### BPP Business School Coursework Cover Sheet

Please use this document as the cover sheet of for the 1st page of your assessment.

Please complete the below table – the grey columns

|  |  |
| --- | --- |
| Module Name | **Applied Modelling and Visualization** |
| Student Reference Number  (SRN) | **BP0274581** |
| Assessment Title | **MAV - Skywards International Airlines Report – CW3 [S]** |

Please complete the yellow sections in the below declaration:

**BP0274581**

Declaration of Original Work:

I hereby declare that I have read and understood BPP’s regulations on plagiarism and that this is my original work, researched, undertaken, completed and submitted in accordance with the requirements of BPP School of Business and Technology.

The word count, excluding contents table, bibliography and appendices, is **2500** words.

Student Reference Number:

Date: **12-03-2024**

By submitting this course work you agree to all rules and regulations of BPP regarding assessments and awards for programs.

Please note that by submitting this assessment you are declaring that you are fit to sit this assessment.

BPP University reserves the right to use all submitted work for educational purposes and may request that work be published for a wider audience.

**Skywards International Airlines Consultancy Report**

**Table of Contents**

1. Introduction 5
   1. Background 5
   2. Purpose Statement 5
   3. Scope 5
   4. Project Plan – CRISP-DM 6
      1. Business Understanding 6
         1. Identify the Business Objectives 6
         2. Define the Scope 6
            1. Non-functional Requirements 6
            2. Functional Requirements 6
         3. Establish Success Criteria 7
      2. Data Understanding 7
         1. Gather Data 7
         2. Explore Data 7
      3. Data Preparation 7
         1. Clean Data 7
         2. Preprocess Data 7
         3. Feature Engineering 7
      4. Modeling 8
         1. Select Modeling Techniques 8
         2. Train Models 8
         3. Evaluate Models 8
      5. Evaluation 8
      6. Deployment 8
         1. Deploy Models 8
         2. Monitor Performance 8
         3. Communicate Findings 8
2. Key Factors that Impact on Passenger ‘satisfaction’ 9
   1. Metadata 9
      1. Customer Information 9
      2. Flight Information 9
      3. Operational Data 9
      4. Geographic Fields 10
      5. Descriptive Metadata 10
   2. Features and Label 10
      1. Features 10
      2. Label 10
3. Tasks 10
   1. ETL – Extract Transform and Load 10
      1. Extract 10
      2. Transform 13
         1. Handling Missing Values 13
         2. Duplicate Rows 14
         3. Remove Outliers 16
         4. Conversion of Non-Numeric Values 19
      3. Load 20
   2. Exploratory Data Analysis 21
      1. Visualization 22
      2. Selection of Analytical Models 27
   3. Predictive Models 27
      1. Decision Tree 27
      2. Random Forest 29
      3. Evaluate the Accuracy of Models 31
         1. Strengths and Weaknesses 31
4. Recommendations 32
   1. Critical Evaluation of Models 32
      1. Loss Function 32
      2. Accuracy Metrics 34
      3. Comparison Table 35
         1. Root Square Mean Error/Mean Square Error 35
      4. Recommended Model 37
   2. Findings 38
      1. Correlational Matrix 38
      2. Heat map 40
      3. Confusion Matrix 42
5. Next Steps 44
6. References 45
7. Appendix 47
8. **Introduction**
   1. **Background**

The Skywards International Airlines Corporation holds a large number of data about passengers as a well-known airline company.

* 1. **Purpose Statement**

This Data Analytics Consultancy Report aims to analyze the data provided by Skywards Airlines Corporation in an Excel form. Specifically, the huge data about satisfied and unsatisfied passengers will be extracted, transformed, and loaded, and then Exploratory Data Analysis on this data will be performed. Moreover, visualization and data modelling techniques will be used to analyze the current situation of Skywards passengers’ data leading to future predictions with the help of analytical models. Lastly, a suitable predictive model will be chosen with the help of different outcomes and future steps will be recommended based on analysis.

* 1. **Scope**

The goal is to design the programming solution to analyse the provided data as follows:

* Uploading the “SKYWARDS\_DATA\_CW3.csv” file in the Google Collab notebook
* Extraction of data
* Cleaning the data
* Exploratory Data Analysis
* Visualization
* Analyzing trends in data
* Implement two Analytical Models
* Critical evaluation models
* Recommending one predictive model
* Graphical illustration of the outcomes finalized from the recommended model
* Future steps
  1. **Project Plan - CRISP-DM**

The methodology CRISP-DM known as “Cross-Industry Standard Process for Data Mining” has been chosen to highlight the project plan of Skywards Airlines data as follows (Nodeh et al., 2020):

* + 1. **Business Understanding**
       1. **Identify the Business Objectives**

This includes understanding Skywards Airlines' objectives to implement data mining techniques to achieve a high number of satisfied customers.

* + - 1. **Define the Scope**

The detailed scope has already been discussed in section 1.3. Moreover, this stage allows the collection of two types of requirements (White, 2010):

* + - * 1. **Non-functional Requirements**

The selected platform for this project is Google Collab because (Pessoa et al., 2018):

* Allows a fully configured environment
* Free of charge
* Works in the same way on different hardware systems
* Easy to detect errors
* Independent of hardware resources

Secondly, Python language is being used because (Nitnaware, 2019):

* Easily readable
* User-friendly language
* Open source
* Compatible with many platforms
* Supports many libraries
  + - * 1. **Functional Requirements**

To design a code in Python that can perform all the steps mentioned in the 1.3 section.

* + - 1. **Establish Success Criteria**

The success of the project will be measured based on future predictions about trends so that the company may make decisions to improve performance.

* + 1. **Data Understanding**
       1. **Gather Data**

The customer data shared by the airline company in CSV format will be utilized in this report for further analysis.

* + - 1. **Explore Data**

The structure and content of the data will be analyzed to:

* Understand the characteristics
* Identify data quality issues
* Gain insights

Further discussed in the ETL stage.

* + 1. **Data Preparation**
       1. **Clean Data**

The missing values, outliers and duplicate rows will be handled to ensure data quality and reliability (Fernandes, 2023).

* + - 1. **Preprocess Data**

The categorical variables will be encoded with scale numerical features.

* + - 1. **Feature Engineering**

This stage involves deriving insights from existing features of data to enhance the predictive power of the models (He et al., 2023).

* + 1. **Modeling**
       1. **Select Modeling Techniques**

The appropriate algorithms for the analysis will be chosen for this report based on the analysis.

* + - 1. **Train Models**

The historical data will be used to train predictive models that can identify patterns and make predictions about future customer behaviour.

* + - 1. **Evaluate Models**

The performance of the models will be accessed to meet the business needs by analysing:

* Accuracy
* Precision
* Recall
* ROC-AUC
  + 1. **Evaluation**
       - Compare Models
       - Validate/interpret results
    2. **Deployment**
       1. **Deploy Models**

This step involves integrating the selected models into Skywards Airlines' operational systems to make predictions and generate insights in real time.

* + - 1. **Monitor Performance**

The performance of deployed models will be monitored. Besides, the models will be updated as needed to maintain their accuracy and relevance.

* + - 1. **Communicate Findings**

The deployment stage also includes sharing the results/insights with stakeholders across the Skywards Airlines organization to inform decision-making and drive business strategy.

1. **Key Factors that Impact on Passenger ‘satisfaction’**
   1. **Metadata**

The SKYWARDS\_DATA\_CW3.csv file represents a plethora of columns about passengers and flight data as follows:

* + 1. **Customer Information**
* **Id:** IDs for the customers in numeric form
* **Gender:** Male or Female characters
* **Satisfied:** ‘Y’: satisfied, ‘N’: unsatisfied.
* **Age:** A numeric field
* **Age Band:** The alphanumeric value of age, such as 25 to 34.
  + 1. **Flight Information**
* **Class:** Text field (i.e. Eco, Eco Plus or Business).
* **Flight Distance/ Flight Distance Rounded:** The distance in numeric form
  + 1. **Operational Data**

The provided criteria are rated on a scale from 0 to 5 (0 indicates poor and 5 indicates excellent):

* Onboard wifi performance
* Departure/Arrival time flexibility
* Seamless online booking
* Gate accessibility
* Grade of dinning-options
* Convenience of Online check-in
* Seat coziness
* Inflight amusement
* Leg space convenience
* Effectiveness of baggage handling
* Check-in efficiency
* Standard of onboard service
* Hygienic conditions
* Delayed minutes of departure
* Delayed minutes of arrival
  + 1. **Geographic Fields**
* **Destination:** Names of countries where flights go.
* **Continent:** Holds text values (i.e. Africa, Asia, etc.).
  + 1. **Descriptive Metadata**

**Type of Travel:** Text values such as Business Travel and Personal Travel.

* 1. **Features and Label**
     1. **Features**

The following features influence the satisfaction level of customers:

* **Passenger attributes:** Id, Age, gender, Age Band.
* **Flight attributes:** Operational data discussed in 2.1.3.
* **Experience attributes:** Food and drink quality, inflight entertainment, Seat comfort, Cleanliness, Inflight wifi service, Ease of Online booking, Online boarding, Leg room service, Baggage handling, Checkin service, Inflight service.
  + 1. **Label**

The label is the targeted variable to predict the satisfaction factor in the case of Skywards data:

* 1: Satisfied
* 0: Dissatisfied

1. **Tasks**
   1. **ETL – Extract Transform and Load**
      1. **Extract**

Firstly, the installation of tabula and pygwalker is done on Google Collab notebook by using the pip command as follows:

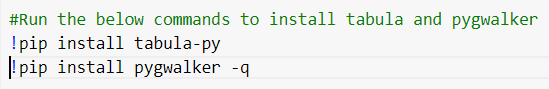


Figure 1.

Afterwards, the important libraries have been imported into the notebook. Their purposes are as follows (Fareez et al., 2020):

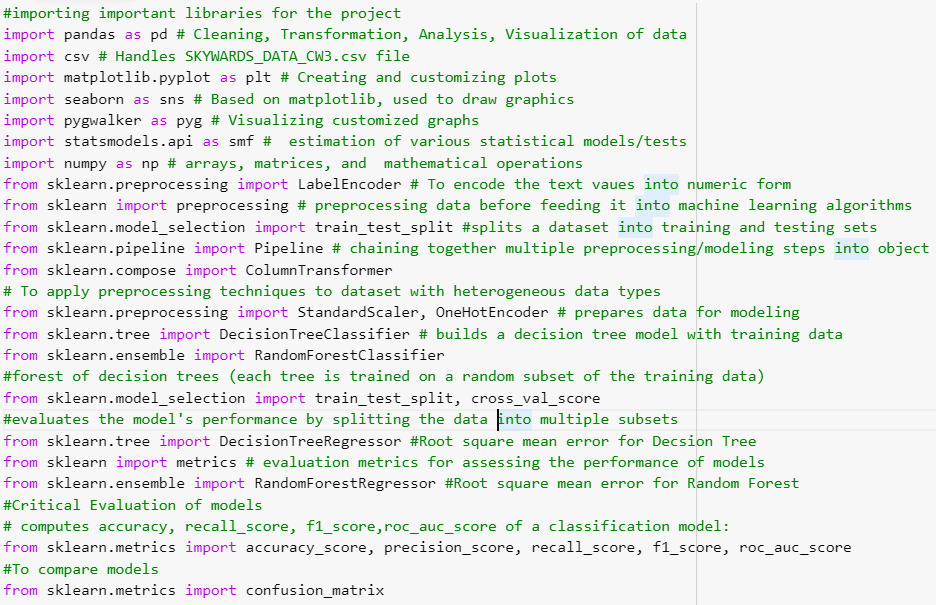


Figure 2.

The commands to import the Excel file and load it in data frame are as follows:

**Input:**

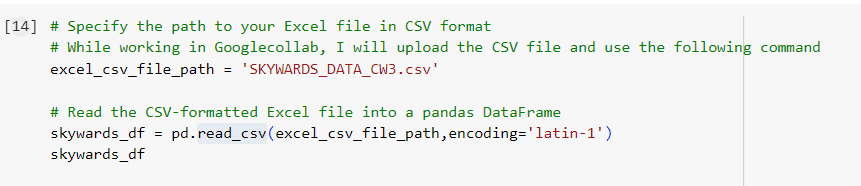


Figure 3.

**Output:**

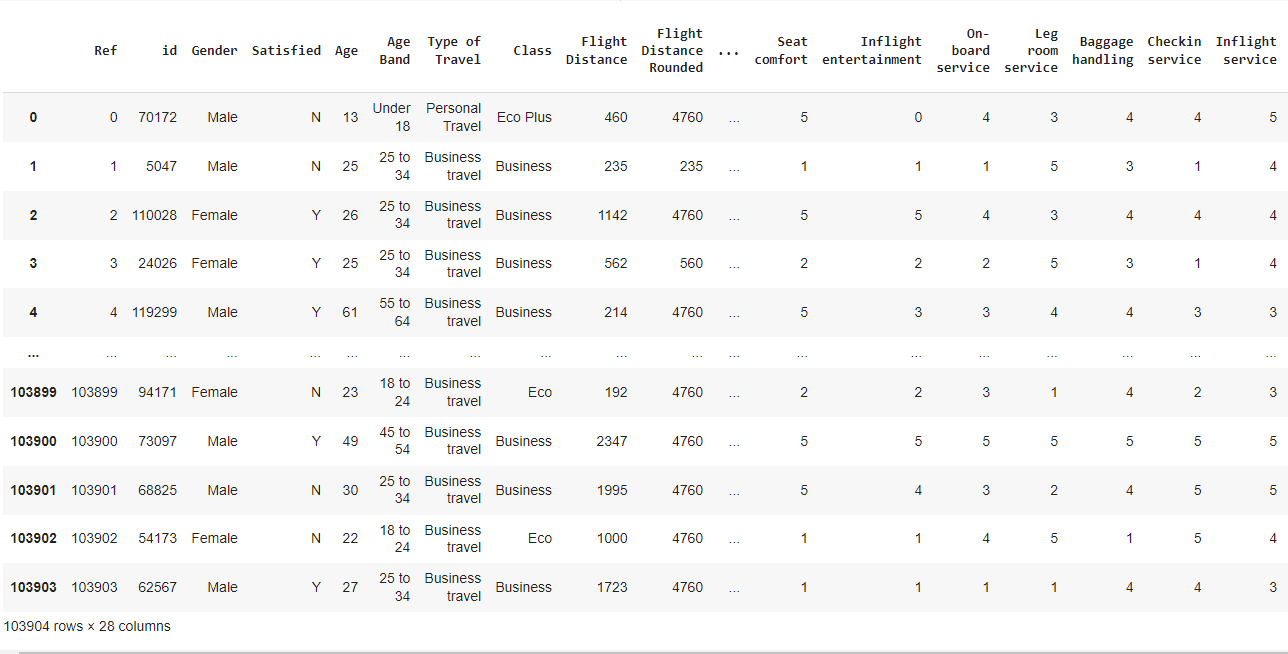


Figure 4.

**Input:**

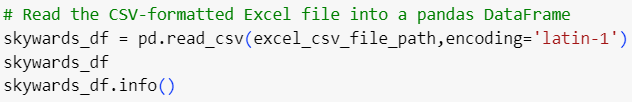


Figure 5.

**Output:**

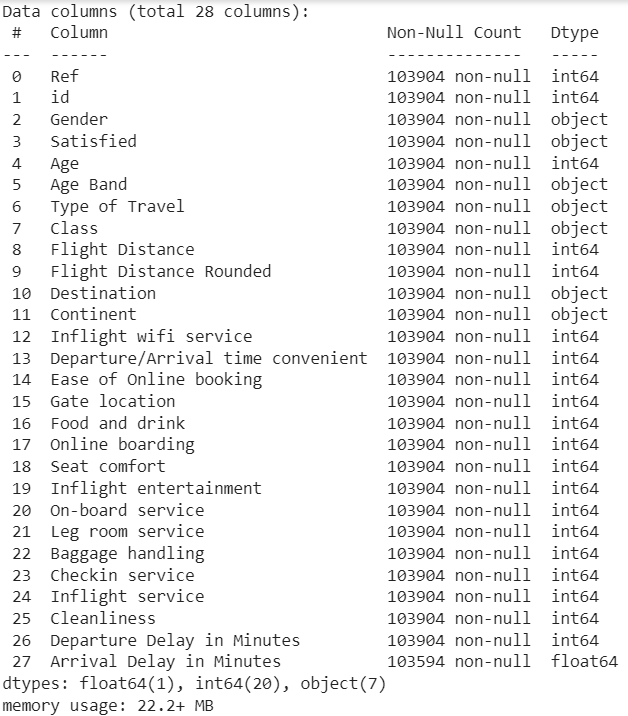


Figure 6.

* + 1. **Transform**
       1. **Handling Missing Values**

The blank fields in the Skywards Airlines customers’ data have been verified by using the isna() function as follows:

**Input:**

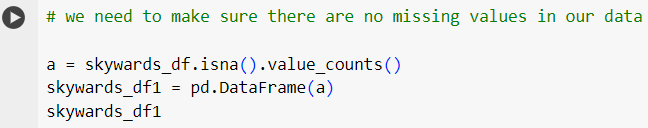


Figure 7.

This shows that there are null values in the last column “Arrival Delay in Minutes”.

**Output:**



Figure 8.

This has been fixed by using the median function as follows:

**Input:**

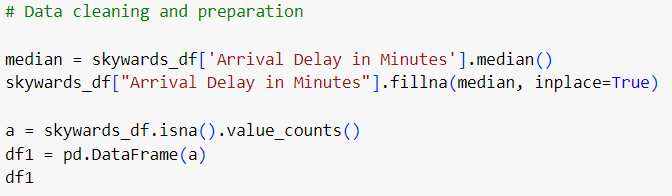


Figure 9.

**Output:**

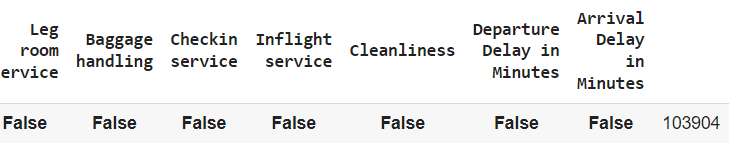


Figure 10.

The output reveals that missing values have been handled successfully.

* + - 1. **Duplicate Rows**

The duplicated() function has been used to verify if there are any duplicated rows in the data frame.

**Input:**

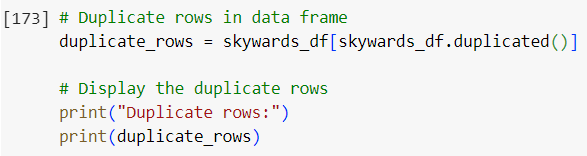


Figure 11.

**Output:**

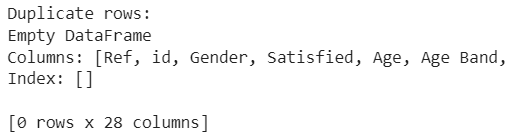


Figure 12.

There were no duplicate rows. However, multiple techniques have been mentioned in the code to remove the duplicate data in future.

**Input:**

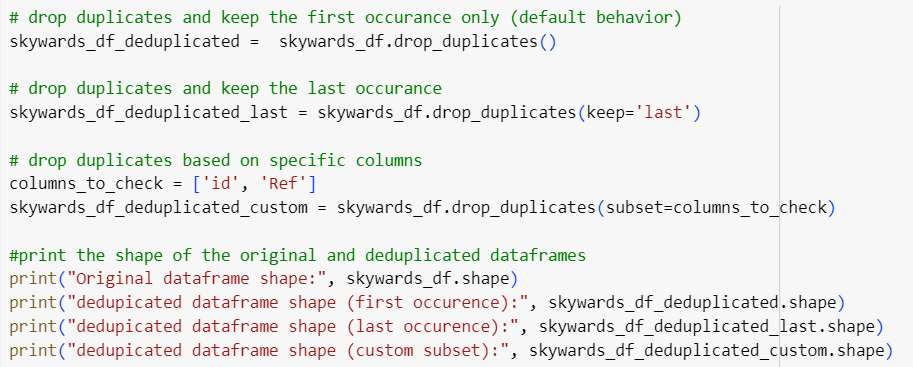


Figure 13.

**Output:**

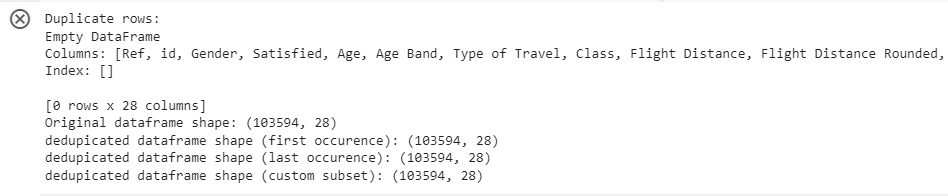


Figure 14.

* + - 1. **Remove Outliers**

Firstly, the data frame has been visualized using boxplot visualization for “Flight Distance” and “Departure Delay in Minutes” parameters as follows:

**Input:**

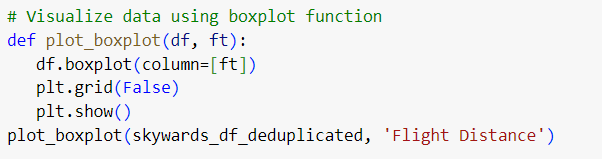


Figure 15.

**Output:**

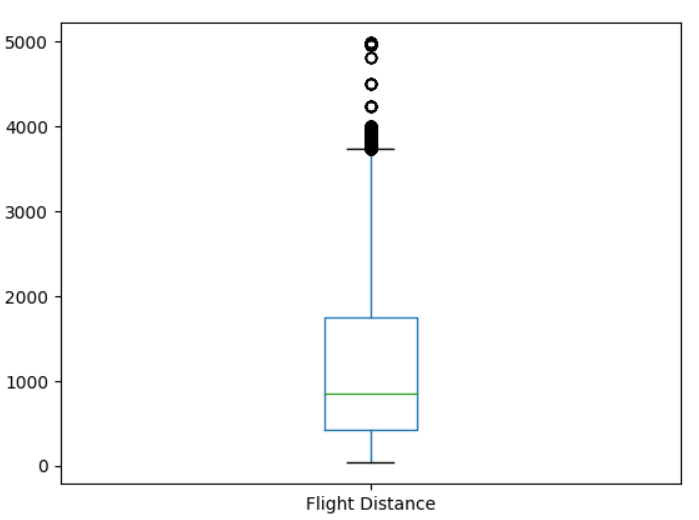


Figure 16.

**Input:**



Figure 17.

**Output:**

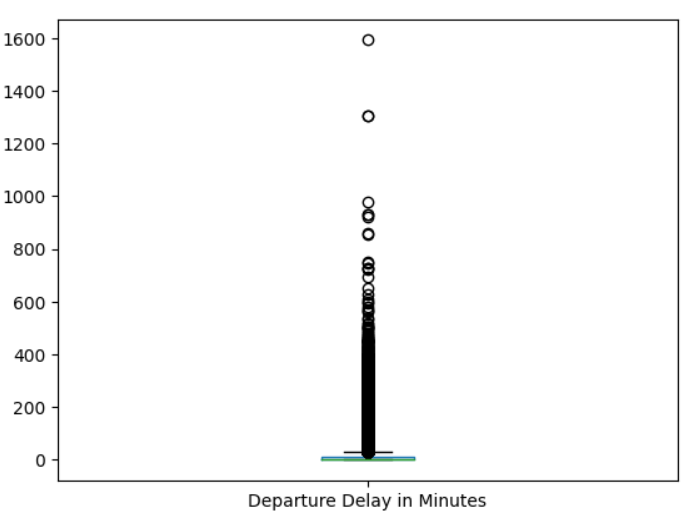


Figure 18.

The following function removes the outliers:

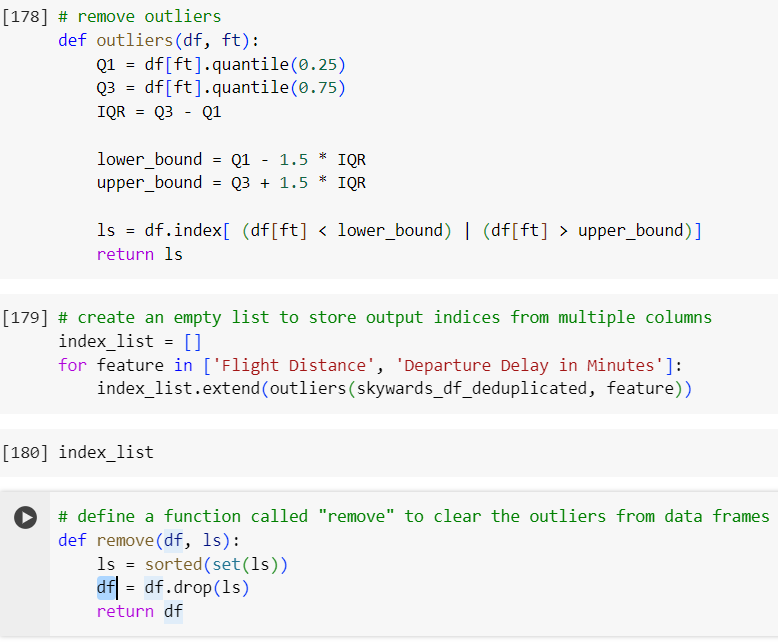


Figure 19.

The shape of the data frame has changed after removing outliers as follows:

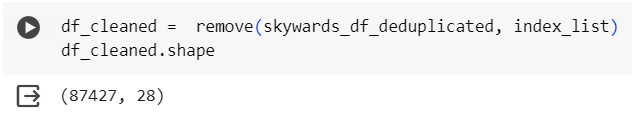


Figure 20.

Besides, the boxplot of departure delay time and flight distance reveals that the outliers have been removed as follows:

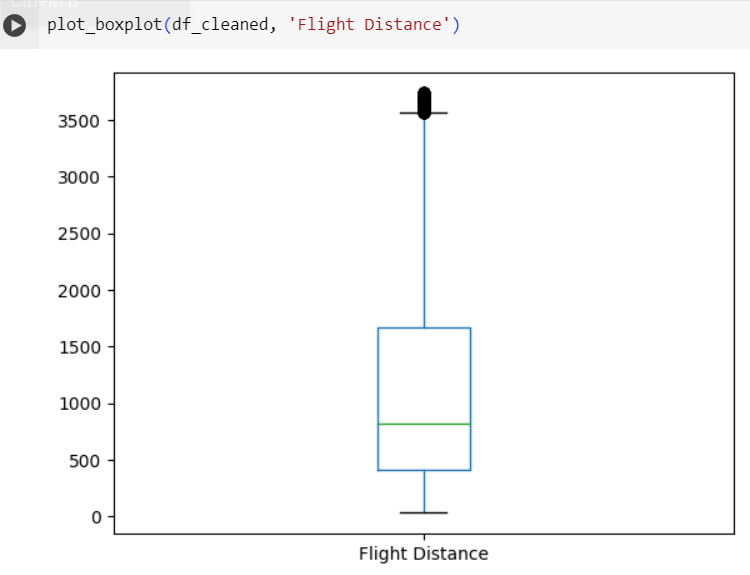


Figure 21.

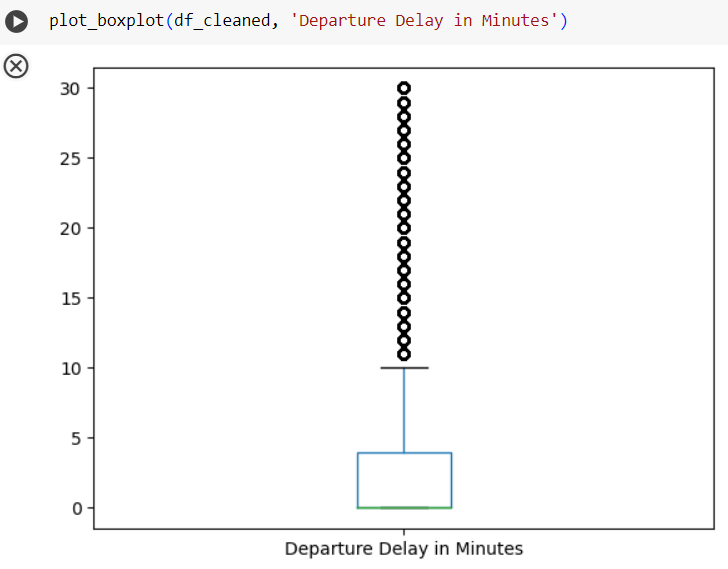


Figure 22.

* + - 1. **Conversion of Non-Numeric Values**

In the skywards data frame, some fields have non-numeric values as below:

**Input:**

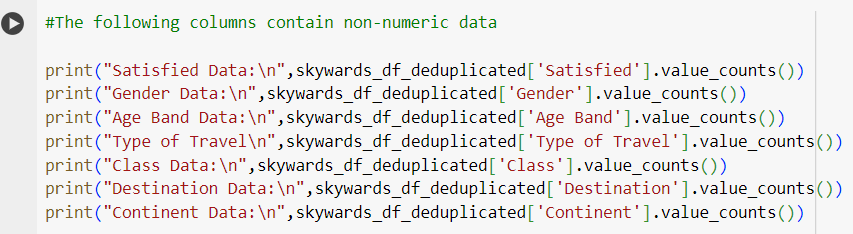


Figure 23.

**Output:**

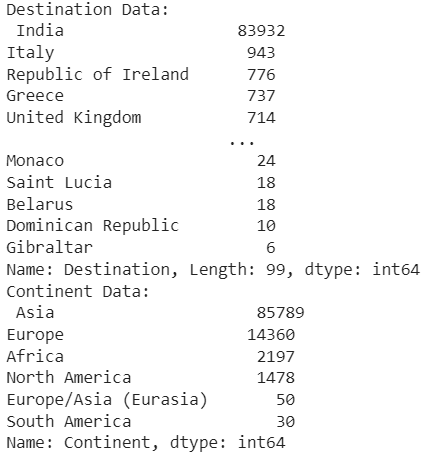
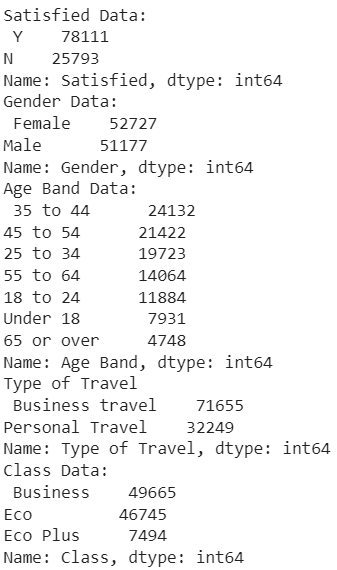


Figure 24.

The LabelEncoder library has been used to encode the text values into numeric form.

**Input:**

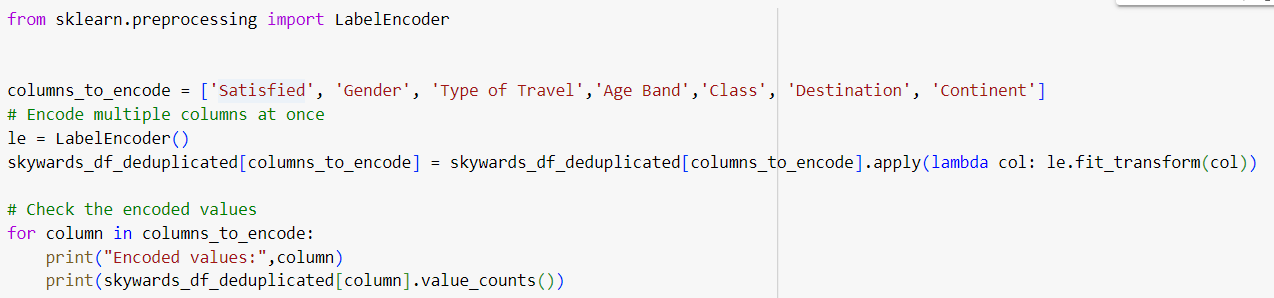


Figure 25.

**Output:**

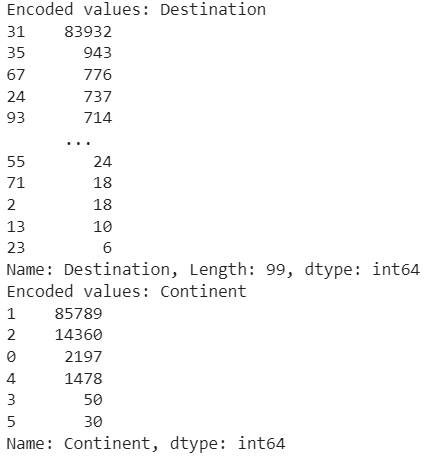
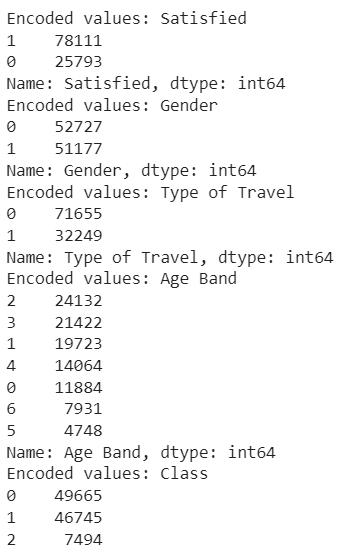


Figure 26.

* + 1. **Load**

After transforming the data, the data is kept as a data frame for further use in Exploratory data analysis and visualization processes.

* 1. **Exploratory Data Analysis**

**Input:**

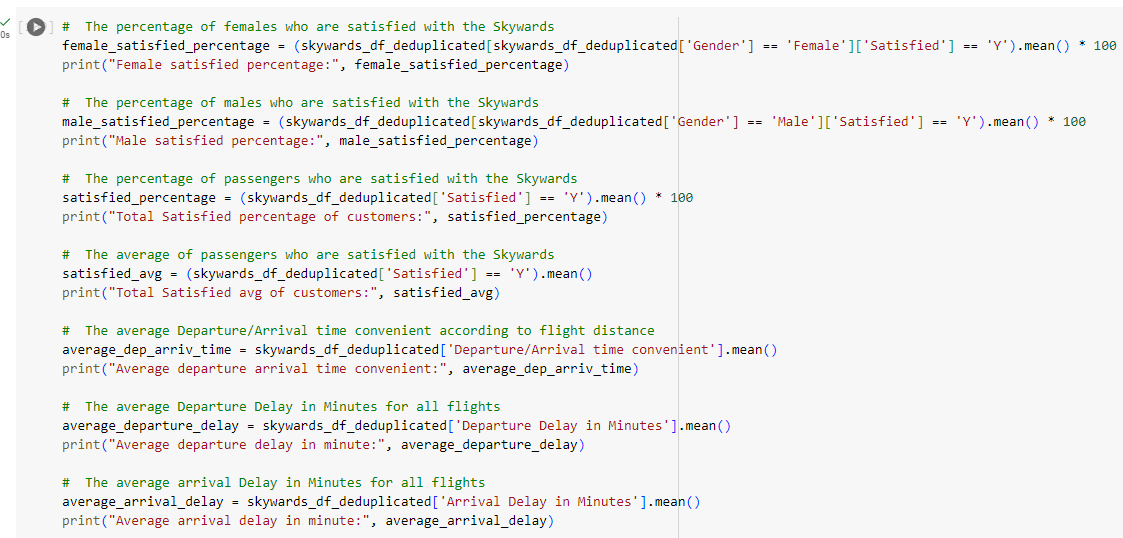


Figure 27.



Figure 28.

**Output:**

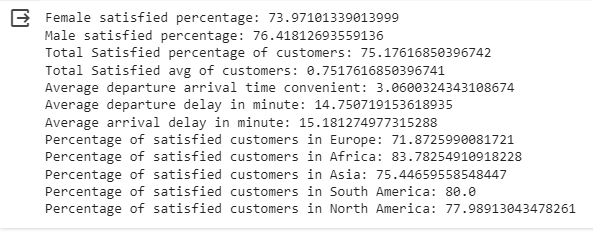


Figure 29.

* + 1. **Visualization**

By using the ‘matplotlib.pyplot’ library, customized plots/charts have been drawn. The following pie chart displays the overall percentage of satisfied and unsatisfied passengers:

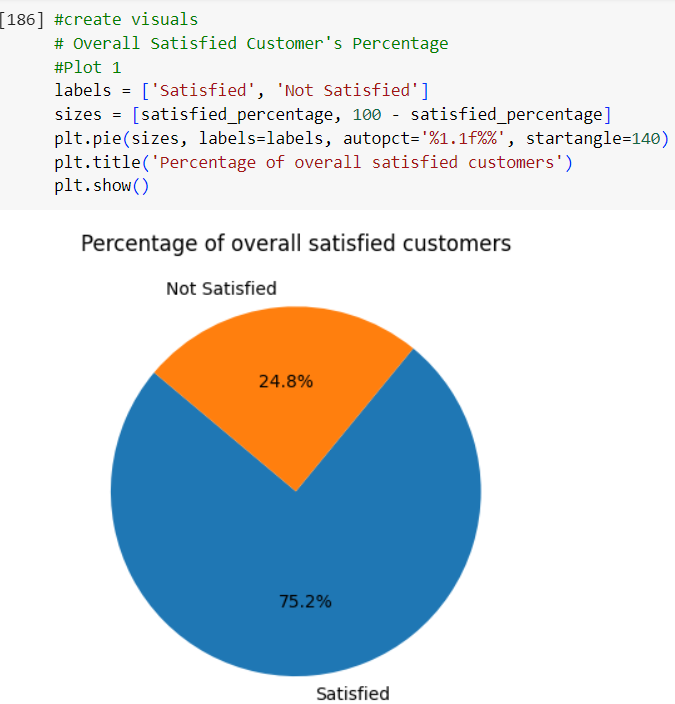


Figure 30.

The number of total men and women satisfied/unsatisfied passengers can be analyzed by using the graph drawn by pygwalker library.

**Input:**

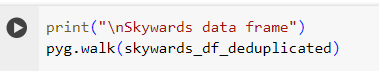


Figure 31.

**Output:**

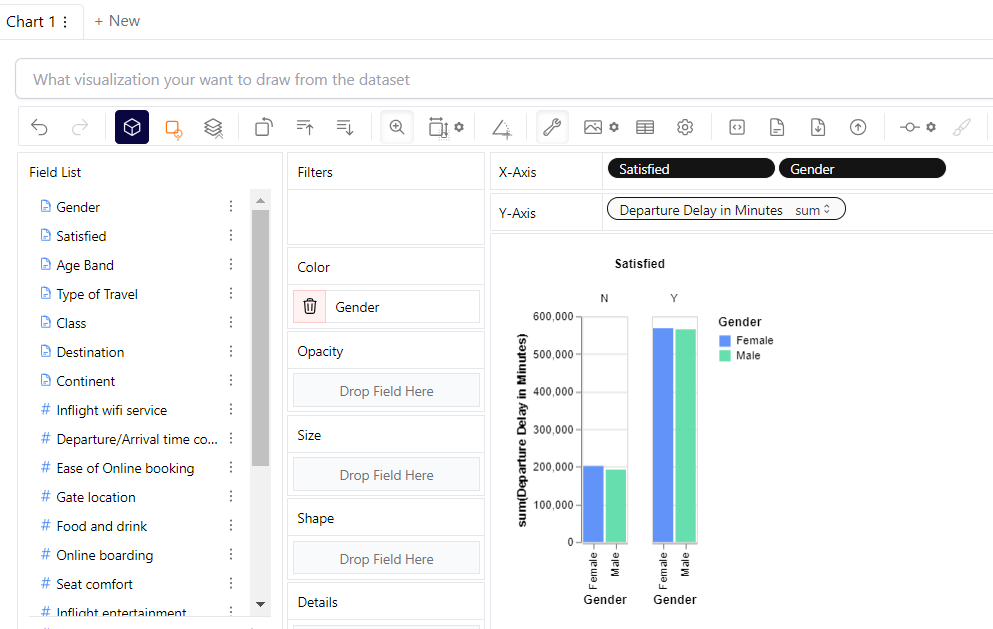


Figure 32.

Similarly, the average arrival delay time can be analyzed as follows:

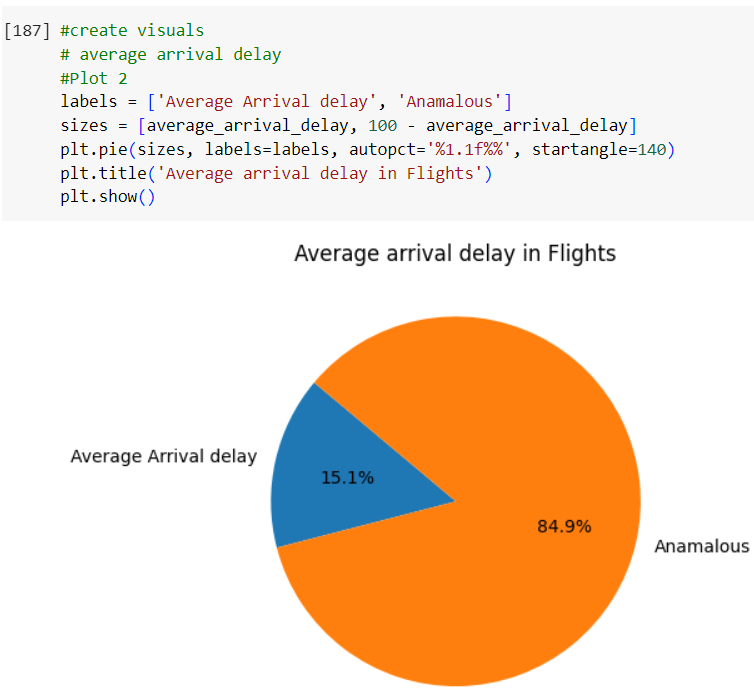


Figure 33.

The following graph represents that high number of satisfied customers is in Africa.

**Input:**

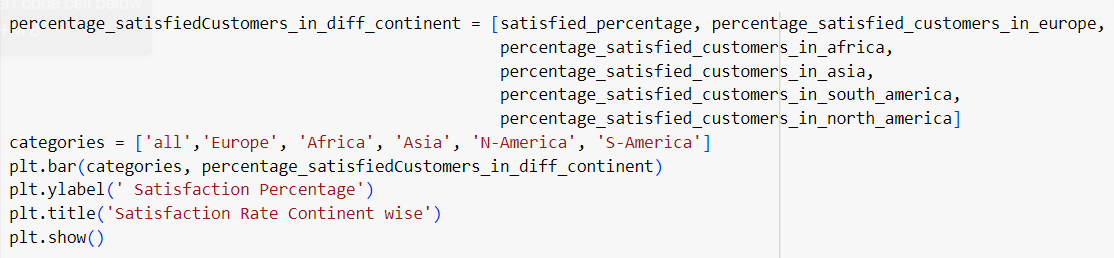


Figure 34.

**Output:**

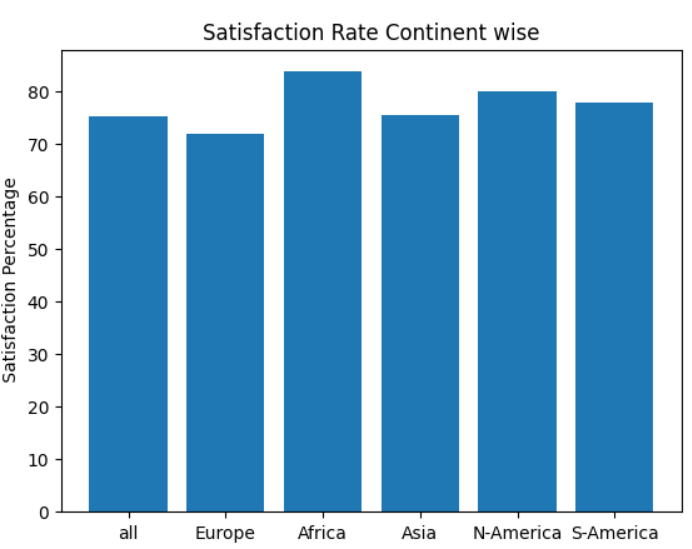


Figure 35.

The following scatter plot shows that the ratio of satisfied customers is higher than the of unsatisfied customers.

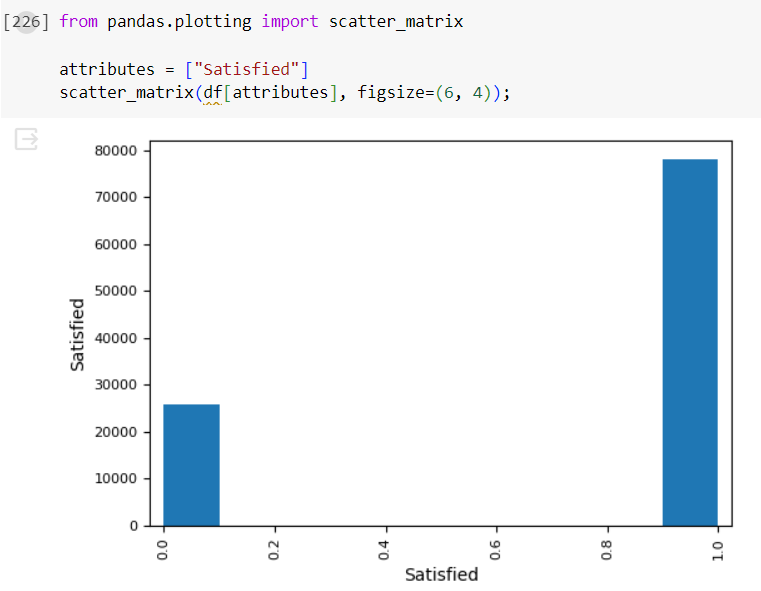


Figure 36.

The following histogram reveals there is independent relationship between the variables/columns.

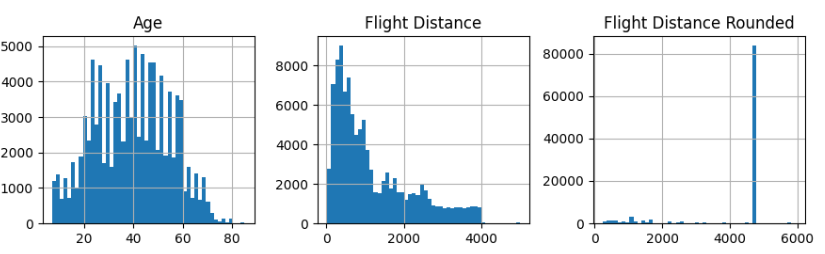


Figure 37.

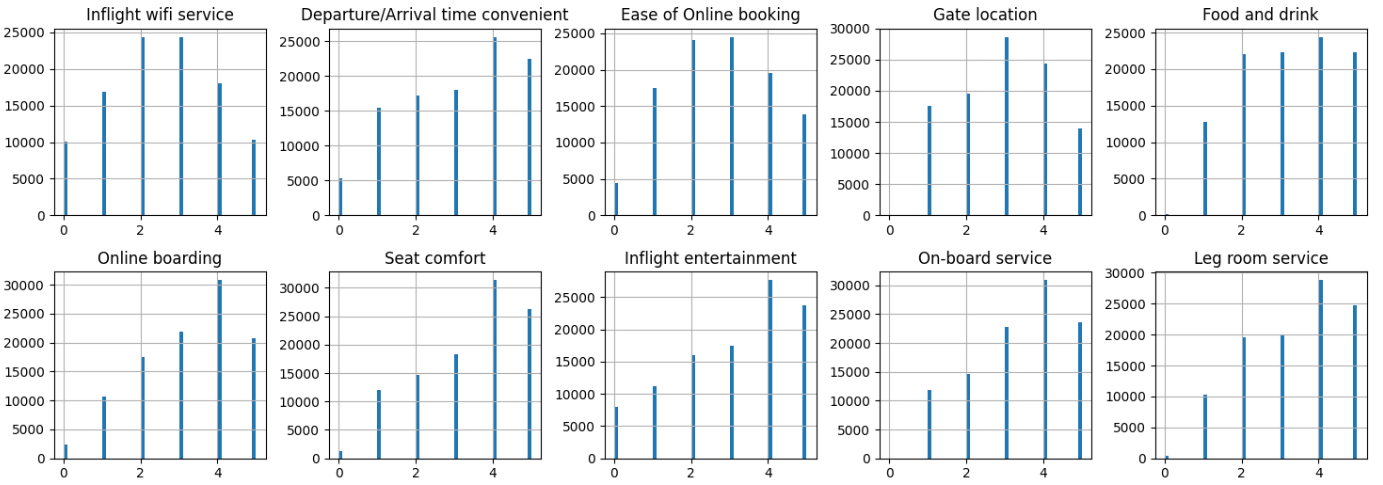


Figure 38.

The scatter chart represents online booking ease for different age groups.

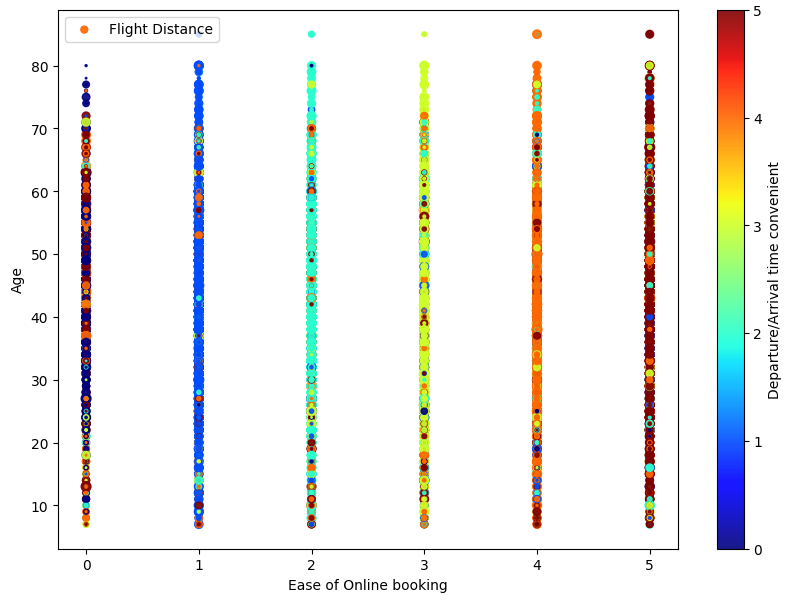


Figure 39.

* + 1. **Selection of Analytical Models**

The visualization and EDA reveal that the data of Skywards Airlines exhibits non-linear relationships between different variables (i.e. the rise or decline in one parameter’s value does not affect the other value), thus, regression models are not suitable, and non-linear models are preferred.

* 1. **Predictive Models**

For the prediction of scale and accuracy of airline data, two analytical models have been selected:

* Decision Tree
* Random Forest
  + 1. **Decision Tree**

The feature scaling has been performed to implement the data analytics model. It involves transforming the features of the dataset so that they all have the same scale (Xu, 2023).

For instance, our input data is at different scales but Flight Distance is in the thousands while Travel Type, class, etc. are just one digit.

**Input:**

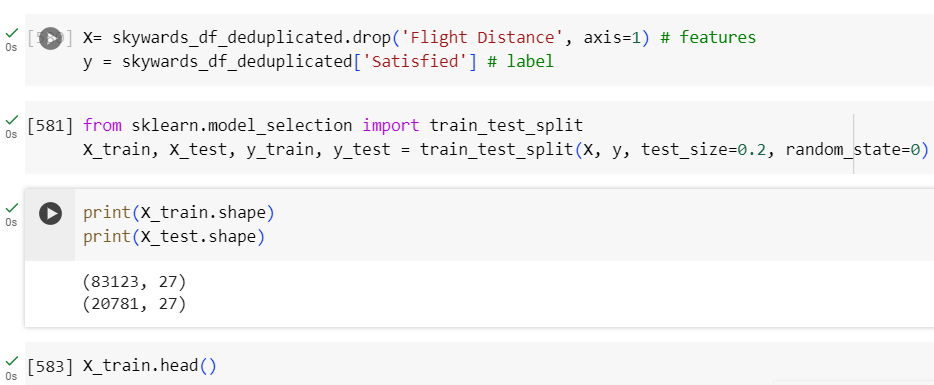


Figure 40.

**Output:**

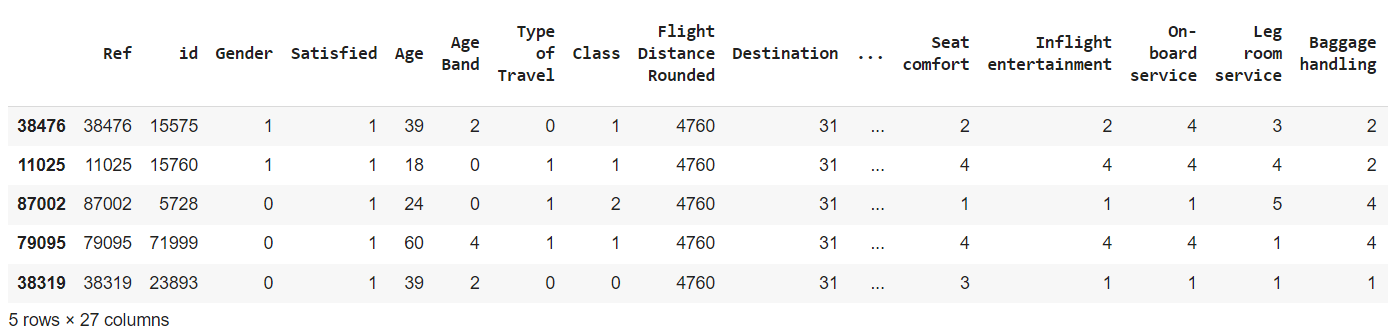


Figure 41.

**Input:**



Figure 42.

Furthermore, the model is tested with the following inputs:

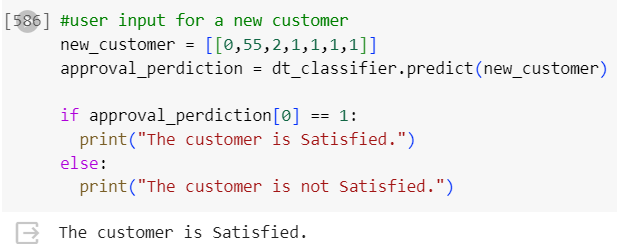


Figure 43.

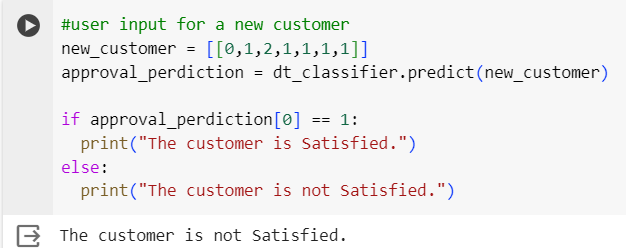


Figure 44.

* + 1. **Random Forest**

Random Forest can be a suitable choice for analyzing Skywards Airlines data because it is suitable for (Chen et al., 2024):

* High-dimensional data
* Classification task
* Scalability
* Interpretability

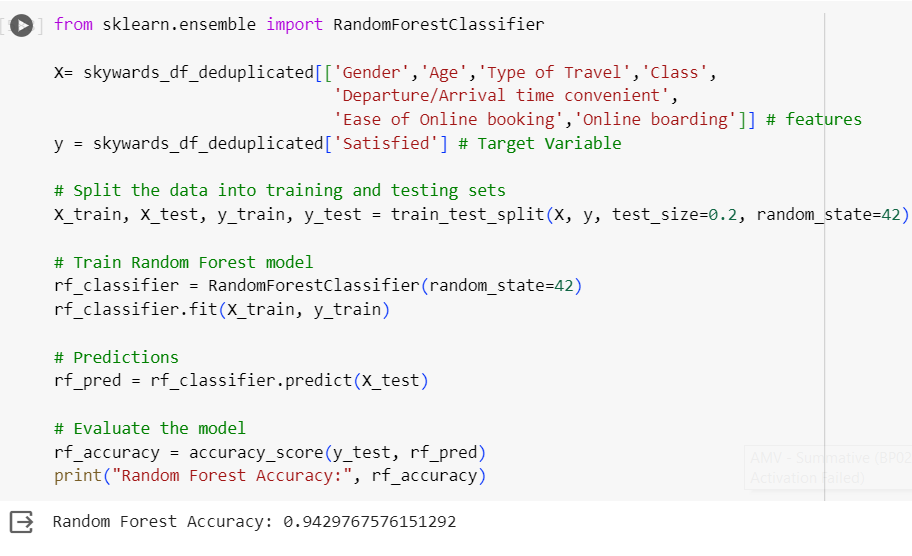


Figure 45.

This model can be tested with different values as follows:

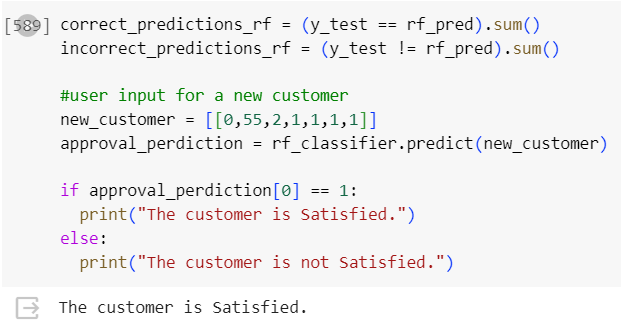


Figure 46.

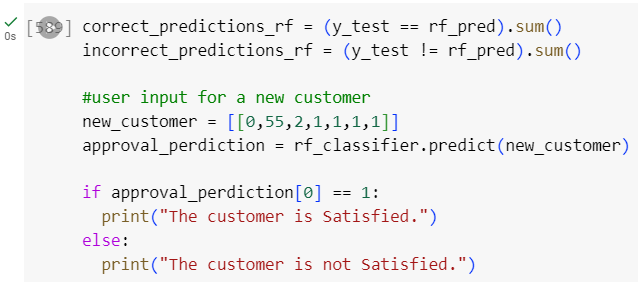


Figure 47.

* + 1. **Evaluate the Accuracy of Models**

The following code provides an unbiased estimate of the model's accuracy by evaluating them on different subsets of the training data.

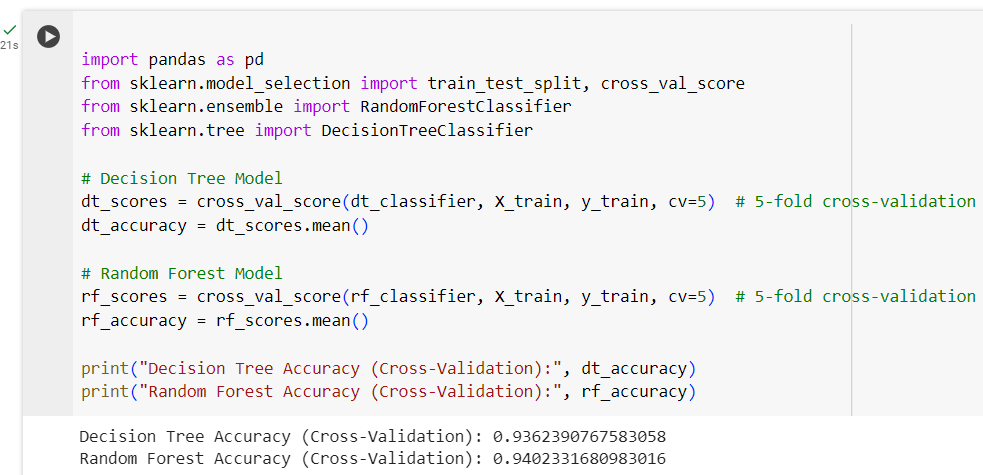


Figure 48.

* + - 1. **Strengths and Weaknesses**

Firstly, Decision Trees are simple, interpretable, and capable of handling non-linear relationships, but they are prone to overfitting (Völker et al., 2024). However, Random Forests address the overfitting issue and generally yield higher accuracy, but they are more complex and computationally expensive (Chen et al., 2024).

According to the current analysis, the accuracy of the Random Forest model is higher and it addresses the overfitting issue (Figure 48). However, some additional factors will be considered to ensure which model to choose.

1. **Recommendations**
   1. **Critical Evaluation of Models**
      1. **Loss Function**

Using accuracy as the chosen loss function allows us to evaluate the performance of both Decision Tree and Random Forest models in terms of their ability to make accurate predictions on unseen data of Skywards Airlines.

**Input:**

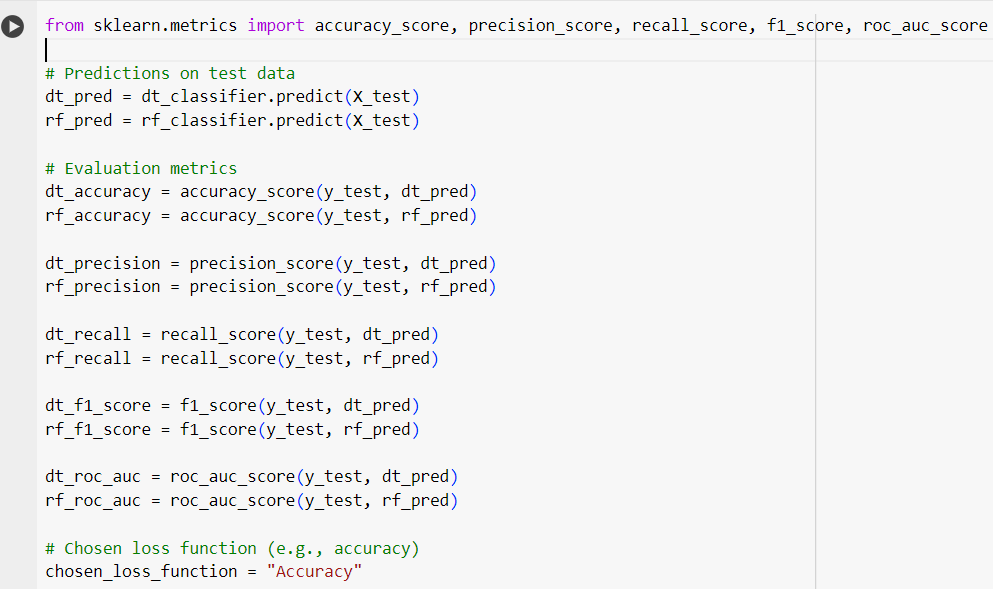


Figure 49.

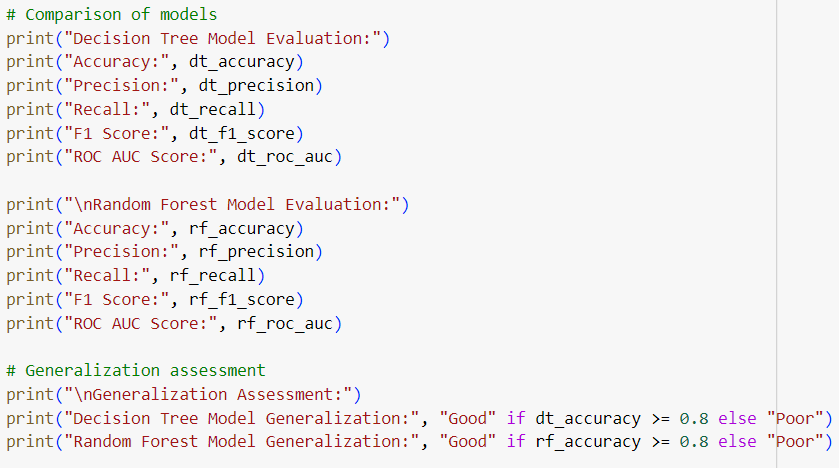


Figure 50.

The above code assesses the generalization of both models based on their accuracy. If the accuracy is greater than or equal to 0.8, it considers the generalization "Good"; otherwise, "Poor".

**Output:**

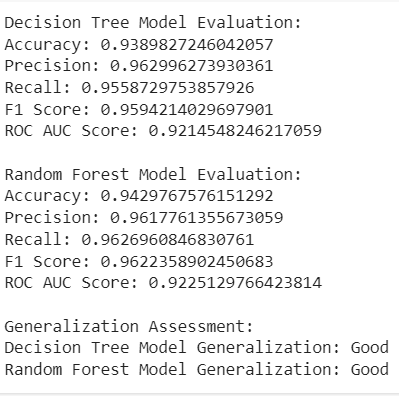


Figure 51.

* + 1. **Accuracy Metrics**

The accuracy metrics calculate the accuracy scores and construct confusion matrices for the Decision Tree and Random Forest models. It then extracts the counts of correct and incorrect predictions from the confusion matrices and creates a data frame and then a comparison table to compare the performance of both models (Epstein et al., 2023).

A comparison of the performance of both models based on accuracy scores and the counts of correct and incorrect predictions is as below:

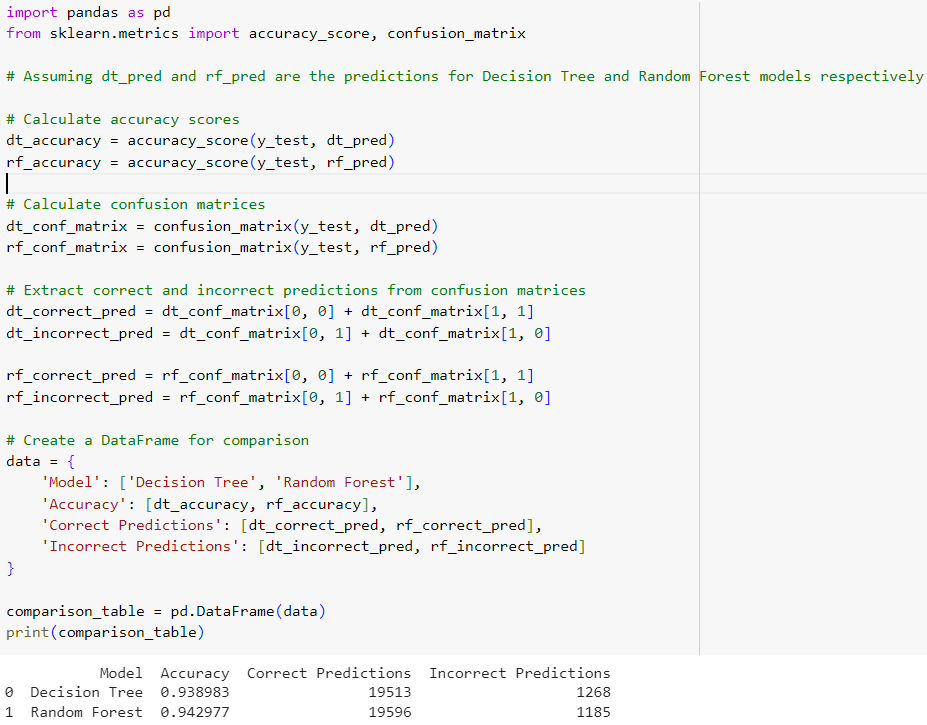


Figure 52.

* + 1. **Comparison Table**

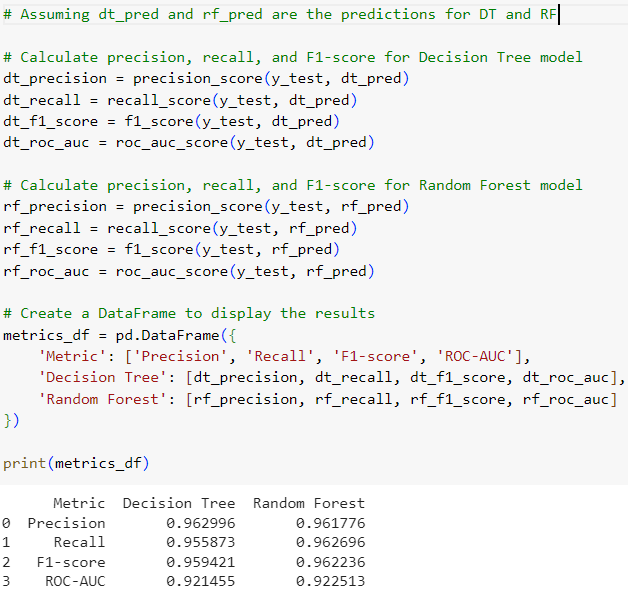


Figure 53.

* + - 1. **Root Square Mean Error/Mean Square Error**

MSE and RMSE quantify the average difference between the predicted values and the actual values of the target variable (Ramdani et al., 2021).

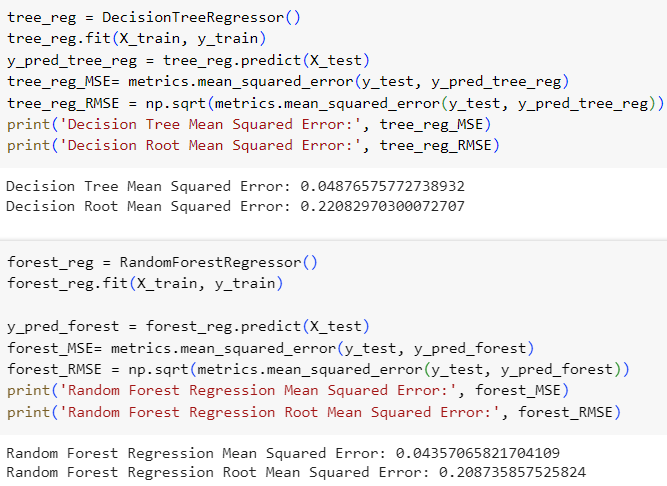


Figure 54.

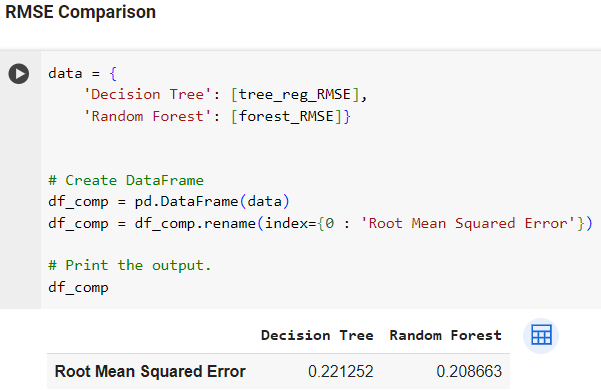


Figure 55.



Figure 56.

The MSE and RMSE values of the Random Forest model are lower than the decision tree. This indicates the better performance of the Random Forest model.

* + 1. **Recommended Model**

Figure 51 reveals that the Random Forest Model has a slightly higher accuracy (0.942) as compared with the Decision Tree Model.

Moreover, there are higher recall, F1 score, and ROC-AUC scores for predicting instances of dissatisfaction, despite the class imbalance in the dataset. However, the precision is almost the same for both models. These metrics indicate that Random Forest can effectively identify and classify dissatisfied customers, which is crucial for reducing the rate of 'satisfaction' (Khan et al., 2024).

Additionally, the other reasons for recommending the Random Forest model are as follows (Fan, 2023):

* Interpretability is not a primary concern and model performance is paramount
* Lower RMSE/MSE
* Computationally intensive model due to the ensemble nature.
* It can accurately predict instances of dissatisfaction that align with the business objective of reducing the rate of 'satisfaction' in the airline's customer base.

Hence, Random Forest's robustness, scalability, and ability to handle class imbalance, coupled with its high-performance metrics and alignment with business objectives, make it a suitable choice for reducing the rate of 'satisfaction' in the airline's customer base (Chen et al., 2024).

* 1. **Findings**
     1. **Correlational Matrix**

**Input:**

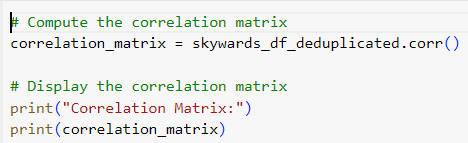
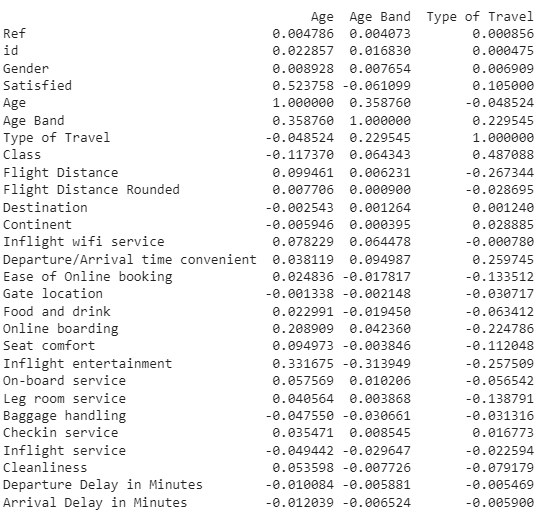
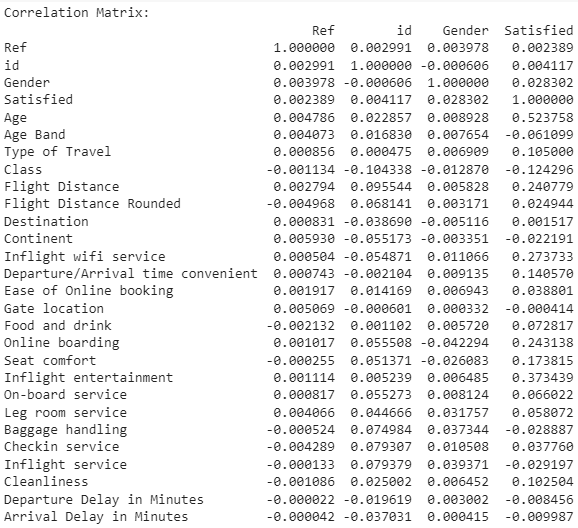
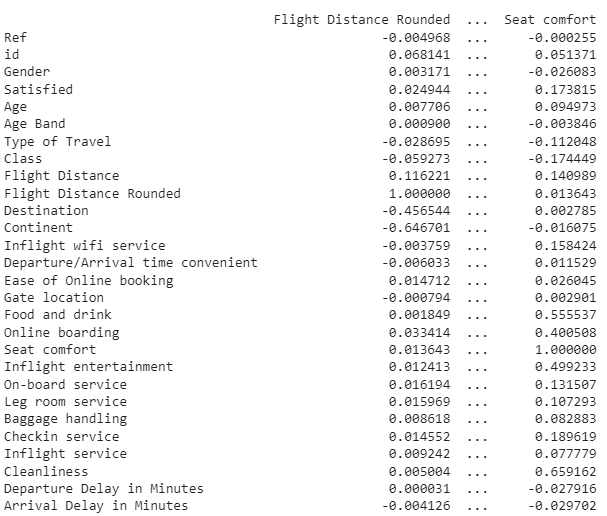
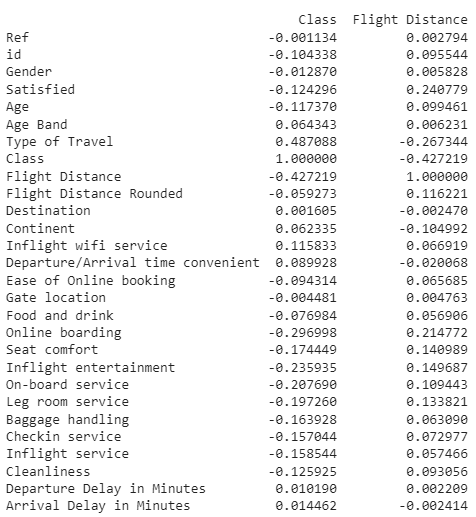


Figure 57.

**Output:**





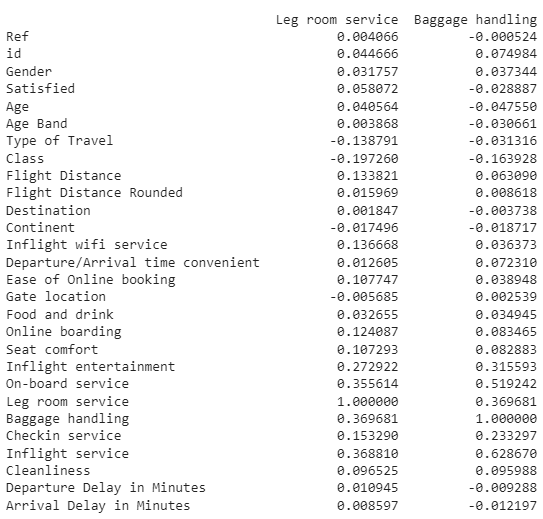
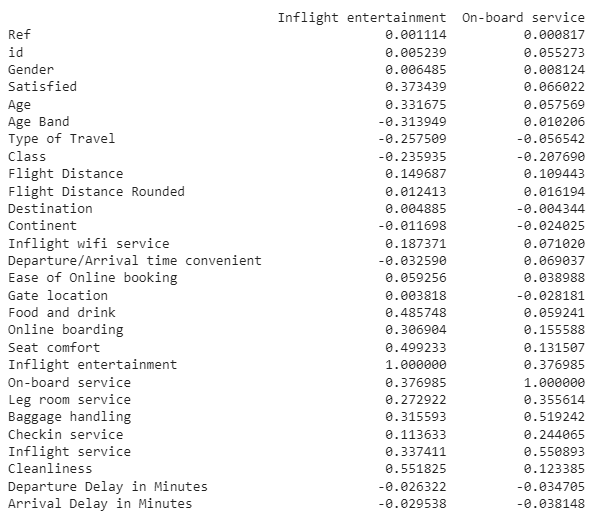


Figure 58.

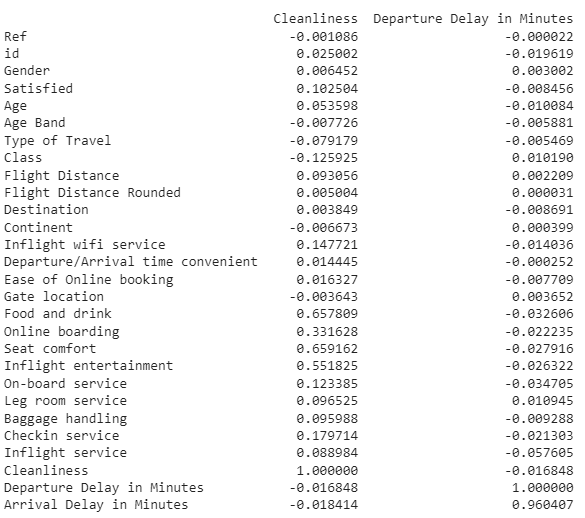
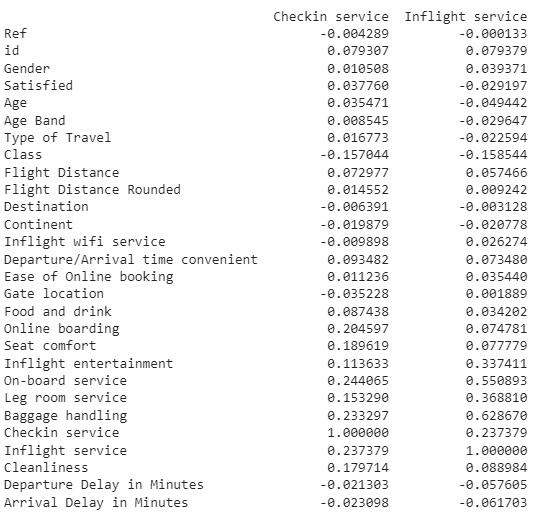


Figure 59.



Figure 60.

* + 1. **Heat map**

**Input:**

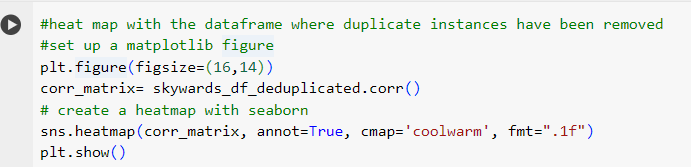


Figure 61.

**Output:**

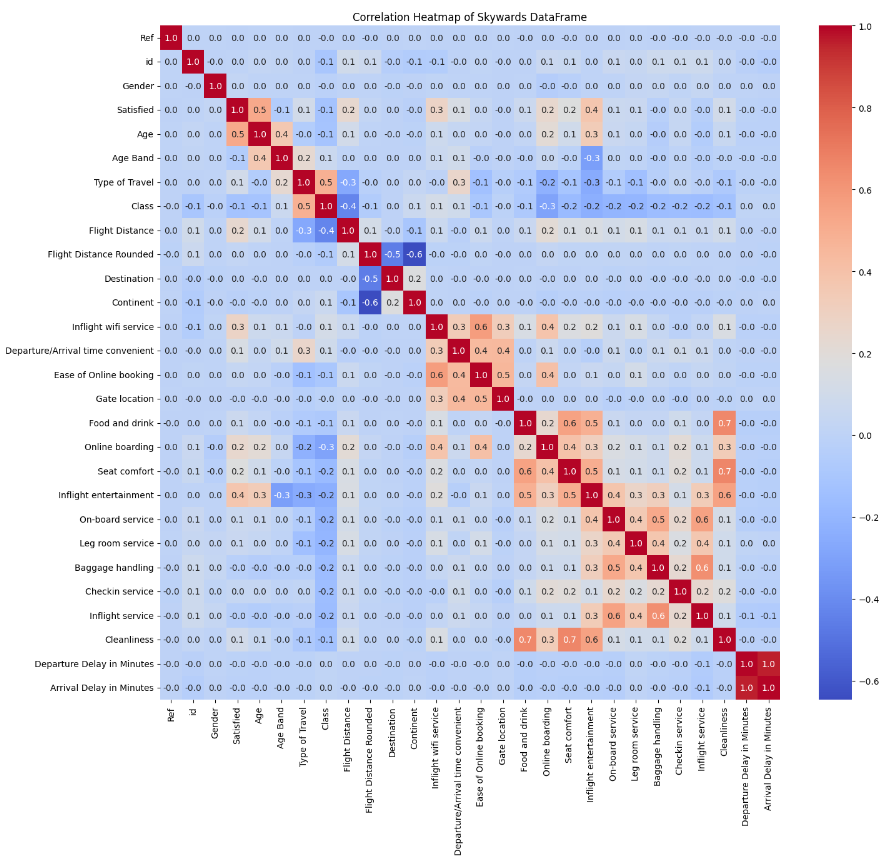


Figure 62.

A heat map of a correlation matrix provides a visual representation of the pairwise correlations between variables in a Skywards dataset (Gu, 2022):

* The intensity of colors in the heat map represents the strength of the correlation between pairs of variables (i.e. Darker colors indicate stronger correlations).
* Positive correlations (where both variables increase) are represented by warm colors, while negative correlations (where one variable increases as the other decreases) are represented by cool colours.
* Higher values on the diagonal suggest better performance.
  + 1. **Confusion Matrix**

**Input:**

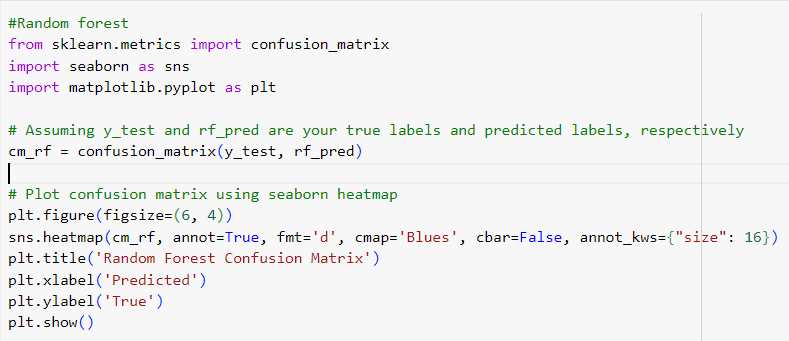


Figure 63.

**Output:**

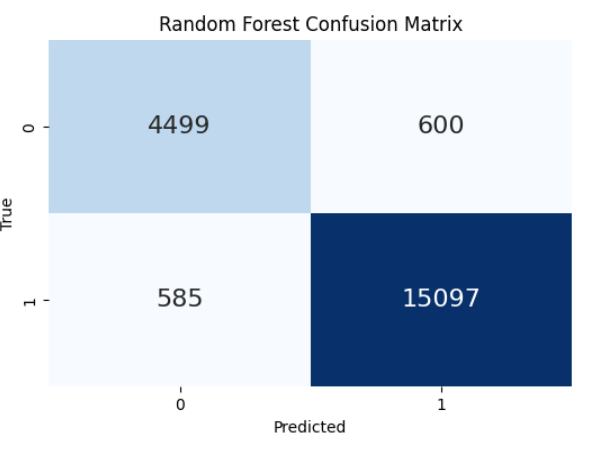


Figure 64.

For the Random Forest Model, 4499 instances are correctly classified as negative, however, 15097 instances are correctly classified as positive by the model. The high true positives of Random Forest Models suggest that the model is performing well in correctly identifying instances of the positive class.

**Input:**

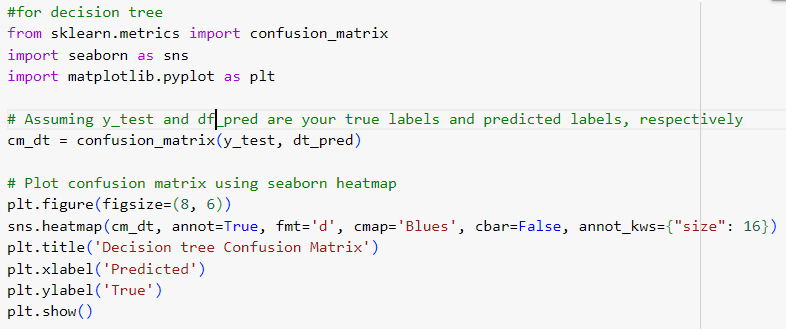


Figure 65.

**Output:**

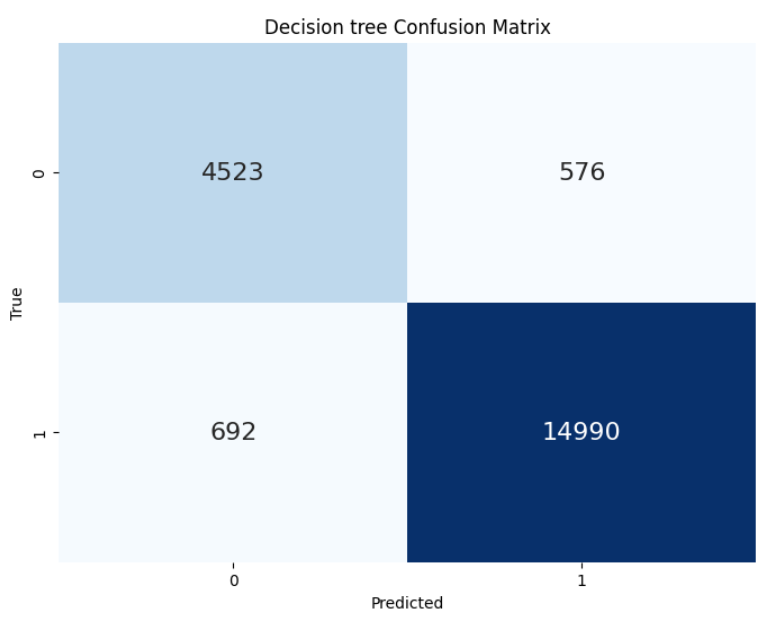


Figure 66.

For the Decision Tree Model, 4523 instances are correctly classified as negative, however, 14990 instances are positively classified.

To conclude, based on the analysis, the Random Forest model is recommended due to its superior performance in accurately identifying dissatisfied customers and its ability to generalize well.

1. **Next Steps**

* **Model Deployment:** Deploy the Random Forest model in a production environment where it can be integrated into the airline's systems for real-time prediction of customer satisfaction. This may involve setting up APIs/deploying the model within existing software infrastructure.
* **Monitoring and Evaluation:** Continuously monitor and evaluate the model's predictions against ground truth data and retrain it as necessary to adapt to changing patterns in customer behavior.
* **Feedback Loop:** Collect feedback from dissatisfied customers to further improve customer satisfaction strategies and interventions.
* **Continuous Improvement:** Iterate the model and associated processes based on feedback/new data. Explore ways to enhance the model's performance, such as incorporating additional features.
* **Ethical Considerations:** Safeguard customer data and prioritize transparency.

By following these steps, the airline can effectively leverage the Random Forest model to reduce the rate of 'satisfaction' and improve customer satisfaction and loyalty (Ding, 2023).

1. **References**

* Chen, J., Wang, X., Lei, F. (2024) ‘Data-driven multinomial random forest: a new random forest variant with strong consistency’, *Journal of Big Data.* DOI: 10.1186/s40537-023-00874-6.
* Ding, H. (2023) ‘Using Random Forest for Future Sea Level Prediction’, *SHS Web of Conferences.* DOI: 10.1051/shsconf/202317403008.
* Epstein, E., Nallapareddy, N., Ray, S. (2023) ‘On the Relationship between Feature Selection Metrics and Accuracy’, *Entropy*. DOI: 10.3390/e25121646.
* Fan, G. (2023) Random Forest Algorithm for Forest Fire Prediction. DOI: 10.1007/978-981-99-4554-2\_15.
* Fareez, M., Thangarajah, V., Saabith, S. (2020) ‘POPULAR PYTHON LIBRARIES AND THEIR APPLICATION DOMAINS’, *International Journal of Advance Engineering and Research Development.*
* Fernandes, A., Koehler, M., Konstantinou, N., Pankin, P., Paton, N., Sakellariou, R. (2023) ‘Data Preparation: A Technological Perspective and Review’, *SN Computer Science.* DOI: 10.1007/s42979-023-01828-8.
* Gu, Z. (2022) ‘Complex Heatmap Visualization’, *iMeta.* DOI: 10.1002/imt2.43.
* He, W., Liu, P., Qian, Q. (2023) Feature Engineering. DOI: 10.1007/978-981-99-3723-3\_4.
* JiravaÚstav, P. (2004) System Development Life Cycle. Available at: <https://www.researchgate.net/publication/44982277_System_development_life_cycle>
* Khan, T., Sadiq, R., Shahid, Z., Su’ud, M. (2024) ‘Sentiment Analysis using Support Vector Machine and Random Forest’, *Journal of Informatics and Web Engineering.* DOI: 10.33093/jiwe.2024.3.1.5.
* Lawanna, A. (2012) ‘The Theory of Software Testing’, *ITS Journal - Intelligent Transportation Systems Journal.*
* Memon, S., Maham, U. (2018) Importance of Software Life Cycle in Development of Quality Software.
* Nitnaware, R. (2019) ‘Basic Fundamental of Python Programming Language and The Bright Future’, *A Peer-Reviewed Journal About.*
* Nodeh, M., CALP, H., [şahin](https://www.researchgate.net/profile/Ismail-Sahin-7?_sg%5B0%5D=3DhN0yKRCi2AWo9by9ukrpHKD16ccpCq5-Ant93vz4z0uLflCjheZI0FrqBLFVfUWcLS0YE.iYoxzgHpzJIr0iR4-0tG0YDRqN9JaafeWn-nbJB05U0lopBrvlByvwmH0_iHavp-EK5a36tUXCu8-tWTTLYboQ&_sg%5B1%5D=W3REr4_DSvRs3ux_kwdkJxhbhJcnHD0W3mswBe-RJOOM_IrWjbf1JpVpVFCnjQOKDyGCGYk._7Bnj3Jx3tNzKUlqwPuhSPDE9tCFiQgJ27UpKCXUZtsbmRPvC2mptCJ-I-H-IFQAGQBAD1Vxly9yTT-MQiz3Zg&_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIiwicHJldmlvdXNQYWdlIjoic2VhcmNoIiwicG9zaXRpb24iOiJwYWdlSGVhZGVyIn19), I. (2020) Analyzing and Processing of Supplier Database Based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) Algorithm. DOI: 10.1007/978-3-030-36178-5\_44.
* Ostan, M. (2003) Unit testing: Test early, test often.
* Pessoa, T., Medeiros, R., Nepomuceno, T., Filho, P. (2018) Performance Analysis of Google Colaboratory as a Tool for Accelerating Deep Learning Applications.
* Ramdani, R., Fanisa, A., Gumilar, S., Sarbini, S. (2021) ‘Root square mean labelling of some graphs obtained from path’, *Journal of Physics: Conference Series*. DOI: 10.1088/1742-6596/1869/1/012143.
* Völker, K., Hümmer, M. (2024) The Decision Tree Procedure. DOI: 10.1007/978-3-031-47206-0\_1.
* Waskom, M. (2021) ‘Seaborn: statistical data visualization’, *The Journal of Open Source Software.*
* White, A. (2010) ‘A review of the impact of requirements on software project development using a control theoretic model’, *Journal of Software Engineering and Applications*, 03(09), pp. 852–857. DOI:10.4236/jsea.2010.39099.
* Xu, J. (2023) ‘Systematic Analysis and Application Prospect of Decision Tree’, *Highlights in Science Engineering and Technology*. DOI: 10.54097/hset.v71i.12687.

1. **Appendix**

* The programming solution of Python is embedded in this project file as follows:



* The link to Google Collab Notebook:

<https://colab.research.google.com/drive/16a2ApuJ77sHDeoNbqjRW0VjOY6PlUrcY?usp=sharing>