Problem Solving** and **Critical Thinking** Skills.

If you are stuck somewhere, feel free to reach out to me or your peers. I can help you understand the question but do note that direct solutions will not be provided. There are many ways to approach this assignment, so be creative and explore different options. The best way to learn is by doing, so dive in and start exploring the data!

Remember:
This assignment is designed to challenge your problem-solving skills and creativity. There is no single "right" answer. The most important thing is to demonstrate your ability to clean, transform, and analyze data using Power BI, and to communicate your findings effectively.

LOs are available for certain steps (HIGHLIGHTED IN RED) to help you solve those steps.
HINTS will help you to solve the whole assignment

You are welcome to use CHAT-GPT or any other RESOURCE but remember, the idea is not to copy paste but UNDERSTAND.

All LOs (Learning Outcomes) must be completed in order to be graded.

I will also drop a voice note on the Whatsapp group to explain the assignment shortly.

Due: **Sunday, 26th May, 2024 (23:59)**

Data Cleaning and Transformation:

The dataset contains various columns with information about flights. Some of these columns need a bit of cleaning and transformation to make them usable for analysis:

LO1: Think about how to tell Power BI that the first row in your dataset isn't actual data, but column labels.

1. **Date and Time**: The dataset includes columns for departure and arrival times, but they are not in the ideal format. You'll need to convert these columns to appropriate data types and extract relevant information (e.g., hours, minutes) for further analysis.

LO2: What data type does Power BI automatically assign to dates?
Is that the most helpful format for later analysis?
Explore the data type conversion options in Power Query.

2. **Duration**: The flight duration information needs some cleanup to make it easier to work with. Look for ways to standardize and convert it into a more useful format.

LO3: Power Query has tools for splitting columns based on specific characters. What character separates hours and minutes?
Consider the appropriate data type for numerical values representing hours and minutes.
Power Query has functions for dealing with missing values. (How though)
You'll want to remove the 'h' and 'm' characters before converting to a number.

3. **Stops**: The information about the number of stops during a flight is not currently numeric. Figure out how to transform this column into a numeric format.

LO4: Explore text manipulation functions in Power Query to remove unnecessary text and convert the values to numbers.

4. Arrival Time: This column seems to have both date and time. How can you isolate just the time portion?

LO5: What character separates the date and time in this column?
Think about how to further break down the time into smaller units for easier analysis.

4. **Other Columns**: Consider if there are any other columns that need to be transformed or cleaned before moving on to analysis.

LO6: Look for options in Power Query to remove columns that aren't currently needed.
How can you convert a currency value to a standard decimal format?
How can you identify errors in Power Query?
If there are many errors, it might signal a deeper issue with your data source.

If there is only one, we have studied in class how to deal with errors that are less than 25% of the original dataset.

Additional Considerations:

- While exploring the data, pay attention to any missing values or outliers. Decide how you want to handle these issues.
- Think about any additional calculations or transformations that could be useful for deeper analysis.

Exploratory Data Analysis (EDA):

Once the data is clean and prepared, start exploring it! Here are some questions to guide your analysis:

1. How does the average ticket price vary across different airlines or destinations?
2. Is there a relationship between the number of stops and the ticket price?
3. Are there any patterns in departure or arrival times? Are certain times of day or days of the week more popular for flights?
4. What other interesting patterns or relationships can you uncover in the data?

Feel free to go beyond these questions and explore any other aspects of the data that spark your curiosity. The goal is to gain a deep understanding of the dataset and uncover potential insights.

Visualizations and KPIs

As a data analyst, your job is not just to explore data but also to think proactively. Can you come up with some key performance indicators (KPIs) that would be relevant for an airline or travel agency? How could these KPIs be used to make data-driven decisions? Let me give you an example to better understand.

Suppose a bank is thinking of opening a new atm in a specific location. How will they decide that they should open an atm? Based on the transaction or maybe the number of residents or maybe the num of competitor banks, or the financial status of people, or maybe umm, based on the frequency of transactions or maybe based on additional costs? There is just so much to consider, no?

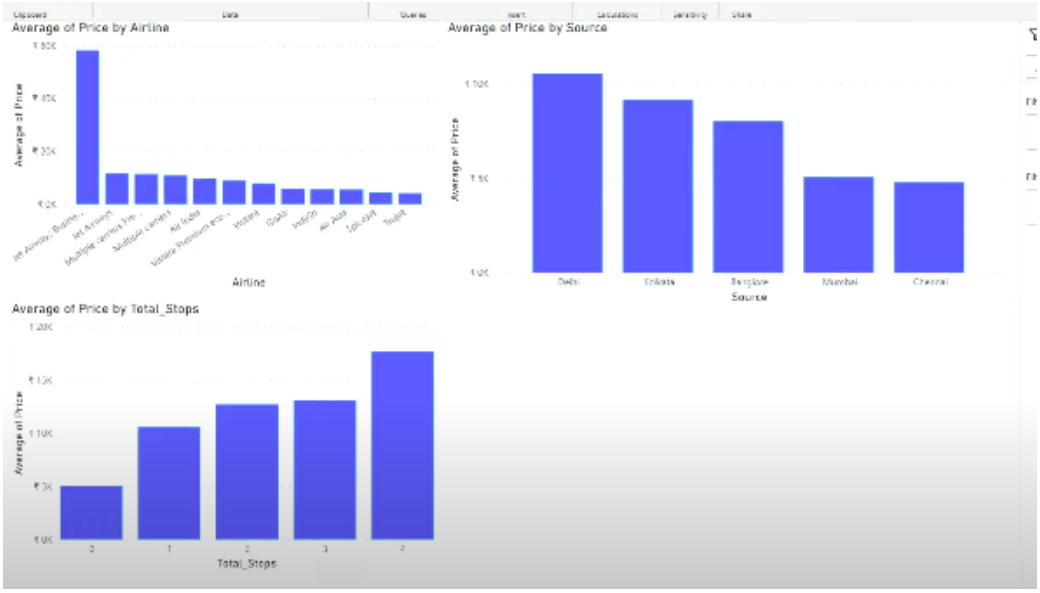
That's where the KPIs come in (Key Performance Indicators). For this, let's create a dashboard with me side by side.

1. Choose a bar chart with Price on y axis and choose average as an aggregate function
2. Put Airline on the x - axis and you can see the average price of the airline.
3. Put Destination on the x- axis and you can see the average price of source, lets say.

Did you figure how we can get different insights by just changing the values on the x-axis?

Now, how about Average Price by Total Stops?

This is the dashboard that I am expecting from you guys.



HINTS

1. Recall the pandas workbook on elections dataset where you split a Full Name based on a blank space and create First and Last Name columns separately. Now, we can do the same thing in Power BI 🙂 How to do that? (Use chat-gpt or your own resources for that)

2. The key is to explore the tools available in Power Query and think step-by-step about how to transform the data into a format suitable for your analysis. Don't be afraid to experiment and try different approaches!

3. Half of your assignment is based on usage of a single button, try to figure out what it is!