Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

# Unlocking Insights into Airline Passenger Satisfaction with Data-Driven Exploratory Mining Approach

Khawaja Qasim Maqbool
Dept. of Computer Science, Bahria University Lahore Campus
Email: qasim@bahria.edu.pk

Maham Fatima Kayani,
Phd Student at Le CNAM Paris. Email: Maham-fatima.kayani@lecnam.net

Muhammad Zulkifl Hasan (Corresponding Author)
Department of Computer Science, Faculty of Information Technology, University of Central Punjab Lahore Pakistan. Email: Zulkifl.hasan@ucp.edu.pk

Muhammad Zunnurain Hussain
Dept. of Computer Science, Bahria University Lahore Campus
Email: Zunnurain.bulc@bahria.edu.pk

Muhammad Atif Yaqub
Department of Computer Science, National College of Business Administration and Economics, Lahore, Pakistan. Email: Atif.yaqub@ue.edu.pk

**Abstract**
This paper thoroughly analyzes an airline passenger satisfaction survey dataset by employing Exploratory Data Analysis (EDA) and machine learning modeling by leveraging Python-based libraries including Pandas, NumPy, Seaborn, Matplotlib and a few others.
Through EDA, we examine diverse aspects of the dataset, encompassing flight attributes, delays, and service ratings. Subsequent utilization of machine learning techniques such as regression, classification, or ensemble methods aim to uncover patterns and predict passenger contentment levels. The integration of EDA and machine learning methods enhance insights, validate findings, and contribute to a comprehensive understanding of factors influencing passenger satisfaction within the aviation sector.

**Problem Statement**
The central problem addressed in this paper is to extract and analyse meaningful insights from the 'Airline Passenger Satisfaction' dataset to deepen the understanding of the factors influencing passenger satisfaction in the airline industry. The challenge is to navigate through the data, cleanse it for accuracy, and perform an extensive visual and statistical examination to determine these factors.
To address this problem, the study employs a range of machine learning models, each with distinct capabilities and evaluation metrics. Linear Regression is used as a baseline for comparison, leveraging its simplicity to establish a linear relationship between features and the target variable. Logistic Regression is deployed for its suitability in binary classification tasks, such as estimating the probabilities of passenger satisfaction levels. The Random Forest Regressor, known for handling unbalanced datasets with high accuracy through its multiple decision trees, and the AdaBoost Regressor, which

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

combines several weak learners into a stronger model, are used for their robustness in predictive performance. These models are rigorously evaluated using metrics appropriate to their type: Mean Squared Error (MSE) and $R^2$ Score for regression models, and the Accuracy Score, Precision, Recall, and F1-score for the classification model. The problem statement underscores the goal of the research, which is to leverage these models to identify and understand the key determinants of passenger satisfaction, thereby contributing valuable insights to the airline industry.

## Introduction

Maintaining and improving passenger satisfaction in the dynamic airline sector is not just a goal, but also essential to its existence and success. Innovative methods for studying consumer preferences, habits, and satisfaction factors have been made possible by the development of data mining techniques. The goal of our research is to use a comprehensive dataset of 129,880 entries that covers a wide range of attributes to dig deeply into the complicated structure of factors which influence airline passenger satisfaction. Air travel was traditionally considered a luxury, but it is now widely used by individuals and groups for mobility. As a result of this shift, client expectations have also changed, becoming more complex and demanding. Airlines now have to make sure that passengers have a pleasant, personalized, reliable travel experience in addition to safe and timely flights. Understanding the several elements that affect a passenger's perception and satisfaction is essential to doing this.

Our dataset at the heart of this project is a rich repository of insights. In addition to travel-related statistics like customer type (first-time or returning), trip type (personal or business), flight class, and flight distance, it also contains demographic information of the passengers, such as age and gender. Additionally, it offers an in-depth analysis of the experiences of the travelers, covering everything from pre-flight features like internet booking and check-in services to post-flight features like leg room service, comfortable seats, clean cabin, good food and drink, in-flight entertainment, and Wi-Fi access. The stats on arrival and departure delays impact and play a critical role in determining the overall travel experience.

The objective of this project is to apply advanced data mining techniques alongside machine learning algorithms to uncover patterns and correlations within this dataset to address important queries such as: Which variables make the biggest effects on how satisfied passengers are? Do different flying classes or populations have significantly varied expectations and satisfaction levels? What is the overall satisfaction level formed by the interactions between the different components of the flight experience?

This study will use a variety of data mining techniques, such as clustering, classification, regression analysis, and association rule mining, to answer the underlying concerns which is a crucial step towards understanding the modern air traveler's psyche. The insights derived from our study have the potential to guide airlines in reshaping their service offerings, leading to improved customer satisfaction, loyalty, and, ultimately, the success of the airline in a highly competitive market.

## Literature Review

### Service Quality

As a major contributor to tourism, international trade, and economic development, the airline business is crucial to the world economy (Ganiyu 2017; Ishutkina and Hansman 2008).

Evaluating and enhancing customer satisfaction is a core aspect in this industry as it is

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

fundamental in the retention of the market and in the provision of future growth. There have been numerous studies conducted with regards to the quality service of airlines. The SERVQUAL model was proposed by Parasuraman et al. (1988) and is regularly used in a multitude of industries including aviation to evaluate its quality services.

In order to highlight the importance of understanding clients' needs and providing them with competent and considerate service, this model's measured ten characteristics were subsequently modified to add two new features; assurance and empathy (Parasuraman et al. 1988). Research in this area has demonstrated how service quality can influence a passenger's choice when selecting an airline (Ostrowski et al., 1993). In addition, Tsaur et al. (2002) and others have also examined various forms of service quality and found that the service quality of tangibility and responsiveness were also important measures (Gilbert , Chen & Chang et al., 2005).

**Customer Satisfaction**

In the airline industry, there are many elements that affect customer satisfaction, which is important for the sustainability and profitability of the airline firm (Kotler et al., 2009). However, because of its subjective nature and the multifaceted structure of service organizations, the concept is complicated (Qin et al., 2010). Research has indicated that several service quality parameters, including baggage handling, pre-, in-, and post-flight services, have an impact on passenger satisfaction in airlines (Archana, 2012). Thus, ensuring satisfaction requires knowing and meeting client expectations.

**Market Segmentation and the RFM Model**

In the airline sector, meeting client preferences and improving service offerings depend heavily on the idea of market segmentation (Khan, 2014). Age, gender, and geography are examples of socio-demographic factors that have been the focus of traditional segmentation techniques.

Due to the shortcomings of these techniques, cluster-derived segmentation strategies based on consumer behavior and purchasing preferences have been developed (Hassan et al., 2005). The Recency, Frequency, and Monetary Value (RFM) model is one useful technique in this field. In order to better understand and segment customers, this model has been modified for the airline business and includes more behavioral factors such as journey purpose, flight intention, class, and frequency (E.-C. Chang Huang et al., 2010).

**Variability in Evaluation Scales for Measuring Airline Passenger Satisfaction**

The literature shows that the evaluation scales that were used to get the opinions of the passengers varied significantly. Hu and Hsiao (2016) indicate that the most popular method employed in the rating customer satisfaction in Chinese studies involves the use of a five-level scale where one can indicate being "strongly dissatisfied" and another being "strongly satisfied". Interestingly, others use a scale of seven indicators which implies that different studies are able to employ various means of measuring customer satisfaction and even judgement (Kuo, 2011).

Apart from the first type, it has also been observed that the five-point scale adequacy includes the use of measures that say, "very poor" to "very good" (Chou et al. 2011). This measurement approach confines one to the construct of satisfaction as opposed to quality. While it can prove to be helpful because it measures the quality of services offered to a passenger, this measure may not be conclusive to the emotional or experiential part of the passenger's satisfaction which is often seen as an attribute that is subjective. The choice of a particular scale in a survey greatly determines the outcomes of the survey in terms of data analysis and interpretation. More complicated scales are expected to return richer data sets but at the same time response patterns would be more elaborate, (Bellizzi et al., 2020).

On the other hand, although a simpler scale may be easier for participants to use, it may miss minor differences in experiences and perceptions. As a result, choosing the appropriate scale demands finding a balance between the importance of precise data and the ease of conducting surveys (Bitner, 1992).

Suki's (2014) study provides a thorough analysis of Malaysian airline service quality, emphasizing the impact of empathy and tangible characteristics of service on customer satisfaction and the subsequent influence on word-of-mouth (WOM) recommendations. The study highlights the need of providing high-quality service delivery in the fiercely competitive airline market and is conducted against the backdrop of Malaysia's expanding airline business, which is crucial to the nation's tourist and economic growth (Suki, 2014).

Data mining decisively identified the links between water quality models and areas such as Artificial Intelligence models which could be replicated in passenger satisfaction datasets and other similar complications (Al Noman et al., 2024). Companies that have adopted AI business intelligence have been seen to utilize information in a manner that supports their business decision processes which is useful to the airlines companies that intend on nurturing their clients (Rimon et al., 2024). The IoT data lakes management has revealed the need for elastic platforms that can cope with large volumes of data which will go a long way in assisting the airline industry with data mining activities (Nuthalapati, 2023). Airlines consider business to be very sophisticated and therefore strategies and decision making is enhanced by insights which come up as a result of machine learning techniques in businesses (Sufian et al., 2024). The projections on the optimization of the AI and Quantum computing suggest such technologies could help in eliminating redundancies in business processes within airlines thus improving passenger experience (Mosaddeque et al., 2024). Airlines are able to address issues of forecasting given that they are able to utilize machine learning techniques in the preparation of various operational activities especially relating to the passenger journey (Nuthalapati et al., 2024). While AI for predictive analytics in load forecasting proves its worth in the real environment, it equally shows great capabilities in the real task of big engagement with the passenger data sources (Ahamed et al., 2024).

Predictive analytics for healthcare has shown how transformative AI can enhance outcomes, offering parallels for improving passenger satisfaction metrics in the airline industry (Tarafder et al, 2024). Architechting data lake-houses in the cloud has underscores the necessity of efficient data storage solutions for complex datasets, crucial for large-scale airline data analysis is (Nuthalapapati, 2024). AI-driven optimization techniques in smart grids further highlight how intelligent systems can improve operational efficiency, providing insights for airlines to optimize service delivery and satisfaction (Ahamed et al, 2024).

The SERVQUAL scale by Parasuraman, Zeithaml, and Berry (1988) outlines three key aspects of service quality that the research is centered around: airline tangibles, terminal tangibles, and empathy. As defined by Norazah in 2013, empathy is compassion extended on an individualized basis towards clients whereas tangibles encapsulate the service design including furnishing, equipment and general atmosphere. Empathy appears to play a very important role in customer satisfaction as other factors such as the point of departure and arrival being well regulated, good flight

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)
connections, and decent service on the flight have been pointed out as critical (Suki 2014).
Apart from being crucial to the global gross domestic product, the airline industry complements a number of other industries like the retail, hospitality and transport sectors (Ganiyu et al. 2008). In this context, the current study focuses on the Iranian airline industry and discusses the issues of competitive pressure and changing needs of the passengers which make it difficult for the airlines to retain their market share (Liou et al., 2011).
Of particular note in the context of Iran is the work of (Tahanisaz and Shokuhyar 2020) which greatly enhances our comprehension of factors that drive satisfaction of airline passengers. These writers combine such innovative methods as ISA and the Kano model and offer insights and recommendations to airline executives on enhancing the quality of service to help them meet the needs and expectations of their customers, thereby increasing their satisfaction and loyalty.
Factors such as the ease of check-in processes and pre-flight interchange communication have a bearing on the perception of the pre-flight experience. These impacts are influenced by the airline's services as well as other elements of the airport that are not directly related to the airline’s operations (Truitt et al, 2006). From their findings, the authors reach the conclusion that investment into the enhancement of pre-flight facilities will probably have an effect on the passengers’ loyalty towards the airlines.
His study suggests that Airlines should also focus on factors weakening pre-flight services and affect the comfort of the passengers and their loyalty like communication to them before the flight and during check-in (Teichert, 2008). The role of pre-flight/in-flight service quality on customer retention in essence underscores the need for airlines to adopt targeted strategies to improve these aspects of service provision. (Etemad-Sajadi et al. 2016) provide a comprehensive explanation for the superior effects of pre-flight versus in-flight service quality on passenger loyalty.
The paper suggests that subsequent studies should investigate the drivers of the relative service quality of the airlines as perceived by the passengers, the factors which come before customer loyalty and the impact of pre- experienced service quality on the clients in the hospitality industry. This investigation adds considerable value to the existing knowledge on the airline service quality and its effect towards customer loyalty.

## Methodology

### Data Collection

The dataset utilized in the study is named “Airline Passenger Satisfaction” and has been obtained with the aim of supporting our research. Data provided by Kaggle embraces a number of attributes capturing the f light patterns of passengers together with their levels of satisfaction. In the data pre-processing stage, the cleaning of the data set was done, and its relevance ensured to the target analysis.
The flowchart is the graphic representation of a typical cycle of data analysis which starts from data acquisition through to data extraction for processing. Where there is an indication of problems such as duplication or missing values, the data is managed such that it is cleaned. Then it is time for exploratory data analysis which helps in developing graphical representations of the data and seeking to identify essential patterns and trends in the data. Some of the machine learning approaches employed in tackling the objectives include regression to estimate values and classification to categorize ordered data. At the end of the project the model outcomes are produced, assessed in regard to

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

their accuracy and appropriate communication methods are employed before the project closes.

**Data Preprocessing**

Cleaning and Preprocessing Steps are as follows:

Duplicate Removal: First and foremost, we addressed this problem of duplicate data since some records were found to be redundant and might affect the analysis.

Dealing with Missing Data: The 'Arrival Delay' parameter was particularly checked for any missing values as well as the entire dataset. Due to the fairly low percentage of missing values, it was determined that records with these missing values would be omitted to safeguard the quality of the dataset.

Encoding: Categorical variables were transformed into a numerical representation by employing the appropriate encoding methodologies so these variables can be utilized in quantitative analysis.

Data Reduction: We have removed irrelevant variables like passenger ID from the database which were one of a kind for each instance and therefore had no significance for the analysis.

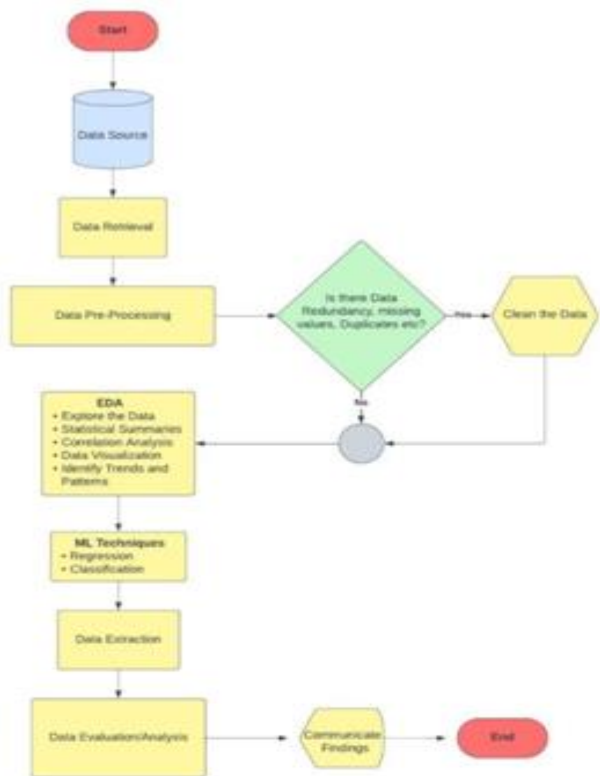



**Fig.1 Flow Chart: Methodology**

**Exploratory Data Analysis (EDA)**

We calculated and evaluated the measures of central tendency, dispersion and shapes of the distribution for the numerical variables to make sense of the descriptive statistics.

1. **Histograms**: We plotted histograms for continuous variables like 'Age' and 'Flight Distance' to visualize their distributions.

2. **Boxplots**: To understand the distribution of satisfaction across different categories, we used boxplots, contrasting passenger satisfaction against variables like 'Age' and 'Flight Distance'.

3. **Correlation Analysis**: A heat map was generated to visualize the correlation between all numerical variables, which helped in identifying potential predictors for passenger satisfaction.

4. **Scatter Plots**: We used scatter plots to investigate the relationship between pairs of variables, notably between 'Departure Delay' and 'Arrival Delay'.

The fourth step in our workflow is concerned with model training, evaluation, and validation of the different machine learning techniques selected for the study. Therefore, this step encompasses two essential components: firstly, training and testing the algorithms/models; and secondly, evaluating their predictive and selection performance. Four machine learning models were constructed and validated:

1. Modelled Linearly: this is a basic model that proposes a direct relationship between the selected features and the outcome. Here it was used only as the benchmark.
2. Logistic Regression: A model describing class membership probabilities. This kind of model is appropriate for two-class classification problems such as predicting the level of satisfaction.
3. Random Forest Regressor: A multiple model consisting of many decision trees and known to predict accurately, especially in cases of unbalanced datasets.
4. AdaBoost Regressor: Also, an ensemble method which utilizes several weak classifiers to build a strong one. It is widely used to boost the outcome of decision trees.

At this stage the models were evaluated with suitable metrics: For regression models (Linear Regression, Random Forest and AdaBoost), Mean squared Error (MSE) and $R^2$ Score were employed. For the classification model (Logistic Regression), Accuracy Score and Classification Report including precision, recall and F1-score were computed.

**Data Mining Techniques and Analysis**

The analysis was performed with the aid of several software tools and libraries including:

• Libraries of Python Language such as pandas for data manipulation and Scikit-learn for machine learning

• Spreadsheet widely used software for prototyping and data visualization

• Jupyter Notebook to document and run the codes during the analysis Using correlation and regression analysis to substantiate the proxies of predictors and their relationship with the customer satisfaction level was our primary focus, this idea helped in determining how various predictor variables' strength and direction worked with satisfaction level, and which variables were more strongly, or more reliably, associated with the level of satisfaction in the context of the assessment model. A lot of information fostering customer satisfaction effects within the given data set in the focus of our

analyses were in this case provided in return.
From the perspective of the first analysis, we proceeded to building and evaluating models which were aimed at predicting the level of customer satisfaction. It should particularly, be pointed out that having established a high correlation between two variables 'Arrival Delay' and 'Departure Delay' raised concerns regarding linearity. In order to deal with the problem, we chose to employ only one of them in the modeling making sure that it will not be redundant or worsen the performance of the model.
Next in the series, the classification algorithms started to consider outlining in our strategy. The goal of these algorithms was to identify the characteristics of satisfied and dissatisfied passengers based on the features that were determined as important upon doing the preliminary exploratory data analysis. We sought to apply classification methods which would cut across these levels against the influencers considered and the expected outcome on passengers' satisfaction levels.
The correlated analysis came first to determine whether there were any associations, then the addressing of the multicollinearity problem through the use of models followed by the consideration of the use of the classification algorithms came in second and this ordered chaining formed a strategy to help understand the levels of passenger satisfaction depending on the determinants of the satisfaction which have been established.

**Ethics and Data Privacy**

During the rigorous process of data mining, our number one priority was to ensure compliance with ethical standards. All measures were taken at all times to protect passenger privacy. This required a policy whereby all identifiable information was neither used nor revealed during the entire process of analysis. The other intent of mooting these ethical issues was to preserve the rainality protection of all people using the 'Airline Passenger Satisfaction' dataset.
The procedure that this study followed also enabled a purposive investigation of the dataset. The analysis of such particulars enabled us to understand the various determinants that relate to passenger satisfaction. After identifying these determinants, we managed to locate the most important areas which the airline could work on in order to enhance the experience of its customers even more.
These results are not mere theoretical propositions, but rather are of practical significance. The actionable insights gained from this wide-ranging study are expected to be a strong basis on which the airline can make its political decisions. By utilizing these conclusions, the business will be able to adjust its tactics, improve the company's operations, and carry out constructive modifications that coincide with the highlighted concern. To conclude, the results of this investigation are meaningful in that they provide helpful suggestions that should improve the quality of services offered and customer satisfaction levels in the airline industry.

**Discussion**

For the purpose of analyzing and gaining knowledge from the dataset, appropriate methods were taken step-by-step in a manner which ensures the data remains consistent and credible. These steps have been explained in detail as follows: Data Acquisition and Preliminary Exploration Beginning with the Loading of 'Airline Passenger Satisfaction' dataset which is required for this analysis and which is assigned to a Data Frame by panda's library, The analysis starts with the loading of the dataset by which the data frame is assigned. The data is viewed in broad sense by pulling out several initial records in order to view a few records. It is crucial as it helps in preparing for more detailed analysis later on. The dataset consist of 129880 rows with 24 columns which is quite large for analysis. Data Cleaning and Preprocessing The stage of data set cleaning and preprocessing is arguably most important when it comes to conducting the analysis. This process consists of:

Identifying Unique Values-There is determining examination of every column for its unique values, which plays a significant role in evaluating the extent of diversity and range of data base over different attributes.

Duplicate Record Management-The dataset is also checked for duplicate instances and the problem of duplicates records is dealt with in order to maintain the integrity of the data. Duplicate records can distort the findings in a manner that inappropriate decisions would be made.

Dealing with Missing Values: The column where some values are missing, for instance 'Arrival Delay', is one of the columns that receive special focus. Out of all 393 cases without 'Arrival Delay', it has been determined that they will be deleted. This step ensure that the dataset and its reliability to be of high standards.

Data Frame Information Review: A first view of the structure of the dataset is done using the info() function, which provides information about the variable types and number of variables which are not null. This information is crucial in formulating how the data would be processed and analyzed further.

*Descriptive Statistics*

The describe() function provides a detailed statistical overview of the dataset. As a part of such summary, averages, medians, standard deviations and ranges of variables are provided for every numeric variable in the dataset. This step is important in consideration of the factors such as the overall distribution and measures of central tendencies of the dataset prior to advanced statistical procedures being implemented.

*Use of Python Libraries*

1. Pandas: This is the primary library for performing a wide spectrum of data manipulation and yearly tourism analysis in Python. In this analysis, pandas is mainly used to pull the dataset from a csv file into a DataFrame. The structure of DataFrame which is provided by pandas enables to easily and conveniently perform cleaning, transforming and analyzing of tabular data sets. Functions such as head(), info(), describe() and dropna() are useful for basic data exploration, summarization and treatment of outliers with other issues around missing values.
2. NumPy: This library is critical for scientific computations using Python. It mainly focuses range of numerical computing which is very vital for any kind of data analysis process. In this analysis, NumPy works well in complementing the function of pandas for instance in performing mathematical computations on the columns of a DataFrame.
3. Matplotlib.pyplot and Seaborn: These libraries are quite important in relation to the aspect of data visualization. Matplotlib.pyplot allows various forms of static plots to be drawn in a high level of detail.

**Visual Exploration of Data**

Visualizations are useful in identifying trends and other relations in the attributes of the dataset:

4. Age Distribution: The age distribution of passengers on the airline service is graphically represented by a histogram with an overlaid Kernel Density Estimate.

**Dialogue Social Science Review (DSSR)**
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

This visualization aids understanding of the proportion of different age groups of the airline customers.

5. Flight Distance Distribution: In the same vein, the flight distances are analyzed by using the means of the histogram. This is significant as it adds to understanding the diversity of the passenger travel patterns and travel preferences on the average distance that they cover in every journey.
6. Departure and Arrival Delay Analysis: The use of scatter plot for both 'Departure Delay' along with 'Arrival Delay' is employed as a means of evaluating the number and intensity of delays since these parameters are crucial for passenger satisfaction rate.
7. Correlation and Comparative Analysis: A heatmap is generated correlating with reporters at USS. More specifically, it was noted that there was a strong relationship between 'Departure Delay' and 'Arrival Delay'. This suggests that on an average these variables are strongly related and such relationships may be helpful in estimating passenger satisfaction.

In addition, boxplots are used to test if 'Age' and 'Flight Distance' have an effect on the level of passenger satisfaction. This relation compares the effects of these variables as to whether or not they are good predictors of satisfaction.

5. Data Refinement for Future Modelling: First, what is noteworthy about the fourth section of this study regarding Predictive Analytics is that it addresses aspects that would have already been implemented comprehensively. In other words, employing the conclusions drawn from the EDA, refinement of the dataset is performed for the purpose of predictive model building in this section.
6. Data Refinement for Future Modelling: In other words, employing the conclusions drawn from the EDA, refinement of the dataset is performed for the purpose of predictive model building in this section. Thus certain columns can be deleted selectively, in this case columns like, "ID," "Age," and "Arrival Delay" are assumed to be useless and selected to be deleted.

**Exploratory Data Analysis (EDA)**

The current report examines in detail Exploratory Data Analysis (EDA), which was performed on the airline passenger satisfaction data set. After EDA was performed, a lot of stages were involved with the aim of fully comprehending the data set and getting useful information from it. The subsequent parts describe all the steps that have been utilized in this EDA process one by one.

**Initial Data Understanding**

• Data Snapshot: The initial step involved the visualization of a subsection of the attribute data which is composed of customers' gender, age, type, and rating service. This mix of categorical and numerical data was meant to give an aggregate picture of the dataset structure.

• Uniqueness Check: The unique mzhe attributes in columns were allowed through the utilization of df.nunique(). This step revealed some unique aspects of the data set, for instance, diversity of ages and number of unique values for gender.

• Duplicates Check: The use of df. duplicated() to check whether there exist duplicate records in the dataset established that the dataset was free of duplicate records thus safeguarding the analysis data.

• Data frame Information: Indeed, this information was simply obtained through the use of df.info() which provided relevant descriptive statistics such as the composition of

**Dialogue Social Science Review (DSSR)**
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)
columns, the number of null values, the size of the data set in terms of unique integer variables and category variables.

• Null Values Check: A further check of the data frame using df.isna(). sum() to assess for the presence of missing data established that a good number of the columns had little data which was missing. There was focus on the 'Departure and Arrival Time Convenience' which indicated that 393 of its respective data points were missing and later remedied the situation by removing the respective rows for such occurrence.

• Descriptive Statistics: The use of set out dummy variables including set out dummy variables including df.describe(). Provided that needed set of the statistical parameters. This greatly lend understanding in the areas of mean, variance and skew of the values in the numerical columns.

**Visualization and Data Enhancement**

- Age Distribution Visualization: The utilization of histograms overlaying kernel density estimates (KDE) effectively visualized customer age distribution, indicating a predominant adult clientele aged between 20 and 60.
- Flight Characteristics Visualization: Visualization of flight distance distribution showed that most flights were short-distanced, likely reflecting the airline's frequent use for shorter trips.
- Delay Insights Visualization: Visualization of departure delays highlighted the airline's positive trend in minimal departure delays, reflecting potential punctuality.
- Missing Values Imputation: Addressing missing values in the 'Arrival Delay' column through mean imputation ensured completeness in the analysis.
- Arrival Delay Distribution Visualization: A visualization of arrival delays mirrored the departure delays' trends, showcasing minimal delays and hinting at potential outliers.

*Categorical Data Analysis*

- Gender and Passenger Satisfaction: A pie chart depicted an equal gender distribution among passengers, signifying a balanced attraction to both genders.
- Passenger Satisfaction Levels: Visual representation indicated that 57% of passengers fell into the neutral or dissatisfied category, highlighting an area for potential improvement.
- Customer Types and Travel Purposes: Analysis revealed a significant majority of returning customers (82%) and a dominance of business-related travel (69%), shaping strategic insights for service enhancements and targeted marketing.

*Variable Relationships and Feature Selection*

- Correlation Analysis: A heatmap visualized correlation coefficients, unveiling relationships between variables, such as departure and arrival delays.
- Scatter Plots and Box Plots: Further analysis through scatter plots and box plots examined relationships between variables, identifying flight distance as a potential predictor of satisfaction while showcasing age's limited influence.
- Feature Selection: Based on the analysis, certain features, including ID, Age, and Arrival Delay, were deemed less predictive and subsequently dropped for further modeling.

The EDA undertaken on the airline passenger satisfaction dataset offered insights into demographics, satisfaction levels, and variable relationships. The findings highlight potential areas for improvement, guiding future modeling efforts to enhance

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

customer satisfaction and operational efficiency. By delving into this study, one can find out what makes certain airlines successful where others fail, how those failing airlines can improve their business and what areas they would need to focus on, both in the short run and the long run.

**EDA Visualization**

To develop a deeper understanding of the dataset at hand, we derived the relationships between them after detailed analysis to help summarize our findings to the readers. For example, the histogram for Age Distribution displays the spread of passenger ages.

- **Bell-Shaped Curve**: The distribution roughly follows a bell-shaped curve, albeit not perfectly symmetrical, suggesting a normal-like distribution of ages among passengers.
- **Mostly Middle-Aged Passengers**: The peak of the distribution is around the 30-40 age bracket, indicating that a significant proportion of the airline's passengers are in their middle age.
- **Wide Age Range**: The distribution covers a wide range of ages from young to elderly passengers, with noticeable frequencies up to the age of 80.
- **Young and Old Passengers**: There's a gradual increase in frequency from the youngest ages up to the peak, followed by a more gradual decline, which suggests that there are fewer very young (children) and older (senior) passengers than there are middle-aged passengers.

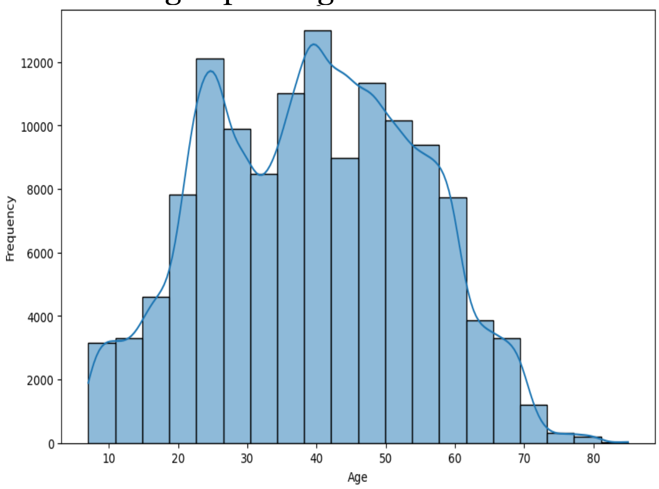



**Fig 2: Age distribution in airline passenger satisfaction**

These age demographics can be valuable for marketing and service adjustments. For instance, if the airline wants to increase patronage among younger or older age groups, they might consider introducing targeted amenities or services that appeal to these age brackets.

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

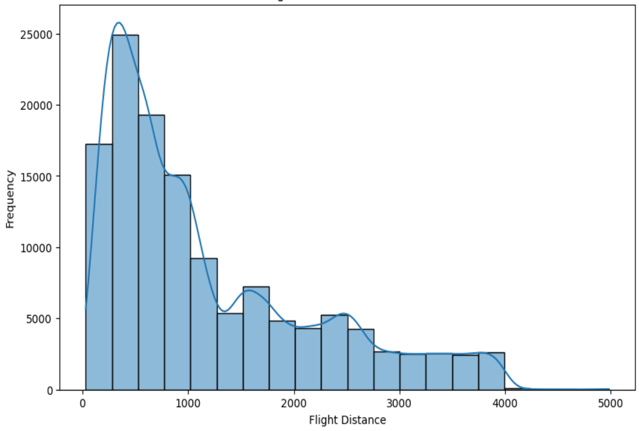



**Fig 3: Flight Distance Distribution in airline passenger satisfaction**

This histogram for Flight Distance Distribution shows the frequency of flights across different distances. Here is what we can interpret from the graph:

- **Right-Skewed Distribution**: The graph is right-skewed, indicating that there are more short-distance flights compared to long-distance flights.
- **Majority Short-Distance**: There is a higher frequency of flights within the 0-1000 km range, which shows that the majority of the flights are short-distance.
- **Long-Distance Flights**: The frequency of flights decreases as the distance increases, which suggests that long-distance flights are less common in this dataset.
- **Outliers**: The presence of bars at the far right of the histogram suggests that there are a few very long-distance flights, which could be considered as outliers.

This distribution could imply several things about the airline's operations, such as a focus on short-haul flights or a customer base that predominantly travels short distances. It may also inform resource allocation, such as crew scheduling and fleet management, which would be different for short versus long-distance flights.

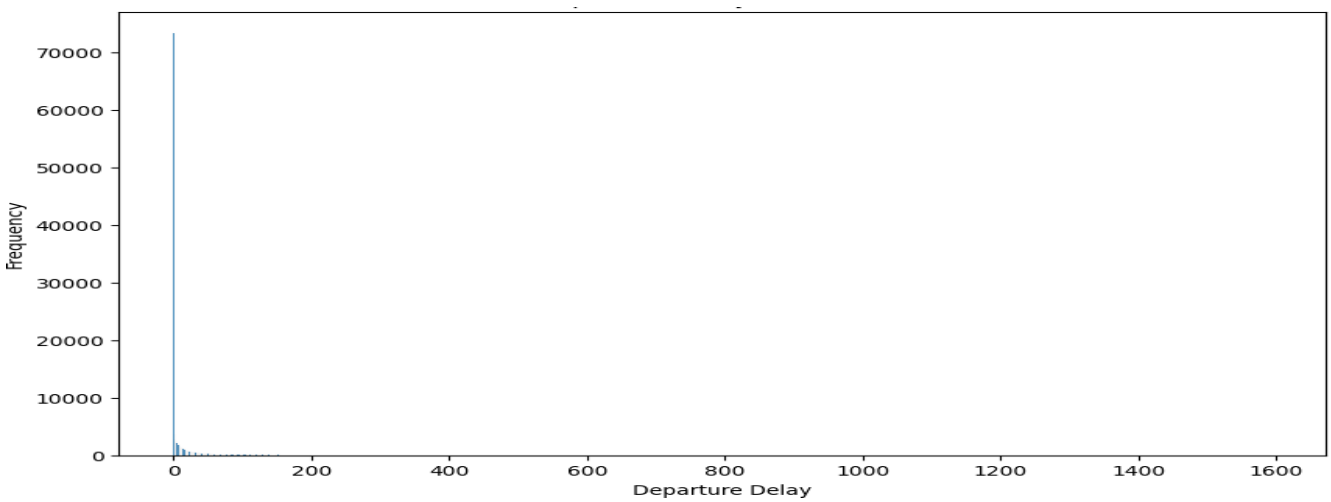



**Fig 4: Departure Delay distribution in airline passenger satisfaction**

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

The Departure Delay Distribution histogram is skewed to the right, indicating that most of the flights leave on time or with a slight delay. The long tail to the right suggests that there are relatively few flights with very long departure delays, but these extreme values can significantly impact passenger satisfaction. Such a distribution often calls for specific analysis of outliers to understand the causes of extreme delays and address them.

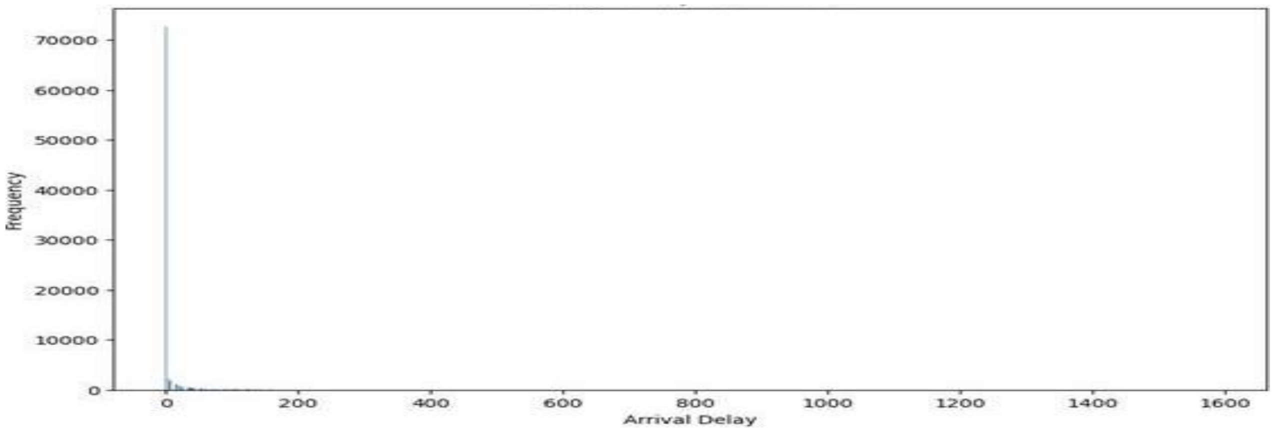



**Fig 5: Arrival Delay distribution in airline passenger satisfaction**

Similarly, the Arrival Delay Distribution histogram is also right skewed. This means that most flights arrive on time or with minor delays, but there are instances of substantial delays. As with departure delays, these outliers are critical to examine, as they can contribute to overall customer dissatisfaction and potentially incur costs to the airline in terms of compensations and disrupted schedules.

**Right-Skewed Distributions**
Both histograms indicate that while most flights are on time or close to on time, delays can sometimes be significant. This skewness is characteristic of operational data in airlines, where most operations run smoothly, but there are occasional significant disruptions.

**Operational Focus**
The presence of delays, even if few, can be a significant pain point for passengers. Airlines may want to investigate the causes of these outliers to improve their operational efficiency and customer service.

**Potential Outliers**
The tails of both distributions highlight the presence of outliers, which could be due to various factors such as weather, mechanical issues, or air traffic control delays.

Identifying and mitigating the root causes of these outliers can improve the overall punctuality of the airline.

**Customer Satisfaction**
Given that these two metrics (departure and arrival delays) are likely to have a strong impact on customer satisfaction, the airline may focus on improving turnaround times and maintaining schedule integrity to enhance the customer experience.

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

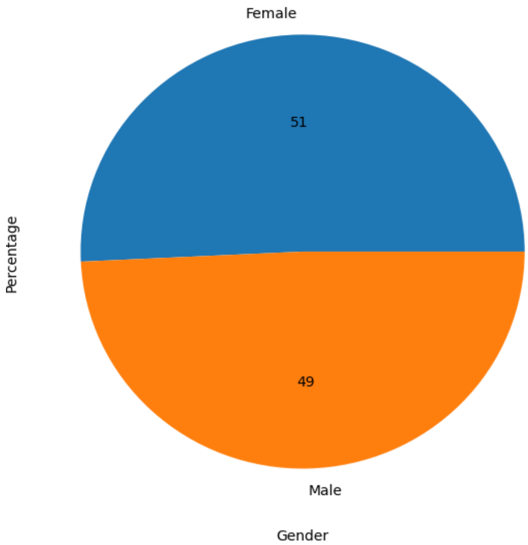
ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

**Data-Driven Decisions**

For predictive modeling, the delay variables might need transformation or segmentation (e.g., categorizing delays into bins) to better capture the relationship between delays and passenger satisfaction. Additionally, understanding the distribution of delays can help in




designing strategies for customer service interventions, like proactive communication about delays.

**Fig 6: Gender distribution in airline passenger satisfaction**

The pie chart displays a nearly even split between male and female passengers, with females slightly outnumbering males, 51% to 49%. This balance suggests that gender is unlikely to be a confounding factor in any analysis of passenger satisfaction and that any marketing or service changes are likely to affect both genders equally. The balanced gender distribution provides an equal representation of perspectives and preferences in any customer satisfaction analysis.

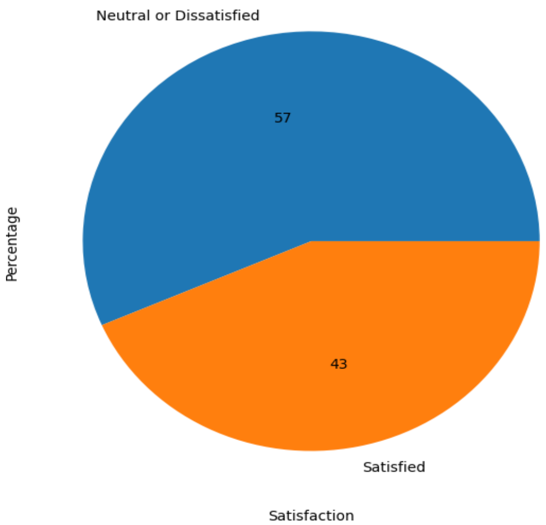



**Fig 7: Satisfaction level of all the passengers in airline in percentages**

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

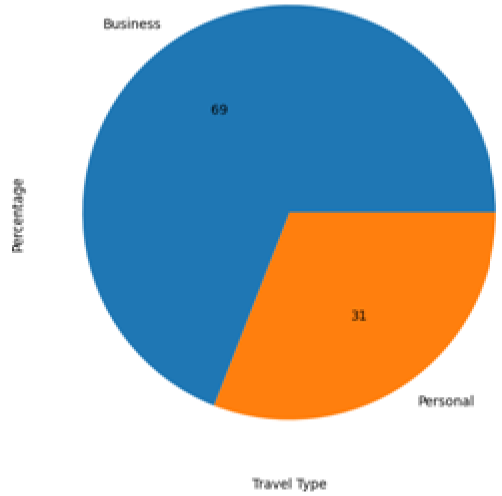
DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

The pie chart shows the proportion of passengers' satisfaction levels, with 57% of passengers being neutral or dissatisfied, and 43% satisfied. This indicates that there is more opportunity for the airline to improve passenger experiences, as most passengers are not fully satisfied with the service. The satisfaction data highlights the need for improvement in passenger services or experiences. The airline may consider investigating the underlying causes of dissatisfaction to improve their services.


**Fig 8: Customers Type Distribution Pie Chart in airline passenger satisfaction**

The chart indicates that most passengers, 82%, are returning customers, while only 18% are flying with the airline for the first time. This could suggest high customer loyalty or a frequent flyer program that effectively retains customers. However, it also indicates that there might be room to grow the customer base by attracting more first-time flyers. With a high percentage of returning customers, the airline seems to have a loyal customer base, which is beneficial for steady revenue. However, strategies to attract new customers may also be needed to expand the market share.



Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

**Fig 9: Travel Type Distribution in airline passenger satisfaction**

The type of travel chart indicates that 69% of travel is for business purposes, while 31% is for personal reasons. This suggests that the airline's customer base is primarily composed of business travelers, which could influence decisions related to flight scheduling, services offered, and loyalty programs. The prevalence of business travel suggests that the airline may need to prioritize services that are valued by business travelers, such as timely departures, work-friendly environments on board, and expedited check-ins. Each of these charts provides actionable insights that can be used to refine the airline's business strategies, improve customer satisfaction, and tailor services to meet the specific needs of their passengers.

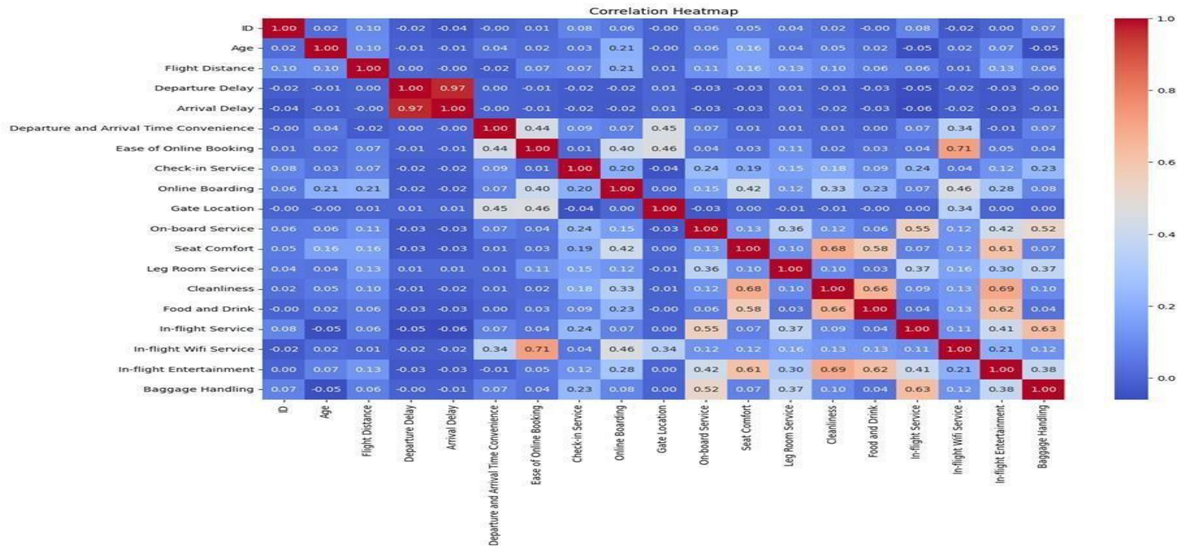



**Fig 10: Correlation Heat map in the context of airline passenger satisfaction**

Each cell in the heat map shows the correlation coefficient between two variables. In the context of airline passenger satisfaction, this heat map can help identify which factors are most closely related to each other, and potentially, to passenger satisfaction.

**Interpretation of the Heat map**

- **Diagonal Values**: The diagonal cells, where the variables intersect with themselves, always have a correlation coefficient of 1, as a variable is perfectly correlated with itself.
- **Correlation Coefficients**: The values range from -1 to 1. A value closer to 1 implies a strong positive correlation, meaning that as one variable increases, the other variable also increases. A value closer to -1 implies a strong negative correlation, meaning that as one variable increases, the other decreases. A value around 0 suggests no correlation.
- **Color Coding**: Typically, warmer colors (like red) indicate a higher positive correlation, and cooler colors (like blue) indicate a higher negative correlation. Neutral colors (like white) indicate no correlation.

**Specific Variables and Their Correlations**

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

Looking at the provided heat map, we can discuss some key correlations related to passenger satisfaction:

- **Departure and Arrival Time Convenience**: High correlation with each other, which is expected as these often go hand-in-hand. If a flight departs conveniently, it often arrives conveniently as well.
- **Ease of Online Booking and Check-in Service**: High positive correlation suggests that passengers who find it easy to book their tickets online also find the check-in service satisfactory.
- **On-board Service, Seat Comfort, and Leg Room Service**: These also have a strong positive correlation with each other. This indicates that passengers' perceptions of these aspects of the flight experience are linked. If a passenger is happy with one, they are likely to be happy with the others.
- **Food and Drink, In-flight Service, and In-flight Entertainment**: These show moderate to strong correlations with each other, suggesting that the overall in-flight experience is interconnected in the passenger's perception.
- **Baggage Handling**: This has a moderate positive correlation with several factors such as Check-in Service and In-flight Service, which implies that good service in these areas might influence the passenger's perception of baggage handling.

**Implications for Airline Passenger Satisfaction**

Understanding these correlations is valuable for improving passenger satisfaction. For example, since on-board service, seat comfort, and leg room are strongly correlated, improving one of these aspects could have a positive effect on the others, thereby enhancing the overall passenger experience.

However, it's important to note that correlation does not imply causation. These relationships can inform hypotheses for further investigation, but they do not necessarily mean that one factor causes changes in another. Additional statistical analysis, such as regression modelling, could be conducted to determine causal relationships.

The airline might prioritize improvements in areas with high correlations to passenger satisfaction. If satisfaction data were included in this heat map, the airline could identify which variables most strongly influence satisfaction and target those for improvement. For instance, if 'In-flight Entertainment' had a highly positive correlation with 'Satisfaction', the airline could focus on enhancing entertainment options to improve overall satisfaction scores.

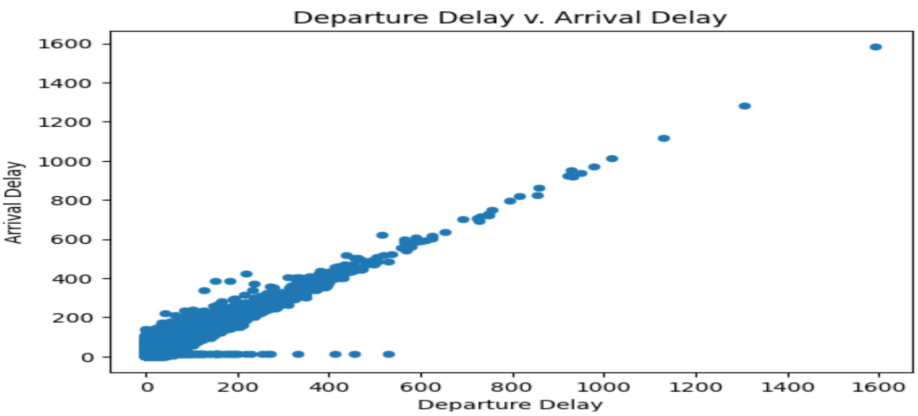



**Fig 11: Relationship between 'Departure Delay' and 'Arrival Delay' in airline**

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

Vol. 2 No. 5 (December) (2024)

**passenger satisfaction**

The scatter plot shows the relationship between 'Departure Delay' and 'Arrival Delay'. A scatter plot with a dense line of points indicates a strong linear relationship, which seems to be the case here. The description mentions a correlation of 0.97, which is very high, signifying that as the departure delay increases, the arrival delay also tends to increase in a proportionate manner.

**Interpretation**

- **High Correlation**: The strong correlation suggests that the two variables are interdependent. In the context of airlines, this could mean that a flight that departs late is also likely to arrive late.
- **Predictive Modelling**: For predictive modelling, this indicates redundancy. Because these variables provide similar information, you might choose only one to avoid multicollinearity in regression models.

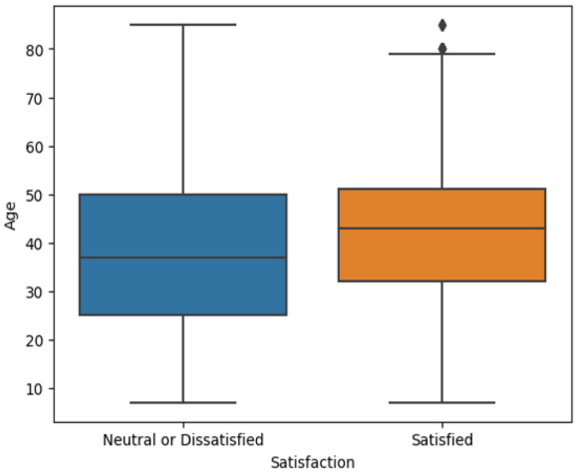
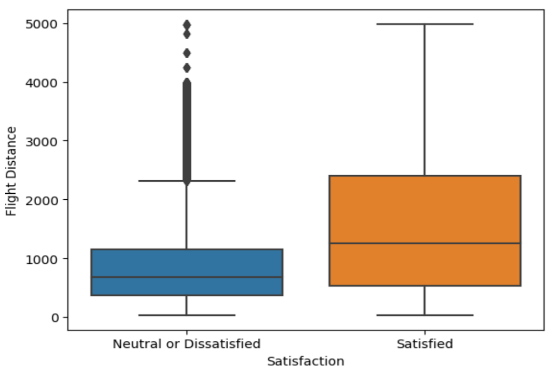



**Fig 12: Comparison of customers satisfaction level against age in airline passenger satisfaction**

The box plot comparing 'Satisfaction' (split into categories like 'Neutral or Dissatisfied' and 'Satisfied') against 'Age' shows that the ages are distributed similarly across both satisfaction categories. The median lines are at a similar level, and the interquartile ranges have substantial overlap. The similarity in distribution suggests that age alone might not be a strong predictor of satisfaction among passengers.



Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

**Fig 13: Comparison of customers satisfaction level against Flight Distance in airline passenger satisfaction**

The box plot here compares 'Satisfaction' with 'Flight Distance'. There seems to be a noticeable difference in the medians of the two categories. The 'Satisfied' category has a higher median flight distance compared to the 'Neutral or Dissatisfied' category.

- **Flight Distance as a Predictor**: The distinct difference in the distribution suggests that flight distance may be related to satisfaction. Perhaps longer flights have more amenities or attract a different type of passenger, resulting in higher satisfaction scores.

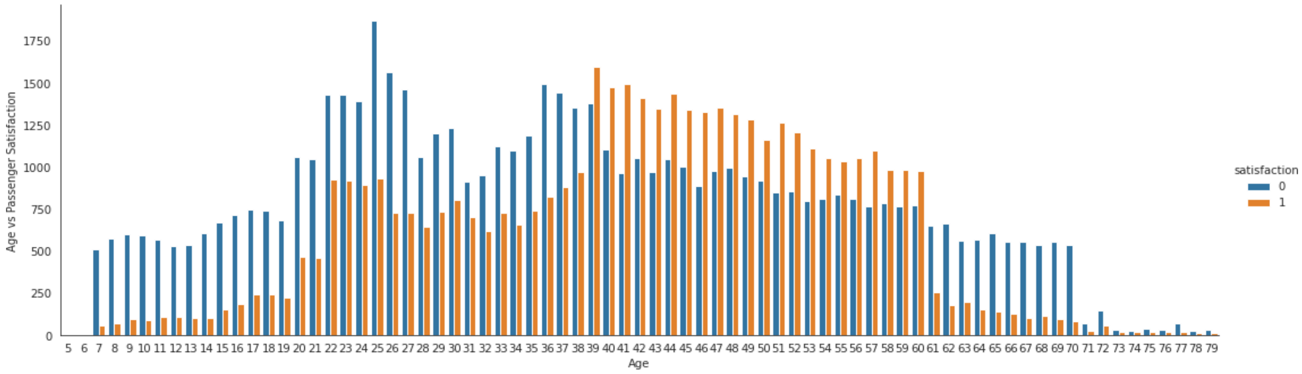



**Fig 14: Dependency check of age on satisfaction level in airline**

The age distribution of passengers is wide-ranging, covering young to older age groups. The satisfaction (depicted by different colors for satisfied and not satisfied) across different age groups seems to be varied. It might suggest that satisfaction is independent of age since there are both satisfied and not satisfied passengers across all age groups. There is no clear trend indicating that a particular age group is satisfied with others.

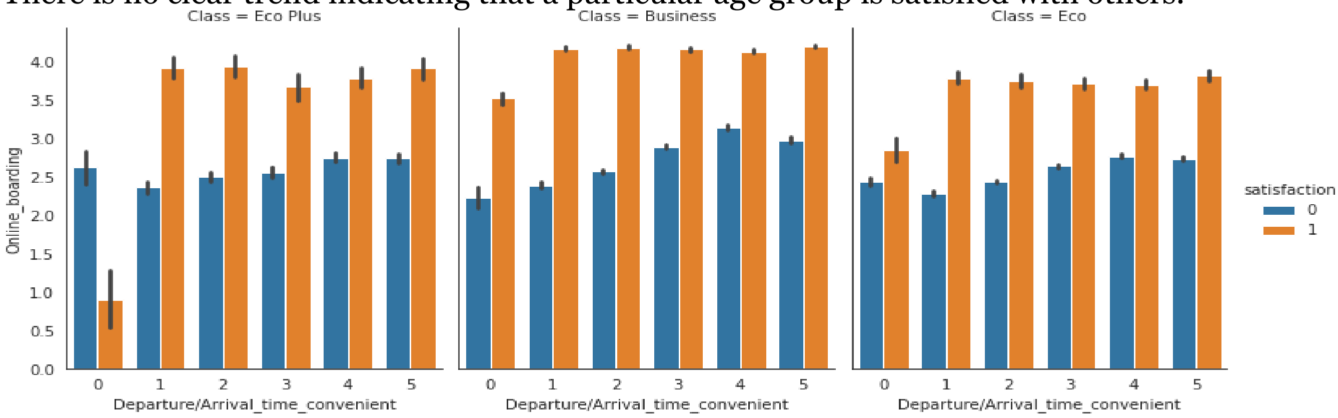



**Fig 15: Convenience of departure/arrival times against passenger satisfaction, with the data separated by class of service (Eco Plus, Business, Eco).**

This bar graph shows the reported convenience of departure/arrival times against passenger satisfaction, with the data separated by class of service (Eco Plus, Business, Eco).

The graph might be showing average ratings for departure/arrival time convenience on a scale (perhaps 0 to 5), with separate bars for satisfied and not satisfied passengers in each travel class. In all classes, passengers who report being satisfied also seem to report

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

higher convenience scores, implying that convenient departure/arrival times are associated with higher satisfaction levels.
This trend is consistent across all travel classes, suggesting that regardless of the class, departure/arrival time convenience is an important factor for satisfaction.

**Fig 16 (a): Descriptive Statistics for each variable that can affect airline passenger satisfaction**

| On-board Service | Seat Comfort | Leg Room Service | Cleanliness | Food and Drink | In-flight Service | In-flight Wifi Service | In-flight Entertainment | Baggage Handling |
|---|---|---|---|---|---|---|---|---|
| 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 |
| 3.383023 | 3.441361 | 3.350878 | 3.286326 | 3.204774 | 3.642193 | 2.728696 | 3.358077 | 3.632114 |
| 1.287099 | 1.319289 | 1.316252 | 1.313682 | 1.329933 | 1.176669 | 1.329340 | 1.334049 | 1.180025 |
| 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 1.000000 |
| 2.000000 | 2.000000 | 2.000000 | 2.000000 | 2.000000 | 3.000000 | 2.000000 | 2.000000 | 3.000000 |
| 4.000000 | 4.000000 | 4.000000 | 3.000000 | 3.000000 | 4.000000 | 3.000000 | 4.000000 | 4.000000 |
| 4.000000 | 5.000000 | 4.000000 | 4.000000 | 4.000000 | 5.000000 | 4.000000 | 4.000000 | 5.000000 |
| 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 |

**Fig 16 (b): Descriptive Statistics for each variable that can affect airline passenger satisfaction**

| | ID | Age | Flight Distance | Departure Delay | Arrival Delay | Departure and Arrival Time Convenience | Ease of Online Booking | Check-in Service | Online Boarding | Gate Location |
|---|---|---|---|---|---|---|---|---|---|---|
| count | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129487.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 | 129880.000000 |
| mean | 64940.500000 | 39.427957 | 1190.316392 | 14.713713 | 15.091129 | 3.057599 | 2.756876 | 3.306267 | 3.252633 | 2.976925 |
| std | 37493.270818 | 15.119360 | 997.452477 | 38.071126 | 38.465650 | 1.526741 | 1.401740 | 1.266185 | 1.350719 | 1.278520 |
| min | 1.000000 | 7.000000 | 31.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 | 0.000000 |
| 25% | 32470.750000 | 27.000000 | 414.000000 | 0.000000 | 0.000000 | 2.000000 | 2.000000 | 3.000000 | 2.000000 | 2.000000 |
| 50% | 64940.500000 | 40.000000 | 844.000000 | 0.000000 | 0.000000 | 3.000000 | 3.000000 | 3.000000 | 3.000000 | 3.000000 |
| 75% | 97410.250000 | 51.000000 | 1744.000000 | 12.000000 | 13.000000 | 4.000000 | 4.000000 | 4.000000 | 4.000000 | 4.000000 |
| max | 129880.000000 | 85.000000 | 4983.000000 | 1592.000000 | 1584.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 | 5.000000 |

These tables summarize descriptive statistics for various variables related to airline passenger satisfaction. The table lists the count, mean, standard deviation, minimum, 25th percentile, median (50th percentile), 75th percentile, and maximum values for each variable. The presence of variables like 'On-board Service', 'Seat Comfort', 'Leg Room Service', 'Cleanliness', and 'Baggage Handling' indicates that these service aspects are being analyzed, likely to understand their influence on overall passenger satisfaction. The statistical measures provided help to understand the distribution and variability of each aspect of service as perceived by passengers.
The EDA process is methodical, addressing key aspects such as data completeness and distribution. By ensuring there are no duplicates and handling missing data carefully, the integrity of the data is maintained, which is crucial for any subsequent analysis or

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

predictive modeling.
The descriptive statistics offer a wealth of information:

- **Age**: The passengers range from young children to the elderly, with a median age around the late 30s, suggesting a mature customer base.
- **Flight Distance**: It shows a wide range of flight distances, indicating the diversity of the trips taken by passengers, from short haul to long-haul flights.
- **Delays**: Both departure and arrival delays exhibit a broad range, with some outliers indicating extreme cases of delay.

The distribution of other service-related variables like 'Ease of Online Booking', 'Gate Location', 'On-board Service', etc., will provide insights into the areas where the airline excels or needs improvement. For instance, if the mean score for 'In-flight Wifi Service' is low, it may indicate customer dissatisfaction in this area.

The EDA results can serve as a guide for targeted improvements and customer satisfaction strategies. For example, if services like 'Leg Room Service' or 'In-flight Entertainment' have lower scores, they could be prioritized for enhancement.

For future modeling, these descriptive statistics will be valuable in feature selection, helping to identify which variables are most influential in predicting passenger satisfaction and should therefore be included in the predictive models.

### Machine Learning Model Evaluation

In the evaluation of machine learning models applied to an airline passenger satisfaction survey dataset, four models underwent scrutiny: Linear Regression, Logistic Regression, Random Forest Regressor, and AdaBoost Regressor. Among these, Logistic Regression demonstrated moderate accuracy, achieving an approximate 66.07% accuracy score. However, deeper analysis revealed challenges in correctly identifying 'satisfied' customers, evident through imbalanced precision and recall metrics for both satisfaction classes. The model displayed a precision of 0.67 and a recall of 0.43 for 'satisfied' instances, suggesting a struggle to capture a significant portion of actual 'satisfied' cases.

**Fig 17: Performance metrics for model**

```
Logistic Regression Metrics:
Accuracy Score: 0.6606867878041269
Classification Report:
              precision    recall  f1-score   support

           0       0.66      0.83      0.74     14723
           1       0.67      0.43      0.53     11253

    accuracy                           0.66     25976
   macro avg       0.66      0.63      0.63     25976
weighted avg       0.66      0.66      0.64     25976


Random Forest Regressor Metrics:
Mean Squared Error: 0.24639502730153187, R2 Score: -0.003487250375826756

AdaBoost Regressor Metrics:
Mean Squared Error: 0.2119083404094793, R2 Score: 0.136965870707363
```

Conversely, the ensemble methods, particularly the Random Forest Regressor, depicted unsuitability for the categorical target variable. This was evidenced by a negative $R^2$

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)
score, indicating poor model performance.

To improve model accuracy and reliability, recommendations were outlined, emphasizing the exploration of additional relevant features, meticulous tuning of model hyper parameters, mitigation of potential class imbalances, and consideration of alternative models like Support Vector Machines or neural networks. Aligning the model choices with the project's objectives and conducting in-depth exploration stand as critical steps toward refining the predictive models and achieving enhanced accuracy relevant to the goals of the airlines.

## Conclusion

In conclusion, the exploratory data analysis and subsequent predictive modelling performed within this paper have produced a detailed understanding of the factors influencing airline passenger satisfaction. The analysis has led to several insights, notably that logistic regression for classification yielded an accuracy of approximately 66.07%, indicating a moderate level of predictive reliability. Although this model demonstrates decent precision and recall, especially in identifying 'not satisfied' customers, there is a notable opportunity for enhancement, particularly in identifying 'satisfied' passengers.

The performance of the Random Forest and AdaBoost regression models was less than optimal, with a negative R^2 score for Random Forest and a weak fit to the data by AdaBoost. These outcomes suggest that regression models may not be best suited for a categorical target variable like 'Satisfaction'. The paper highlights the potential for improved model performance through feature engineering, model tuning, adjusting for class balance, and exploring alternative modelling techniques such as Support Vector Machines or neural networks.

These findings underscore the importance of aligning model selection with the objectives of the analysis and the nature of the data. With these predictive insights, the paper significantly contributes to the realm of customer service within the airline industry. It serves as a catalyst for more focused research, particularly in areas that can substantially enhance the passenger experience. The results and methodologies applied offer a comprehensive resource for industry stakeholders looking to invest in data-driven strategies to uplift customer satisfaction and loyalty. The future direction of this research opens avenues to refine modelling techniques and incorporate a broader spectrum of data, ultimately advancing the understanding of passenger needs and preferences.

## References

1. Tahanisaz, S., & shokuhyar, S. (2020). Evaluation of passenger satisfaction with service quality: A consecutive method applied to the airline industry. Journal of Air Transport Management, 83, 101764. https://doi.org/10.1016/j.jairtraman.2020.101764

2. Ganiyu, R.A., 2017. Customer satisfaction and loyalty: a study of interrelationships and effects in the Nigerian domestic airline industry. Oradea Journal of Business and Economics 2 (1), 7–20.

3. Ishutkina, M., Hansman, R.J., 2008. Analysis of interaction between air transportation and economic activity. report. In: The 26th Congress of ICAS and 8th

AIAA ATIO, 8888.

4. Parasuraman, A. B. L. L., Zeithaml, V. A., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. 1988, 64(1), 12-40.

5. Ostrowski, P. L., O'Brien, T. V., & Gordon, G. L. (1993). Service Quality and Customer Loyalty in the Commercial Airline Industry. Journal of Travel Research, 32(2), 16-24. https://doi.org/10.1177/004728759303200203

6. Tsaur, S.-H., Chang, T.-Y., & Yen, C.-H. (2002). The evaluation of airline service quality by fuzzy MCDM. Tourism Management, 23(2), 107–115. https://doi.org/10.1016/s0261- 5177(01)00050-4

7. Gilbert, D., & Wong, R. K. C. (2003). Passenger expectations and airline services: a Hong Kong based study. Tourism Management, 24(5), 519–532. https://doi.org/10.1016/s0261- 5177(03)00002-5

8. Chen, F.-Y., & Chang, Y.-H. (2005). Examining airline service quality from a process perspective. Journal of Air Transport Management, 11(2), 79–87. https://doi.org/10.1016/j.jairtraman.2004.09.002

9. Liou, J. J. H., & Tzeng, G.-H. (2007). A non-additive model for evaluating airline service quality. Journal of Air Transport Management, 13(3), 131–138. https://doi.org/10.1016/j.jairtraman.2006.12.002

10. Radović-Marković, M., Shoaib Farooq, M., & Marković, D. (2017). Strengthening the resilience of small and medium-sized enterprises. Review of applied socio-economic research, 345-356.

11. Chang, EC., Huang, SC. & Wu, HH. Using K-means method and spectral clustering technique in an outfitter's value analysis. Qual Quant 44, 807–815 (2010). https://doi.org/10.1007/s11135- 009-9240-0

12. Bellizzi, M. G., Eboli, L., & Mazzulla, G. (2020). An online survey for the quality assessment of airlines' services. Research in Transportation Business & Management, 100515. https://doi.org/10.1016/j.rtbm.2020.100515

13. Norazah, M. S. (2013). Customer satisfaction with service delivery in the life insurance industry: An empirical study. Jurnal Pengurusan, 38, 101–109.

14. Malaysia, Tourism (2013). Malaysia receives a record 25 million tourist arrivals last year. http://corporate.tourism.gov.my/mediacentre.asp?page=news_desk&news_id=763 (accessed 25-09-2012)

15. Liou, J. J. H., Hsu, C.-C., Yeh, W.-C., & Lin, R.-H. (2011). Using a modified grey relation method for improving airline service quality. Tourism Management, 32(6),

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

1381–1388. https://doi.org/10.1016/j.tourman.2011.01.013

16. Ganiyu, R. (2017). Customer Satisfaction and Loyalty: A Study of Interrelationships And Effects In Nigerian Domestic Airline Industry. http://ojbe.steconomiceuoradea.ro/wp-content/uploads/2017/03/OJBE_vol-21_01_Ganiyu.pdf

17. Abdullah Al Noman, Md Tanvir Rahman Tarafder, S. M. Tamim Hossain Rimon, Asif Ahamed, Shahriar Ahmed, and Abdullah Al Sakib, "Discoverable Hidden Patterns in Water Quality through AI, LLMs, and Transparent Remote Sensing," *The 17th International Conference on Security of Information and Networks (SIN-2024)*, Sydney, Australia, 2024, pp. 259–264.

18. S. M. T. H. Rimon, Mohammad A. Sufian, Zenith M. Guria, Niaz Morshed, Ahmed I. Mosaddeque, and Asif Ahamed, "Impact of AI-Powered Business Intelligence on Smart City Policy-Making and Data-Driven Governance," *International Conference on Green Energy, Computing and Intelligent Technology (GEn-CITy 2024)*, Johor, Malaysia, 2024.

19. A. Nuthalapati, "Building Scalable Data Lakes For Internet Of Things (IoT) Data Management," *Educational Administration: Theory and Practice*, vol. 29, no. 1, pp. 412–424, Jan. 2023, doi:10.53555/kuey.v29i1.7323.

20. M. A. Sufian, Z. M. Guria, N. Morshed, S. M. T. H. Rimon, A. I. Mosaddeque, and A. Ahamed, "Leveraging Machine Learning for Strategic Business Gains in the Healthcare Sector," *2024 International Conference on TVET Excellence & Development (ICTeD-2024)*, Melaka, Malaysia, 2024.

21. A. I. Mosaddeque, Z. M. Guria, N. Morshed, M. A. Sufian, A. Ahamed, and S. M. T. H. Rimon, "Transforming AI and Quantum Computing to Streamline Business Supply Chains in Aerospace and Education," *2024 International Conference on TVET Excellence & Development (ICTeD-2024)*, Melaka, Malaysia, 2024.

22. S. B. Nuthalapati and A. Nuthalapati, "Accurate Weather Forecasting with Dominant Gradient Boosting Using Machine Learning," *Int. J. Sci. Res. Arch.*, vol. 12, no. 2, pp. 408–422, 2024, doi:10.30574/ijsra.2024.12.2.1246.

23. A. Ahamed, N. Ahmed, J. I. Janjua, Z. Hossain, E. Hasan, and T. Abbas, "Advances and Evaluation of Intelligent Techniques in Short-Term Load Forecasting," *2024 International Conference on Computer and Applications (ICCA-2024)*, Cairo, Egypt, 2024.

24. M. T. R. Tarafder, M. M. Rahman, N. Ahmed, T.-U. Rahman, Z. Hossain, and A. Ahamed, "Integrating Transformative AI for Next-Level Predictive Analytics in Healthcare," *2024 IEEE Conference on Engineering Informatics (ICEI-2024)*, Melbourne, Australia, 2024.

25. A. Nuthalapati, "Architecting Data Lake-Houses in the Cloud: Best Practices and Future Directions," *Int. J. Sci. Res. Arch.*, vol. 12, no. 2, pp. 1902–1909, 2024,

**Dialogue Social Science Review (DSSR)**
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)

doi:10.30574/ijsra.2024.12.2.1466.

26. A. Ahamed, M. T. R. Tarafder, S. M. T. H. Rimon, E. Hasan, and M. A. Amin, "Optimizing Load Forecasting in Smart Grids with AI-Driven Solutions," *2024 IEEE International Conference on Data & Software Engineering (ICoDSE-2024)*, Gorontalo, Indonesia, 2024.

27. Ishutkina, M., & Hansman, R. J. (2008, November). Analysis of Interaction between Air Transportation and Economic Activity. In The 26th Congress of ICAS and 8th AIAA ATIO (p. 8888).

28. Tsafarakis, S., Kokotas, T., & Pantouvakis, A. (2018). A multiple criteria approach for airline passenger satisfaction measurement and service quality improvement. Journal of Air Transport Management, 68, 61–75. https://doi.org/10.1016/j.jairtraman.2017.09.010

29. Nadiri, H., Hussain, K., Ekiz, E. H., & Erdoğan, S. (2008). An investigation on the factors influencing passengers' loyalty in the North Cyprus national airline. The TQM Journal, 20, 265-280. doi:10.1108/17542730810867272

30. Truitt, L., & Haynes, R. (1994). Evaluating service quality and productivity in the regional airline industry. Transportation Journal, 33(4), 21-32.

31. Rendeiro, R., & Cejas, M. (2006). Tourism service quality begins at the airport. Tourism Management, 27, 874-877. doi:10.1016/j.tourman.2005.05.005

32. Etemad-Sajadi, R., Way, S. A., & Bohrer, L. (2016). Airline Passenger Loyalty. Cornell Hospitality Quarterly, 57(2), 219–225. https://doi.org/10.1177/1938965516630622

33. De Meyer, C. F., & Mostert, P. G. (2011). The influence of passenger satisfaction on relationship formation in the South African domestic airline industry. South African Journal of Business Management, 42(4), 79-87.

34. Hussain, R. (2016). The mediating role of customer satisfaction: evidence from the airline industry. Asia Pacific Journal of Marketing and Logistics, 28(2).

35. Hong, A. C. Y., KHAW, K. W., CHEW, X., & YEONG, W. C. (2023). Prediction of US airline passenger satisfaction using machine learning algorithms. Data Analytics and Applied Mathematics (DAAM), 8-24.

36. Kumar, S., & Zymbler, M. (2019). A machine learning approach to analyze customer satisfaction from airline tweets. Journal of Big Data, 6(1), 1-16.

37. Sezgen, E., Mason, K. J., & Mayer, R. (2019). Voice of airline passenger: A text mining approach to understand customer satisfaction. Journal of Air Transport Management, 77, 65-74.

38. An, M., & Noh, Y. (2009). Airline customer satisfaction and loyalty: impact of in-

Dialogue Social Science Review (DSSR)
www.journalforeducationalresearch.online

ISSN Online: 3007-3154
ISSN Print: 3007-3146

DIALOGUE SOCIAL SCIENCE REVIEW

Vol. 2 No. 5 (December) (2024)
flight service quality. Service Business, 3, 293-307.

39. Chow, C. K. W. (2014). Customer satisfaction and service quality in the Chinese airline industry. Journal of air transport management, 35, 102-107.

40. Ahmad, M. B., Wasay, E., & Ullah, S. (2012). Impact of employee motivation on customer satisfaction: Study of airline industry in Pakistan. Interdisciplinary Journal of Contemporary Research in Business, 4(6), 531-539.

41. Teichert, T., Shehu, E., & von Wartburg, I. (2008). Customer segmentation revisited: The case of the airline industry. Transportation Research Part A: Policy and Practice, 42(1), 227-242.

42. Graham, D. R., Kaplan, D. P., & Sibley, D. S. (1983). Efficiency and competition in the airline industry. The Bell Journal of Economics, 118-138.

43. David Mc A, B. (2013). Service quality and customer satisfaction in the airline industry: A comparison between legacy airlines and low-cost airlines. American journal of tourism research, 2(1), 67-77.

44. Borenstein, S., & Rose, N. L. (2014). How airline markets work… or do they? Regulatory reform in the airline industry. In Economic regulation and its reform: What have we learned? (pp. 63-135). University of Chicago Press.