M4.04 - Big Data and Artificial Intelligence

Project Work: Airline Customer Analytics

Winter Semester 2021/22

Munich University of Applied Sciences
Faculty of Business Administration
Am Stadtpark 20
81243 München

Andreas Birnkammer
Nina Ge
Chi-Yuan Lee
Timon Leuchtmann

Table of Contents

Table of Figures

Table of Tables

Link to Jupyter Notebook

https://colab.research.google.com/drive/1zjENfBsUqFqb-nFz3Z7fev_f0-w8G8qj?usp=sharing

1. Introduction

Customer satisfaction is very crucial for airlines as unhappy or unsatisfied customers naturally mean fewer passengers and less profit. Therefore, it is important for the customer to have an excellent experience every time they travel. The passenger's experience is not just the flight itself, but also everything from ticket purchasing on the airline's website or mobile app to the boarding process. To understand the different factors that can influence customer satisfaction, this report examined a US airline passenger satisfaction survey dataset.

For this report, the Cross-Industry Standard Process for Data Mining (CRISP-DM) is used to help with the data analysis. This survey contains 23 variables such as type of travel, departure/arrival time, customer type etc. The correlations between the different variables are analysed with the help of Google Collab and the programming language python. First, the data is understood and then prepared and cleaned. There are two main questions relevant to this analysis. First, whether customer satisfaction can be predicted and, second, whether it is possible to find groups of similar customers. Next, different modelling methods are applied to answer the two main business questions about customer satisfaction prediction and clustering. Consequently, the relevant variables for satisfaction are identified and different customer groups are characterized. Furthermore, business interpretation and recommendations of the findings are provided for each question. Lastly, the conclusion will summarize the report's findings of determining the key drivers for customer satisfaction and improving passengers flight experience. Figure 1 depicts the analysing process.

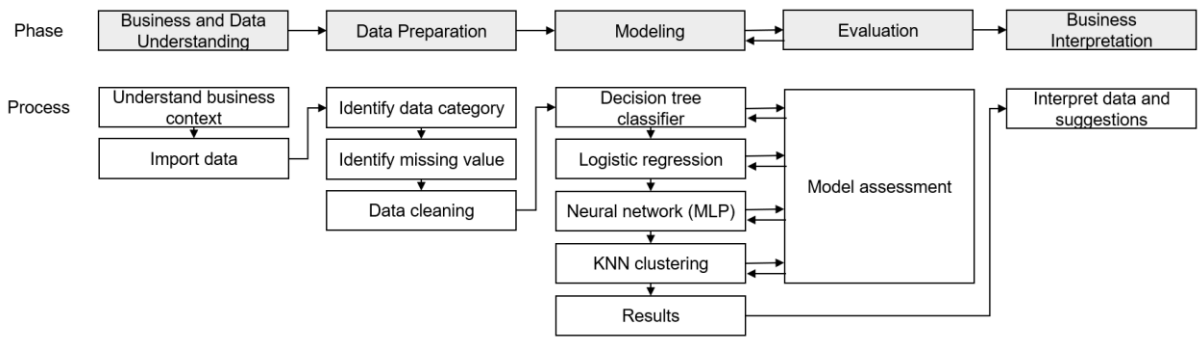

Figure 1 Analyzing Process

2. Goal of the Report

A deep analysis of the conducted survey provides useful information for understanding the correlations between customer satisfaction and the current services offered by the airline. The

objective of this report is to understand what leads to higher customer satisfaction and what kind of customer groups can be defined. With this knowledge, the overall satisfaction can be improved by adapting the services according to the report's outcome. This information can be used for a more targeted advertisement by identifying the customer groups and helping enhance the service palette of the airline.

3. Data Pre-processing

In this chapter, several topics will be introduced: (1) data preparation, (2) data cleaning, and (3) data overview, which help to understand the original data and transform it to the correct form for mathematical analysis. Data preparation refers to the data understanding of the raw data such as data type and data category. It is an essential before data cleaning. Second, data cleaning is the process to organize and unifying the data for further analysis. Lastly, in data interpretation, the scatters and histograms are applied to visualize the data.

3.1 Data Preparation

In short, data preparation can be defined as the process that an analyst understands the raw data and transforms the raw data into the requirement for algorithm models. In data preparation, identifying the different data scales in the dataset is a must. The process of data preparation includes data category recognition, data unification, and data-type recognition. Thus, the data condition is important for applying the analysis technique correctly. Data can be divided into two main categories: nonmetric data and metric data. Nonmetric data include nominal scales and ordinal scales. Specifically, nominal scales also called categorical scales which descript the sample characteristics, such as gender, social status, occupation, or religion. Ordinal scales indicate the numerical scale with a non-quantitative method. On the other hand, metric data includes interval scales and ratio scales. Interval scales and ratio scales provide a more precise measurement such as distance, time, temperature, and weight.

First, to understand the data category, Jupyter notebook was used to process the data. Then, data were categorized into different data categories based on the column attributes, namely nominal scales (e.g. gender, customer type, age, type of travel, class, and satisfaction), ordinal scales (from inflight wifi service to cleanliness) and ratio scales (e.g. flight distance, departure delay in minutes, and arrival delay in minutes). Through this process, the whole dataset has been identified as a two-dimension raw dataset with 103,924 rows and 25 columns.

Second, the data unification process was applied. In this procedure, the unique values from each column were identified. For example, in “gender”, the unique value includes male, female, diverse and “na”. The benefits of this procedure are to highlight the columns that need to be transformed from string to numerical values and to identify the string format. Therefore, the function called “unique_detector” was designed to execute this task. “Customer type” and “type of travel” have errors in their format, which might cause transformation errors. To be more specific, some data use lowercase and others use capital letters such as “Business travel” and “Business Travel”. Moreover, missing values are also a factor that can influence the data application. The missing values are displayed as “na” and can be found through the python code print(df.isnull().sum()).

Third, the data type needs to be understood for each column. The results show that the data types of “gender”, “customer type”, “type of travel”, “class”, and “satisfaction” are object values, while other columns are numerical values. Last but not the least, in the ordinal scales from “inflight WIFI service” to “cleanliness” the scale goes from 1 to 5 and 0 represents “Not applicable”. There are around 8,190 rows with zeros that can be interpreted as missing values. Reasons for this could be that people did not fill out certain sections or they for instance did not do the online booking. Therefore, these rows with zero have been dropped.

3.2 Data Cleaning

After the preparation was completed, the raw data is modified. First, all the data with any missing values are deleted. Next, the unwanted value in “gender” and “satisfaction” are deleted. Thus, “na” and “no answer” from “gender” and “satisfaction” are removed as well. The value with 0 in ordinal scales are also removed. Moreover, the letter format in “customer type”, “type of travel”, and “class” are corrected. The column called “Unnamed: 0” is the ID for each column. However, this is not useful for data analysis and mathematical calculation. Thus, this column is dropped as well. To make sure the data has been changed successfully, the “unique_detector” function is used again.

Creating a new column with the accumulation of relative values might be helpful for further analysis in terms of feature engineering. With this consideration, “departure delay in minutes” and “arrival delay in minutes” are accumulated into a new column named “total delay in minutes”. Another column called “flight distance classification” is created as a new classification feature. There are divided into three groups short-, mid-and long-haul based on

the flight distance. This can help the data analysis later to get a better understanding of the data. After deleting the rows with missing values, 95,419 rows remain in the dataset.

3.3 Data Overview

3.3.1 Descriptive Statistics

The descriptive statistics indicates the overview of the dataset. It helps the analyst to understand the abstract of the raw data before further analysis. According to descriptive statistics, the mean age is 39.8 years old. The standard deviation is 15 which means around 68 percent of passengers range around 33 to 47 years old and 95 percent of passengers range between 18 to 62 years old. The mean flight distance is 1,222 km with the 999 as one standard deviation. As for the flight delay, the average departure delay is 14.9 minutes, where the standard deviation is 38.3 minutes. On the other hand, the distribution of arrival delay is similar, where the mean is 15.3 minutes with 38.8 minutes as a standard deviation. Table 1 shows the descriptive statistics of the dataset.

	Age	**Flight Distance**	**Departure Delay in Minutes**	**Arrival Delay in Minutes**
Average	39.8	1,222.3	14.9	15.3
Standard Deviation	15.0	999.6	38.2	38.8
Min	7	31	0	13
Max	85	4,983	1,592	1,584

Table 1: Descriptive Statistics

3.3.2 Customers Rating Overview

After the data adjustment, the data has been visualized with diagrams. The values - “satisfied” and “neutral or dissatisfied” have been transformed into 1 and 0 respectively. Then, the histogram was adapted to demonstrate the mean of each questionnaire question with satisfied customers and dissatisfied customers. Figure 1 depicts the distribution of customer feedback.

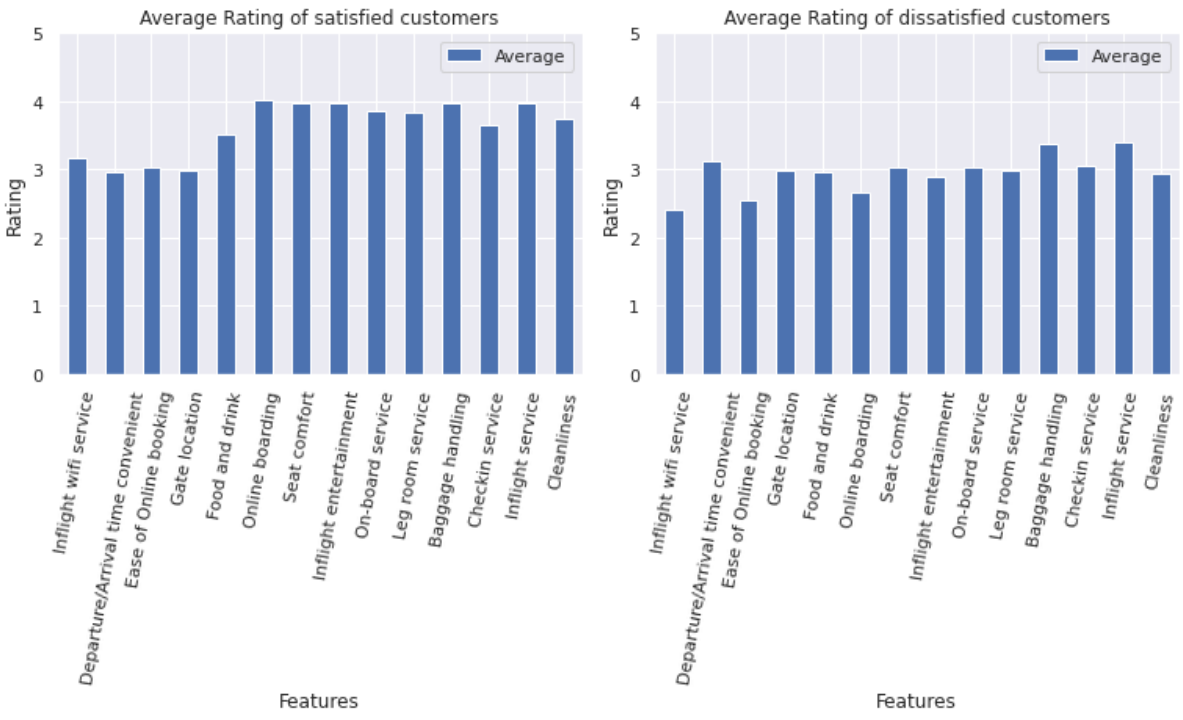

Figure 2: Customer Average Rating

In Figure 2, compared to dissatisfied customers, satisfied customers had a better overall rating in each criterion. However, the company has only limited capacity and budgets for service improvement, which is impossible to correct all the services at once. In addition, the summary will be too general if one concludes that all the services need to be improved. In the following report, the critical factor for customer satisfaction will be determined.

The highest score of satisfied customers is “online boarding”, whereas the lowest score of satisfied customers is “departure/arrival time convenient”. On the other hand, from the dissatisfied customers' feedbacks, the relevant lower scores are “inflight WIFI service”, “ease of online booking” and “online boarding”. Especially, the difference of “online boarding” between satisfied customers and dissatisfied customers is around 1.4. This indicates that “online boarding” might be a determinant factor in customer satisfaction. Another interesting finding in the diagram is that the ratings of “departure/ arrival time convenient” and “inflight service” from either satisfied customers or dissatisfied customers are similar. Based on this finding, it can be concluded that “departure/ arrival time convenient” and “inflight service” can be relatively less decisive than other services.

3.3.3 Correlation Matrix

Correlation represents the relations between two variables. The correlation value ranges from –1 to 1, where 1 indicates the strong positive correlation, -1 indicates the strong negative correlation, and 0 indicates no correlation. Figure 3 depicts the correlation matrix.

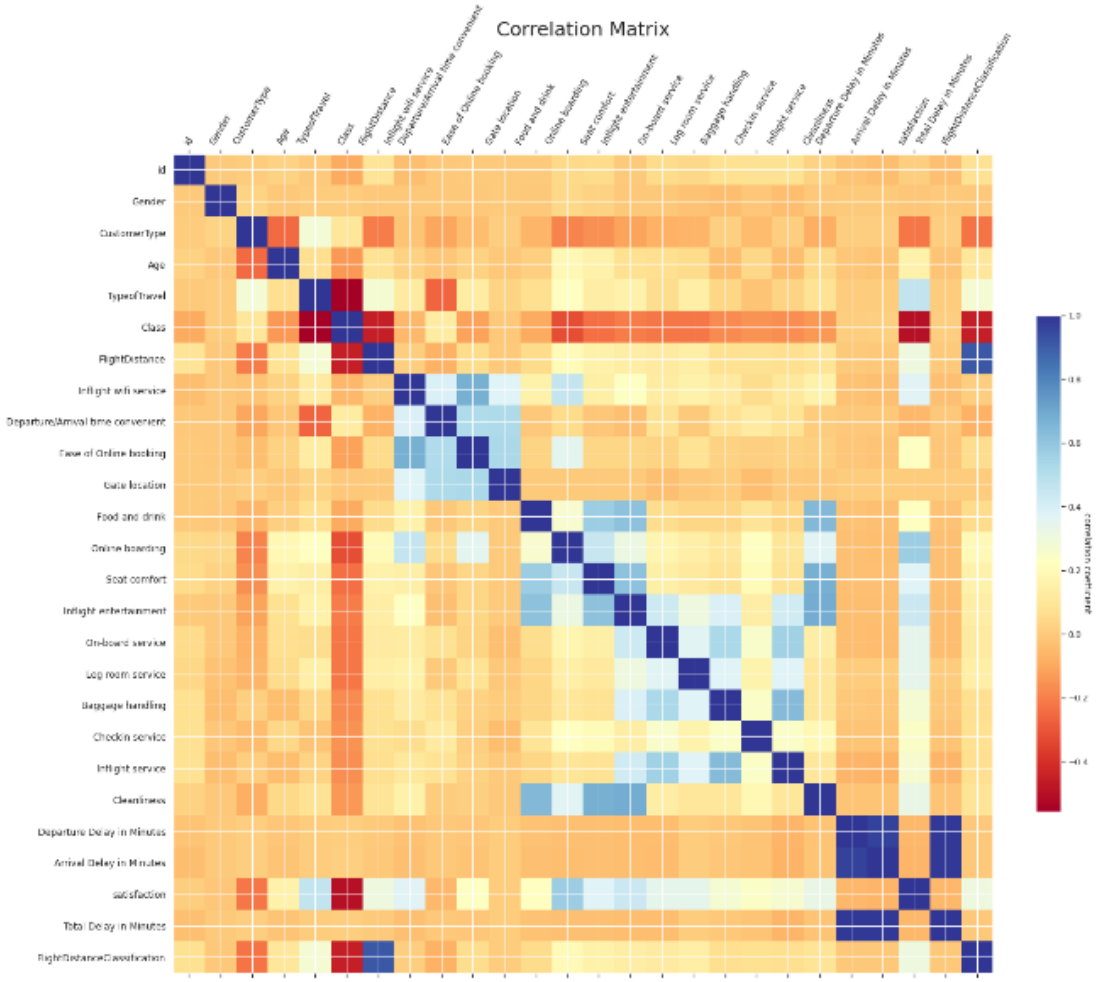

Figure 3 Correlation Matrix

The results show that “cleanliness” has a high positive correlation with “food and drink”, “seat comfort”, and “inflight entertainment”; “inflight WIFI service” has a high positive correlation with “ease of online booking”. On the other hand, “satisfaction” has a strong negative correlation with “class”, which means that the passenger in business class has more satisfaction than passengers in economy class. Interestingly, “satisfaction” has a fine positive correlation with “online boarding”, which was the same conclusion in the previous paragraph. The matrix can be found in the Google collab.

4. Customer Satisfaction Prediction

This section analyses the airline data regarding current and future passenger satisfaction to examine the reasons for dissatisfaction and satisfaction among the customers. Furthermore, based on the analysis recommendations for improvement of future operations are provided.

4.1 Data Analysis

Moving forward, the cleaned data is processed to analyse current customer satisfaction as well as predict passenger satisfaction for future operations and improvement. In order to identify relevant factors and variables influencing the passenger's flight experience, a decision tree, where the expected values of competing alternatives are calculated, was created. Thereby, all remaining values of the database are considered in the coding process. Based on the written code the following decision tree (figure 4) was built with a depth of four.

Figure 4: Decision Tree with a Depth of 4

Subsequently, eight factors appeared to be considered as the most influential parameters regarding customer satisfaction. These parameters are online boarding, inflight wifi service, type of travel (personal or business travel), inflight entertainment, travel class (business, eco, or eco plus), gate location, legroom service, and customer type (loyal or disloyal customer). Therefore, these drivers need to be considered further to predict passenger satisfaction and provide proper recommendations regarding the improvement process of the airline. Moreover, besides the additional factor of cleanliness, age has been determined as a key driver for passenger satisfaction as illustrated in the following chart. Figure 5 depicts the distributions.

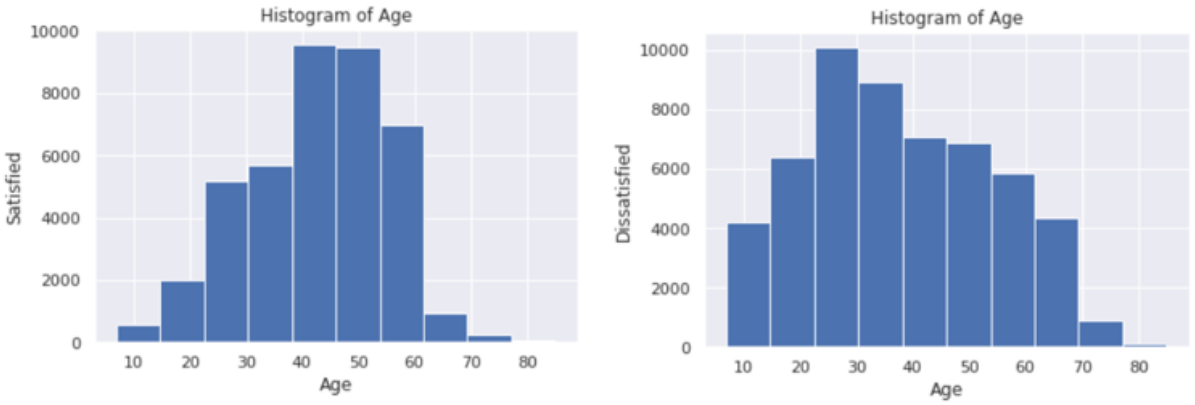

Figure 5: Age Histograms of Satisfied and Dissatisfied Customers

Regarding the connection of age and satisfaction, it is evident that age plays a vital role as the younger generations are more dissatisfied with the current situations than passengers that are between 40 and 60 years old. Hence, this clear difference needs to be considered in the prediction analysis.

Furthermore, based on the ten parameters mentioned above plus the total delay in minutes, a logistic regression model was trained and tested for the prediction of customer satisfaction. Thereby, a training accuracy of 0.8879 and a test accuracy of 0.8873 was detected, which proves a high precision of the model's results and subsequently a high relevance of the parameters regarding the overall satisfaction. Moreover, to illustrate the model's accuracy and relevance a confusion matrix was created, which is depicted in Table 2. Each row of the matrix represents the instances in an actual class while each column represents the instances in a predicted class.

		Predicted Label	
		Negative (N)	Positive (P)
Actual Label	Negative (N)	**9,892** (True Negatives)	**1,073** (False Positives)
	Positive (P)	**1,081** (False Negatives)	**7,038** (True Positives)

Table 2: Confusion Matrix

This matrix displays the model's number of rightfully and falsely predicted results of a separate test data set and based on the results, it can be concluded that the created model is fairly accurate and delivers reliable findings for further examinations.

Another more advanced methodology is to use Principal Component Analysis (PCA) to select critical features and create the weighted value, then, insert it into the logistic regression to calculate it once again. The training accuracy is then 0.8897 and the test accuracy is 0.8879, which are very close to the previous model without PCA. In the next step, to train a more accurate model, a neural network was introduced to predict satisfaction by fitting PCA. Neural network, a network-looking algorithm, is capable of learn the important features from the dataset and amplified the features. The results are quite successful as the test accuracy increased to 0.945, which is more precise than applying the logistic regression. Conclusively, if the airline aims to apply a more in-depth model, a neural network would be a further option.

4.2 Business Interpretation and Recommendation

Resulting from the analysis and determination of the most influential parameters, measures for improving passenger satisfaction can be concluded.

4.2.1 Type of Travel & Travel Class

This part elaborates on the possible impact of the type of travel and travel class on passenger satisfaction. The type of travel is classified in business and personal travel. As in figure 6 displayed, the passengers who are travelling for business are more satisfied with the current situation than people who take the plane for private reasons (e.g., vacation trips). However, business travellers are typically more used to the flight travel system as they fly more frequently than private persons. In addition, the probability of business travellers booking a business class ticket is higher compared to a private person doing so. Business travellers normally do not have to spend private money as their employers are paying for the plane tickets.

Moreover, as the services that are offered by the airlines differ between the travel classes, passengers in the business class are more likely to be satisfied. The services in business class are more comprehensive than in the eco or eco plus class. Therefore, these two factors are highly correlated, and it can be concluded that passengers with a high familiarity with the overall flying system (e.g., booking, check-in, airport infrastructure, inflight services) and/or buyers of business class tickets are more likely to be satisfied.

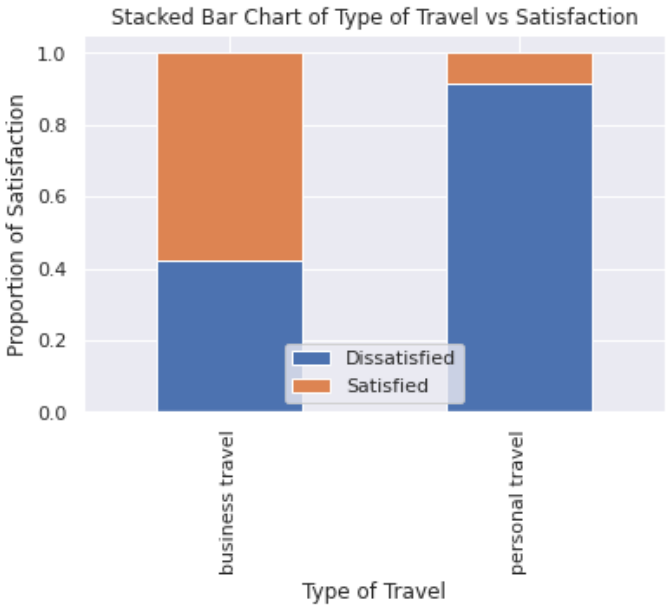

Figure 6: Type of Travel vs Satisfaction

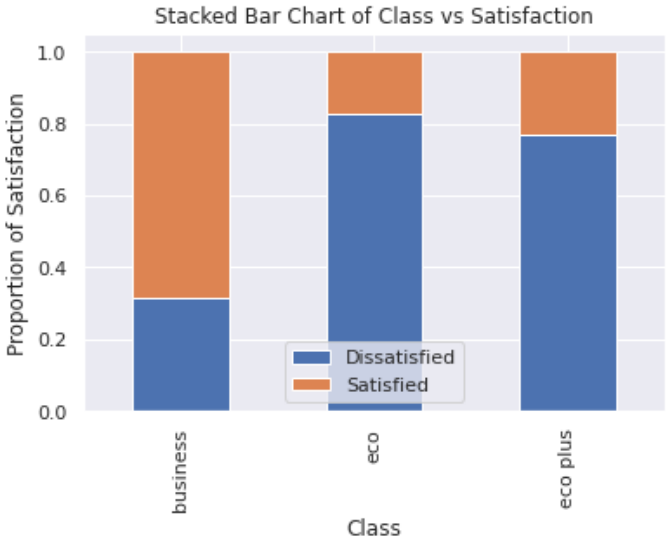

Figure 7: Travel Class vs Satisfaction

However, to reach the dissatisfied or neutral passengers, in this category the personal travellers and buyers of eco or eco plus tickets, it is advisable to ease the overall flying process by accompanying the customer during the whole process. Therefore, the airline should consider analysing the general customer journey and outline possible weak spots that could cause problems or incomprehension for non-frequent travellers. Furthermore, the current services offered in the eco or eco plus class need to be examined for improvement in alignment with current pricing. More detailed recommendations regarding the improvement of the inflight services and the departure process are provided in the subsequent sections.

4.2.2 Customer Type

During the conduction of the survey, two types of customers have been identified: disloyal and loyal customers. In general, loyal customers, who are familiar with the airline's procedures and services are more likely to be satisfied than disloyal customers, because they already know what standards they can expect from the airline. Nevertheless, the ratio of satisfied and dissatisfied customers in this category is nearly equal (see figure 8), which indicates that even the loyal customers are not always pleased with the airline's offerings. However, there is no further information regarding this issue available, so it cannot be clearly determined what impacts a loyal customer negatively. Theoretical reasons can be the number of business clients, who frequently use this airline or the mileage program. Disloyal customers, on the other hand, can be people who booked this airline due to individual circumstances (e.g., best price for a specific flight) rather than loyalty. This kind of customer is more difficult to satisfy, because they have no prior connection to the company.

Figure 8: Customer Type vs Satisfaction

The primary goal in this section is to examine possibilities to keep the loyal customers and gain disloyal or temporary customers as permanent passengers. Therefore, the airline is tasked to focus more on building and maintaining customer loyalty by extending its customer relationship management. One option would be partnering with other brands/companies to expose and offer existing customers more extensive services. For example, they can collaborate with car rental companies or hotels to offer services beyond their own boundaries. Another option would be

the creation of a strong customer engagement by creating marketing content that is as exciting, entertaining, and informative as it is promotional, developing a brand voice that feels authentic and approachable, or trying different mediums such as videos, infographics, and social media content to find out to what the customers respond. In any case, the airline should introduce a strong customer-centric focus and develop a more intensive customer relationship management.

4.2.3 Age

As previously discussed, age also plays an essential role in defining customer satisfaction. The perception of what is noticed as good service differs between the generations as can be concluded by Figure 9 while passengers who are between 40 to 60 years old are more satisfied with the current service level of the airline, younger passengers are more likely to be dissatisfied as they, for example, expect more comprehensive services.

Figure 9: Age vs Satisfaction

Regarding the business implications of this analysis, the airline needs to further examine the demographics of its customers and follow current trends and behaviour to create more profound customer profiles. For instance, the airline needs to collect customer feedback, build personas, and review their customer journey map.

4.2.4 Gate Location & Online Boarding

Furthermore, the additional parameters gate location and online boarding have been identified as substantial drivers for customer satisfaction. Nonetheless, the location of the departure or arrival gates is a variable that cannot be influenced by the airline as the flight coordination is processed by the individual airports and their applied system. However, the airline can provide

its customers with a more detailed description and insight of the structure of the approached airports by offering, for example, an app that provides detailed information and maps of the individual airports.

Moreover, regarding the factor of online boarding, which is under the sole responsibility of the airline operator, it is advisable to ease the process as much as possible to avoid questions and lack of understandings. In addition, the airline should offer the option to download the boarding pass on mobile phones for a quicker boarding procedure.

4.2.5 Inflight Services

Under this section, the determined influencing factors inflight entertainment, inflight wifi service, leg room service, and cleanliness are addressed. The inflight entertainment service is highly valued by customers, because they are spending much time with it during the flight, especially during long-haul flights. Therefore, it is advisable to offer an extensive entertainment package by offering a good selection of movies, tv-series and music. The airline can partner with streaming services like Amazon or Netflix to provide a comprehensive entertainment experience. Additionally, a stable Wi-Fi service with a high download rate is demanded by many passengers. Nowadays, business travellers, as well as private travellers, are used to having constant access to the digital world. Businesspeople, for example, use the time during the flight to do some works, which is often only possible with access to the internet. Private persons, in addition, want to use their phones, tablets, and laptops to play games, listen to music, or watch movies on their own devices.

Furthermore, the leg room is also highly valued by the passengers. However, the definition of an adequate leg room depends on the customer's height and length of legs. Tall people with long legs are more likely to be dissatisfied with the given space than smaller people. A measure for this issue could be to conform to the industry standard of providing leg room space. Furthermore, if the airline aims to satisfy all passengers in this category it needs to broaden the leg room space by reducing the number of rows in the airplane. This will decrease the profit per flight but perhaps increase the company’s image as it gets known for its comfort. However, the airline would then need to compensate for this by increasing the price per seat which could lead to a loss of customers. Consequently, the airline needs to weigh out the benefits and advantages of restructuring airplanes seating arrangements.

Lastly, the cleanliness of the planes needs to be discussed as it is an essential point for customer satisfaction. Besides the overall understanding of cleanliness, the airline's perception of cleanliness should also include the presentation of seat areas, tables, carpets, cabin panels, and washrooms. A proper cleaning procedure or protocol needs to be adopted in order to provide a high level of hygiene. In addition, since the beginning of the Covid-19 pandemic, the cleaning practices of airlines took a centre stage in the consciousness of travellers. Every plane needs to be cleaned properly after each flight to ensure a comfortable level of cleanliness.

5. Customer Clustering

5.1 Data Analysis

After identifying relevant factors for predicting customer satisfaction and building a reliable prediction model, clustering is used to answer the question of whether it is possible to find groups of similar customers and how to characterize these groups.

Clustering, also known as cluster analysis, is an unsupervised type of learning that helps for this goal because we don’t know what we are looking for. It is defined as the process of recognizing the natural and homogeneous groups within data (Aljarah et al. 2021, p. 2). The goal of clustering is to have k-clusters in which the objects have high similarity in comparison to one another but are very dissimilar to objects in other clusters (Raschka and Mirjalili 2021, p. 377). In other words, the intra-cluster similarity should be high and the inter-cluster similarity low. According to this, a cluster is a collection of data objects that are similar to each other. Quantifying the similarity or dissimilarity among data is achieved through the utilization of proximity measures such as distance measures.

Different methods are existing, e.g. hierarchical, partitional, probabilistic clustering, that vary in the approach of measuring similarity (HajKacem et al. 2019, p. 1). For this report k-means is used, which is the most fundamental partitional clustering method (HajKacem et al. 2019, p. 2; Aljarah et al. 2021, p. 5). For this method, a centre represents a cluster and each data point is assigned to a cluster based on a Euclidean distance measure. This method performs exclusive clustering, which means that each object belongs only to one cluster. It creates an initial partitioning and improves by iterative relocation of each data object while calculation and improving the mean value of the clusters.

Before running the clustering algorithm on our dataset, the ID column is deleted as it is not useful for characterizing different clusters. Furthermore, the clustering method is sensitive to outliers and the ideal dataset should consist of objects with the same scaling (Aljarah et al. 2021, p. 7).

For the k-means method, the number of clusters must be given as an input variable for the model. The “elbow method” in the clustering part of the Jupiter notebook (Figure XY) finds the best number of clusters. In this method, a certain number of clusters is defined, so adding another cluster does not allow for significantly better modelling of the data. For this purpose, the percentage of variance explained by the clusters is plotted against the number of clusters (Bholowalia and Kumar 2014, p. 18 f.). The algorithm clusters the data several times with varying i (number of clusters). For the percentage of variance, the algorithm calculates the WCSS (within-cluster sum of the square), which is the sum of squared distance between each point and the centroid in a cluster.

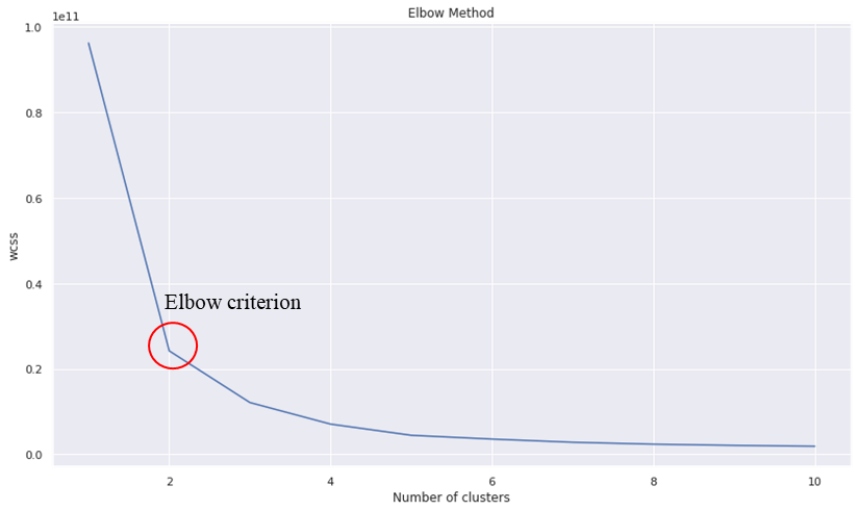

Figure 10: Elbow Diagram

As shown in Figure 10, the first cluster will add much information but already at the second cluster the information gain drops. The “elbow criterion” is fulfilled at two clusters because the slope of the graph changes the most. Increasing the number of clusters is not leading to better information and new clusters are very close to already existing clusters. The results with additional clusters are presented in more detail in the critical reflection of the next section. Because no other input variables are necessary, the “KMeans” model with four clusters is defined and fitted with the data (df_new). In this part, no train-test-split is necessary because of

the unsupervised type of learning. It is not helpful to train the cluster model on one data set and apply it to an unknown data set to see if the clusters predicted for the unknown data are correct. There is no way to verify correctness because, in this type of learning, the outcome is not known in advance.

After fitting the KMeans-model, clusters for the data are predicted and a list (cluster) is created with the results. This list contains the individual cluster predictions (0 and 1) ordered by the customer entries. Before proceeding with the results, the length of the dataset (df_new) is compared with the length of the prediction results, to check whether the prediction was successful for all entries. The total length is also used for creating a new dataset (cluster_df) that includes the customers and the predicted clusters (Figure 10). The code below shows how the table is created. It is based on the index of the overall dataset and the results of the cluster prediction. The total length is used to define the total rows for the two columns.

```
#create a table with customer and cluster
cluster_df = pd.DataFrame(list(zip(df_new.index,model.predict(df_new))),
columns=['Customer','Cluster_predicted']) [0:95419]
cluster_df.head()
```

	Customer	Cluster_predicted
0	0	0
1	1	0
2	2	0
3	3	0
4	4	0

Figure 11: Structure of cluster_df

A left-join function combines the overall dataset (df_new) and the cluster dataset in the new dataset “merged_df”. In this dataset are 68,963 customer entries that are allocated to cluster zero and 26.456 entries to cluster one.

Cluster	Data Objects	Proportion
0	68,963	72,2%
1	26,456	27,7%
Total	**95,419**	**100,0%**

Table 3: Customer Clusters

To interpret the different clusters and answer the question of how to characterize the groups, two new data sets are created with the following code, containing only the entries for each cluster.

```
#create a data frame for each cluster
df_new_cluster0 = df_new_cluster[df_new_cluster.Cluster_predicted == 0]
df_new_cluster1 = df_new_cluster[df_new_cluster.Cluster_predicted == 1]
```

Before moving on to the data interpretation, the cluster model was validated by testing another model (model2). The purpose of this test was to see if a less complex dataset would lead to a different cluster prediction. For this, a new dataset (df_new2) was used. It contains only descriptive characteristics about the different customers. The "Elbow Method" also identifies two clusters as the best input to the KMeans model, and after fitting the model to the new data set, the cluster prediction was the same as the first cluster model. Therefore, the first model is used for cluster interpretation.

5.2 Business Interpretation and Recommendation

Finally, to answer the question of whether it is possible to find groups of similar customers and how to characterize these groups, the different datasets for each cluster are analysed. In python, the function .describe( ) already provides many KPIs for each column in the dataset. For the business interpretation, the mean values are very informative. To visualize them with bar charts, a for loop is used to get the relevant column headings and the corresponding mean values in two different lists.

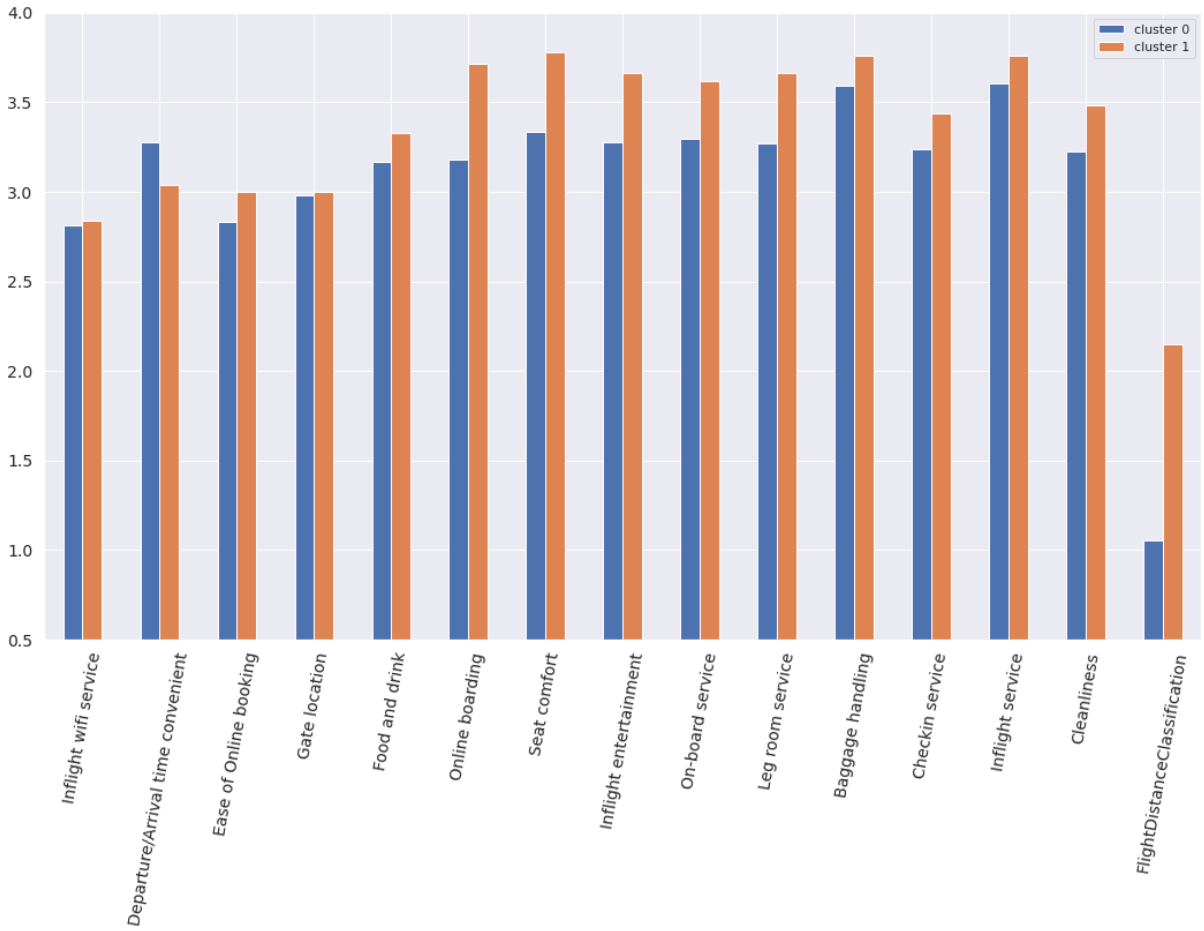

Figure 12: Mean Values of Rating Categories for Different Clusters

It can be seen that for some answer categories, customers in the different clusters answered similarly and for other answers, the evaluation is rather different. In general, customers in cluster one are more satisfied as ratings in all categories except for "Departure /Arrival time convenient" is higher than in cluster zero. To better interpret what types of customers are in each cluster and what makes them different, an analysis of the descriptive factors can be used. These are "Gender", "Customer Type", "Age", "Type of Travel", "Class" and "Flight Distance Classification".

5.2.1 Gender

In terms of gender, there is no difference between the two clusters. Male and female are equally distributed in both clusters. Thus, for the airline, gender cannot be used to advertise to different gendered customer groups. Since a total of only 7 customers indicated gender as diverse, it is not possible to see in the visualization to which clusters they belong.

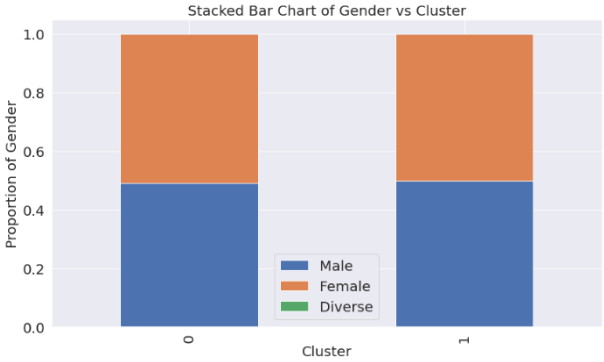

Figure 13: Proportion of Genders in Clusters

5.2.2 Customer Type

The customer type is slightly different for the clusters. 21 % of the customers in cluster zero are disloyal customers, whereas almost all customers in cluster one are loyal customers. Since 83.9% of all customers in the data set are loyal, most of the customers in both clusters are loyal but the analysis shows that almost all disloyal customers are in cluster zero.

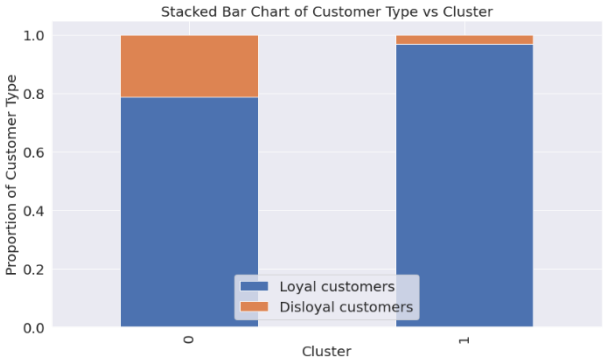

Figure 14: Proportion of Customer Types in Clusters

5.2.3 Age

As far as age is concerned, the diagrams show that most younger customers (between 20 and 40 years of age) are found in cluster zero. The age difference is also reflected in the mean values. For cluster zero, the average age is 38.87 years and for cluster one, it is 42.27 years.

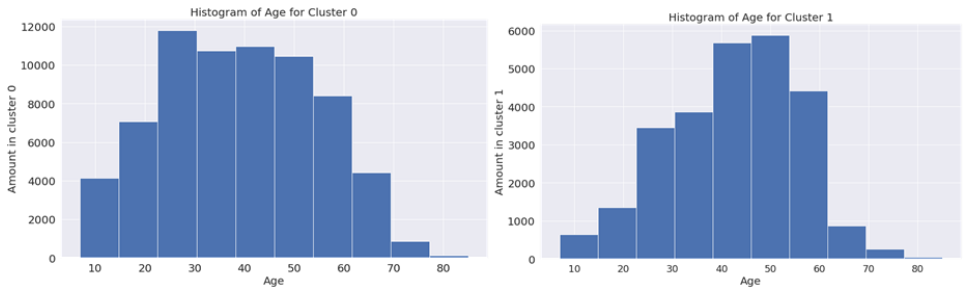

Figure 15: Age Histograms for the Clusters

5.2.4 Type of Travel

For Type of Travel, the majority in both clusters are business travellers. However, in cluster zero, almost 40 % of the customers are personal travellers. Therefore, when advertising to personal travellers, the airline should focus on the characteristics of cluster zero. Since the majority of customers are business travellers, the airline has the potential to attract more personal travellers to its service.

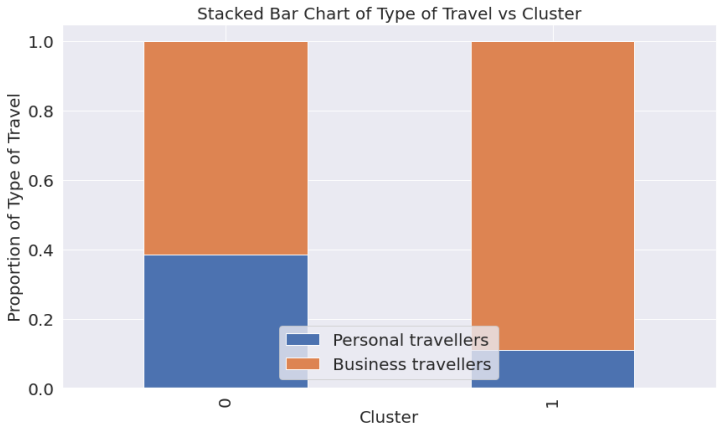

Figure 16: Proportion of Types of Travel in Clusters

5.2.5 Travel Class

Similar to the “Type of Travel”, the class indicates, that the share of businesspeople is higher in cluster one. Almost 90 % of the customers in cluster one booked a flight in the business class. In cluster zero, nearly 70% of customers were in either the eco-class or eco-plus class.

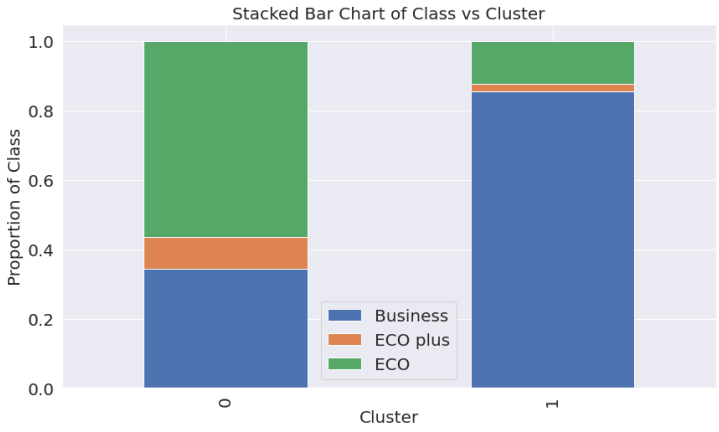

Figure 17: Proportion of Classes in Clusters

5.2.6 Flight Distance Classification

Finally, the flight distance classification shows that almost all customers in cluster zero are short-haul travellers and cluster one contains medium-haul and long-haul travellers. This difference is useful for identifying the customer group that uses the airline's more profitable service.

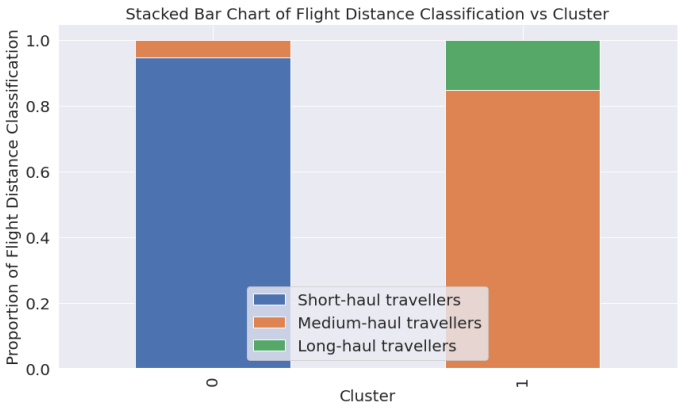

Figure 18: Proportion of Flight Distance Classifications in Clusters

All this information can be summarized in the following customer profiles, which describe similar preferences:

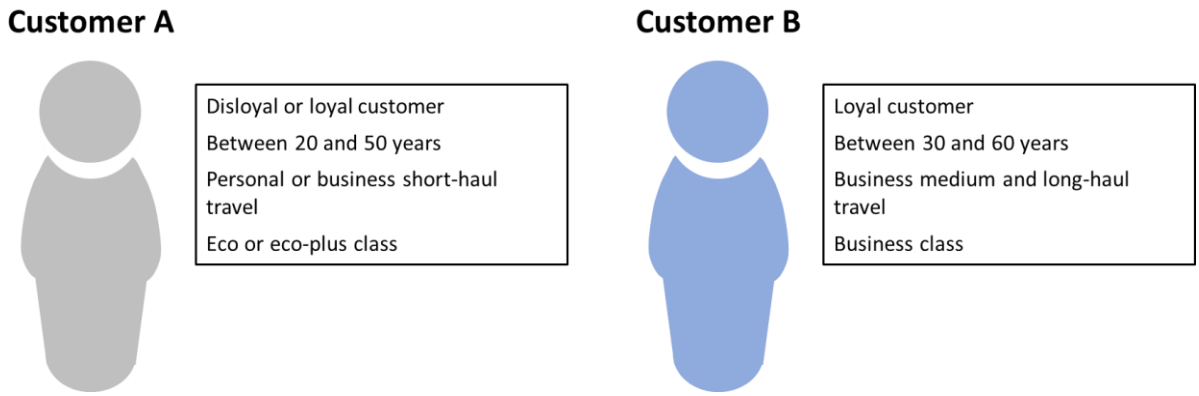

Figure 19: Customer Profiles Representing 2 Clusters

Customer A is a disloyal customer between the ages of 20 and 50. The customer uses the airline for personal short-haul flights and books eco or eco-plus seats. Also, some loyal business travellers have the same preferences as Customer A, but this can be explained by the fact that the majority of customers in the data set are business travellers. Customer B is a loyal customer slightly older than Customer A. The customer can be characterized as a business traveller, booking business seats for medium and long-haul flights.

With this information, the airline could advertise the different customer groups more effectively. Targeted advertising that addresses clearly defined groups of potential customers with only relevant content is a very successful marketing approach that reduces wastage and is cost-effective. The rating of Customer B is higher than Customer A for all categories except “Departure / Arrival time convenient”. One of the reasons could be that business class travellers are generally more satisfied because they enjoy better service. If the survey was conducted during or immediately after the flight, comfort would influence the results more.

Nevertheless, the results are important because customers evaluate the various services concerning the higher price of business seats. As "Departure/Arrival time convenient" has received a low rating in the comparison, this characteristic seems to be particularly relevant for Customer B. The airline could use this information to improve it and advertise accordingly. The advertising could also be targeted to the age of the customers. Age is a useful indicator of certain preferences such as entertainment and online services for younger customers or special food and beverages for older customers. The information about the current customers can also be used to address new potential customers. Because the majority of the customers are business travellers, has the potential to attract personal travellers by flying to new travel destinations and advertising the airline for vacation customers.

In order to get more detailed information about the customers and because the elbow method was not very clear, the cluster model was also tested and analysed for three and four clusters. The results of the rating analysis already show that clusters one and two, as well as clusters zero and three, are very similar.

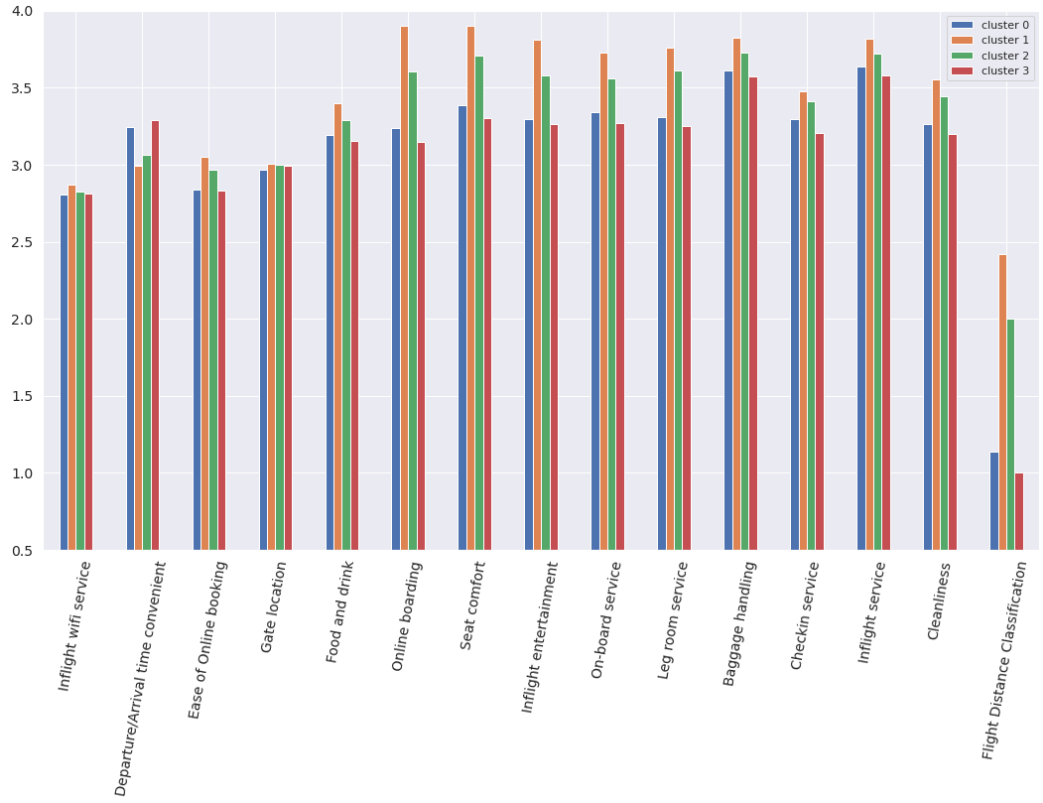

Figure 20: Mean Values of Rating Categories for 4 Clusters

Furthermore, the analysis of the descriptive characteristics shows that four customer profiles are derived, where two are also similar in each case. Two clusters are therefore sufficient for deriving relevant marketing measures.

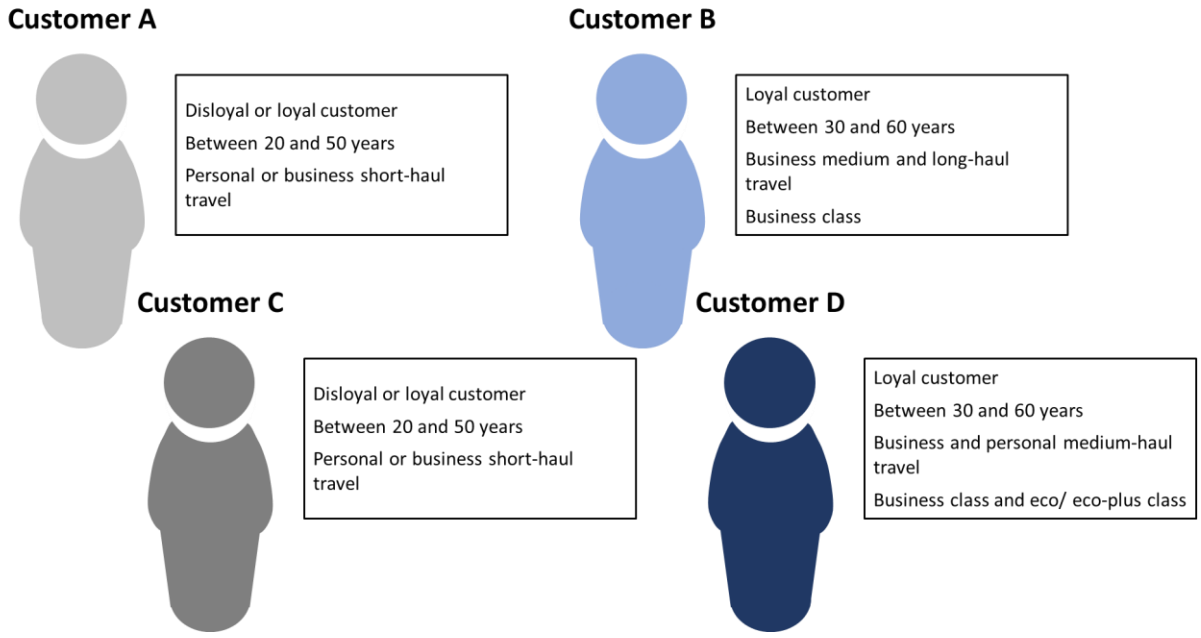

Figure 21: Customer Profiles Representing 4 Clusters

6. Conclusion

To conclude, predicting customer satisfaction and clustering passenger groups inherits a substantial number of variables that need to be considered for a proper analysis of current data for future improvements. Thereby, ten parameters have been identified and examined to provide profound recommendations for possible areas of service improvement. Based on the gathered insights, the airline, for instance, needs to extend its customer relationship management and review its customer journey map. Additionally, through the informed creation of customer profiles further marketing strategies and issues can be concluded.

Lastly, as the survey is from pre-covid times, the parameters of customer satisfaction may change, and further research is required. Moreover, this survey includes only feedback from passengers of a US airline and can, therefore, not necessarily be applied to other countries.

7. References

Aljarah, I., Habib, M., Faris, H. and Mirjalili, S. (2021) 'Introduction to Evolutionary Data Clustering and Its Applications', in Evolutionary Data Clustering: Algorithms and Applications, Algorithms for Intelligent Systems, Singapore: Springer Nature Singapore Pte Ltd, pp. 1–23. doi: 10.1007/978-981-33-4191-3_1.

European Union (2021) Air Passenger Rights, Available at: https://europa.eu/youreurope/citizens/travel/passenger-rights/air/index_en.htm#upgradeAndDowngrade (Accessed on: 12.01.2022).

HajKacem, M.A.B., Ben N'Cir, C.-E. and Essoussi, N. (2019) 'Overview of Scalable Partitional Methods for Big Data Clustering', in Clustering Methods for Big Data Analytics Techniques, Toolboxes and Applications, Cham, Switzerland: Springer Nature Switzerland AG, pp. 1–24. doi: 10.1007/978-3-319-97864-2_1.

Raschka, S. and Mirjalili, V. (2021) Machine Learning mit Python und Keras, TensorFlow 2 und Scikit-learn : Das umfassende Praxis-Handbuch für Data Science, Deep Learning und Predictive Analytics', mitp, 2021. ProQuest Ebook Central.

Sammut, C. and Webb, G. I. (2017) Encyclopedia of Machine Learning and Data Mining, Springer, Boston, MA. doi: https://doi.org/10.1007/978-1-4899-7687-1