SKYHACK

HACKATHON

2K20/MC/049   Gaurav Jain

TABLE OF CONTENTS

EXECUTIVE SUMMARY

In the initial stages of this project, I started by importing fundamental libraries such as NumPy and Pandas to manage our datasets efficiently. With a focus on data exploration, I utilized techniques like "df.info()" and "df.value_counts()" to gain a comprehensive understanding of the dataset's characteristics, including unique values and their frequencies. Moving forward, I did data preprocessing by filling in missing values, detecting and eliminating outliers, and lastly creating more descriptive categories to enhance data interpretation. Next, I split the dataset based on the two questions found in the "question_type" column, thereby refining the dataset's relevance. Subsequently, we segmented the data further into tables for satisfied and dissatisfied customers using the "satisfaction_type" column. Finally, I created a new data frame that compiled unique values and their frequencies from each column, calculating their effects on customer satisfaction and streamlining the dataset by removing less impactful factors. These steps laid the foundation for a more focused and visual analysis of United Airlines' food and beverage services' that I did using google spreadsheet (excel) customer feedback.

UNDERSTANDING THE PROBLEM

- Root Cause for F&B Satisfaction:
    - Identify the underlying reasons behind in food and beverage (F&B) satisfaction scores specifically during the summer months and analyze data to find the key factors influencing F&B satisfaction during this period.
- Analysis of Survey Comments:
    - Gain insights from customer feedback by analyzing survey comments and understand the major themes and issues related to F&B that customers are complaining about.
- Coding Skills Showcase:
    - Demonstrate technical proficiency in SQL, Python, or R to manipulate and summarize data effectively and write queries or scripts to extract, transform, and summarize data at various levels for meaningful conclusions.
- Storytelling and Initial Recommendations:
    - Present the analysis findings cohesively in a narrative format and tell a story using data to convey the factors impacting F&B satisfaction provide initial recommendations based on insights.

APPROACHING THE DATASETS

Step 1: Data Understanding

- Importing essential libraries such as 'numpy' and 'pandas.'
- Beginning by comprehensively understanding the dataset using the dataset description to gain insights into its context and purpose and also examination of the types of values and frequency of such values present in each column.

Step 2: Data Cleaning

- Interpreting and addressing missing values using appropriate methods.
- Identifying and dealing with outliers to ensure data integrity.
- Creating more descriptive categories to enhance data interpretability.
- Removing redundant or unnecessary columns that do not contribute to the analysis.

APPROACHING THE DATASETS

Step 3: Data Segmentation

- Split the cleaned dataset into two distinct tables based on the "question_type" column.
- This separation helps us focus on each question separately since the "score" column now contains answers from either question in different tables

Step 4: Further Segmentation

- Within one of the segmented tables, further split the data based on "satisfaction_type."
- This results in two additional tables—one for satisfied customers and one for dissatisfied customers.
- This is useful for fine data manipulation and data extraction

APPROACHING THE DATASETS

Step 5: Creating Factors Affect DataFrame

- Create a new dataframe to store unique values and their frequencies from each column of the previously segmented dataset.
- For each column in the original dataset:
    - Calculate the frequency of each unique value across the entire dataset.
    - Calculate the frequency of each unique value separately for satisfied and dissatisfied customers.
    - Add these frequencies to the new dataframe.
    - Also add the percent of each unique value with respect to the number of entries in that column
- Using all values entered in this dataframe, calculate the effect of the occurrence of each unique value and its positive or negative effects on the customer satisfaction
- Lastly, remove all factors that have very less effect on the consumer satisfaction

ANALYSING THE RESULTS

Consumer Satisfaction Vs Origin Code
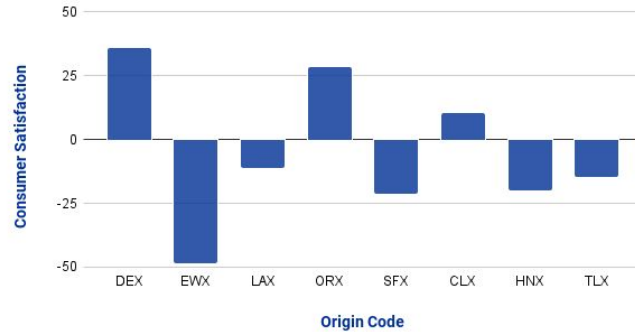
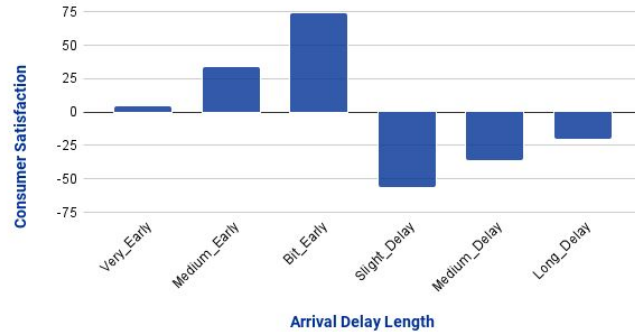**Consumer Satisfaction Vs Destination Code**

Using the previously generated table, I extracted the impact of each factor and represented it visually through bar charts that showcase the relationship between customer satisfaction with Food and Beverages (F&B) and select relevant factors from categories.

In these particular charts:

- There are some airports that appear both times, for example:
    - EWX: this airport seems to decreases the F&B satisfaction relatively a lot
    - SFX: similar to EWX, this also decreases F&B satisfaction but not as much
    - DEX: this airport generally increases the F&B satisfaction by a bit

ANALYSING THE RESULTS

Consumer Satisfaction Vs Arrival Delay Length

**Consumer Satisfaction Vs Cabin Type**
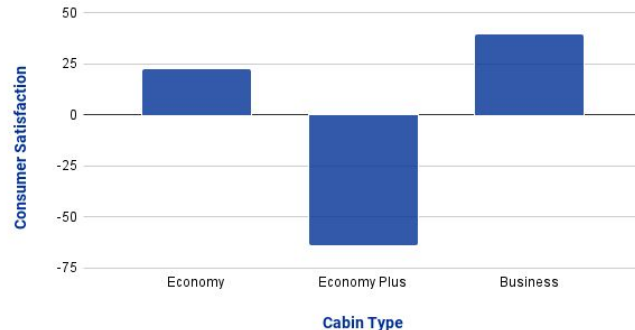
In the first chart:

- We can visually see a very reasonable relationship:
  - If delay, then dissatisfaction and vice versa
- However, a long-delay would cause more dissatisfaction than slight delay. The reason why this is not the case is because the calculation also takes into account the frequency of such cases. Since the algorithm focuses more on more frequent factors of consumer satisfaction, the rare cases are less considered

In the second case:

- It is interesting to see that those who pay for economy plus are less satisfied than those who get economy or business class. Other cabin types are near 0, hence not considered
  - This may be due to the fact that those in economy plus are expecting a bigger difference in the treatment than in economy class which is not truly the case.

ANALYSING THE RESULTS

Consumer Satisfaction Vs Number Of Leg

Consumer Satisfaction Vs Hub / Spoke Departure
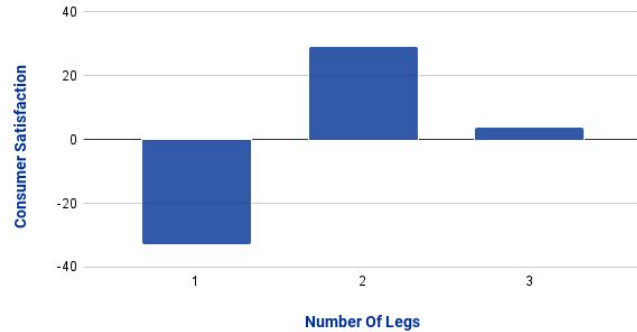
In the first chart:

- Unexpectedly, those without a stop in between are the most dissatisfied whereas those who have stop(s) are more satisfied, it can be assumed that this is due to the fact that many 1 leg flights are short in duration and some don't have meal time resulting in dissatisfaction
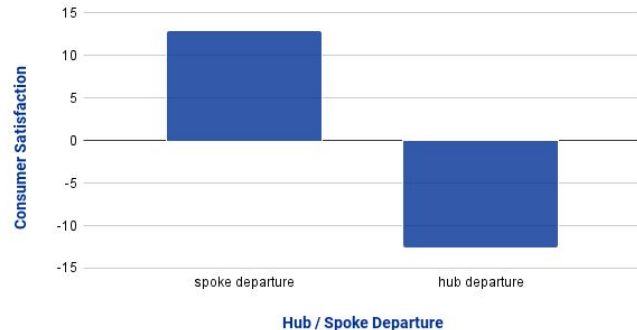
In the second chart:

- The hub departures are those leaving from a larger airport having many facilities, hence one would assume the F&B services here would be satisfactory but the data seem to show otherwise.

It is also worth noting the scale that each chart is showing as they are all ranging differently

ANALYSING THE RESULTS

Consumer Satisfaction Vs Flight Type

Consumer Satisfaction Vs Equipment Type
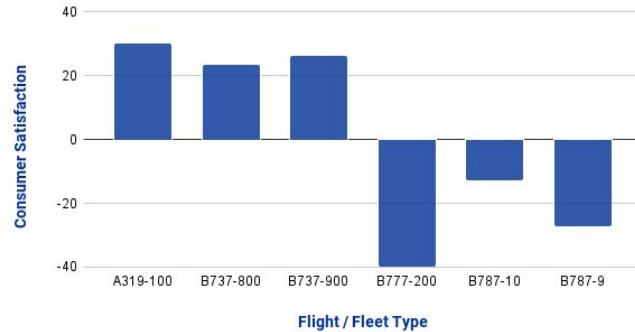
In the first chart:

- We can clearly see that some of the fleet / airplane types are preferred over the others, namely:
    - A319-100, B737-800, B737-900
    - It would be smarter to try to acquire more of these fleet types to increase the customer satisfaction
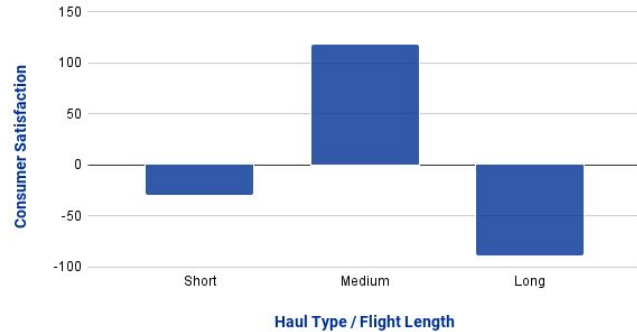
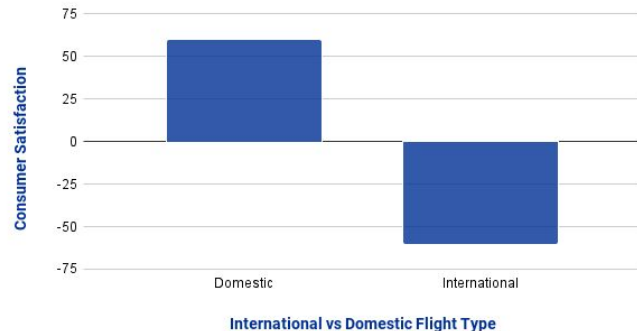Similarly in the second chart:

- We can that even for equipment type, there are some that are better as they increase consumer satisfaction, like:
    - 19F, 73Y, 37K, 19G

ANALYSING THE RESULTS

Consumer Satisfaction Vs Haul Type

**Consumer Satisfaction Vs International / Domestic**


In these charts:

- We can see an understandable dissatisfaction when it comes to long flights (which is also reflected in the second chart as most international flight tent to be longer) as at this point we can assume that the journey itself is creating more dissatisfaction than the F&B service
- The short journeys are also slightly in dissatisfaction due to the possibility that they may not serve food at all
- However we can see that the customers in medium and domestic flights are super satisfied with their F&B services, paying attention the first chart, the medium length has even broken the 100 unit barrier

ANALYSING THE RESULTS

Consumer Satisfaction Vs Seat Factor Band %
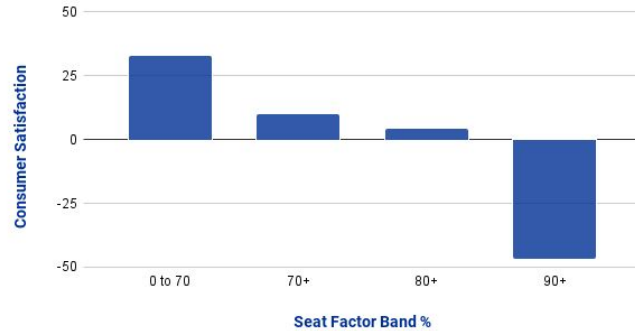
Consumer Satisfaction Vs Media Provider
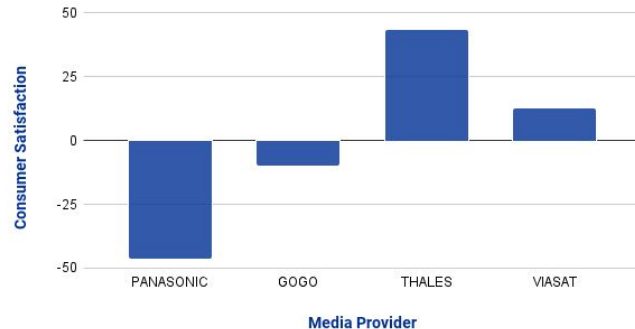
In the first chart:

- Flight that are filled to the brim also cause customer dissatisfaction as even your personal space gets cramped
- We can assume that the flight attendants are busy hence your food also may get delay leading to high dissatisfaction
- And we can see this hypothesis work when we consider the converse of this situation

In the second chart:

- It is interesting that even a media provider can change a customers F&B experience for example Parasonic causes a lot of dissatisfaction while Thales is causing satisfaction.

ANALYSING THE RESULTS

Consumer Satisfaction Vs Loyalty Program Level

In this chart:

- It is fascinating to see that those who are not a member of united are way more satisfied than those who are a member of the loyalty program.
- It can be assumed that those who are non-member are not frequent-flyers meaning that the high satisfaction is coming not from F&B but the flying experience itself.

PLANNING FOR THE FUTURE

With more time available, I would have proceeded by taking the (first question) segmented table from inflight satisfaction dataset and applying various machine learning algorithms to assess which one best predicts customer satisfaction or dissatisfaction.

I would then select the most effective algorithm and apply it to the other half of the inflight satisfaction dataset, which contains responses to the second question and details about the types of food and beverages (F&B). The result from this model prediction would give me which customers were satisfied or not and combining that with the data of which customer has ordered which item, in the end I would have a table showing which item causes satisfaction or not based on all factors. This approach would have allowed us to discern which specific types of F&B contribute to customer satisfaction and which do not.

PLANNING FOR THE FUTURE

Subsequently, I would have used natural language processing (NLP) on the second dataset, which contains customer reviews of F&B. This analysis would involve sentiment analysis to identify which F&B items are well-received and which are criticized. And then if my hypothesis holds true, the ratio of positive to negative sentiment for each F&B category would align with the results obtained from the machine learning model. Throughout this process, I would have also noted common themes recurring in customer reviews for further drawing conclusions.

Finally, I would have conducted data analysis on the planned versus consumed F&B counts available in the inventory dataset to gain insights into consumption patterns and potentially shortlisted some items that are preferred over the others. Then I would compare this data to the one we got from ML and NLP models to see one more time that the items shortlisted would not only have a larger ratio of satisfaction from ML model predictions, but we would also see a higher ratio of positive sentiment towards the frequently ordered items from the NLP model.

PROJECT LINK

Hackathon Link:

https://skyhack.hackerearth.com/

Project Links:

- GitHub Link: https://github.com/Gaurav7877/SkyHack_Hackathon
- Google Spreadsheet Link: SkyHack Hackation Excel Sheet Link (G-Drive)

THANK YOU