**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning |
| **Assessment Title:** | What factors most impact short-haul dissatisfaction? |
| **Lecturer Name:** | Dr Muhammad Iqbal |
| **Student Full Name:** | Group:  Ana Isabel Nieves Barcenas  Bárbara Azevedo Pereira  Daniela Daia  Vicente Rubio |
| **Student Number:** | 2022455 - Ana Isabel Nieves Barcenas  2022310 - Bárbara Azevedo Pereira  2017207 - Daniela Daia  2022178 - Vicente Rubio |
| **Assessment Due Date:** | 27 November 2022 |
| **Date of Submission:** | 27Th November 2022 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

**What factors most impact**

**short-haul dissatisfaction?**

****

by,

Ana Isabel Nieves Barcenas

Bárbara Azevedo Pereira

Daniela Daia

Vicente Rubio

**Higher Diploma in Science in Data Analytics for Business Strategic Thinking**

Dr Muhammad Iqbal

CCT College Dublin, Ireland.



## **Abstract**

This analysis is based on a classification data set research of over 120,000 airline passengers' satisfaction. It will study what factors are highly correlated to dissatisfaction with short-haul passengers.

It will describe the motivation of the chosen data, a description of the business problem and an explanation of the project goal. It will present the characterisation of the data by applying Exploratory Data Analyses (EDA), filling in the missing values, observing the outliers by plotting boxplots, and using the feature selection model to extract the influencing factors of passenger dissatisfaction by using machine learning accuracy. Afterwards, cross-validation techniques will be applied by using machine learning approaches, hyperparameters and a comparison between the chosen model. And finally, the interpretation and explanation of the results obtained based on different classification models.

Keywords: a*irlines, passenger satisfaction, machine learning models, CRISP-DM.*

## **Table of Contents**

[**Abstract** 3](#_Toc120447798)

[**Table of Contents** 4](#_Toc120447799)

[**Table of Figures** 5](#_Toc120447800)

[**Introduction** 6](#_Toc120447801)

[**Business Understanding** 7](#_Toc120447802)

[**Data Understanding** 8](#_Toc120447803)

[**Data Preparation** 17](#_Toc120447804)

[**Modelling** 21](#_Toc120447805)

[**Deployment** 25](#_Toc120447806)

[**Extra Contents** 26](#_Toc120447807)

[Roles and responsibilities 26](#_Toc120447808)

[Team Project management 26](#_Toc120447809)

[**Reference List** 28](#_Toc120447810)

## **Table of Figures**

[Figure 1. CRISP-DM Methodology. 6](#_Toc120442801)

[Figure 2. Data dictionary 8](#_Toc120442802)

[Figure 3. Required libraries 9](#_Toc120442803)

[Figure 4. Head and shape of the dataset. 9](#_Toc120442804)

[Figure 5. The function .info() 10](#_Toc120442805)

[Figure 6. Number summary (describe). 10](file:////Users/danidaia/Downloads/CA%20ML%20Final%20Draft.docx#_Toc120442806)

[Figure 7. Pairplot of the dataset numerical variables 11](#_Toc120442807)

[Figure 8. Histogram of the dataset numerical variables 12](#_Toc120442808)

[Figure 9. Categorical variables summary statistics 12](#_Toc120442809)

[Figure 10. Bar chart of the dataset categorical variables 13](#_Toc120442810)

[Figure 11. Skewness distribution 13](#_Toc120442811)

[Figure 12. Satisfaction level bar chart 14](#_Toc120442812)

[Figure 13. Outliers boxplot 15](#_Toc120442813)

[Figure 14. Heatmap 15](#_Toc120442814)

[Figure 15. Finding missing values 17](#_Toc120442815)

[Figure 16. missing values graph 17](#_Toc120442816)

[Figure 17. Short-distance flights 18](#_Toc120442817)

[Figure 18. Dropping columns 19](#_Toc120442818)

[Figure 19. Encoded data set 20](#_Toc120442819)

[Figure 20. Sparse data check-up 20](#_Toc120442820)

[Figure 21. Separating independent from dependent variables 21](#_Toc120442821)

[Figure 22. Accuracy Scores with different splits 21](#_Toc120442822)

[Figure 23. Cross Validation Scores 23](file:////Users/danidaia/Downloads/CA%20ML%20Final%20Draft.docx#_Toc120442823)

[Figure 24.Randon Forest Confusion Matrix 24](#_Toc120442824)

[Figure 25.Logistic Regression Cofusion Matrix 25](#_Toc120442825)

[Figure 26.KNN Confusion Matrix 25](file:////Users/danidaia/Downloads/CA%20ML%20Final%20Draft.docx#_Toc120442826)

[Figure 27.Roles and Responsibilities 28](#_Toc120442827)

[Figure 28. Project Management: Trello Board 29](#_Toc120442828)

[Figure 29.Team's effort 29](file:////Users/danidaia/Downloads/CA%20ML%20Final%20Draft.docx#_Toc120442829)

## **Introduction**

Whether travelling to nearby destinations or even the farthest corners of the world, the keywords in choosing airlines are to travel quickly, comfortably, and safely. Airlines are constantly offering new destinations, and over time more companies have entered the market, offering passengers choices, competitiveness, and affordability.

In a highly competitive environment, the aviation industry stands to develop from a transport role to a service. Improving service quality is essential for competitiveness and ensures sustainable and healthy development. Therefore, airlines should promptly investigate passenger satisfaction and overall satisfaction with various services to understand the quality of existing services.

This report will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. CRISP-DM provides a complete model for a Data management project. The project is divided into six phases: Business understanding, Data Understanding, Data Preparation, Modelling, Evaluation, and Deployment (Shearer, 2000). Code available at [*GitHub*](https://github.com/IsabelNieves/Machine-Learning-CA.git)*.* The life cycle is shown in Figure 1.

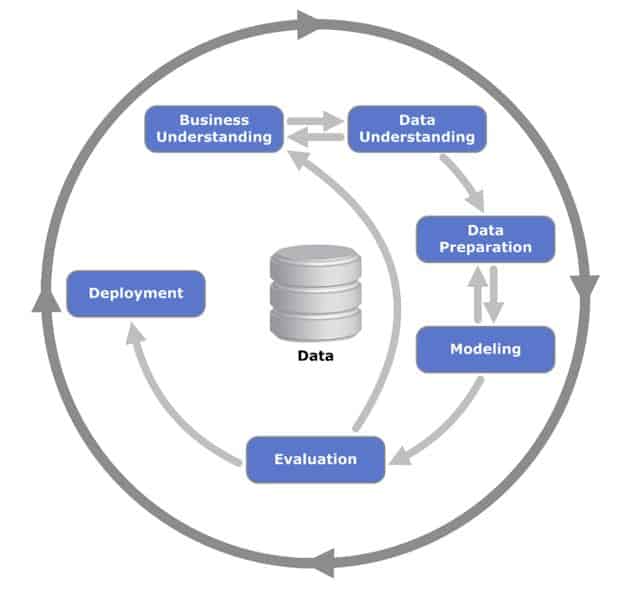
**

Figure 1. CRISP-DM Methodology.

## **Business Understanding**

It is known that the airline industry is one of the fastest transportation sectors in the world; in that regard, Bart (2000) argues that traces of the strategic developments and the strategic responses of the airline players have had a profound impact on the shape and direction of the industry. These include the deregulation of the sector, the nature and extent of competition, the emergence of brand/differentiation-based competition, and airline alliance developments, strategies and their implications (Williams & Naumann, 2011).

Based on the above challenges, this study focuses on the full-service passenger information and satisfaction survey results. This analysis aims to evaluate different machine learning algorithms and determine the most suitable algorithm for classifying customer short-haul flight dissatisfaction. This analysis also aims to ascertain and highlight the most critical variables in determining customer dissatisfaction for a better insight into the issues. Finally, this study is a reference for airlines to use customer evaluation-driven service methodologies to improve their competitiveness.

## **Data Understanding**

According to Han, Kamber and Pei (2011), data characterisation is a summarisation of the variables and factors of a target course of information, such as simple data summaries based on statistical measures and plots and other strategies. For this project, the dataset chosen is Airline Passenger Satisfaction, available on [Kaggle](https://www.kaggle.com/datasets/mysarahmadbhat/airline-passenger-satisfaction), and whose information is related to customer satisfaction scores from over 120,000 airline passengers, including additional information about each passenger, their flight, and type of travel, as well as their evaluation of different factors like cleanliness, comfort, service, and overall experience.

A data dictionary is used to catalogue and communicate the structure and content of data and provides meaningful descriptions for individually named data objects (Wertz, 1993).

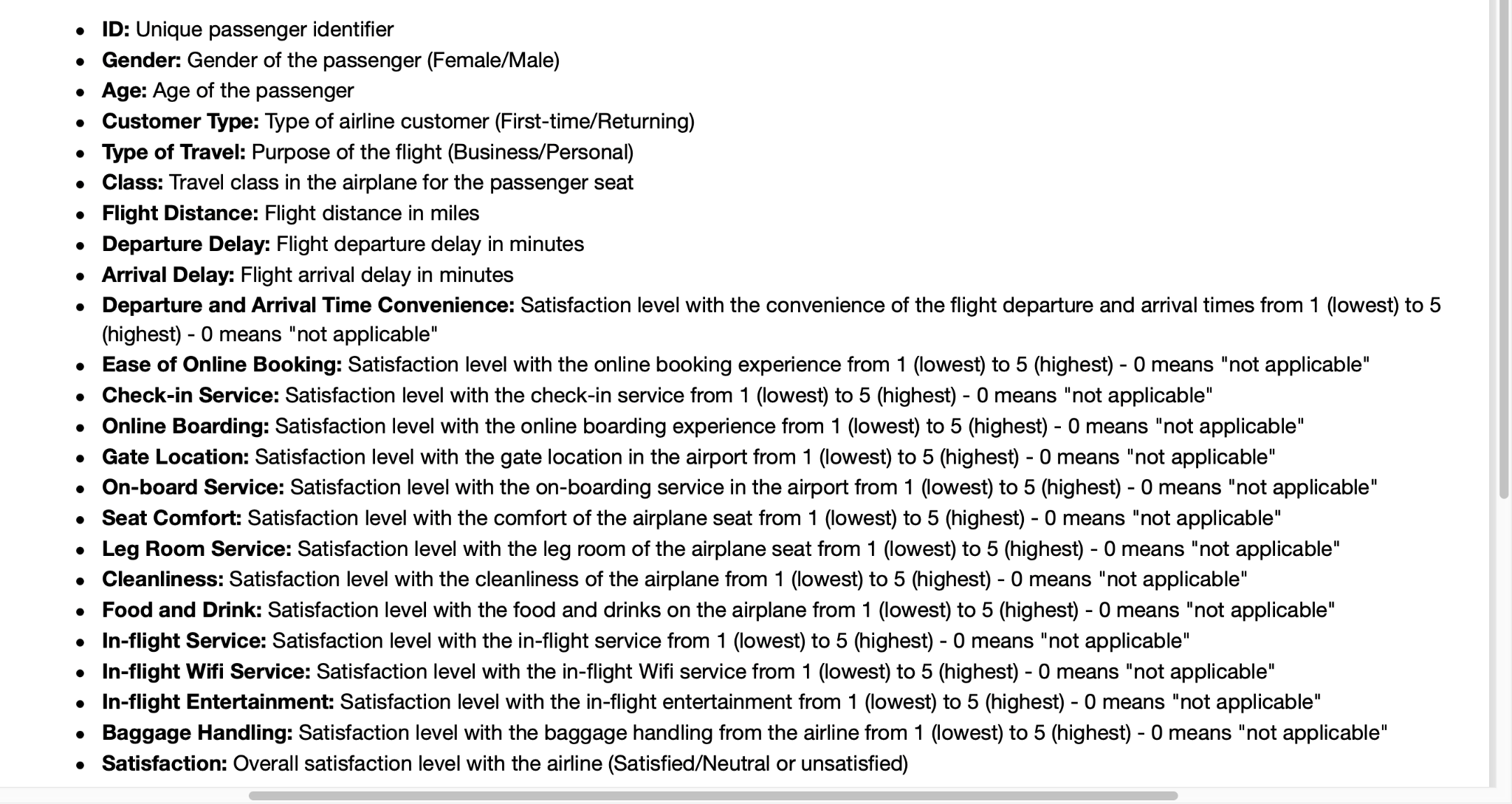
****

Figure 2. Data dictionary

We import the required packages such as Pandas, Seaborn, Numpy, Matplotlib, Math, Missingno, Sklearn and a warning filter initial installation, allowing us to run all the analyses in our Colab notebook code. With Colab, we can import an image dataset, train an image classifier, and evaluate the model (Google Colaboratory, 2022).



Figure 3. Required libraries

In this session, it will be imported the raw data set to find relevant information from this data by identifying the Predictor (Input) and Target (output) variables (Ray, 2019). After loading, by using the “head” function, we can see the five rows of the data set and with the “.shape” function, it is possible to see that the data size consists of 129880 rows and 24 columns (The pandas development team, 2020).

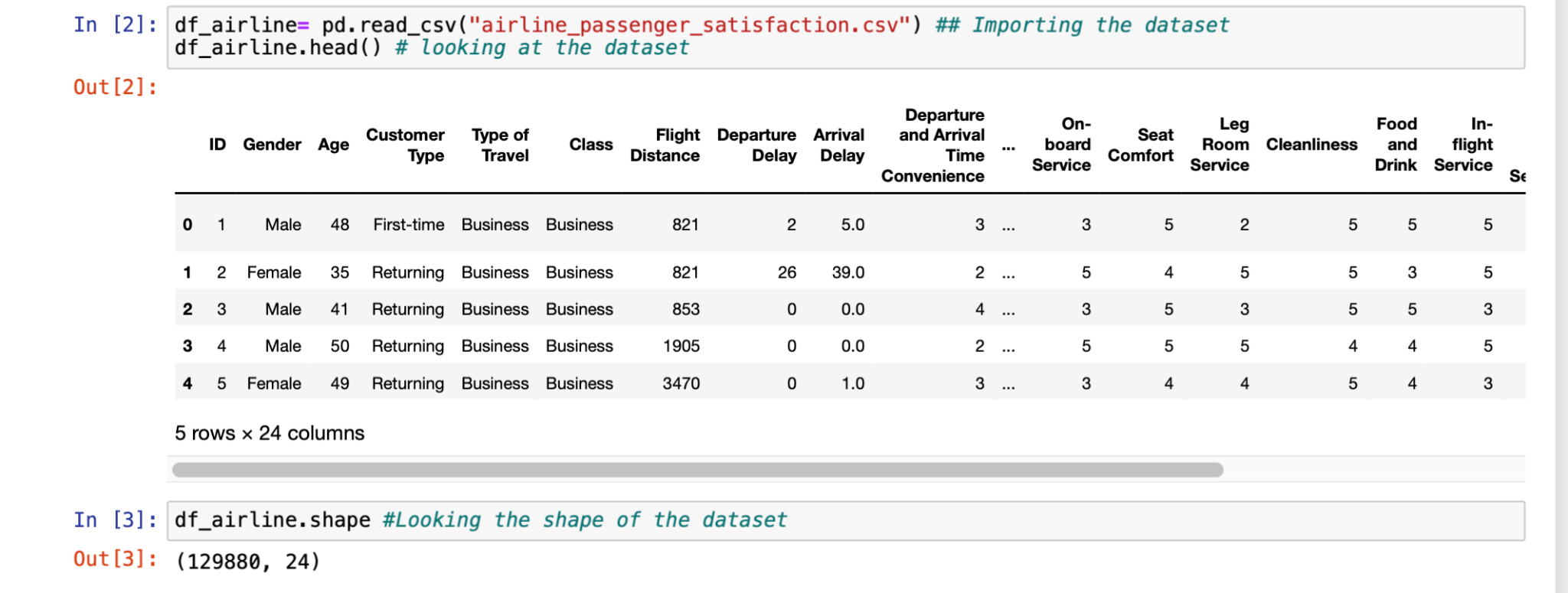


Figure 4. Head and shape of the dataset.

The function info() shows more details about the dataset, such as shape, type of variables and memory used:

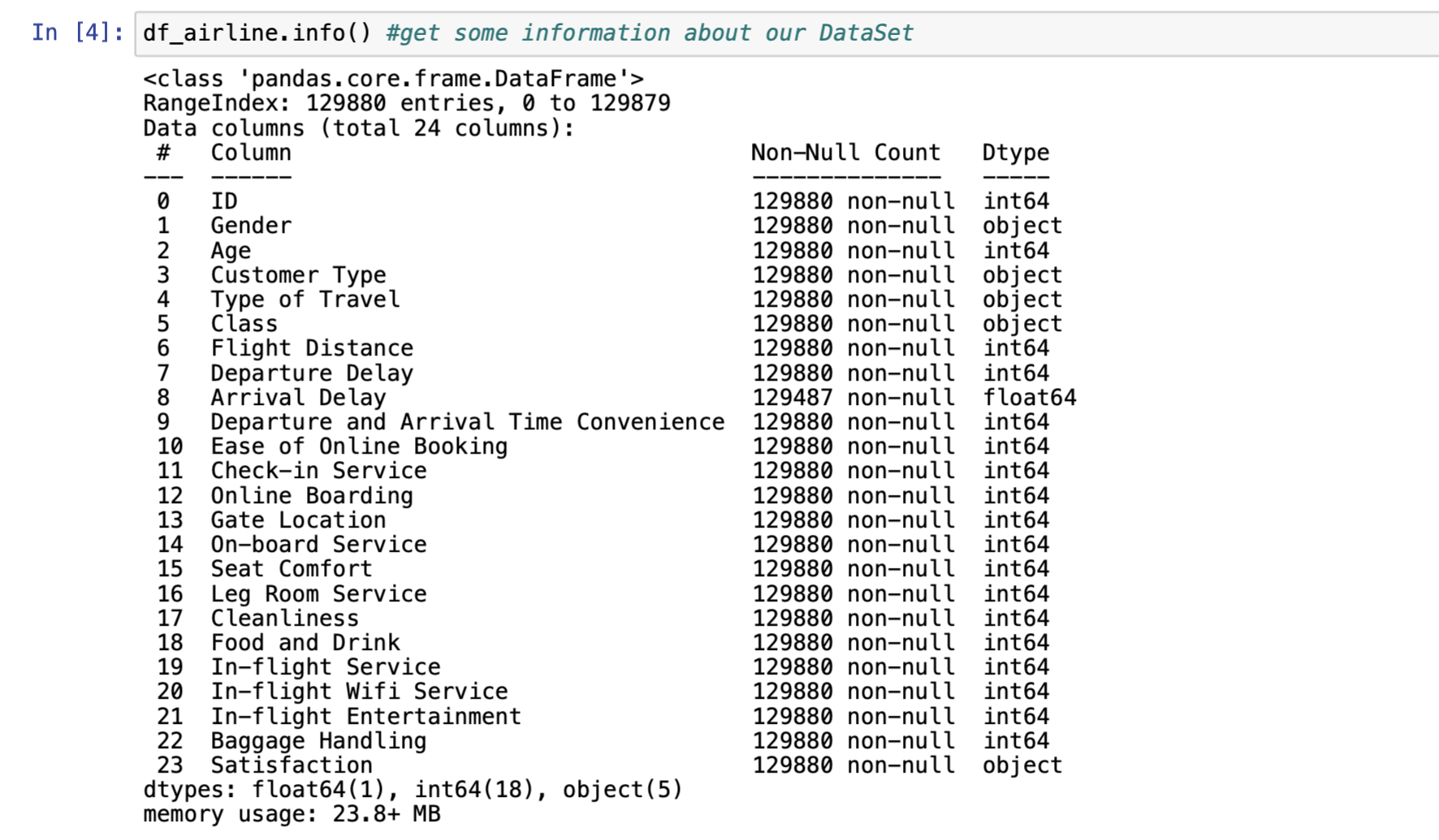


Figure 5. The function .info()

The next step consisted of obtaining the summary statistics for the numerical values in the data frame using the “.describe” function, which is responsible for generating descriptive statistics that summarise the central tendency, dispersion and shape of a dataset’s distribution, excluding NaN values (McKinney, 2017).

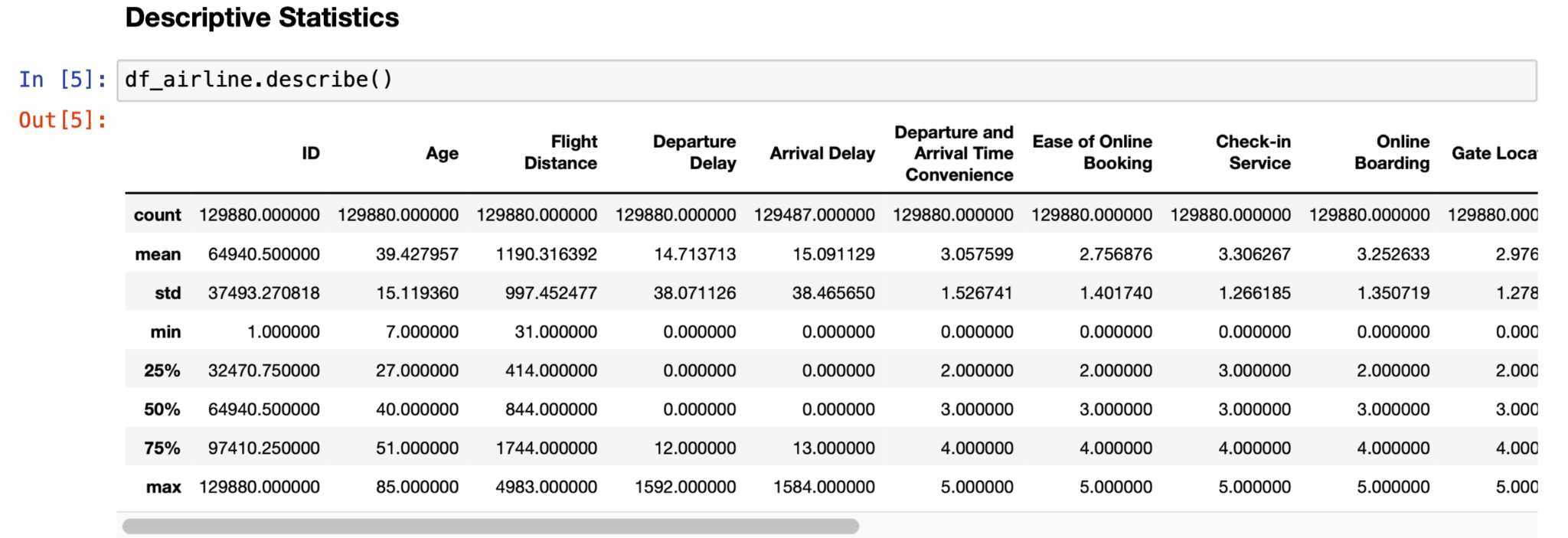


Figure 6. Number summary (describe).

The measures of dispersion evaluate how distributed the collected data are. They are standard deviation, variation and interquartile range (The pandas development team, 2020).

As it is known, a data set is very spread out when the standard deviation value is high. In this case, the standard deviation result is lower than the mean and thus is asymmetrical distribution. The plots above show the distribution of each variable in the dataset:

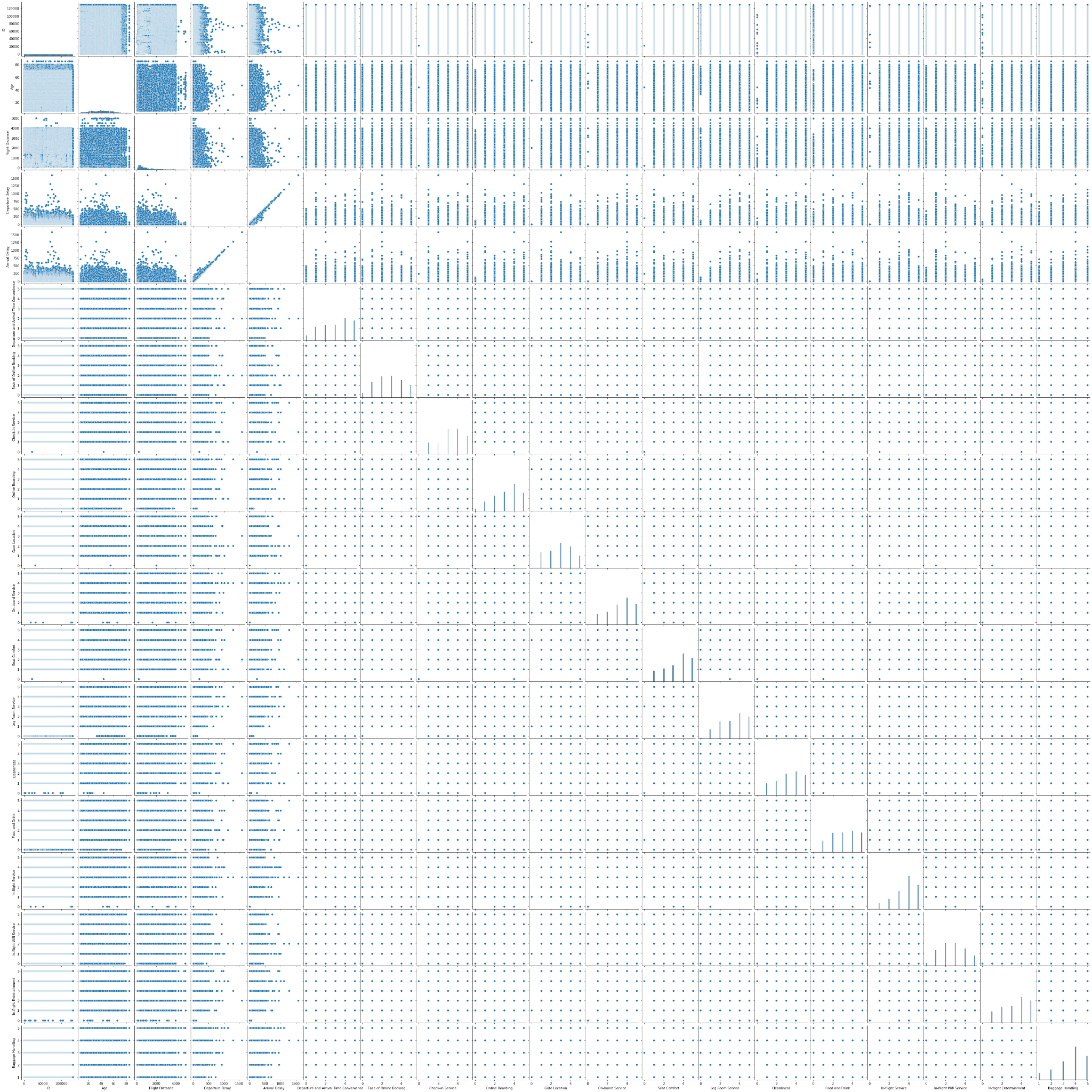


Figure 7. Pairplot of the dataset numerical variables



Figure 8. Histogram of the dataset numerical variables

Below, it can be seen the categorical summary statistics with their unique values and bar chart plot:

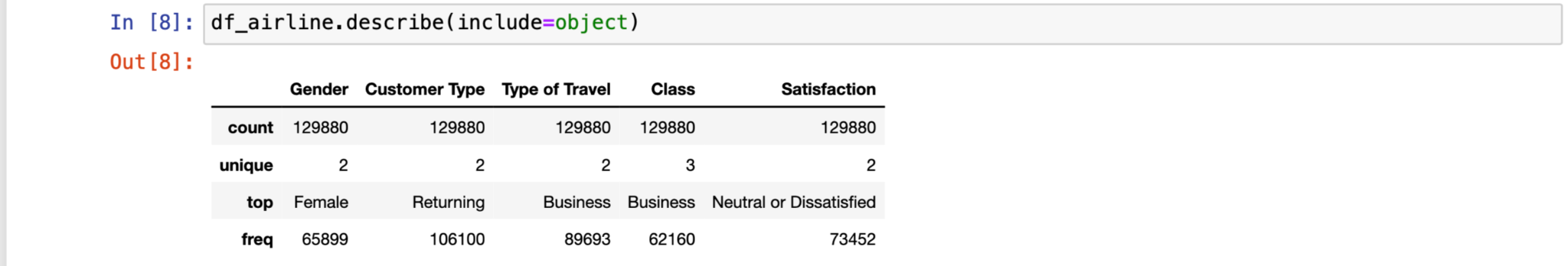


Figure 9. Categorical variables summary statistics

From the bar chart below, it can be seen that most of the customer type is not travelling for the first time. They mostly travel for business and are neutral or dissatisfied regarding their overall service experience.

Chart, bar chart

Description automatically generated

Figure 10. Bar chart of the dataset categorical variables

The function below shows the skewed distribution for the numerical variables, which is noted to likely have a negative skew.

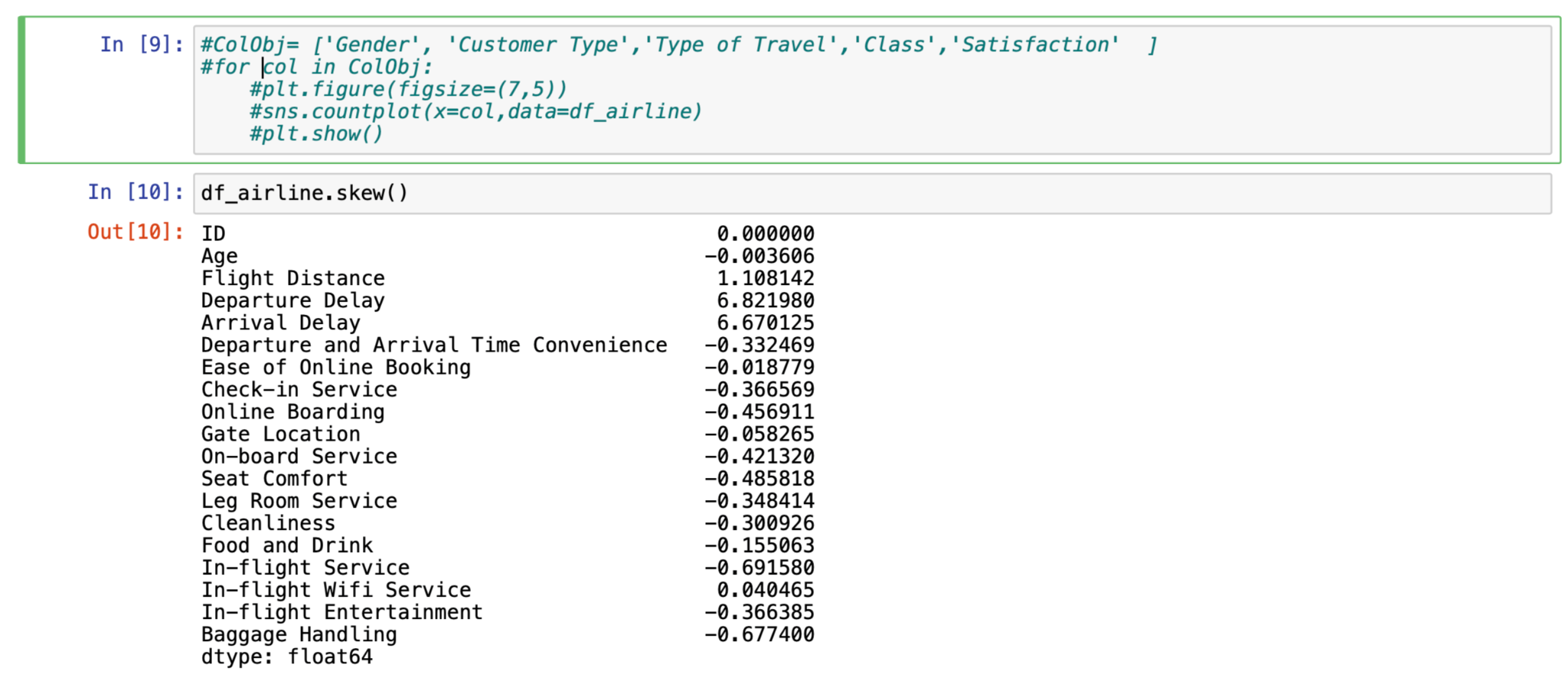


Figure 11. Skewness distribution

The plot below shows the distribution of the variables according to their level of satisfaction

classified by “Neutral” or “Dissatisfied: in blue colour and “Satisfied’: in green colour:

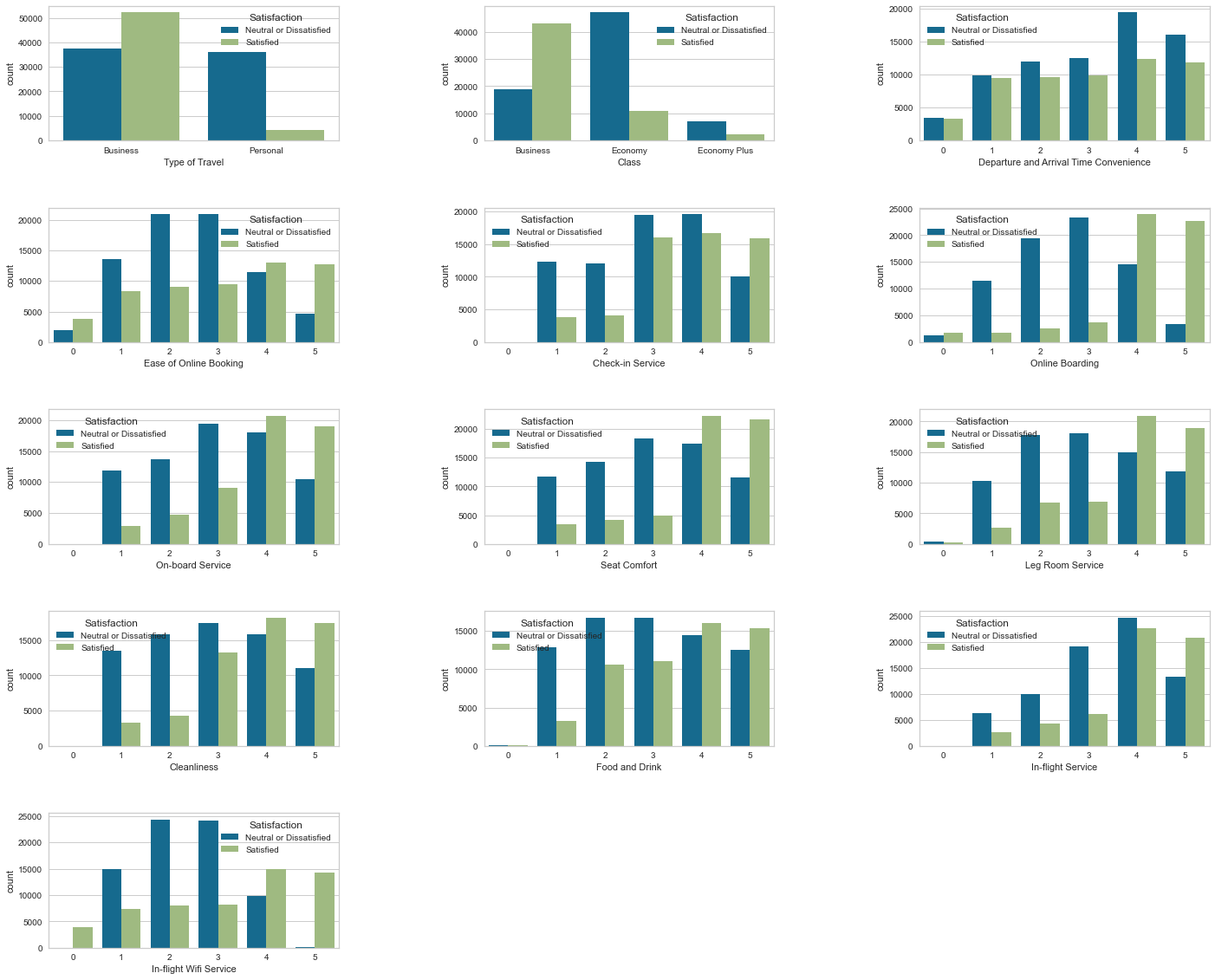


Figure 12. Satisfaction level bar chart

Here it can be seen that passengers travelling for Business propose are satisfied with the services provided and for Personal propose passengers are dissatisfied. Business class passengers are satisfied, and Economy class are not. Passengers, in general, are not satisfied with the arrival and departure times, the online booking system, check-in services and so on.

According to the boxplot below, it can be seen the presence of outliers in 3 variables. “Flight Distance” has short flights from 31 to long flights of 4983 nautical miles. The same happens with "Departure Delay" and "Arrival Delay": some flights are delayed by only a few minutes, while in some particular cases, it can be more than 24 hours, while some are not delayed at all.

Briefly, outliers are extreme values that are significantly from the overall pattern of values in a dataset (Ramalho, 2015). It is an observation that lies an abnormal distance from other values in a random sample from a population, and it will be discussed further about them.

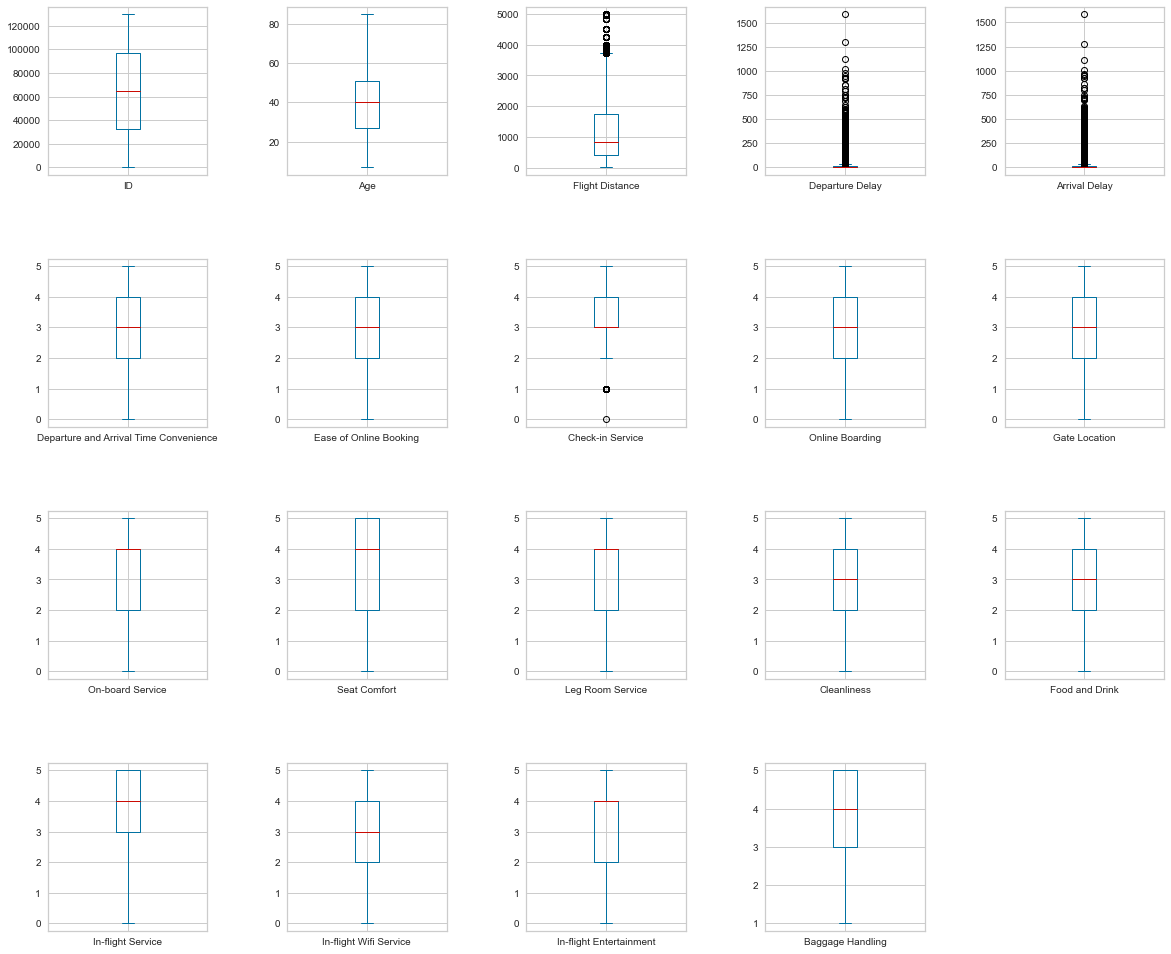


Figure 13. Outliers boxplot

Now, the correlation between the variables for this dataset will be analysed:

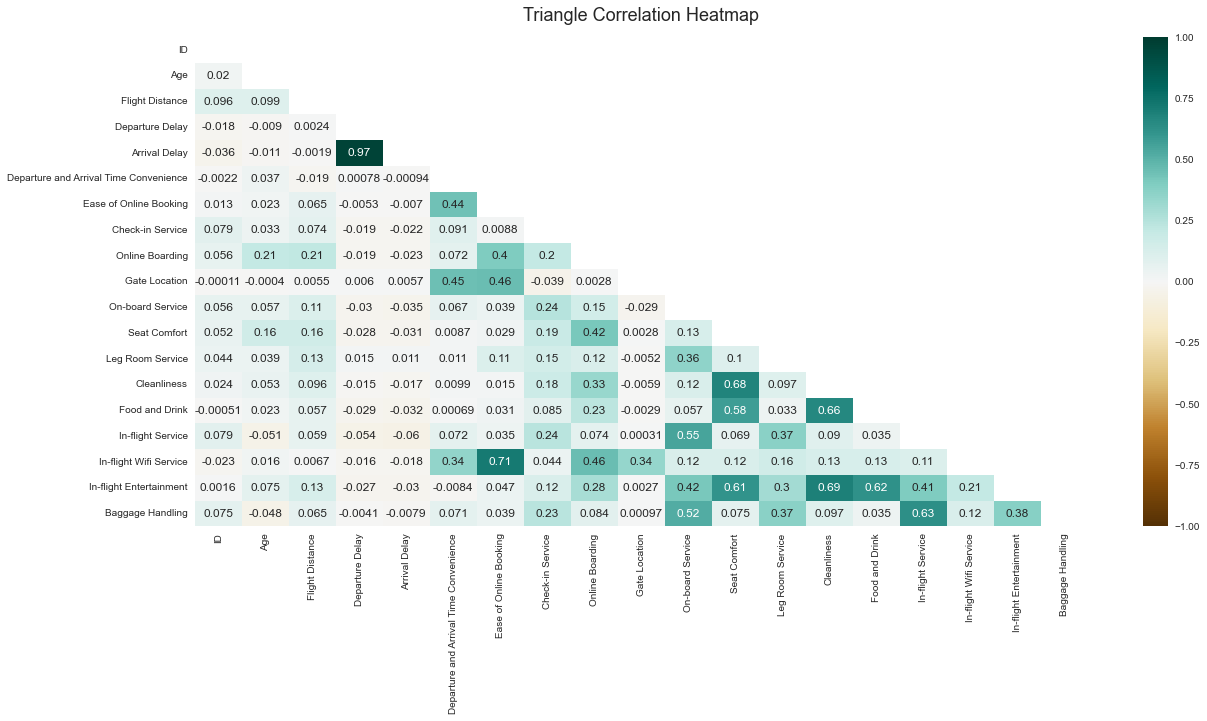


Figure 14. Heatmap

As seen in the heatmap above, there is a strong correlation between the “Arrival Delay” variable and the “Departure Delay “one. Between “In-flight-service” with the “Easy of Online Booking” and also with the “Cleanliness” and the “Seat Comfort”.

Now that all the data has been presented and analysed, it is time to prepare the data for further training and application of the Machine Learning models to find the principal factors of passenger dissatisfaction with short-haul flights.

## **Data Preparation**

Data preparation is an essential step that consists of cleaning, constructing, integrating and formatting the data set. And the first step is to identify the missing values further and analyse the best practice to deal with them (Little and Rubin, 2019).

After applying the function below, the number of 393 missing values in the “Arrival Delay” variable can be seen. It is insignificant compared to the number of data contained in our dataset. In that sense, the team decided to impute them with the median because, according to our EDA, it does not follow a normal distribution. Rather, it tends to have a positive skew (6.670125). We can also see outliers that could considerably affect our mean.

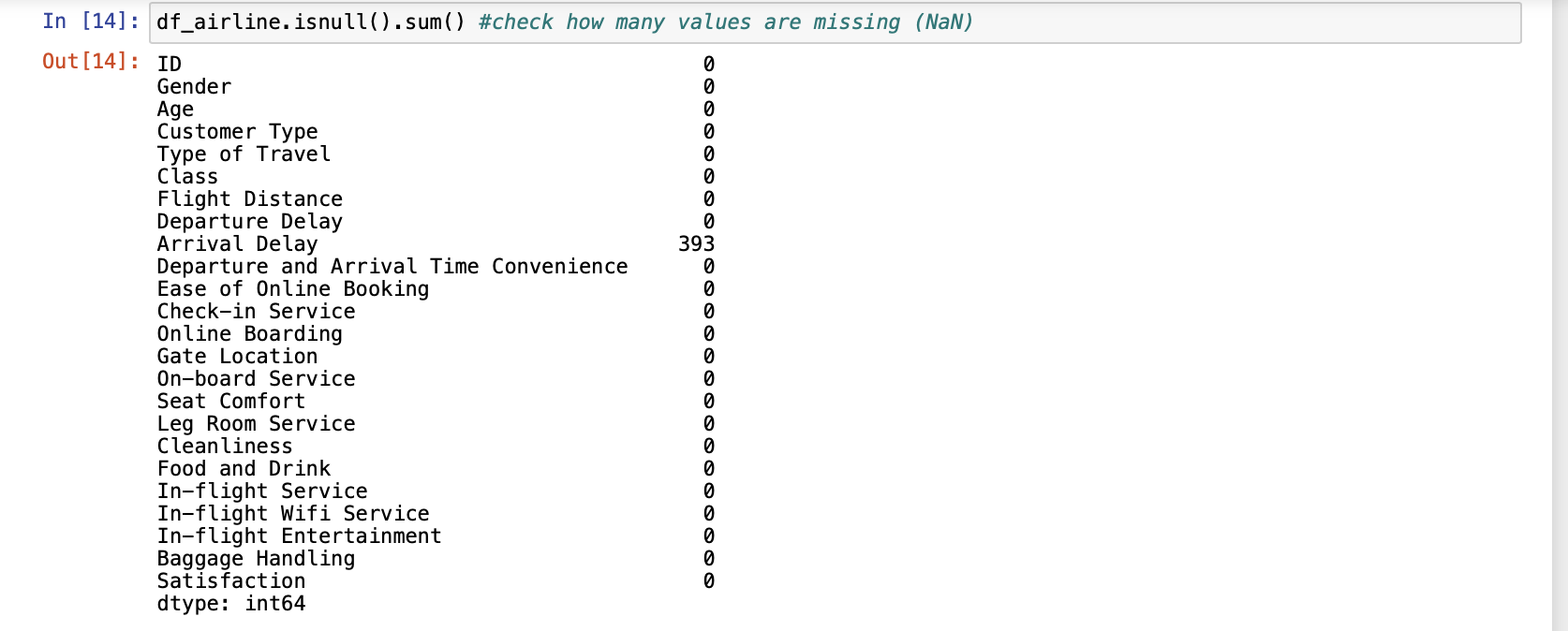


Figure 15. Finding missing values

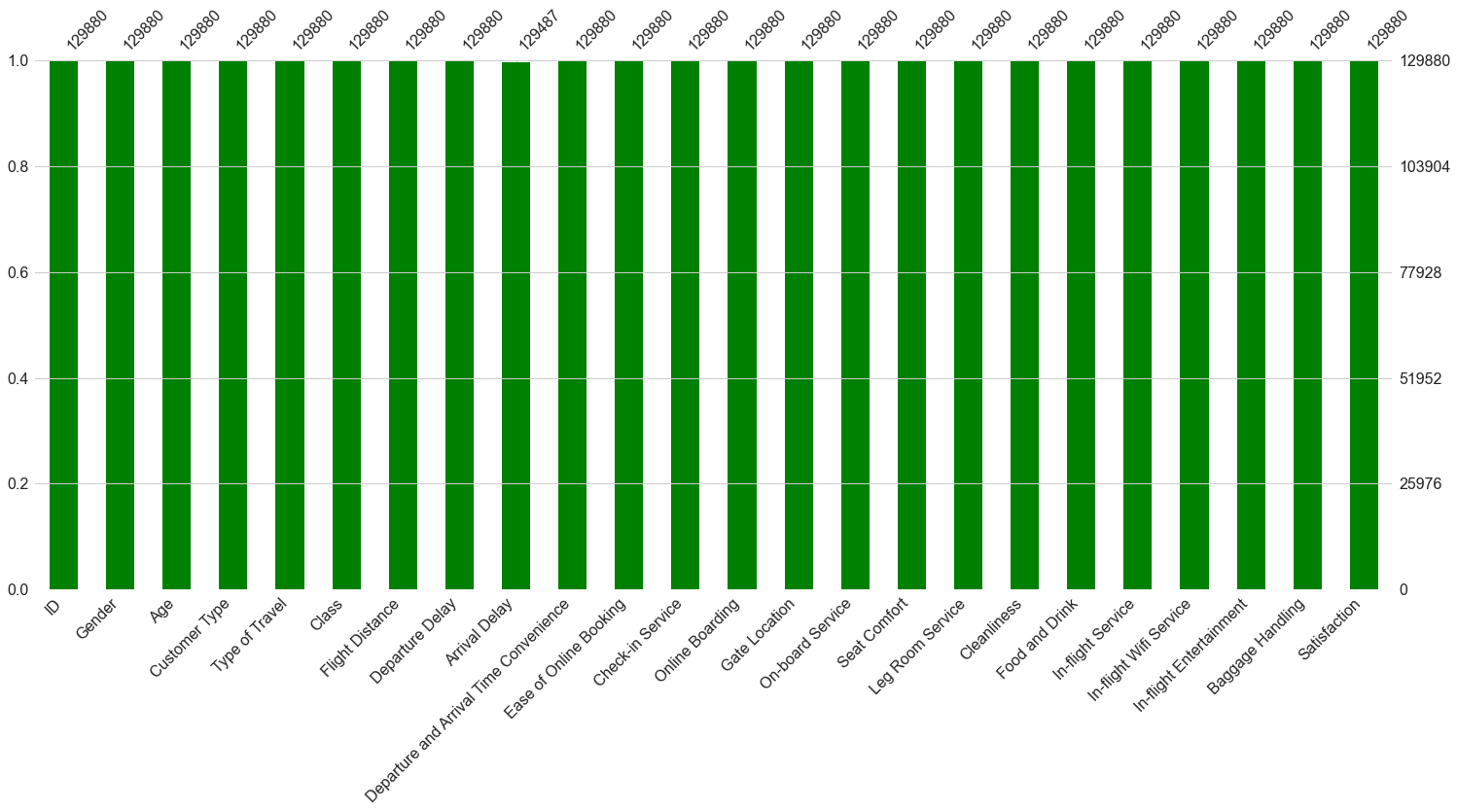


Figure 16. missing values graph.

The next step is Feature engineering. According to our scenario, we want to know which factors impact the dissatisfaction of short-distance flight passengers; that is why we need to convert our categorical values to numerical values to classify the flights according to their distance.

For this, we will rely on the aviation rules that classify them into short, medium and long distance. We will focus on short-haul flights. A short-distance trip is known to have a space of less than or equal to 800 nautical miles (InsureMyTrip,2021). After running the code below, our data set has a length of 46495 rows (before, it had 129880, as previously seen in the Data Understanding session).

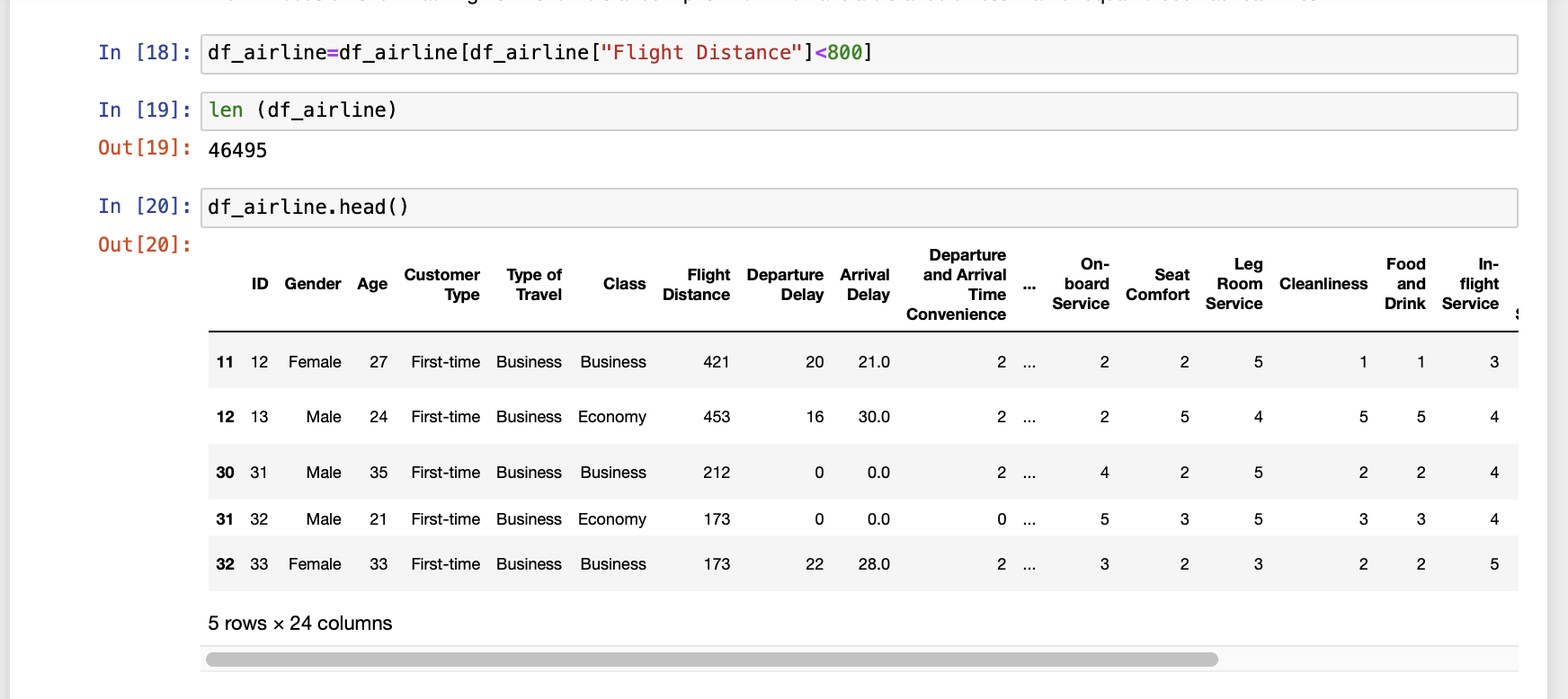


Figure 17. Short-distance flights

After that, we will drop the columns that have no relevance to this analysis: ‘ID’, “Gender”, and “Age”:

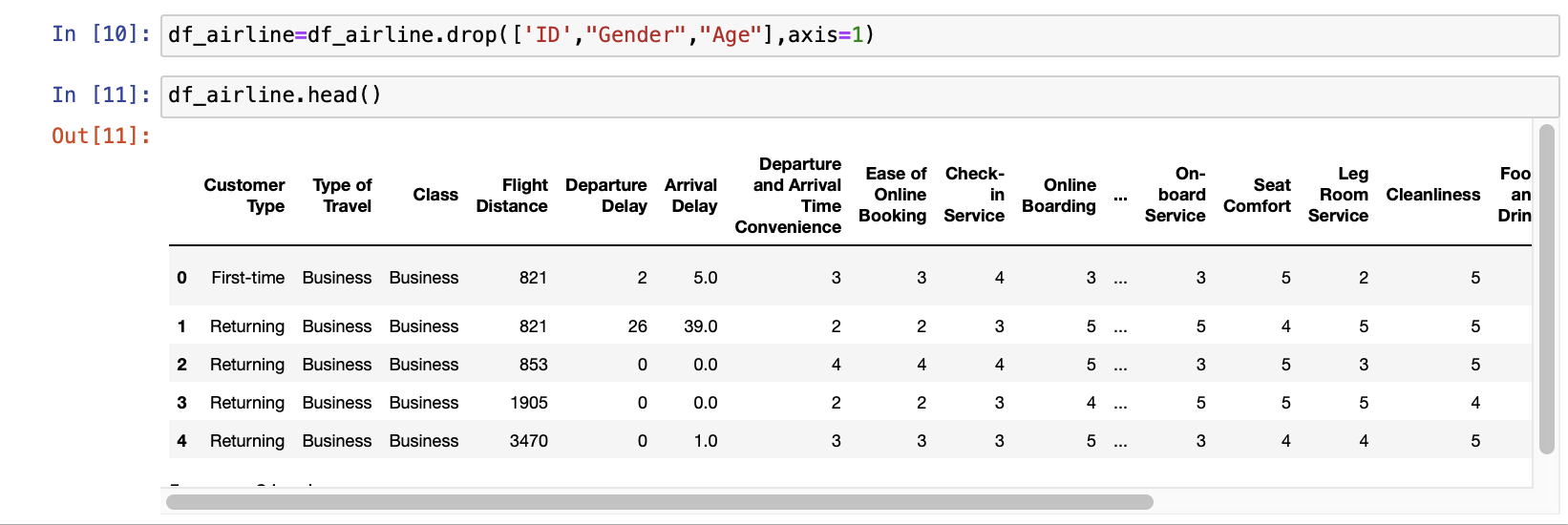


Figure 18. Dropping columns

Now is the time to encode our data set. In general, encoding is converting data from one form to another. That means if data contain a categorical variable, it just has to encode it to the numbers before fitting the data into the model (Brownlee, 2020).

* Class label: “**Satisfaction”:** Neutral or Dissatisfied :0, Satisfied: 1
* **The “Customer Type”** variable is as follows: First-time :0, Returning: 1
* **The “Type of Travel”** variable is as follows: Business: 0, Personal: 1
* **“Class”** variable as: Business: 0, Economy: 1, Economy Plus: 2

Now the data set is encoded. In other words, all the variables contain numerical numbers; from now, machine models can be applied.

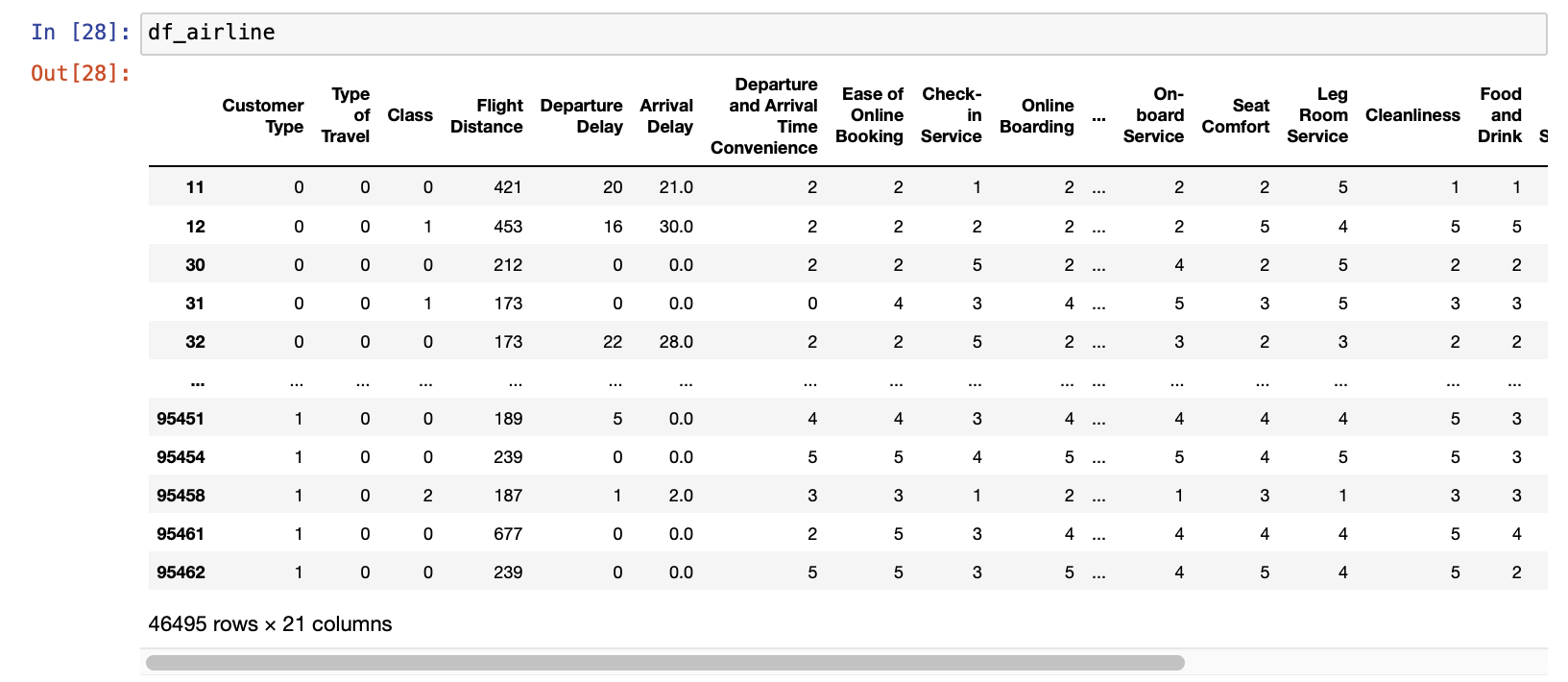


Figure 19. Encoded data set

But before, we will check if the data set has sparse data. According to Brownlee (2020), Sparse is a dataset with high zero values that can cause problems like over-fitting in the machine learning models and several other issues. However, our data set is not sparse, as seen below:

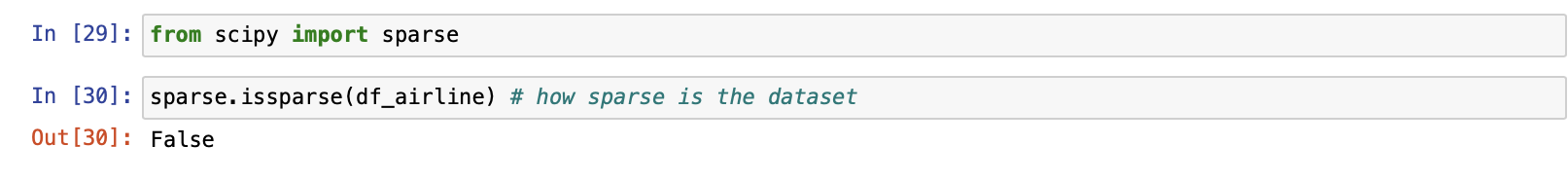


Figure 20. Sparse data check-up

Now is the time to split the dataset into independent and dependent (“Satisfaction”) variables in other to further train our models and find our answer.

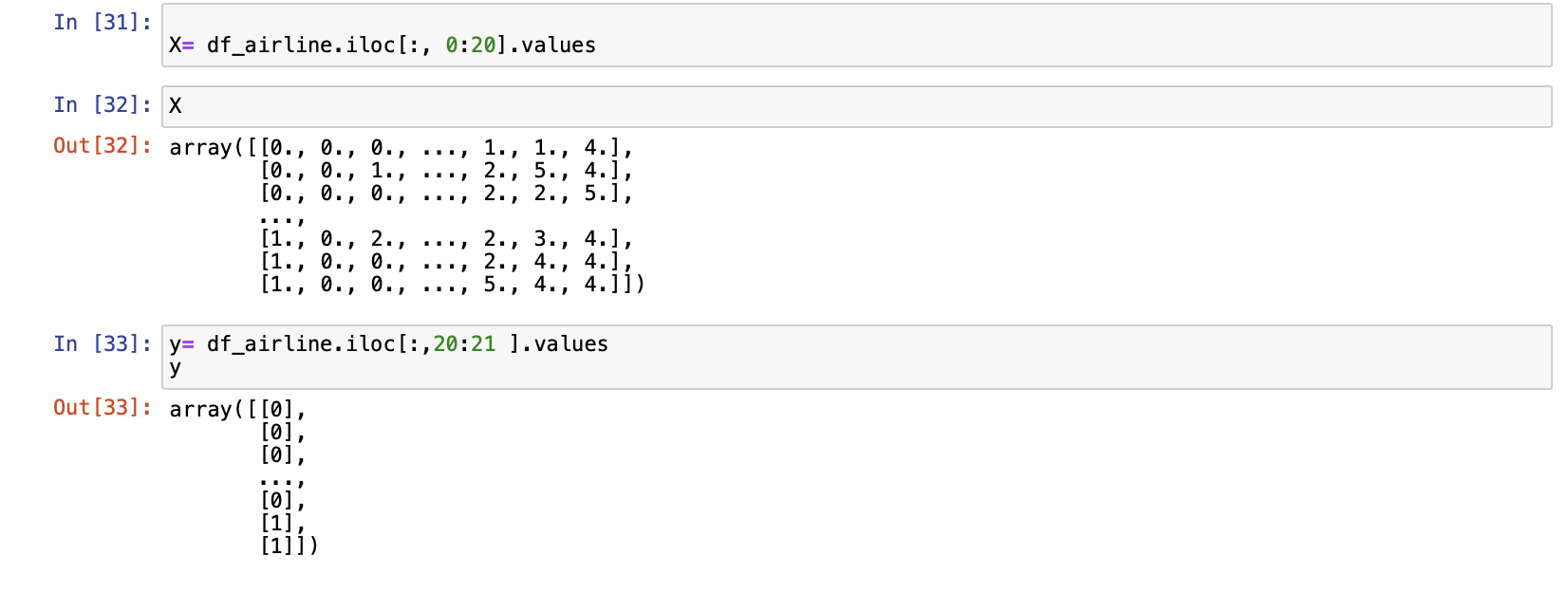


Figure 21. Separating independent from dependent variables

Finally, we Split the data set into different percentages for training and testing to check each of its accuracies: 20%, 15%, 10% and 5%:

**Split 1:** With 20%, we got an accuracy of: 0.6607162060436607

**Split 2:** With 15%, we got an accuracy of:0.6607162060436607

**Split 3:** With 10%, we got an accuracy of: 0.6606451612903226

**Split 4:** With 5%, we got an accuracy of: 0.6606451612903226

We will train and split our model with 20% accuracy based on the above results.

Chart, bar chart

Description automatically generated

Figure 22. Accuracy Scores with different splits

## **Modelling**

Different classification algorithms were tested for the modelling step. The training dataset performed well overall, showing an accuracy between 79% and 95%. Some challenges were encountered during the process, such as the division of the data set, as the system needed to behave more effectively in the face of long machine learning processes.

To have a reliable training process without failures or loss of quality, the DummyClassifier model was applied as a baseline. This baseline model was used to contextualise the results of the training models and improve understanding of the data. Although this model has low predictive power, it is considered adequate and may need to act as a guide to compare with other complex models in the ML project.

Raul Garreta et al. (2017) agree that testing complex models against simple models are important, and dummy estimators provide this. The usefulness of this model can be easily illustrated with a fraud example when 5% can be considered a warning. More complex models can lead to misleading accuracy without detecting the fraud at hand, and dummy models filter this kind of situation better.

Another challenge found was the application of the hyperparameters in just one code. Therefore, a more in-depth study was necessary so that there was a manual configuration of the algorithm so that the results were assertive and reached the objective of this parameter, which is the optimisation of the training models.

After analysing models’ performance through cross-validation techniques in the hyperparameter toward accuracy results, it was decided to use the following models: RandomForest, Logistic Regression, and KNeighborsClassifier. A brief description of each model and a confusion matrix with the performance are given below for a better understanding:

*Chart, bar chart

Description automatically generated*

Figure 23. Cross Validation Scores

**Random Forest:** Following the same supervised line of the model above, a set of trained random decision trees is built through a combination of models that increase the accuracy of the final result. According to Sarkar and Natarajan (2019), the random forest minimises overfitting because it tends to give more accurate results.

Chart, treemap chart

Description automatically generated

Figure 24.Randon Forest Confusion Matrix

**Logistic Regression:** According to Samir Madhavan (2015), the logistic function is very useful and can take any value from negative infinity to positive infinity and return a value between 0 and 1. Therefore, it can be interpreted as a probability.Considered as a primary method, it is also a fundamental resource, fast, simple, and easy to understand the results that use the sigmoid function, with values either 0 or 1, which is mainly used for binary classifications.

Chart, treemap chart

Description automatically generated

Figure 25.Logistic Regression Cofusion Matrix

**K Neighbor Classifier:** Also considered simple and one of the most used models for classification and regression problems. Its purpose is to use the nearest neighbours method, classifying them through the calculated distance. It is more correctly used when the data set has a low number of features because this model tends to overfit, causing the curse of dimensionality.

Chatterjee (2021) states that KNN is usually evaluated in the following aspects: ease of interpretation that can show a better visualization for stakeholders, time to calculate the output in a way that tends to get the results quickly, and power to predict accuracy once the data set is not impacted by outliers. Chatterjee (2021) also affirms that the results have a chance to vary depending on the chosen distance measurement method. The most common distance measure for this algorithm is the Euclidean distance.

Chart, treemap chart

Description automatically generated

Figure 26.KNN Confusion Matrix

**Evaluation**

Evaluate results

Review process

Determine next steps : Depending on the results of the assessment and the process review, the project team decides how to proceed. The team decides whether to finish this project and move on to deployment, initiate further iterations, or set up new data mining projects. This task includes analyses of remaining resources and budget, which may influence the decisions.

Considering that the main objective of this study is to understand the factors that most impact the dissatisfaction of passengers on short air flights, then, because of the resolution of previous training, it is decided that the model that performed better was Random Forest. This model does not tend to overfit, so the accuracy shows that it was well-trained. To confirm the validity of the model, Cross-Validation was applied to bring the result close to reality.  
  
According to Refaeilzadeh et al.(2009), Cross-Validation is a statistical method to evaluate and compare algorithms that are in the learning process dividing them into two parts, one part is used to train, and the other part is used to predict and validate the model.

Until here we have 1332 + 496 words

**The** graph below **shows** that the model performs well without **underfitting** or **overfitting after cross-validation.**

BAR PLOT COMPARING RANDOM FOREST BEFORE AND AFTER CROSS-VALIDATION

Regardless of the accuracy results being significant in the modelling algorithms, other factors influenced this decision-making in the application of the model to the evaluation stage, such as the K Neighbors Classifier (KNN) is an algorithm sensitive to the presence of outliers, and this could cause an overfitting of the model, so it was considered not to apply it in this step and choose Random Forest algorithm.

## 

## **Deployment**

Reviewing this project was more challenging than expected due to XXXX, with a machine-learning model being a novelty for the whole group. However, the results are very satisfactory regarding its complete development, and the group looks forward to implementing the following steps.

The analysis results show that the XXXX model selects a feature subset containing xxx variables. The classification prediction model’s random forest after the XXXX feature selection offers the best classification performance. Finally, combined with the four critical variables extracted by xxx and logistic regression, further discussion is carried out, and suggestions are given for airlines to improve passenger satisfaction.

Plan deployment

Task

This task takes the evaluation results and determines a strategy for deployment. If a general procedure has been identified to create the relevant model(s), this procedure is documented here for later deployment. Our practical experience tells us that it makes sense to consider the ways and means of deployment during the business understanding phase as deployment is crucial to the success of the project. This is where predictive analytics really helps to improve the operation side of your business.

Output

• Deployment plan – summarise the deployment strategy including the necessary steps and how to perform them.

page1image47850880

Plan monitoring and maintenance

Task

Monitoring and maintenance are important issues if the data mining result becomes part of the day-to-day business and its environment. The careful preparation of a maintenance strategy helps to avoid unnecessarily long periods of incorrect usage of data mining results. In order to monitor the deployment of the data mining result(s), the project needs a detailed monitoring process plan. This plan takes into account the specific type of deployment.

Output

• Monitoring and maintenance plan– summarise the monitoring and maintenance strategy, including the necessary steps and how to perform them.

page1image47844544

Produce final report

Task

At the end of the project, the project team writes up a final report. Depending on the deployment plan, this report may be only a summary of the project and its experiences (if they have not already been documented as an ongoing activity) or it may be a final and comprehensive presentation of the data mining result(s).

Outputs

* •  Final report – this is the final written report of the data mining engagement. It includes all of the previous deliverables, summarising and organising the results.
* •  Final presentation – there will also often be a meeting at the conclusion of the project at which the results are presented to the customer.

Review project

Task

Assess what went right and what went wrong, what was done well and what needs to be improved.

Output

• Experience documentation – summarise important experience gained during the project. For example, pitfalls, misleading approaches, or hints for selecting the best suited data mining techniques in similar situations could be part of this documentation. In ideal projects, experience documentation also covers any reports that have been written by individual project members during previous phases of the project.

## **Extra Contents**

## Roles and responsibilities

This research was formed by a multidisciplinary team to meet the established goals and perform well even within diversities founded during its development. The roles were described as follows:



Figure 27.Roles and Responsibilities

## Team Project management

The team decided to work with Trello for the project management as it is a free, simple, and easy-to-use collaboration tool that enables us to organise and track the project on the board; it was a great tool to help to accomplish the task on time required, as seen in our board below:

Graphical user interface, application

Description automatically generated

Figure 28. Project Management: Trello Board

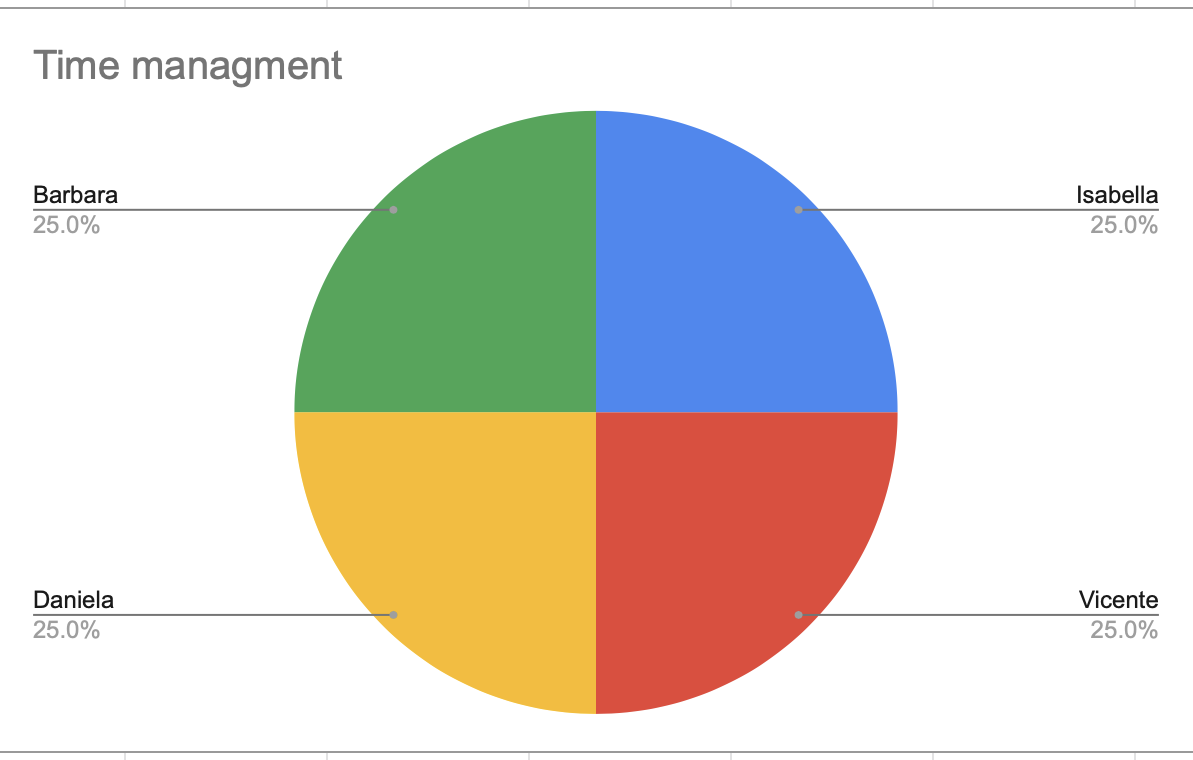
Finally, the effort of the team is represented in the pie chart below

Figure 29.Team's effort

## Challenges faced

Daniela Daia:

As a group, I felt that misinterpreting what was being discussed led to a lack of understanding and poor communication with some individuals, causing conflicts.

As a group, people must understand that working in a team may require negotiation and compromise.

As an individual, I faced challenges with the google collab tool for the first time using it, also, in building the report, I needed to install Office as google drive was no longer supporting 2 people working simultaneously. Overall, in the end, we learnt about each other's difficulties and improved our communication leading to success in the project delivery.

## **Reference List**

Bart, C.K. (2000) The Relationship between Mission and Innovativeness in the Airline Industry: An Exploratory Investigation. London: Int. J. Technology Management.

Brownlee, J. (2020) Data Preparation for Machine Learning: Data Cleaning, Feature Selection and Data Transforms in Python. New York: Jason Brownlee.

Chatterjee, I. (2021) Machine Learning and Its Application: a Quick Guide for Beginners. London: Bentham Science Publishers.

InsureMyTrip (2021), Travel Tips for Long Haul Flights. InsureMyTrip Available at: <https://www.insuremytrip.com/travel-advice/travel-tips/long-haul-flights/#:~:text=While%20no%20international%20standard%20definition,any%20flight%20length%20in%20between> [Accessed 20 November 2022].

Google Colaboratory (2022) Tutorials. Available at: <https://research.google.com/colaboratory/faq.html> [Accessed 26 November 2022].

Han, J., Kamber, M. and Pei, J. (2011) Data Mining Concepts and Techniques. 3rd Edition. Burlington: Morgan Kaufmann Publishers.

IMB, (2021) IBM Documentation. Available at: <https://www.ibm.com/docs/en/spss-modeler/saas?topic=dm-crisp-help-overview> [Accessed 20 November 2022].

Little, R., Rubin, D. (2019) Statistical analysis with missing data, third edition. United States of America: Wiley.

McKinney, W (2017) Python for Data Analysis. 2nd ed. Beijing: O’Reilly Inc.

Madhavan, Samir (2015) Mastering Python for Data Science: Explore the World of Data Science through Python and Learn How to Make Sense of Data. Birmingham: Packt Publishing Limited.

Nair, A. (2022) Baseline Models: Your Guide for Model Building. Medium. Available at: towardsdatascience.com/baseline-models-your-guide-for-model-building-1ec3aa244b8d#. [Accessed 20 November 2022].

Navlani, A.(2019) Python Logistic Regression Tutorial with Sklearn & Scikit. Datacamp.com. Available at: www.datacamp.com/tutorial/understanding-logistic-regression-python.[ Accessed

21 November 2022].

Navlani, A. (2018) KNN Classification Tutorial Using Sklearn Python. Datacamp.com.

Available at: [www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn](http://www.datacamp.com/tutorial/k-nearest-neighbor-classification-scikit-learn)

[Accessed 22 Nov. 2022].

Ramalho, L. (2015) Fluent Python: clear, concise and practical programming. 1st Edition. United States of America: O’Reilly Media Inc.

Raúl Garreta, Moncecchi, G., Hauck, T. and Hackeling, G. (2017) Scikit-learn: Machine Learning simplified. Birmingham, UK: Packt Publishing.

Refaeilzadeh, P., Tang, L. and Liu, H. (2009) Cross-Validation. Available at: <http://leitang.net/papers/ency-cross-validation.pdf> [ Accessed 26 November 2022].

Sarkar, D. and Vijayalakshmi, N. (2019) Ensemble Machine Learning Cookbook Over 35 Practical Recipes to Explore Ensemble Machine Learning Techniques Using Python. Birmingham Packt Publishing.

The panda's development team (2022) Python Data Analysis Library. [Pandas Documentation. Available at: <https://pandas.pydata.org/docs/> [Accessed 22 November 2022].

Wertz, C. J. (1993) The Data Dictionary Concepts and Uses. 2nd Edition. New York: QED Information Sciences, Inc.

Williams, P. & Naumann, E. (2011) Customer satisfaction and business performance: A firm-level analysis. J. Serv. Market. Available at: <https://www.researchgate.net/publication/235317930_Customer_satisfaction_and_business_performance_A_firm-level_analysis> [Accessed 20 November 2022].