Laboratory assignment

Component 2

Authors: Patricia Moga, Ciobanu Sergiu-Tudor
Group: 6

January 13, 2026

1 Data Analysis

1.1 Analysis of the features used in learning

1.1.1 Correlation

Two correlation matrices were generated to analyze feature relationships using the Pearson Correlation Coefficient which measures linear dependence between two variables. All non-numerical features (nominal/ordinal) were first transformed into numerical representations such that a diverse relationship could be analyzed. The target feature, 'Satisfaction', was excluded for this initial feature to feature dependency check.

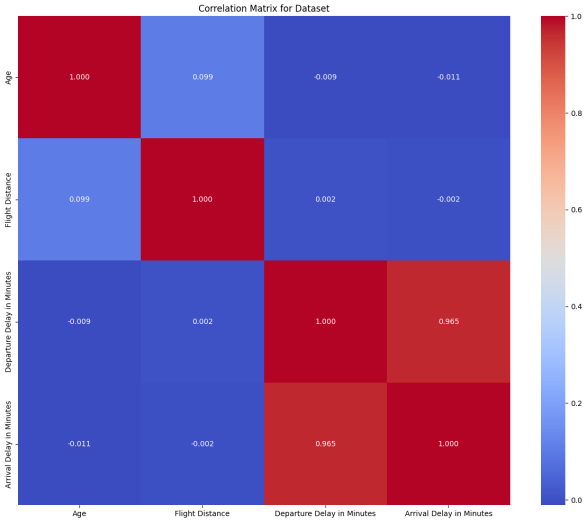


Figure 1: Correlation Matrix for Only Quantitative Features

Both matrices revealed one extremely strong linear relationship between *Departure Delay in Minutes* and *Arrival Delay in Minutes* that was observed to be nearly perfect ($|r| > 0.965$).

This near-perfect linear relationship is further visualized in the scatter plot (Figure 2), where the data points tightly follow the identity line. It was determined that the dependence of the arrival time delay on the departure time delay is more or less linear. This indicates that if the departure is delayed by a certain amount of time, the landing will most likely be delayed as well by about the same amount of time, taking into consideration that the aircraft does not significantly accelerate in flight to catch up the lost time.

The comprehensive matrix, which included the transformed categorical and ordinal features, revealed several 'important' inter-dependencies. Going into the service ratings, the

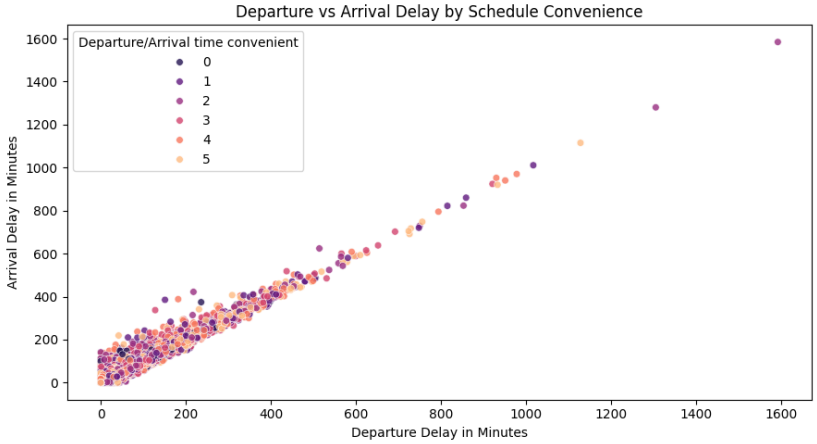


Figure 2: Scatter Plot of Departure vs. Arrival Delay by Schedule Convenience

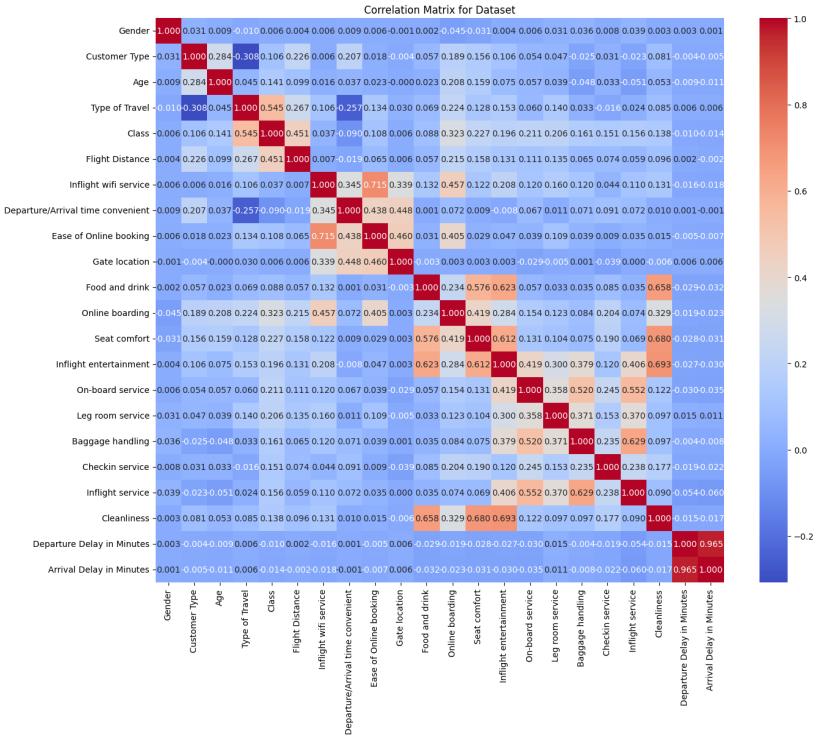


Figure 3: Comprehensive Correlation Matrix for All Transformed Features

strong positive correlation between features *Ease of Online booking* and *Inflight wifi service* suggests that satisfaction with the technological aspects of the passenger experience is highly consistent across the journey. The strong positive correlations ($|r| > 0.5$) among service ratings like *Seat Comfort*, *Food and Drink*, *Inflight Entertainment*, *Inflight Service*, and *Cleanliness* primarily highlight a General Satisfaction Factor, rather than the independent relevance of each feature.

A strong association ($|r| > 0.545$) was observed between *Type of Travel* and *Class*. Passengers booking Business Class tend to be associated with Business Travel. Conversely, passengers booking Economy or Economy Plus classes are strongly associated with Personal Travel.

1.1.2 Independence

The Chi-Squared Test of Independence is a statistical hypothesis test utilized to determine whether two categorical or nominal variables are related. This analysis focused on 14 ordinal feedback features and 4 nominal passenger/flight features (*Gender*, *Customer Type*, *Type of Travel*, *Class*); all discrete and continuous numerical features were excluded. The test was conducted on all possible pairs of these variables to ascertain statistical associations. The analysis confirmed that three independence relationships existed in the tested subset, all involving *Gender*. Focusing on the pairing with the highest χ^2 value, *Gender* and *Ease of Online booking* ($\chi^2 \approx 8.658$), the calculated statistic was found to be located before the critical region (as visualized in Figure 4).

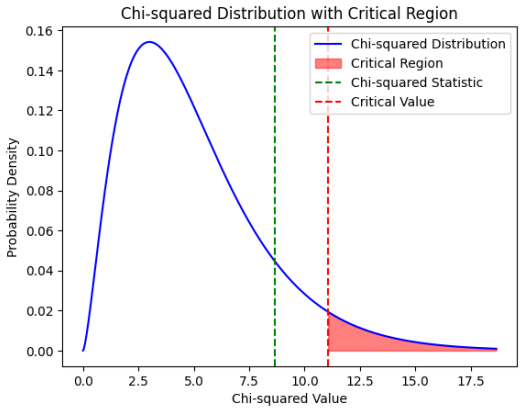


Figure 4: Chi-Squared Distribution for *Gender* and *Ease of Online Booking*

This position resulted in a p-value greater than 0.05, confirming the independence of the features. Furthermore, two supplementary plots were displayed, visually representing the actual versus expected passenger counts, which exhibited minimal distributional differences, reinforcing the conclusion that *Gender* has limited statistical influence on this service rating.

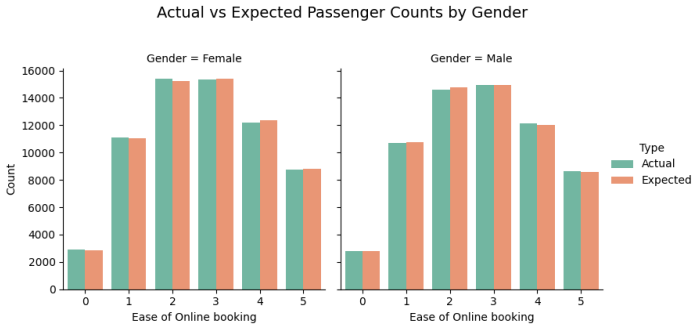


Figure 5: Actual vs. Expected Counts for Gender and Ease of Online Booking

The specific results for the tested Gender pairings are documented in the table below:

Table 1: Chi-Squared Test Results for Feature Independence

Feature 1	**Feature 2**	**χ^2 Statistic**	***p*-value**	**DOF**	**Hypothesis**
Gender	*Inflight wifi service*	6.922	0.226	5	independent
Gender	*Ease of Online booking*	8.658	0.124	5	independent
Gender	*Inflight entertainment*	4.247	0.514	5	independent

1.1.3 Feature Importance

The analysis of feature importance was conducted across all input features against the binary target variable, *Satisfaction*, by employing three distinct statistical tests tailored to the feature-target type. For the numerical input against the categorical target, the Pearson's Correlation was utilized, revealing a generally weak lienar relationship. Although *Flight Distance* showed the strongest correlation ($r_{pb} \approx 0.3$) and *Age* a moderate one ($r_{pb} \approx 0.13$), the *Delay* features exhibited negligible correlation (approximately -0.05), indicating a limited linear predictive value.

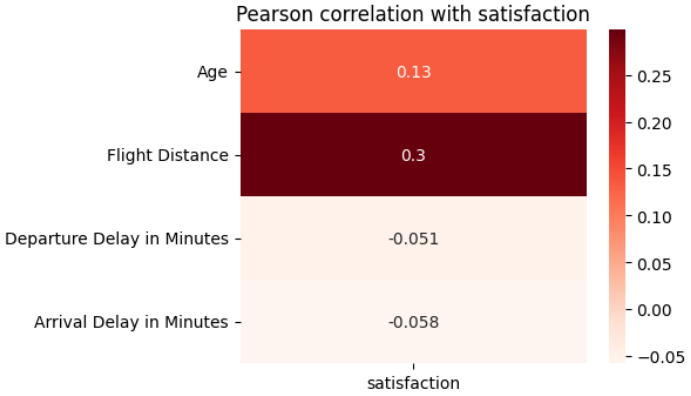


Figure 6: Pearson's Correlation between numerical inputs and categorical output

The statistical analysis of the categorical features against the binary target revealed that while all features were found to be statistically dependent ($p < 0.05$), only two features registered a p-value that was not computationally truncated to zero. The results of the Chi-Squared Test (χ^2) for these two specific features are provided below, detailing their specific test statistics p-values. The following table isolates the features where the statistical significance was calculated to be greater than 0:

Table 2: Chi-Squared Test Results for Features with P-Value > 0

Feature	**χ^2 Statistic**	**P-Value**	**DOF**	**Hypothesis**
Gender	17.0668	3.6089×10^{-5}	1	dependent
Departure/Arrival time convenient	599.5951	2.4738×10^{-127}	5	dependent

The Analysis of Variance (ANOVA) test was utilized to evaluate the structural dependence of categorical features against the binary target. The F-statistic acted as a rank metric, confirming that every tested categorical feature was highly dependent on the satisfaction outcome ($p \approx 0$). The results showed that *Departure/Arrival time convenient* (F-Statistic: 1599.5) and *Gender* (F-Statistic: 17) were the most influential structural determinants. This high F-statistic value confirms that the variation in satisfaction levels is best explained by the differences between the travel purpose groups and the service classes

1.2 Analysis

1.3 Data statistics

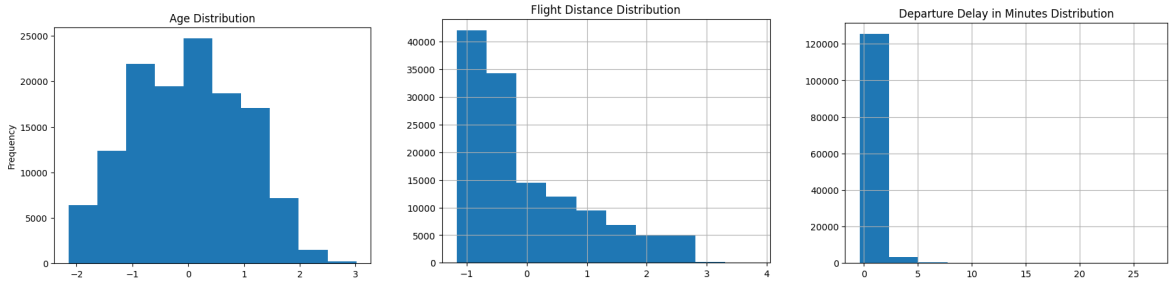
As the concepts of statistics for analyzing data such as mean, variance, standard deviation do not apply for nominal or ordinal data since these types of data are measured on a scale with only a few possible values, they make sense for continuous data. These data are measured on a scale with many possible values. Therefore, metrics detailing the mean, median, variance,

and standard deviation were calculated exclusively for the four numerical features present in our dataset: *Age*, *Flight Distance*, *Departure Delay in Minutes*, and *Arrival Delay in Minutes*. This summary is crucial for analyzing central tendency, dispersion, and identifying potential anomalies. The core statistics are presented below, providing a general overview of the value distribution and spread:

Table 3: Statistical Summary of Numerical Features

Metric	**Age**	**Flight Distance**	**Departure Delay**	**Arrival Delay**
Count	129,487	129,487	129,487	129,487
Mean	39.43	1190.21	14.64	15.09
Median (50%)	40.00	844.00	0.00	0.00
Std. Dev.	15.12	997.56	37.93	38.47
Min	7.00	31.00	0.00	0.00
Max	85.00	4983.00	1592.00	1584.00

The *Age* feature has a mean of 39.43 and a low standard deviation ($\sigma \approx 15.12$, variance 228.54), indicating the data is tightly grouped around the average. Its distribution is well-behaved, closely following the normal distribution curve, which is visible in the provided histogram.



(a) Age (b) Flight Distance (c) Departure Delay

Figure 7: Distribution of the numerical features: *Age* (Normal Distribution), *Flight Distance* (High Variability), and *Departure Delay in Minutes* (Highly Right-Skewed).

The *Flight Distance* feature has a mean of 1190.2 and a high standard deviation ($\sigma \approx 997.56$, variance 995,120.17), reflecting the wide range of journey lengths. This large spread is expected and confirms high variability in the type of flights present in the dataset, which is reflected in its corresponding distribution plot. A small percentage of outliers (long-distance flights) were identified at 2.1987% of the data.

The most critical observations concern the Delay features (*Departure Delay in Minutes* and *Arrival Delay in Minutes*). The mean departure delay is 14.64 minutes, but the median and 25th percentile are both 0.0, indicating that most flights were on time. The presence of extreme outliers is confirmed by the maximum value reaching 1592.0 minutes. This difference between the mean and median shows the data is right-skewed; a small number of extremely long delays pulls the mean far from the median.

This skewness results in a high standard deviation ($\sigma \approx 37.93$ for Departure Delay). Critically, the percentage of extreme outliers is high: 13.88% for Departure Delay and 13.51% for Arrival Delay.

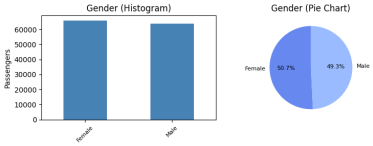
1.4 Data distribution

This section analyzes the distribution of passengers across various categories and services, providing a clear profile of the airline's customers and their rating patterns.

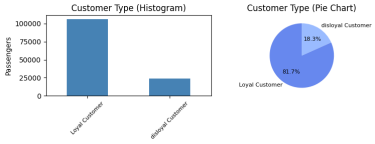
The sample is characterized by near equal representation of men and women, as demonstrated by the comparison plot. However, the vast majority of the airline's patronage comes from repeat customers, indicating a high level of passenger loyalty. Furthermore, most clients fly for business reasons, and approximately half of all passengers choose Business Class seats, which is visually confirmed by the class distribution plot. Regarding service satisfaction, the airline generally performs well: more than 60% of passengers were satisfied (rated 4-5) with the baggage transportation service, and over 50% of passengers found their seating comfort and legroom acceptable.

Figure 8: Distribution of Passenger Categories and Core Service Ratings

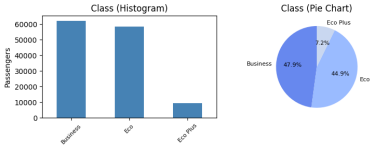
(a) Gender Comparison



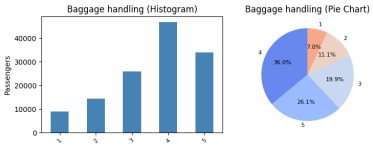
(b) Customer Loyalty



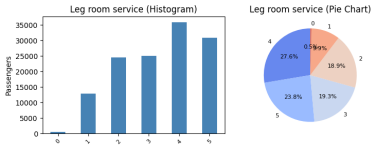
(c) Class Distribution



(d) Baggage Handling Rating



(e) Leg Room Rating

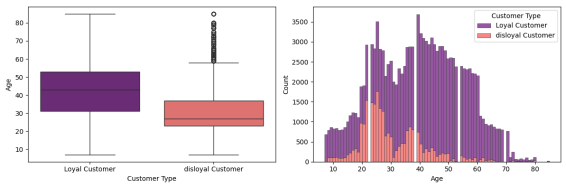


A closer inspection of demographics and travel patterns reveals key relationships. The box diagram analysis shows that regular customers typically fall between the ages of 30 and 50, with an average age slightly exceeding 40. In contrast, non-regular customers have a slightly younger and narrower age range, averaging just under 30 years old.

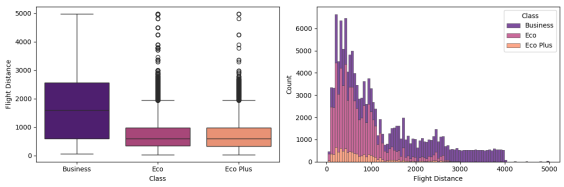
Looking closely at passenger age and loyalty, the data draws a sharp distinction: the airline's regular customers are generally older, with most falling into the 30 to 50 age bracket and an average age just over 40. In contrast, non-regular, or disloyal, customers are noticeably younger, clustered between 25 and 40, and averaging slightly under 30.Analyzing travel habits and expectations, two clear patterns emerge based on flight length. First, the longer the flight distance, the more likely customers are to fly Business Class—it's clear that people pay for comfort and service on long-haul trips. Second, and relatedly, the further a passenger travels, the more satisfied they tend to be with in-flight facilities like entertainment and legroom. When customers are stuck on a plane for many hours, their appreciation for extra space and quality distractions goes up significantly.

Figure 9: Customer Behavior and Service Expectation Plots

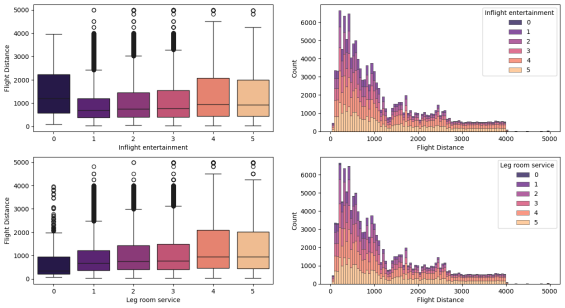
(a) Customer Type vs. Age



(b) Flight Distance vs. Class



(c) Distance vs. Inflight Services



1.5 Data Visualization

Due to the fact that our data are non-linear because involve more complex boundaries to separate classes as well as being described both by quantitative and qualitative variables, the PCA method would not be suitable for the representation. That's why the algorithm that works the best for this type of visualization is UMAP. The resulting UMAP graph of the data takes a very interesting shape. MAP is a non-linear dimensionality reduction technique that excels at maintaining the local and global structure of the data, making it ideal for visualizing complex, manifold-like distributions in lower dimensions.

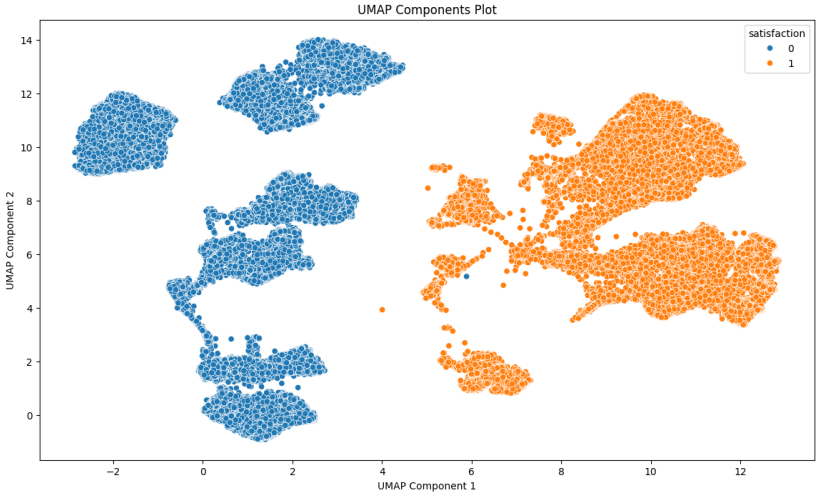


Figure 10: 2D UMAP Projection of Airline Passenger Data. The distinct clusters indicate strong separability between satisfaction classes.

Almost all of the data points are very well distinguished, indicating that the major underlying classes (Satisfied vs. Neutral/Dissatisfied) are highly separable in the feature space. This suggests the implemented classification model should achieve high accuracy.

A key area of interest is the overlap near the middle of the graph. This region represents

the central decision boundary where passenger features are not distinct enough.