Analysis of Customer Satisfaction Survey for an Airline using Statistical and Machine Learning Techniques

Morteza Layegh Mirhosseini – student id: 220003166

Abstract:

Data science has become increasingly important in today's business landscape as it allows for more efficient and effective exploration of data. Besides, data analysis helps businesses to understand their customers better, improve customer targeting and sales, and find the best problem-solving strategies. In this report, we aim to analyze a customer satisfaction survey for an unknown airline in order to identify areas that need improvement to increase customer satisfaction. Using statistical techniques, we have identified groups of passengers with the highest dissatisfaction rates, and have also identified services that potentially can cause an unpleasant experience for passengers. Additionally, we have used a machine learning model to provide the airline with a strategy to prioritize service improvements. Our analysis has revealed that passengers under 40 and those traveling in economy class have high dissatisfaction rates, and improving online services is the most effective strategy for increasing passenger satisfaction.

Keywords—survey analysis, customer satisfaction analysis, machine learning

I. Introduction

The global financial climate requires precise resource management. Investigating customer satisfaction has become more common because providing high-quality service is necessary for airline companies to stay competitive[1]. In order to gain an advantage and improve business performance, it is necessary to focus on service quality and passenger satisfaction as key areas of focus[2].

One of the tools that companies can provide to measure their performance from a customer standpoint is surveys, specifically customer satisfaction surveys[3]. The main reason a survey analysis is crucial is that it enables us to infer more general information about our target customers. By analyzing a customer satisfaction service, we can investigate which cross-section of our target customers are reacting differently to a service/product. Therefore, we would be able to make more strategic decisions about our business. The benefit of survey data analysis is that it allows us to constantly work on enhancing our services and tracking results to stay one step ahead of the competition[4].

II. Analytical Questions

The aviation industry is a multi-billion dollar [5] and highly competitive industry. A comprehensive analysis of a customer satisfaction survey could help an airline to detect its weak spots in terms of services. Such analysis also allows for the investigation of the socio-economic characteristics and demographics of the airline's customers. With this knowledge, airlines can allocate well-targeted resources to improve key services, gaining an advantage and ultimately leading to more satisfied passengers.

With a focus on analyzing an unknown airline survey, this research aims to answer the following questions:

1. Which groups of passengers do not have a pleasant experience with the airline in general?
2. What are the most important airline services regarding the satisfaction rate of passengers?
3. What is the optimum strategy to improve services in order to increase the satisfaction rate?

III. Data

The dataset used for this analysis is from Kaggle (Airline Passenger Satisfaction)[6]. It contains 103904 rows and 25 columns. Each row corresponds to one passenger, and each column to a specific feature.

Table I. shows the different features in our data set and the corresponding values. Our dataset contains six features that allow us to group dissatisfied and satisfied passengers based on their gender, age, flight class, type of travel, flight distance, and type of customer. Some of these features will allow us to identify the demographic or socio-economic characteristics of the passengers. We use these features to answer the first research question.

TABLE I. FEATURES AND THEIR CORRESPONDING VALUES

feature	values
Unnamed: 0	Irrelevant
id	Irrelevant
Gender	male,female
Customer Type	Loyal, disloyal
Age	[7 to 85]
Type of Travel	Personal, Business
Class	Eco, Eco-Plus, Business
Flight Distance	[31 to 4983]
Inflight wifi service	(0: not rated; 1-5)
Departure/Arrival time convenient	(0: not rated; 1-5)
Ease of Online booking	(0: not rated; 1-5)
Gate location	(0: not rated; 1-5)
Food and drinks	(0: not rated; 1-5)
Online boarding	(0: not rated; 1-5)
Seat comfort	(0: not rated; 1-5)
Inflight entertainment	(0: not rated; 1-5)
On-board service	(0: not rated; 1-5)
Leg room service	(0: not rated; 1-5)
Baggage handling	(0: not rated; 1-5)
Checkin service	(0: not rated; 1-5)
Inflight service	(0: not rated; 1-5)
Cleanliness	(0: not rated; 1-5)
Departure Delay in Minutes	(0: not rated; 1-5)
Arrival Delay in Minutes	(0: not rated; 1-5)
Satisfaction	satisfied, neutral/dissatisfied

Furthermore, the dataset contains 15 features on an ordinal scale that relate to different services of the airline. Each passenger has rated each of these services from 1 to 5. A key consideration is to not be misled by the integer values of these service features. In fact, features that are on an ordinal scale, i.e., ratings from 1 to 5, are actually qualitative features characterized by ordered categorical responses. [3]. Therefore, we will treat the service features as ordinal qualitative data. These features will provide information to answer our second research question.

The final column we have is the *satisfaction* feature, which indicates whether each passenger was neutral/disatisfied or satisfied with their flight. Fig.1 shows the proportion of people who were satisfied with their flight is 44%, while the proportion of those who were neutral or dissatisfied is 56%. The satisfaction feature is our target label and helps us answer our third question. For the sake of simplicity, we will assume that passengers who are neutral/dissatisfied with the service are dissatisfied overall.

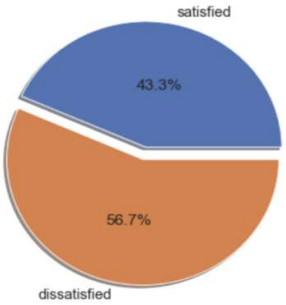

Fig. 1. The percentage of satisfied and dissatisfied passengers

Regarding missing values in our dataset, the *Arrival Delay in Minutes* feature has some missing values, with a total of 393 missing values. Given the relatively small size of these missing values compare to the size of our dataset, we decided to exclude them from our analysis.

IV. Analytical Approach

In order to answer the research questions, we followed these steps.

- Pre-processing the dataset, including cleaning, dropping irrelevant features, handling missing values, etc.
- Deriving new features using KDE (Kernel Density Estimation) curves
- Finding the groups of passengers with the highest dissatisfaction rate using contingency tables and 100% stacked bar charts
- Finding weak services using contingency tables and 100% stacked bar charts
- Fitting a random forest model to groups of passengers with the highest dissatisfaction rate to find out which service feature should be improved to increase the satisfaction rate.

V. Analysis

In this section, we will answer report questions in detail.

A. *Question 1; Which groups of passengers do not have a pleasant experience with the airline in general?*

To answer this question, we need to find the dissatisfaction rate of different groups of passengers. We can use contingency tables, such as *Pandas crosstab* [7], and plot the results using 100% stack bar charts[8] for this task. We start our analysis by finding the flight class which has the highest dissatisfaction rate:

1) The difference between flight classes regarding satisfaction rate

The airline offers three classes: Eco Plus, Business Class, and Eco. Fig.2 shows that only in Business Class most people find the service desirable. On the other hand, a significant percentage of people who choose Eco and Eco Plus are dissatisfied with the service quality, 81%, and 75% respectively. These numbers suggest that the airline is only focused on offering good services in Business Class. This may be justified, as nearly half of the passengers choose Business Class as it is shown in Fig.3. However, the dissatisfaction rate related to the Eco class is concerning (81%) as 45% of passengers use this class (Fig.3).

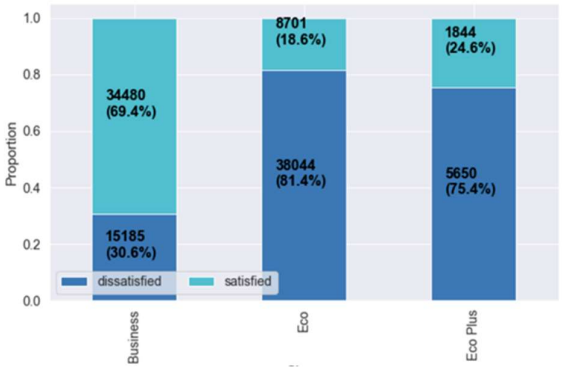

Fig. 2. Dissatisfaction rate in different flight classes

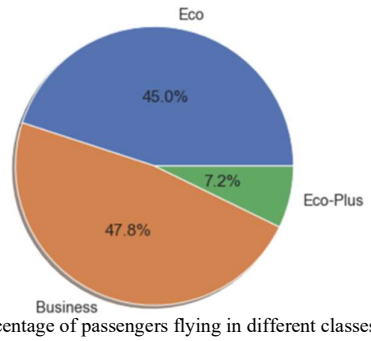

Fig. 3. The percentage of passengers flying in different classes

2) The difference between age groups regarding satisfaction rate

There are too few samples for each age to calculate the dissatisfaction rate for each individual value. For instance, there might be only a few 80-year-old passengers, and it would not be meaningful to calculate the dissatisfaction rate for this age individually [9]. Therefore, using the thresholds derived from the KDE distribution curves in Fig 4, we will categorize passengers into three age groups: young (less than 40), middle-aged (between 40 and 60), and senior (more than 60). This would allow us to calculate the satisfaction rate for each age group.

Looking at Fig. 5, the only age group with a satisfaction rate of more than 50% is the middle age group. It is possible that this group of people generally belongs to the high socio-economic class, and therefore can afford business class, which might explain why they are more likely to be satisfied with their flight. On the other hand, approximately 66% of young people are dissatisfied with the service. This is concerning as young passengers are accounted for nearly 50% of the passengers as it is shown in Fig. 6.

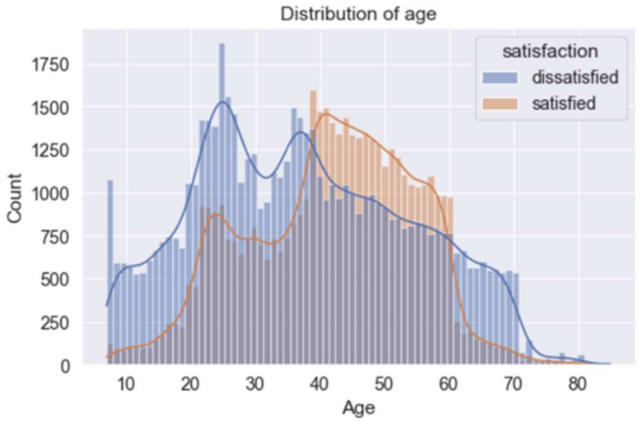

Fig. 4. Distribution of passengers with different ages and Kernel Density Estimation curves (KDE)

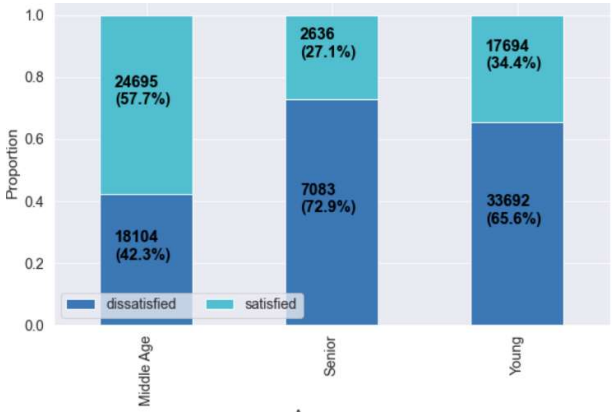

Fig. 5. Dissatisfaction rate in different age groups

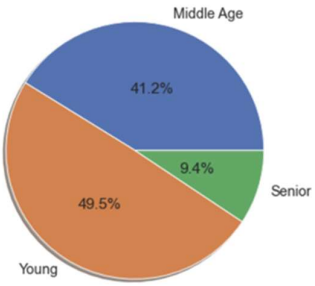

Fig. 6. The percentage of passengers in each age group

3) The difference between flight distances regarding satisfaction rate

We propose creating another feature called "distance_group" using the KDE distribution curve shown in Fig.7. We categorize passengers into two groups: those who traveled less than 1500 kilometers and those who traveled more than 1500 kilometers.

Fig. 8 indicates that the airline has a lower satisfaction rate for flights with distances shorter than 1500 km. Since nearly 70% of the airline's flights are in this category as it is shown in Fig. 9, it is important to investigate the reasons for dissatisfaction in this group.

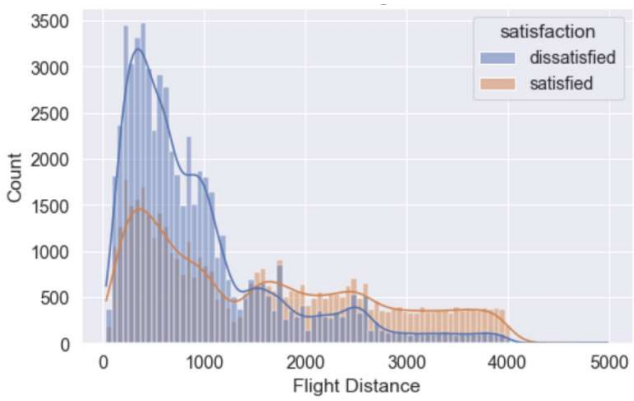

Fig. 7. Distribution of different flight distances

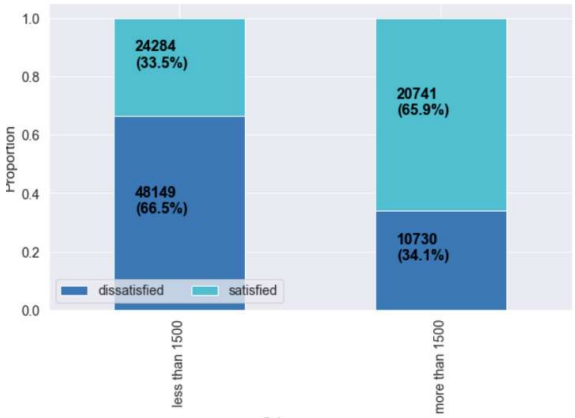

Fig. 8. Dissatisfaction rate in different flight distances

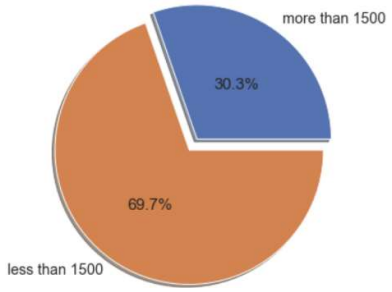

Fig. 9. The percentage of passengers flying more than 1500 km and less than 1500 km

4) The difference between flight reasons regarding satisfaction rate

Fig.10 shows that personal travels are accounted for nearly one-third of our dataset, but nearly 70% of passengers are traveling for business reasons. On the other hand, Fig.11 shows that nearly 90% of those who traveled for personal reasons are dissatisfied with the service, whereas only 58% of those who traveled for business are dissatisfied. This suggests that the allocation of airline resources may be prioritizing the provision of a better service for business travelers.

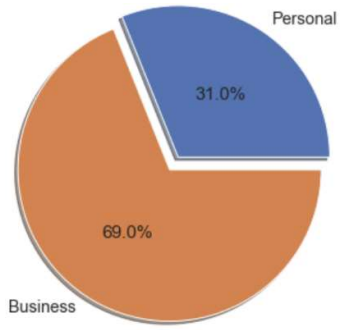

Fig. 10. Percentage of passengers flying for Business or Personal reasons

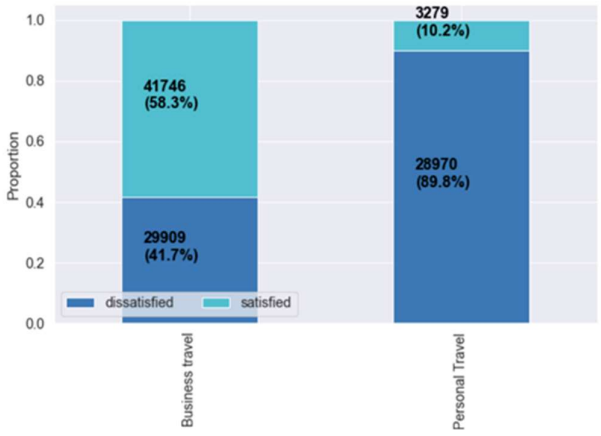

Fig. 11. Dissatisfaction rate in different Types of Travel

B. *Question 2: What are the most important airline services regarding the satisfaction rate of passengers?*

After finding those groups of passengers with the highest dissatisfaction rate, now our analysis focus to find the most important service features that can have a significant effect on satisfaction rate.

After plotting the dissatisfaction rate for each ordinal rating value of all service features (all figures are not included in the report), we used a visual analytics approach to identify the most important features in terms of satisfaction rate. We looked for the service features in which the dissatisfaction rate showed the biggest drop as the rating increased from 1 to 5. This approach allowed us to identify the following service features as those that had the greatest potential to affect customer satisfaction.

- Online boarding
- Inflight WIFI service
- Inflight Entertainment

By looking at Fig. 12, 13, and 14, we can see that the satisfaction rate for online boarding, Wi-Fi, and inflight entertainment increases as ratings increase from 1 to 5. A rating of 3 seems to be a turning point for customer satisfaction, with a significant increase in satisfaction rate as ratings move from 3 to 4 or 5. It is also notable that almost 100% of those who rated the Wi-Fi service 5 are satisfied with their flight.

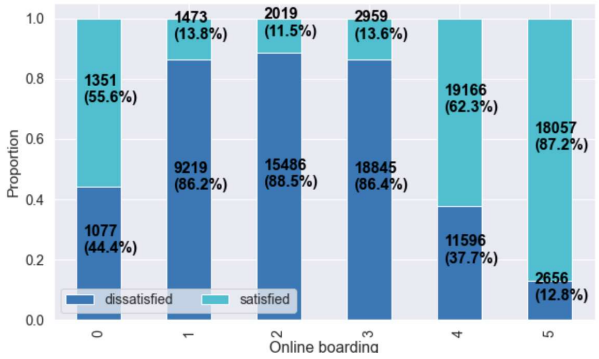

Fig. 12. Dissatisfaction rate across each rating scale for online boarding service

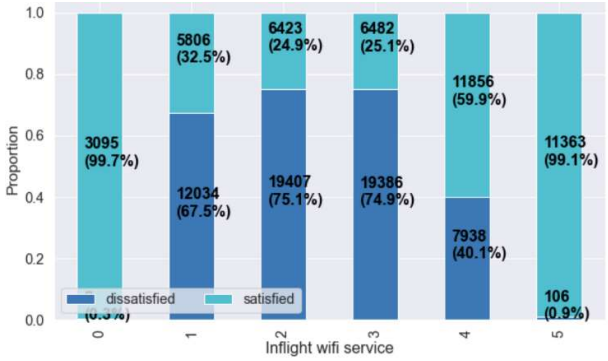

Fig. 13. Dissatisfaction rate across each rating scale for inflight wifi service

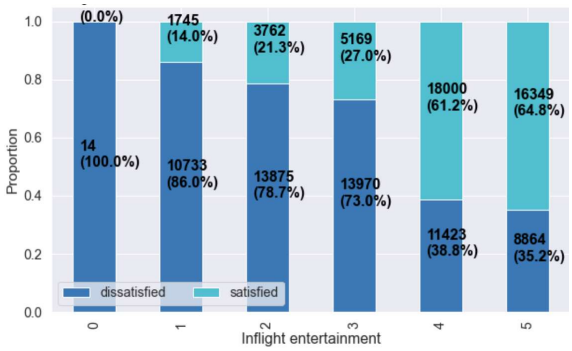

Fig. 14. Dissatisfaction rate across each rating scale for inflight entertainment

C. *Question 3: What is the optimum strategy to improve services in order to increase the satisfaction rate?*

To increase the overall satisfaction rate, the airline should be equipped with a reliable strategy. We can, also, use a random forest model to find the most important service features regarding customer satisfaction. The feature with the highest importance will do the best job of splitting the data with the lowest Gini impurity value[10].

Fig.15 shows that online boarding, inflight Wi-Fi service, class, type of travel, and inflight entertainment are the top five important features. Interestingly, our prior analysis of service features indicated that online boarding, inflight Wi-Fi service, and inflight entertainment were the most important factors for customer satisfaction, and our model appears to validate these findings. With a reasonably high accuracy of 96%, we can rely on this model for decision making.

In the final stage of our analysis, we focus on two groups of customers with high dissatisfaction rates: young customers (aged under 40) and those who traveled in Eco class. We trained a random forest model on these groups, and Fig.16 and Fig 17 show the feature importance results of these models. We can see similar results for these groups. The top important features are online boarding, inflight Wi-Fi service, and ease of online booking. This indicates the importance of improving online services for airlines to increase customer satisfaction.

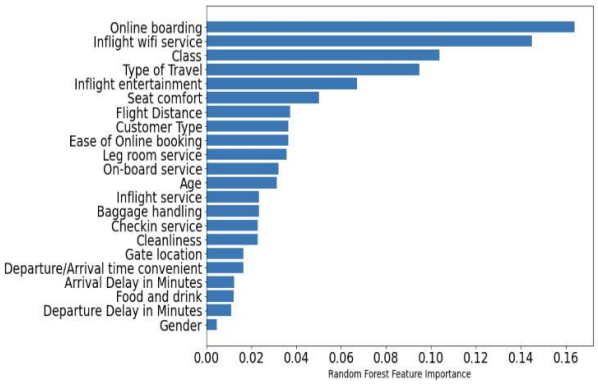

Fig. 15. Random Forest feature importance for all the features in the dataset

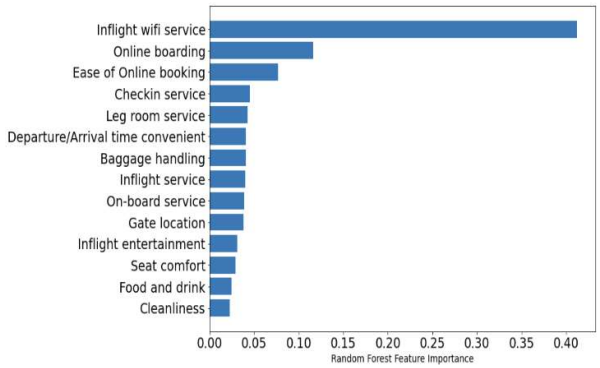

Fig. 16. Random Forest feature importance for passengers flying in Eco class

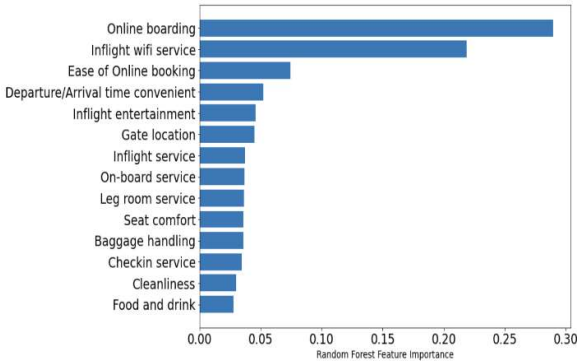

Fig. 17. Random Forest feature importance for the young group of passengers in the dataset

VI. Findings and Reflection

According to our information, the following groups of passengers are dissatisfied:

- 66% of passengers aged younger than 40 (young category) - this group accounts for 50% of passengers
- 90% of passengers who traveled for personal reasons - this group accounts for 31% of passengers
- 81% of passengers who traveled with Eco class flights - this group accounts for 31% of passengers
- 66% of those who traveled less than 1500 km - this group accounts for 70% of passengers

These numbers answer our first research question, which was what are the groups of passengers who had unpleasant experience in their flights. The airline should pay special attention to these groups and find the reasons behind their dissatisfaction in order to improve the overall satisfaction rate.

Our analysis of service quality features shows that online boarding, Wi-Fi, ease of online booking, and inflight entertainment are the most important factors regarding satisfaction rate. Passengers who rated these services less than 4 are more likely to be dissatisfied with their flight.

The random forest analysis validated our findings in the previous stage. Also, this model can help the airline to find which services should be prioritized for improvement. To do this, we trained the model on groups of passengers with a high dissatisfaction rates, including young customers and those

who traveled in Eco class. We found that the airline should improve online services, including Wi-Fi, online boarding, and ease of online booking, to increase the satisfaction rate for young passengers and those who travel in Eco class.

VII. Further work

The satisfaction of passengers is currently measured using a binary value that classifies them as either satisfied or neutral/dissatisfied. However, this method does not allow us to distinguish between neutral and dissatisfied passengers. If the satisfaction feature included a neutral value, we could conduct a more comprehensive analysis. Additionally, the data set could be improved by defining a range for the satisfaction level, such as a scale from 1 to 100. This would allow us to use regression analysis and provide more detailed insights.

Also, the feature *Type of Travel* is not clearly defined by the dataset. However, this feature represents a group of passengers which has a very high dissatisfaction rate. This feature may define the reason for the travel, which is assumed in this report. However, by making this feature more clear, we can conduct a more comprehensive analysis.

References

1. Park, J.-W., R. Robertson, and C.-L. Wu, *The effect of airline service quality on passengers' behavioural intentions: a Korean case study.* Journal of Air Transport Management, 2004. **10**(6): p. 435-439.
2. Li, W., et al., *A hybrid approach based on fuzzy AHP and 2-tuple fuzzy linguistic method for evaluation in-flight service quality.* Journal of Air Transport Management, 2017. **60**: p. 49-64.
3. Eboli, L. and G. Mazzulla, *An ordinal logistic regression model for analysing airport passenger satisfaction.* EuroMed Journal of Business, 2009.
4. Frampton, S. *A Beginner's Guide to Survey Data Analysis and Data Collection*. 2020; Available from: https://www.chattermill.com/blog/survey-data-analysis#how-to-analyze-survey-data.
5. *IATA reveals latest outlook for airline industry financial performance*. Available from: https://www.internationalairportreview.com/news/164767/airline-industry-201-billion-loss-iata-reveals-imrpoved-results-covid-19/#:~:text=The%20passenger%20business%20will%20contribute,to%20%24378%20billion%20in%202022.
6. Klein, T. *Airline Passenger Satisfaction*. Available from: www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction.
7. Andrienko, N., et al., *Visual analytics for data scientists*. 2020: Springer.
8. Muralidhar, K. *100% stacked charts in Python*. 2021; Available from: https://towardsdatascience.com/100-stacked-charts-in-python-6ca3e1962d2b.
9. Chris, D.s.w. *Kaggle Titanic – Data Analysis*. Available from: https://datasciencewithchris.com/kaggle-titanic-data-analysis/.
10. Płoński, P. *Random Forest Feature Importance Computed in 3 Ways with Python*. 2020; Available from: https://mljar.com/blog/feature-importance-in-random-forest/.

Word Counts

Section	Word count
Abstract	147
Introduction	161
Analytical Questions	134
Data	323
Analytical Questions and Analysis	997
Findings	253
Further Work	141