32130 Fundamentals of Data Analytics

ASSIGNMENT 2:
DATA EXPLORATION AND PREPARATION

Student: 25422229 Thu Thuy Nguyen

Teacher: Dr Maoying Qiao

Session: Spring 2025

TABLE OF CONTENTS

TASK 1A. INITIAL DATA EXPLORATION

1. IDENTIFY THE ATTRIBUTE TYPE OF EACH ATTRIBUTE IN THE DATASET

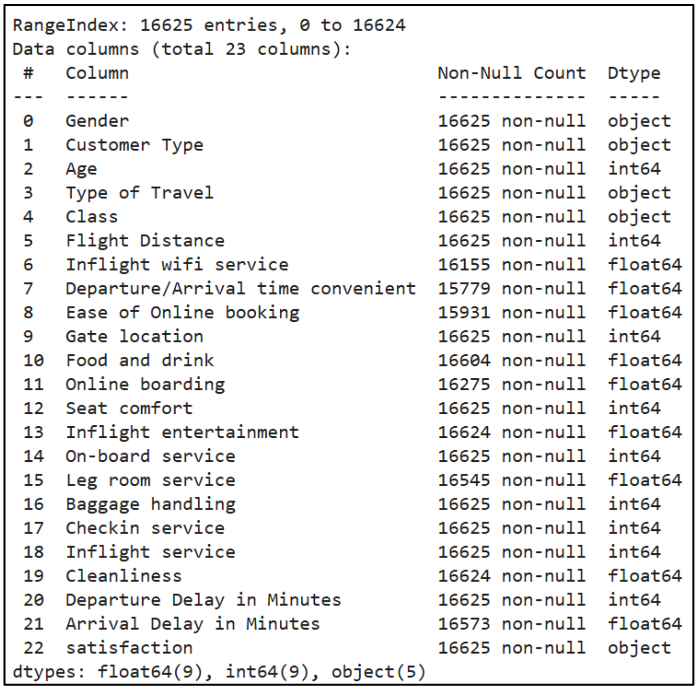
The dataset has 23 qualities; however, the ID attribute is excluded because it is simply a random integer with no relation to customer satisfaction.

No.	Name of attribute	Attribute type	Explanation
1	Gender	Qualitative, nominal	Gender of the passengers (Female, Male). They cannot be arranged in any particular order, and they cannot be ranked.
2	Customer Type	Qualitative, nominal	The customer type (Loyal customer, disloyal customer). No ordering between "Loyal" and "Disloyal", just different classes.
3	Age	Quantitative, ratio	The actual age of the passengers so it can have a true zero (age 0 = newborn, absence of age).
4	Type of Travel	Qualitative, nominal	Purpose of the flight of the passengers (Personal Travel, Business Travel). Just labels for different travel purposes.
5	Class	Qualitative, ordinal	Travel class in the plane of the passengers (Business, Eco, Eco Plus). There is an order between them (Business > Eco Plus > Eco) but they cannot be measured.
6	Flight distance	Quantitative, ratio	The flight distance of this journey. It is numeric, continuous, and supports all arithmetic operations. and has a true zero (0 distance = no travel). So, it is a Ratio attribute.
7	Inflight wifi service	Qualitative, ordinal	Satisfaction level of the inflight Wi-Fi service. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
8	Departure/Arrival time convenient	Qualitative, ordinal	Satisfaction level of Departure/Arrival time convenient. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
9	Ease of Online booking	Qualitative, ordinal	Satisfaction level of online booking. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
10	Gate location	Qualitative, ordinal	Satisfaction level of Gate location. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
11	Food and drink	Qualitative, ordinal	Satisfaction level of Food and drink. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.

12	Online boarding	Qualitative, ordinal	Satisfaction level of online boarding. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
13	Seat comfort	Qualitative, ordinal	Satisfaction level of Seat comfort. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
14	Inflight entertainment	Qualitative, ordinal	Satisfaction level of inflight entertainment. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
15	On-board service	Qualitative, ordinal	Satisfaction level of On-board service. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
16	Leg room service	Qualitative, ordinal	Satisfaction level of Leg room service. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
17	Baggage handling	Qualitative, ordinal	Satisfaction level of baggage handling. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
18	Check-in service	Qualitative, ordinal	Satisfaction level of Check-in service. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
19	Inflight service	Qualitative, ordinal	Satisfaction level of inflight service. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
20	Cleanliness	Qualitative, ordinal	Satisfaction level of Cleanliness. There is an order between them (from 1 to 5) but they are not measurable, so they are Ordinal attribute.
21	Departure Delay in Minutes	Quantitative, ratio	Minutes delayed when departure. It is numeric, continuous, and has a true zero (0 minutes delay = no delay). Furthermore, it can be compared meaningfully. So, it is Ratio attribute.
22	Arrival Delay in Minutes	Quantitative, ratio	Minutes delayed when Arrival. They are Ratio as they have same reasons as Departure Delay.
23	Satisfaction	Qualitative, ordinal	Airline satisfaction level (Satisfaction, neutral or dissatisfaction). The categories have a clear ranking (Dissatisfaction < Neutral < Satisfaction) but they are not measurable differences between levels. Hence, Ordinal.

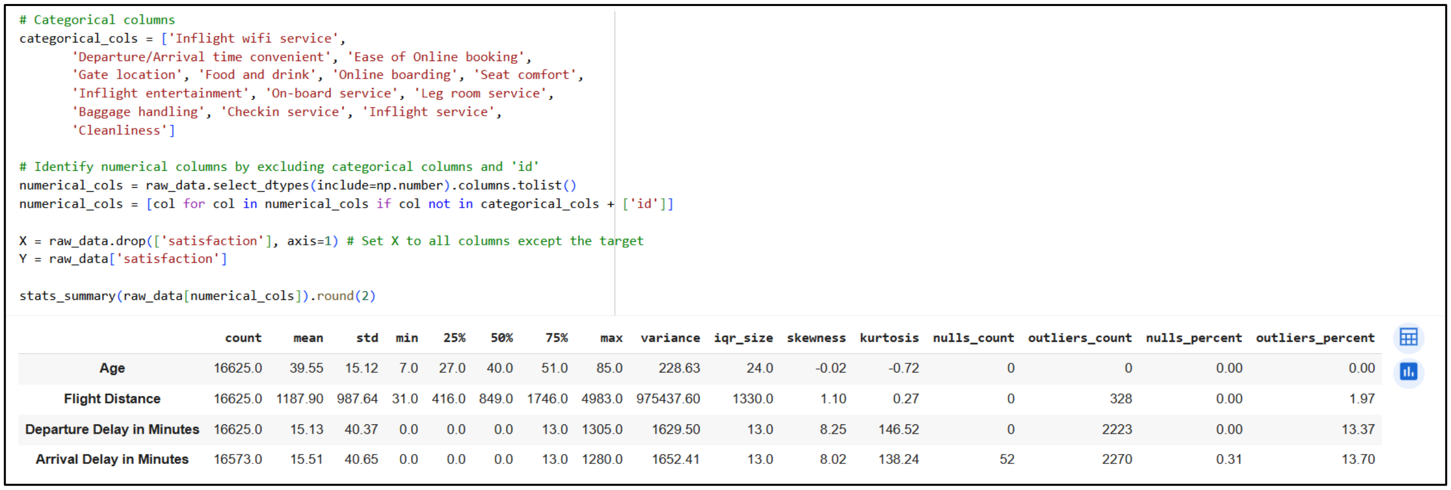
2. IDENTIFY THE VALUES OF THE SUMMARIZING PROPERTIES FOR THE ATTRIBUTES

In this dataset, service ratings range from 1 to 5, while a value of 0 indicates "Not Applicable," meaning the passenger did not use or experience that service. Since 0 does not reflect a level of satisfaction, it would not be meaningful to analyze it alongside valid ratings. Therefore, we treat 0 as a missing value to avoid biasing the results. This simplifies the dataset and ensures the analysis focuses only on actual passenger feedback.



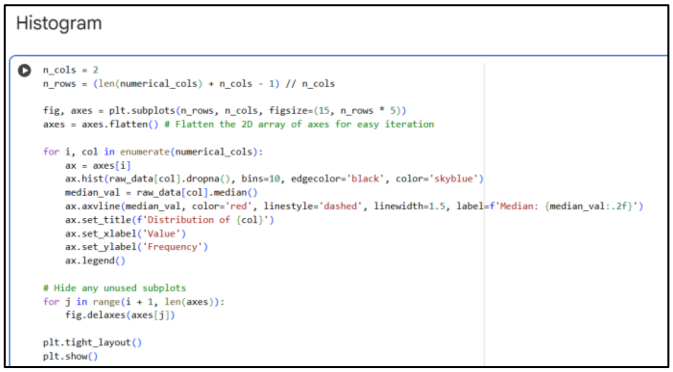
```
RangeIndex: 16625 entries, 0 to 16624
Data columns (total 23 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   Gender                             16625 non-null  object
 1   Customer Type                      16625 non-null  object
 2   Age                                16625 non-null  int64
 3   Type of Travel                     16625 non-null  object
 4   Class                              16625 non-null  object
 5   Flight Distance                    16625 non-null  int64
 6   Inflight wifi service              16155 non-null  float64
 7   Departure/Arrival time convenient  15779 non-null  float64
 8   Ease of Online booking             15931 non-null  float64
 9   Gate location                      16625 non-null  int64
 10  Food and drink                     16604 non-null  float64
 11  Online boarding                    16275 non-null  float64
 12  Seat comfort                       16625 non-null  int64
 13  Inflight entertainment             16624 non-null  float64
 14  On-board service                   16625 non-null  int64
 15  Leg room service                   16545 non-null  float64
 16  Baggage handling                   16625 non-null  int64
 17  Checkin service                    16625 non-null  int64
 18  Inflight service                   16625 non-null  int64
 19  Cleanliness                        16624 non-null  float64
 20  Departure Delay in Minutes         16625 non-null  int64
 21  Arrival Delay in Minutes           16573 non-null  float64
 22  satisfaction                       16625 non-null  object
dtypes: float64(9), int64(9), object(5)
```

For classification, the dataset's attributes can be grouped into five main categories. **Passenger Profile** (Age, Gender) captures demographic influences on satisfaction. **Travel Characteristics** (Flight Distance, Type of Travel, Class) reflect the type of trip. **Loyalty status** (Customer Type) is a strong predictor. **Service Quality Attributes** cover passenger interactions with the airline, divided into *Pre-flight and Ground Services* (booking, check-in, boarding, baggage, gate location), *Inflight Services* (seat comfort, leg room, Wi-Fi, food, entertainment, cabin service, cleanliness), and *Schedule Convenience* (departure/arrival time convenient). **Operational Performance** (Departure Delay, Arrival Delay) is a key determinant of passenger satisfaction. This structure helps isolate which dimensions of the passenger experience are most strongly associated with satisfaction outcomes.



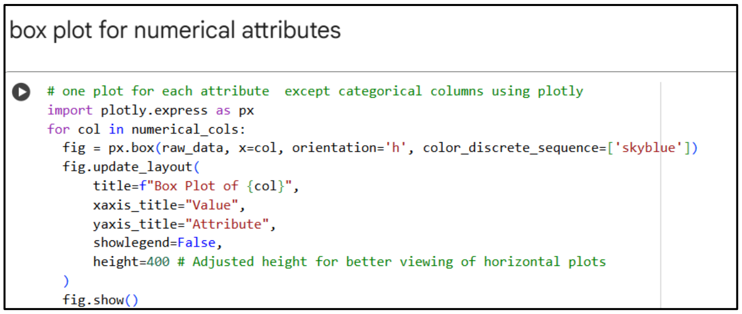
```
# Categorical columns
categorical_cols = ['Inflight wifi service',
       'Departure/Arrival time convenient', 'Ease of Online booking',
       'Gate location', 'Food and drink', 'Online boarding', 'Seat comfort',
       'Inflight entertainment', 'On-board service', 'Leg room service',
       'Baggage handling', 'Checkin service', 'Inflight service',
       'Cleanliness']

# Identify numerical columns by excluding categorical columns and 'id'
numerical_cols = raw_data.select_dtypes(include=np.number).columns.tolist()
numerical_cols = [col for col in numerical_cols if col not in categorical_cols + ['id']]

X = raw_data.drop(['satisfaction'], axis=1) # Set X to all columns except the target
Y = raw_data['satisfaction']

stats_summary(raw_data[numerical_cols]).round(2)
```

	count	mean	std	min	25%	50%	75%	max	variance	iqr_size	skewness	kurtosis	nulls_count	outliers_count	nulls_percent	outliers_percent
Age	16625.0	39.55	15.12	7.0	27.0	40.0	51.0	85.0	228.63	24.0	-0.02	-0.72	0	0	0.00	0.00
Flight Distance	16625.0	1187.90	987.64	31.0	416.0	849.0	1746.0	4983.0	975437.60	1330.0	1.10	0.27	0	328	0.00	1.97
Departure Delay in Minutes	16625.0	15.13	40.37	0.0	0.0	0.0	13.0	1305.0	1629.50	13.0	8.25	146.52	0	2223	0.00	13.37
Arrival Delay in Minutes	16573.0	15.51	40.65	0.0	0.0	0.0	13.0	1280.0	1652.41	13.0	8.02	138.24	52	2270	0.31	13.70



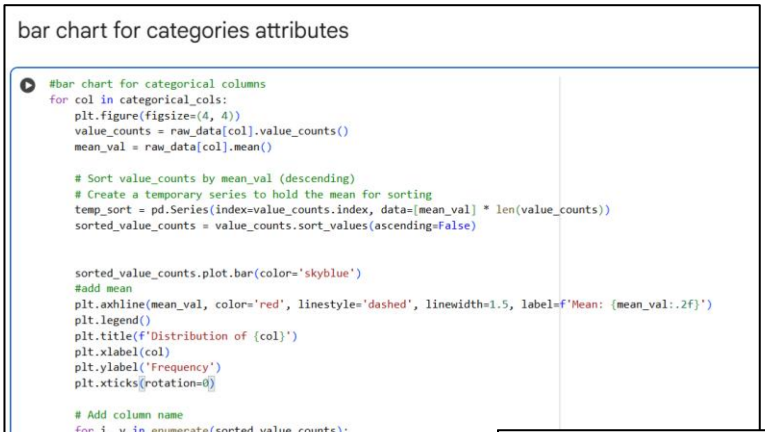
Histogram

```
n_cols = 2
n_rows = (len(numerical_cols) + n_cols - 1) // n_cols

fig, axes = plt.subplots(n_rows, n_cols, figsize=(15, n_rows * 5))
axes = axes.flatten() # Flatten the 2D array of axes for easy iteration

for i, col in enumerate(numerical_cols):
    ax = axes[i]
    ax.hist(raw_data[col].dropna(), bins=10, edgecolor='black', color='skyblue')
    median_val = raw_data[col].median()
    ax.axvline(median_val, color='red', linestyle='dashed', linewidth=1.5, label=f'Median: {median_val:.2f}')
    ax.set_title(f'Distribution of {col}')
    ax.set_xlabel('Value')
    ax.set_ylabel('Frequency')
    ax.legend()

# Hide any unused subplots
for j in range(i + 1, len(axes)):
    fig.delaxes(axes[j])

plt.tight_layout()
plt.show()
```



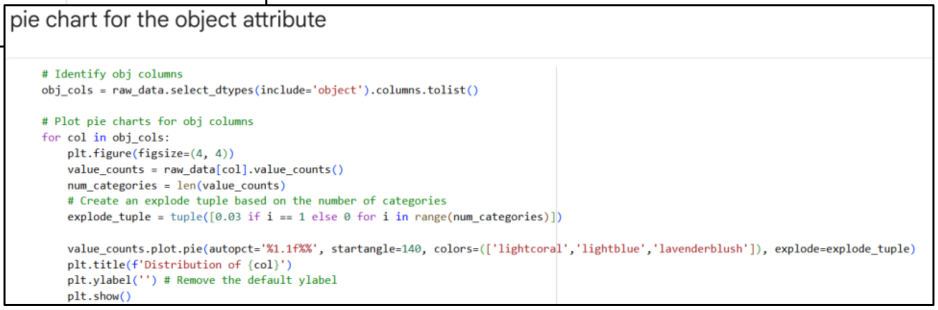
box plot for numerical attributes

```
# one plot for each attribute  except categorical columns using plotly
import plotly.express as px
for col in numerical_cols:
  fig = px.box(raw_data, x=col, orientation='h', color_discrete_sequence=['skyblue'])
  fig.update_layout(
      title=f"Box Plot of {col}",
      xaxis_title="Value",
      yaxis_title="Attribute",
      showlegend=False,
      height=400 # Adjusted height for better viewing of horizontal plots
  )
  fig.show()
```



bar chart for categories attributes

```
#bar chart for categorical columns
for col in categorical_cols:
    plt.figure(figsize=(4, 4))
    value_counts = raw_data[col].value_counts()
    mean_val = raw_data[col].mean()

    # Sort value_counts by mean_val (descending)
    # Create a temporary series to hold the mean for sorting
    temp_sort = pd.Series(index=value_counts.index, data=[mean_val] * len(value_counts))
    sorted_value_counts = value_counts.sort_values(ascending=False)


    sorted_value_counts.plot.bar(color='skyblue')
    #add mean
    plt.axhline(mean_val, color='red', linestyle='dashed', linewidth=1.5, label=f'Mean: {mean_val:.2f}')
    plt.legend()
    plt.title(f'Distribution of {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.xticks(rotation=0)

    # Add column name
    for i, v in enumerate(sorted_value_counts):
        plt.text(i, v, str(v), ha='center', va='bottom')
    plt.show()
```

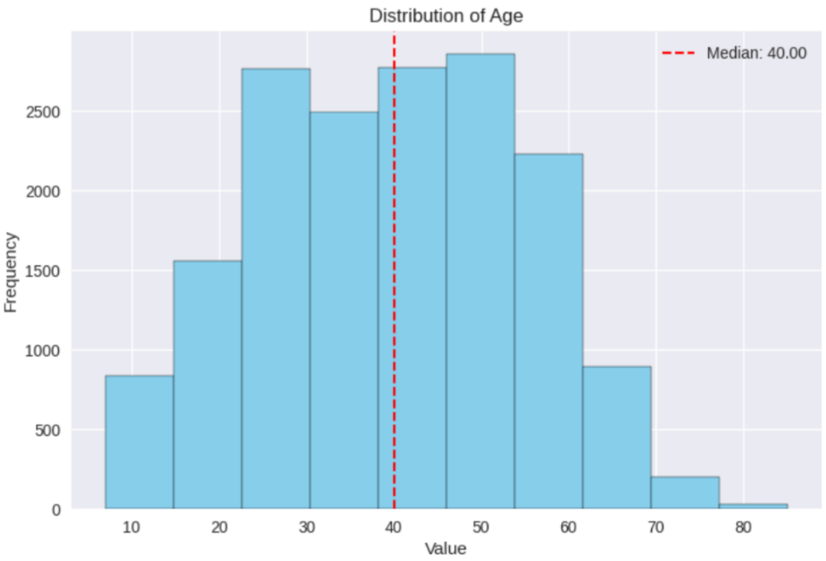


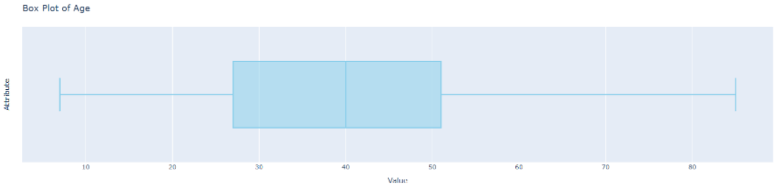
pie chart for the object attribute

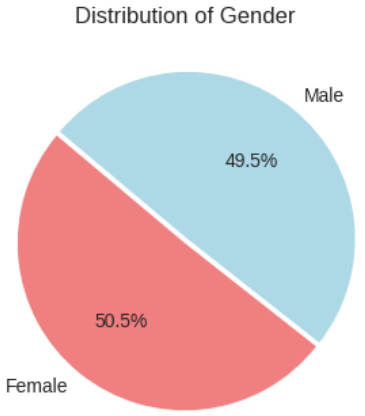
```
# Identify obj columns
obj_cols = raw_data.select_dtypes(include='object').columns.tolist()

# Plot pie charts for obj columns
for col in obj_cols:
    plt.figure(figsize=(4, 4))
    value_counts = raw_data[col].value_counts()
    num_categories = len(value_counts)
    # Create an explode tuple based on the number of categories
    explode_tuple = tuple([0.03 if i == 1 else 0 for i in range(num_categories)])

    value_counts.plot.pie(autopct='%1.1f%%', startangle=140, colors=(['lightcoral','lightblue','lavenderblush']), explode=explode_tuple)
    plt.title(f'Distribution of {col}')
    plt.ylabel('') # Remove the default ylabel
    plt.show()
```

2.1 Passenger Profile (Age, Gender)

Age has a mean of 39.55 years, spanning from 7 to 85. The standard deviation (15.12) and variance (228.63) indicate a moderate spread across the passenger population. The distribution is nearly symmetric (skewness = –0.02), with a median of 40 and an interquartile range (IQR) of 24, showing that most passengers fall in the young-to-middle-aged group.



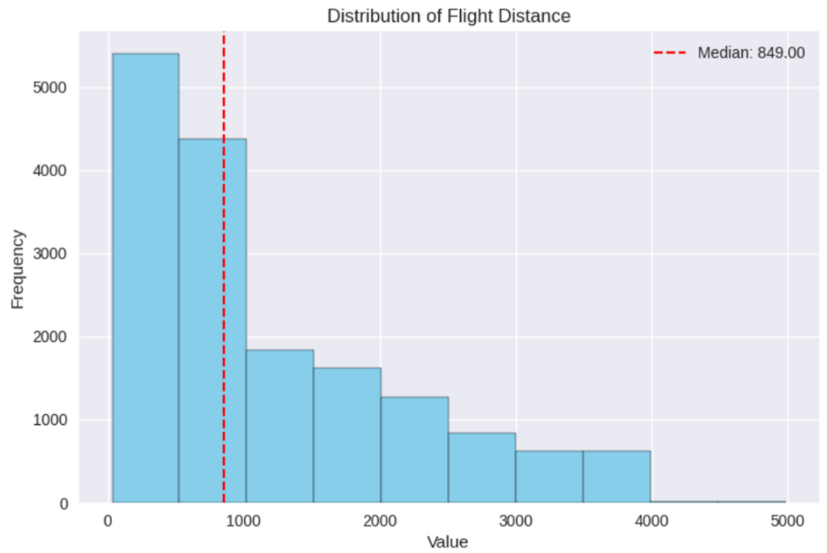


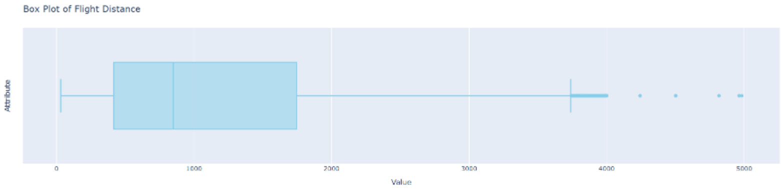


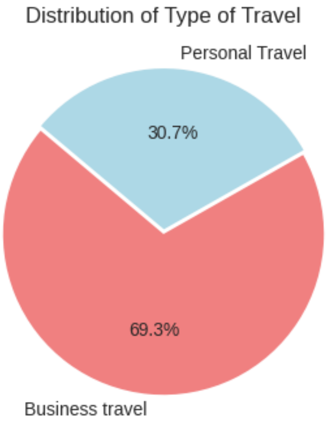
There are two categories for **Gender**: female (8397) and male (8228), with the former being slightly more feminine.

2.2 Travel Characteristics (Flight Distance, Type of Travel, Class)

Flight Distance shows an average of 1187.9 km, ranging from very short flights (31 km) to long-haul journeys (4983 km). The distribution is strongly right-skewed (skewness = 1.10), with a median of 849 km and IQR of 1330 km. The large variance (975,437.6) and standard deviation (987.64) reflect the wide variation in travel distances, with about 1.97% of cases identified as extreme long-haul outliers.

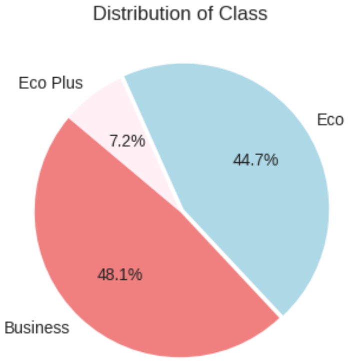






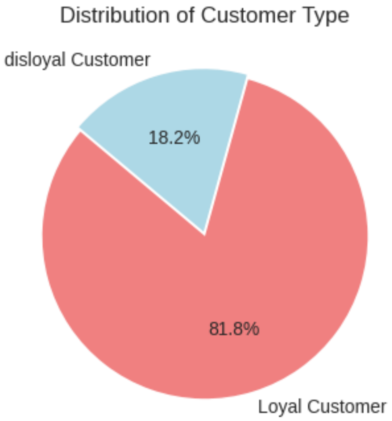
Personal Travel and Business Travel are the two main **Type of Travel**, with Business Travel dominating at 69.3%, meanwhile Personal Travel is 30.7%.

The attribute **Class** is categorized into three different categories: Eco (7435), Eco Plus (1194), and Business (7996), with occurrences of 44.7%, 7.2%, and 48.1% correspondingly, indicating that Business contributes the most.



2.3 Customer Type

Customer Type includes Loyal (13064) and Disloyal (3021), with Loyal distribution taking up 81.8% of the total, while customers who are not loyal represent only 18.2%.



2.4 Pre-flight & Ground Services (Ease of Online Booking, Check-in Service, Gate Location, Baggage Handling, Online Boarding)

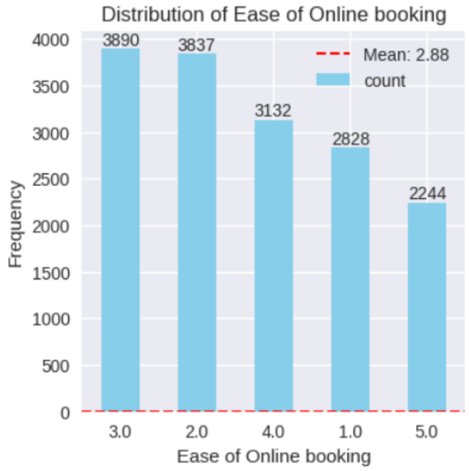
Passenger ratings for **Ease of Online Booking** show a core tendency towards 2 (3837) and 3 (3890), which account for over half of all responses. This pattern indicates a significant level of disagreement or ambivalence over the digital booking process. The small number of 5-star reviews demonstrates that, while some customers find the process efficient, many face usability concerns or challenges that limit satisfaction. The mean score is 2.88, suggesting that the booking experience is below average.

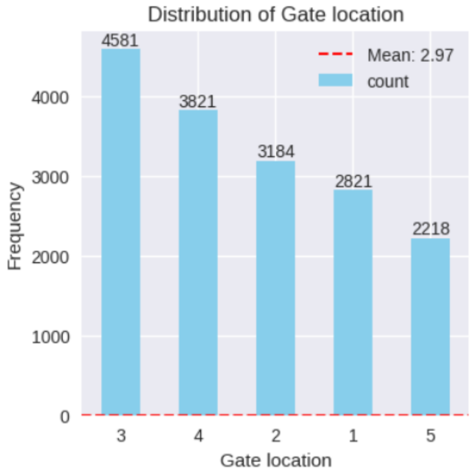
Gate Location has a high concentration of scores between 3 and 4 with average of 2.97 indicating neutral experiences. However, the relatively high number of 1-star evaluations (2,821) suggests that gate accessibility and convenience are major sources of annoyance for many travelers. This feature appears to be influenced by airport infrastructure as well as airline operations, making it a significant yet complex factor in customer happiness.

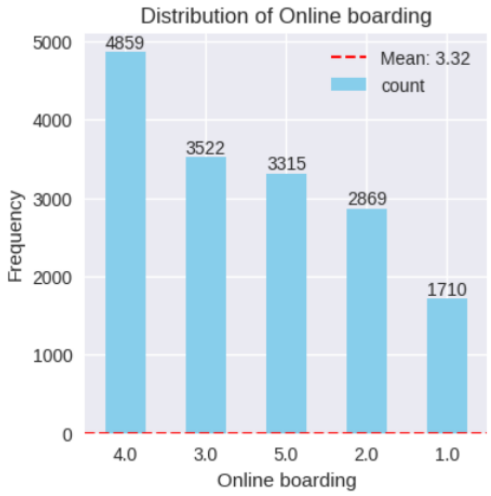
Online Boarding averages 3.32, with ratings from 1 to 5. This indicates that most passengers perceive the process as efficient and convenient. The relatively small number of low ratings suggests that issues in this area are rare, making it a strength in the customer journey.

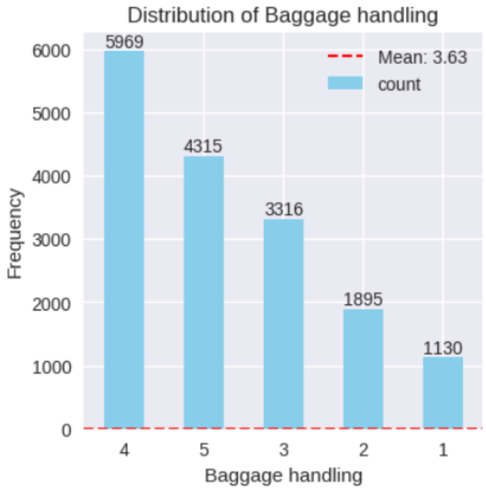
Baggage Handling is one of the strongest-performing operational areas, with an average of 3.63. Very few passengers reported dissatisfaction, indicating that the airline's baggage processes are reliable and efficient. This consistency makes baggage handling a significant strength in overall passenger experience

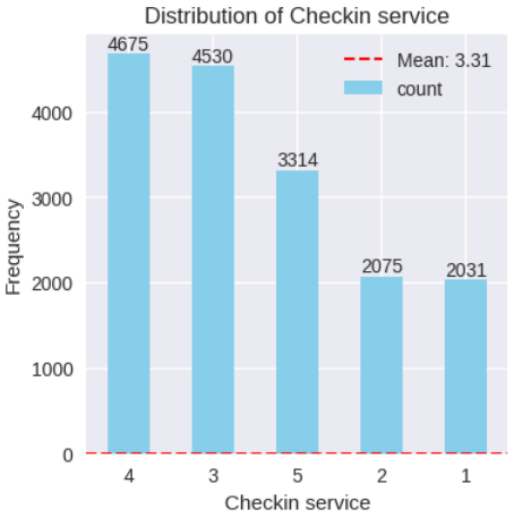
Check-in Service ratings average 3.31, with most passengers assigning values of 3 or 4. However, compared to other service dimensions, this attribute shows a wider spread, with substantial counts at both low (1–2) and high (5) ends. This reflects variability in passenger experience, potentially influenced by differences in staffing, queue management, or check-in technology.


Distribution of Ease of Online booking


Distribution of Gate location


Distribution of Online boarding


Distribution of Baggage handling


Distribution of Checkin service

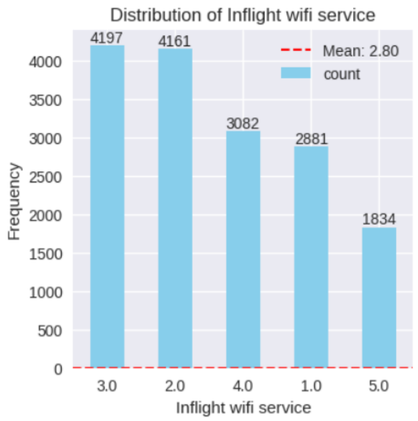
2.5 Inflight Services (Inflight Wi-Fi Service, Food and Drink, Seat Comfort, Inflight Entertainment, On-board Service, Leg Room Service, Inflight Service, Cleanliness)

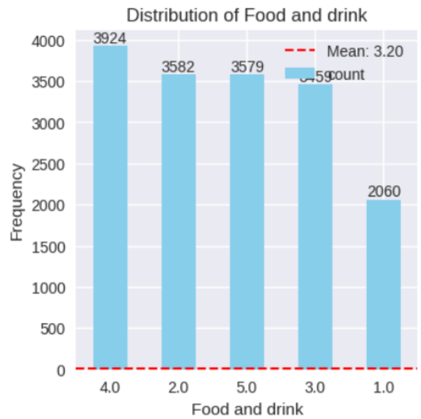
Inflight Wi-Fi Service is the weakest performing attribute, with most passengers rating it only 2 or 3, and the mean of 2.80, indicating dissatisfaction experiences. The relatively low number of 5-star ratings (1,834) suggests that only a minority of customers consider the service highly satisfactory.

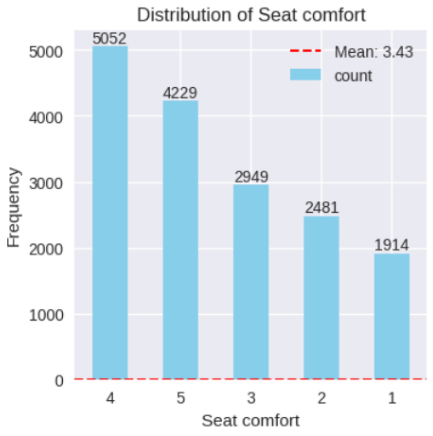
Food and Drink records a mean of 3.20. Many passengers provided ratings of 4 and 5, demonstrating that many travelers view the catering services positively. At the same time, a substantial portion rated food and drink poorly (1 or 2), which suggests inconsistency in quality, presentation, or availability. This variability points to a service area where improvements could have a strong impact on overall satisfaction

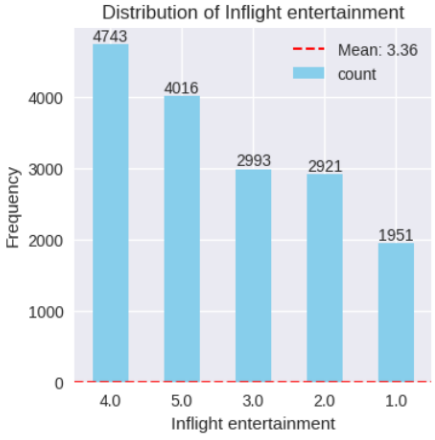
Seat Comfort shows relatively high ratings, with most passengers awarding 4 or 5. However, a significant proportion gave scores of 1 or 2, highlighting that discomfort remains a concern for a substantial minority of passengers. The average score of 3.43 suggests that, while comfort is satisfactory for many, it is still a significant source of disappointment for others.

Inflight Entertainment receives strong ratings, with the majority of passengers giving scores of 4 or 5. This suggests that the variety, quality, and accessibility of entertainment options meet or exceed expectations for many travelers. The relatively small share of negative ratings indicates that entertainment is a clear service strength. The mean score of 3.36 supports this conclusion, confirming a generally positive perception.







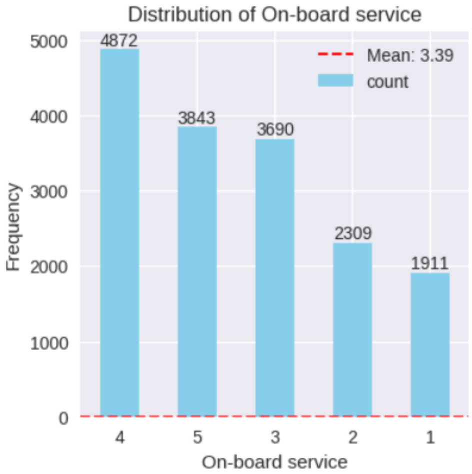


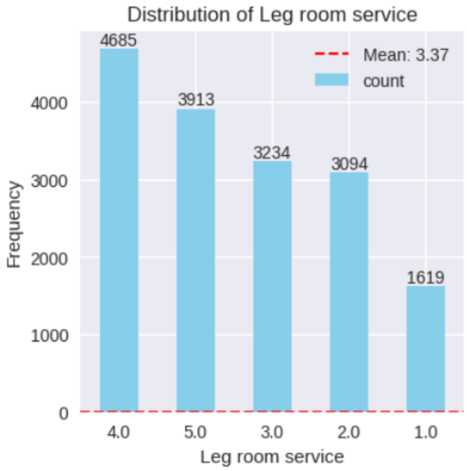
On-board Service emerges as a high-performing attribute, with most ratings in the 4–5 range and a mean of 3.39. This reflects positively on cabin crew interactions, responsiveness, and overall service quality. While some passengers remain dissatisfied, the distribution suggests that most customers experience consistent and professional service delivery.

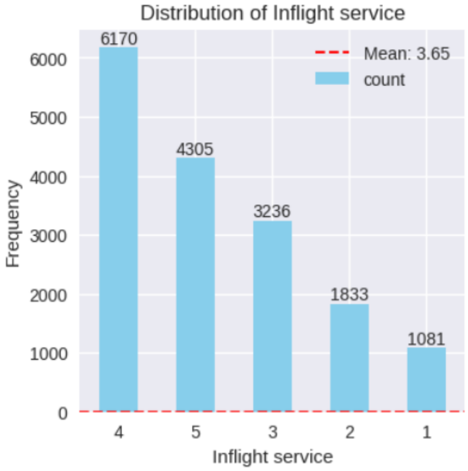
Leg Room Service exhibits a mean of 3.37. While many rated the attribute highly (4 and 5), a similarly large number rated it 3 or below. This indicates a persistent issue with space allocation, especially for economy-class passengers, and suggests that comfort is not experienced equally across different customer segments

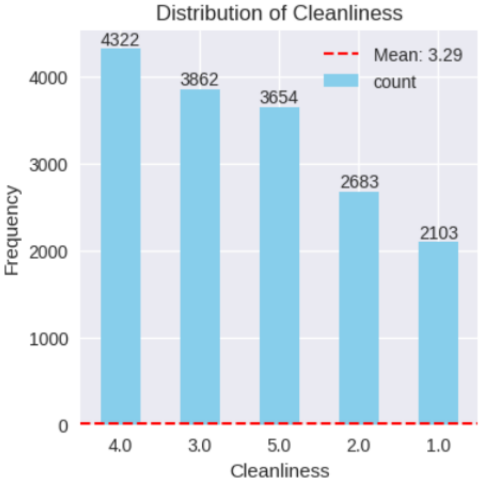
Inflight Service has the highest average among services, at 3.65, and most passengers rate it at 4 or 5. This suggests that the quality of assistance, hospitality, and attentiveness during flights is consistently high. Low dissatisfaction levels indicate this as a core strength of the airline, contributing positively to overall satisfaction.

Cleanliness receives a majority of positive ratings, with 4 and 5 making up the largest share. However, the relatively high frequency of low scores (1 and 2) suggests that service consistency may be an issue. While many passengers find the cabin environment clean and well-maintained, others perceive lapses, which may negatively affect perceptions of service quality. The mean score of 3.29 highlights this inconsistency, showing overall moderate satisfaction.


Distribution of On-board service

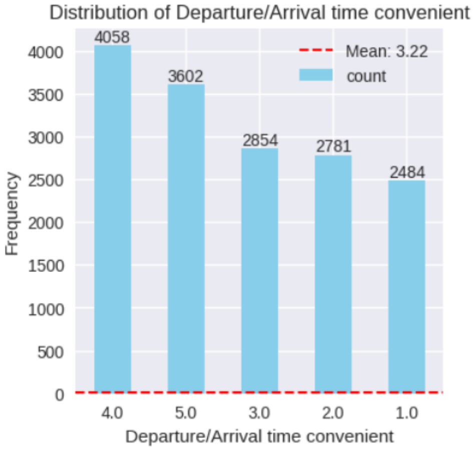

Distribution of Leg room service


Distribution of Inflight service


Distribution of Cleanliness

2.6 Schedule Convenience (Departure/Arrival time convenient)

Departure/Arrival Time Convenient exhibits a mean of 3.22, with the majority of passengers assigning 4 or 5. This suggests that scheduling is generally considered acceptable to very good. However, a significant proportion of ratings below 3 (approximately 30%) indicates that punctuality and timing convenience remain areas where passenger experiences vary considerably.

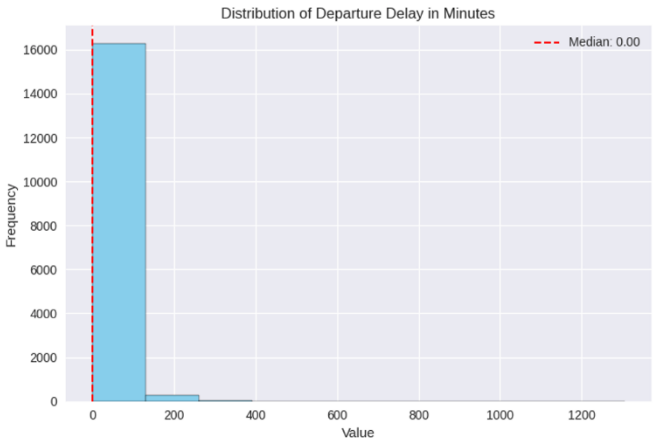


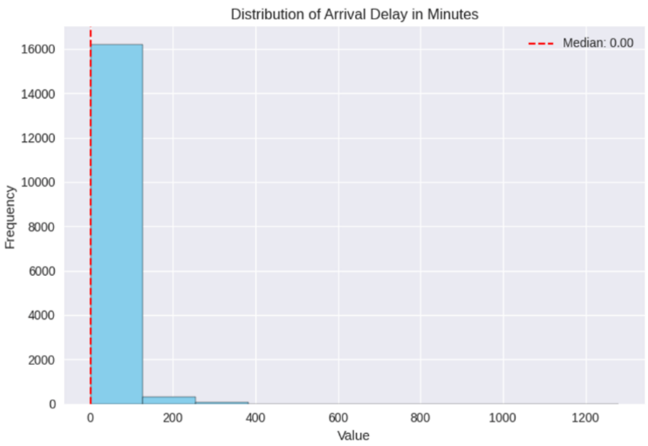
2.7 Departure Delay in Minutes, Arrival Delay in Minutes

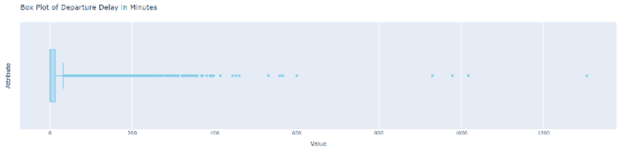
Departure Delay in Minutes and **Arrival Delay in Minutes** are both highly concentrated near zero, with most flights experiencing minimal delays. However, their distributions show long positive tails reflecting occasional extreme disruptions. The skewness is strongly positive, indicating that while most delays are short, a minority of very long delays extend the upper range. Importantly, both variables exhibit high kurtosis, meaning a sharp peak around zero combined with heavy tails. This highlights that delays are usually small, but when they occur, they can be a critical factor influencing passenger satisfaction.

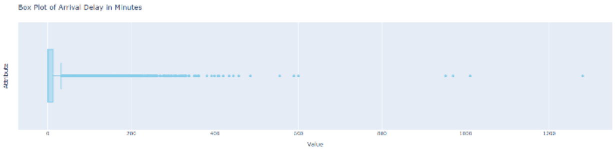
Departure Delay in Minutes averages 15.1 minutes, with values ranging from 0 to extreme cases of 1305 minutes. The distribution is heavily right-skewed (skewness = 8.25), with a median of 0 and IQR of 13, showing that most flights depart on time or with minimal delay, while a few experience very long delays. The large variance (1629.5) highlights the high variability.

Arrival Delay in Minutes has a mean of 15.5 minutes, spanning 0 to 1280 minutes. Similar to Departure Delay, the distribution is strongly right-skewed (skewness = 8.02), with a median of 0 and IQR of 13. The variance (1652.4) reflects high variability due to extreme outliers, though 0.3% of values are missing and 13.7% are flagged as outliers.

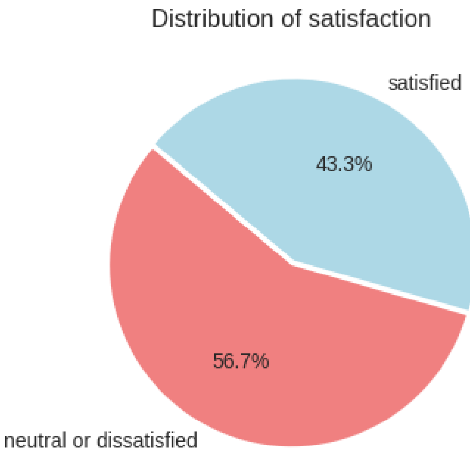








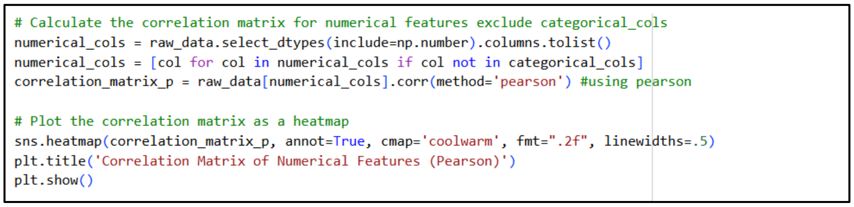
2.8 Satisfaction



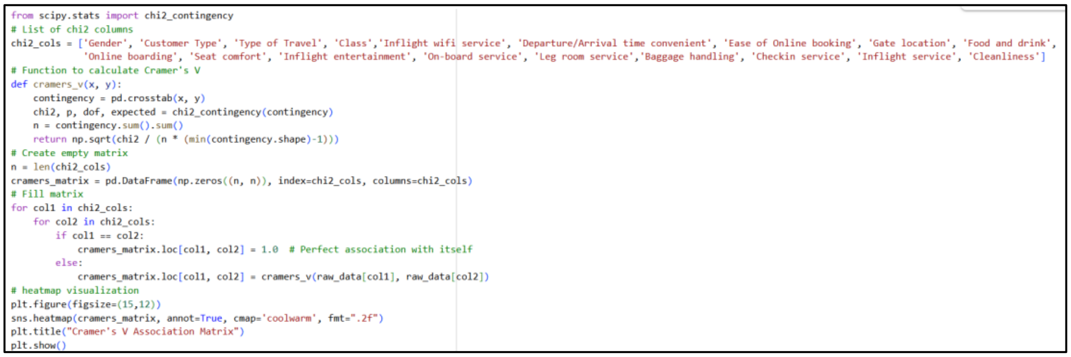
The metric sorted **Satisfaction** into two groups: Satisfied and Neutral or Dissatisfied. The first group had 43.3% (7204) of the total, and the second group had 56.7% (9421), which shows that Neutral or Dissatisfied was the most common group.

3. EXPLORE MULTIPLE ATTRIBUTES RELATIONSHIP OF THE DATASET

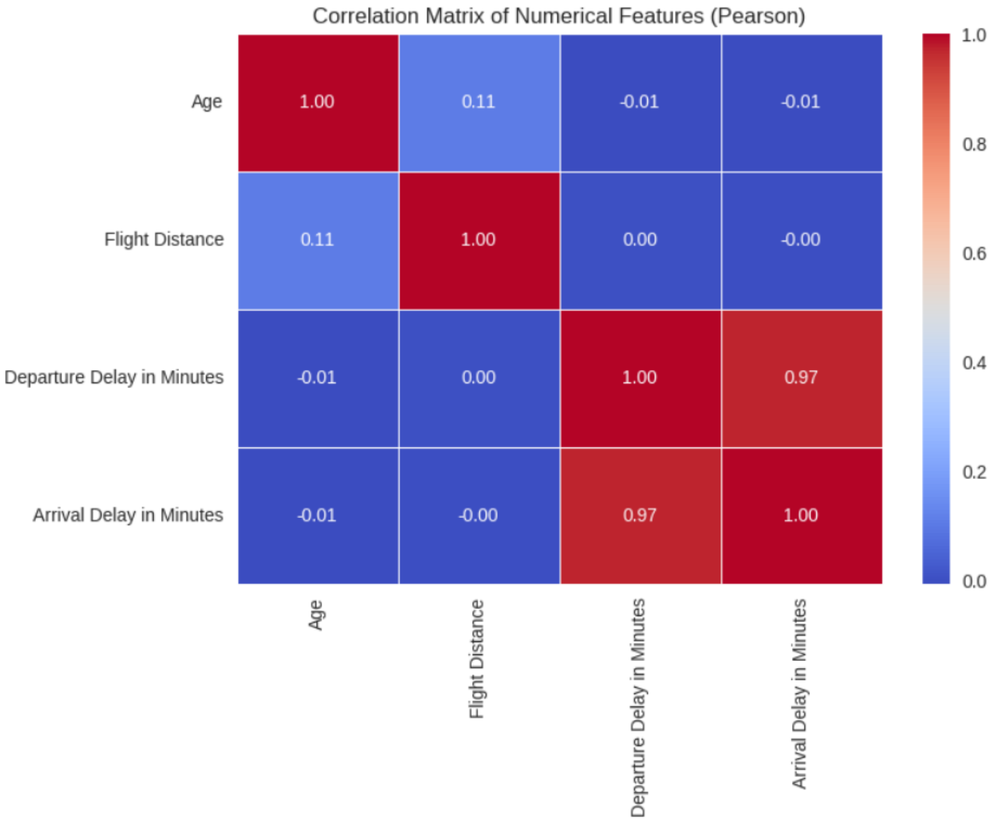
To explore relationships between attributes, we applied a dual approach. Numeric variables were analyzed using a correlation matrix, which highlights linear associations between service ratings, delays, and satisfaction. Categorical variables were instead tested with the Chi-Square test of independence to measure association strength and then visualize by a Cramér's V matrix to show how strong the relationship. This ensures each attribute type is handled with an appropriate statistical method, providing a more reliable view of which factors are correlated with passenger satisfaction.

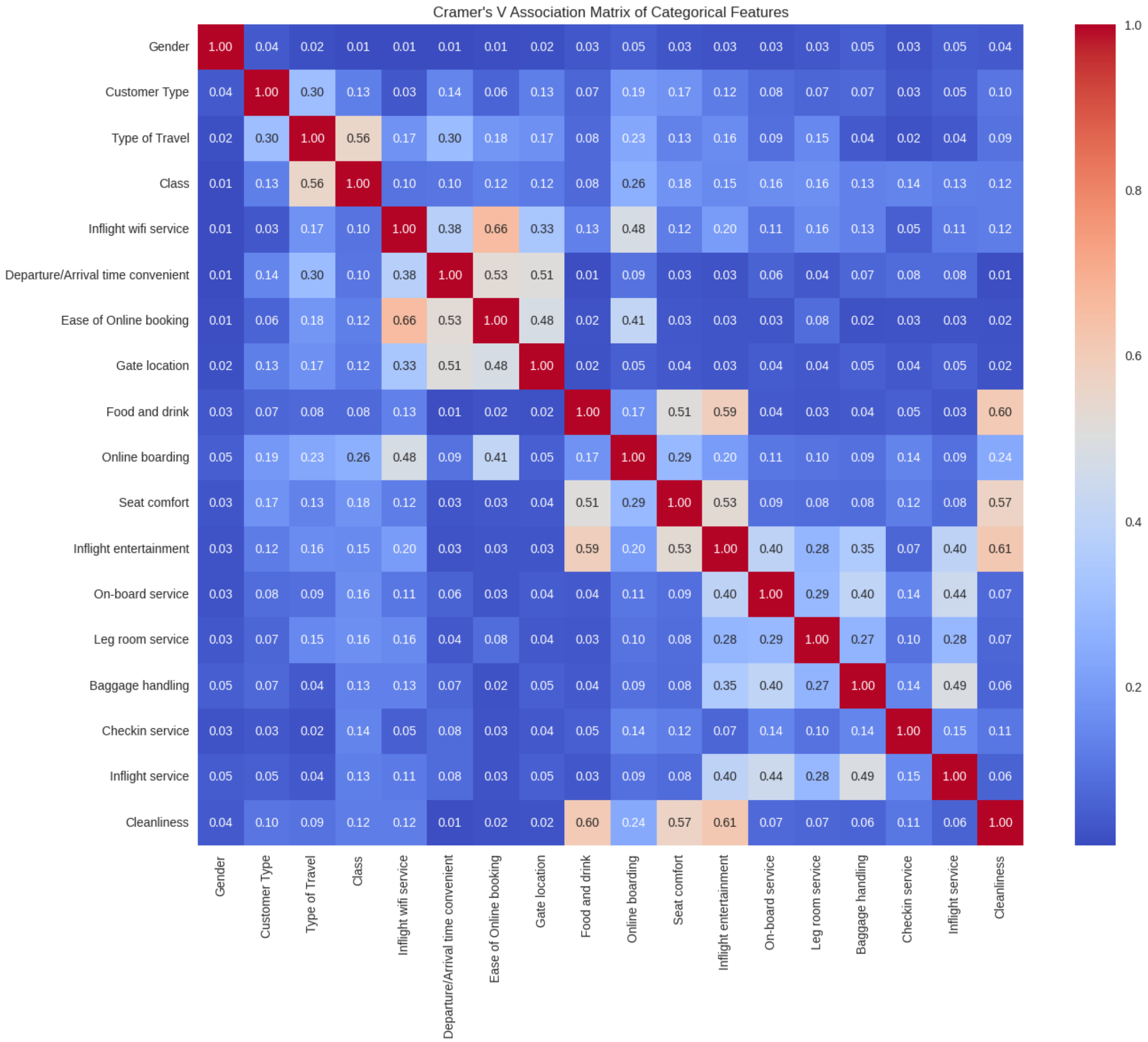


```
# Calculate the correlation matrix for numerical features exclude categorical_cols
numerical_cols = raw_data.select_dtypes(include=np.number).columns.tolist()
numerical_cols = [col for col in numerical_cols if col not in categorical_cols]
correlation_matrix_p = raw_data[numerical_cols].corr(method='pearson') #using pearson

# Plot the correlation matrix as a heatmap
sns.heatmap(correlation_matrix_p, annot=True, cmap='coolwarm', fmt=".2f", linewidths=.5)
plt.title('Correlation Matrix of Numerical Features (Pearson)')
plt.show()
```



```
from scipy.stats import chi2_contingency
# List of chi2 columns
chi2_cols = ['Gender', 'Customer Type', 'Type of Travel', 'Class','Inflight wifi service', 'Departure/Arrival time convenient', 'Ease of Online booking', 'Gate location', 'Food and drink',
             'Online boarding', 'Seat comfort', 'Inflight entertainment', 'On-board service', 'Leg room service','Baggage handling', 'Checkin service', 'Inflight service', 'Cleanliness']
# Function to calculate Cramer's V
def cramers_v(x, y):
    contingency = pd.crosstab(x, y)
    chi2, p, dof, expected = chi2_contingency(contingency)
    n = contingency.sum().sum()
    return np.sqrt(chi2 / (n * (min(contingency.shape)-1)))
# Create empty matrix
n = len(chi2_cols)
cramers_matrix = pd.DataFrame(np.zeros((n, n)), index=chi2_cols, columns=chi2_cols)
# Fill matrix
for col1 in chi2_cols:
    for col2 in chi2_cols:
        if col1 == col2:
            cramers_matrix.loc[col1, col2] = 1.0  # Perfect association with itself
        else:
            cramers_matrix.loc[col1, col2] = cramers_v(raw_data[col1], raw_data[col2])
# heatmap visualization
plt.figure(figsize=(15,12))
sns.heatmap(cramers_matrix, annot=True, cmap='coolwarm', fmt=".2f")
plt.title("Cramer's V Association Matrix")
plt.show()
```





Cramer's V Association Matrix of Categorical Features

	Gender	Customer Type	Type of Travel	Class	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink	Online boarding	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness
Gender	1.00	0.04	0.02	0.01	0.01	0.01	0.01	0.02	0.03	0.05	0.03	0.03	0.03	0.03	0.05	0.03	0.05	0.04
Customer Type	0.04	1.00	0.30	0.13	0.03	0.14	0.06	0.13	0.07	0.19	0.17	0.12	0.08	0.07	0.07	0.03	0.05	0.10
Type of Travel	0.02	0.30	1.00	0.56	0.17	0.30	0.18	0.17	0.08	0.23	0.13	0.16	0.09	0.15	0.04	0.02	0.04	0.09
Class	0.01	0.13	0.56	1.00	0.10	0.10	0.12	0.12	0.08	0.26	0.18	0.15	0.16	0.16	0.13	0.14	0.13	0.12
Inflight wifi service	0.01	0.03	0.17	0.10	1.00	0.38	0.66	0.33	0.13	0.48	0.12	0.20	0.11	0.16	0.13	0.05	0.11	0.12
Departure/Arrival time convenient	0.01	0.14	0.30	0.10	0.38	1.00	0.53	0.51	0.01	0.09	0.03	0.03	0.06	0.04	0.07	0.08	0.08	0.01
Ease of Online booking	0.01	0.06	0.18	0.12	0.66	0.53	1.00	0.48	0.02	0.41	0.03	0.03	0.03	0.08	0.02	0.03	0.03	0.02
Gate location	0.02	0.13	0.17	0.12	0.33	0.51	0.48	1.00	0.02	0.05	0.04	0.03	0.04	0.04	0.05	0.04	0.05	0.02
Food and drink	0.03	0.07	0.08	0.08	0.13	0.01	0.02	0.02	1.00	0.17	0.51	0.59	0.04	0.03	0.04	0.05	0.03	0.60
Online boarding	0.05	0.19	0.23	0.26	0.48	0.09	0.41	0.05	0.17	1.00	0.29	0.20	0.11	0.10	0.09	0.14	0.09	0.24
Seat comfort	0.03	0.17	0.13	0.18	0.12	0.03	0.03	0.04	0.51	0.29	1.00	0.53	0.09	0.08	0.08	0.12	0.08	0.57
Inflight entertainment	0.03	0.12	0.16	0.15	0.20	0.03	0.03	0.03	0.59	0.20	0.53	1.00	0.40	0.28	0.35	0.07	0.40	0.61
On-board service	0.03	0.08	0.09	0.16	0.11	0.06	0.03	0.04	0.04	0.11	0.09	0.40	1.00	0.29	0.40	0.14	0.44	0.07
Leg room service	0.03	0.07	0.15	0.16	0.16	0.04	0.08	0.04	0.03	0.10	0.08	0.28	0.29	1.00	0.27	0.10	0.28	0.07
Baggage handling	0.05	0.07	0.04	0.13	0.13	0.07	0.02	0.05	0.04	0.09	0.08	0.35	0.40	0.27	1.00	0.14	0.49	0.06
Checkin service	0.03	0.03	0.02	0.14	0.05	0.08	0.03	0.04	0.05	0.14	0.12	0.07	0.14	0.10	0.14	1.00	0.15	0.11
Inflight service	0.05	0.05	0.04	0.13	0.11	0.08	0.03	0.05	0.03	0.09	0.08	0.40	0.44	0.28	0.49	0.15	1.00	0.06
Cleanliness	0.04	0.10	0.09	0.12	0.12	0.01	0.02	0.02	0.60	0.24	0.57	0.61	0.07	0.07	0.06	0.11	0.06	1.00

3.1 Delay Time in Minutes

Departure Delay vs. **Arrival Delay (0.97, Robust Positive Correlation)**

- A nearly perfect linear relationship exists as departure delay increases, arrival delay also rises almost proportionally. Flights that depart late almost always arrive late as well, showing delays cascade forward.
- This operational inefficiency is one of the most direct drivers of dissatisfaction among passengers, especially for business and time-sensitive travelers.
- The scatter distribution would likely show points clustering tightly along the diagonal, reflecting consistency in delay propagation.

- Outliers may exist where long departure delays do not fully translate into arrival delays (e.g., flights making up time in the air) or where minor departure delays cascade into disproportionately longer arrival delays (e.g., congested airports, air traffic control issues).

This relationship highlights the operational dependency between departure punctuality and arrival performance. The near-perfect correlation (0.97) means passengers interpret departure punctuality as a signal of overall flight reliability. In other words, when flights leave late, passengers expect to arrive late, and their dissatisfaction increases accordingly. In contrast, guaranteeing departures on time can considerably increase satisfaction scores because it leads to punctual arrivals. Airlines should prioritize reducing departure delays as a key lever to improve overall satisfaction ratings.

3.2 Customer Classification

Class vs **Type of Travel** (0.56, Moderate-Strong association): This result suggests that passenger class is not independent of travel purpose. For example, business travelers dominate the population in the higher classes (Business/First), while leisure travelers are more commonly associated with Economy class. This relationship highlights how socio-economic segmentation and travel intent jointly shape passenger distribution across service classes. From a commercial perspective, the findings reinforce the idea that airlines can anticipate customer needs and expectations based on the interaction of class and travel type.

3.2 Digital and Logistic Convenience

This cluster includes **Inflight wifi service, Ease of Online booking, Online boarding, Departure/Arrival time convenient**. The attributes share strong associations, representing the digital and logistical aspects of travel.

- **Ease of Online Booking** vs. **Wifi** (0.66, Strong association), **Wifi vs Online Boarding** (0.48, Moderate association): Passengers satisfied with the online booking and online boarding process also report better wifi ratings, reflecting how digital services reinforce each other.

- **Ease of Online Booking** vs. **Online Boarding** (0.41, Moderate association): A smooth booking process often coincides with smoother boarding experiences.

- **Ease of Online booking** vs. **Departure/Arrival time convenient** (0.53, Moderate-Strong association): Customers who find booking easy also perceive flight schedules as more convenient, suggesting a effect where initial satisfaction colours later judgments

- **Gate location** vs. **Departure/Arrival time convenient** (0.51, Moderate-Strong association): Passengers who report convenient schedules also tend to view gate locations more favorably, indicating how logistical factors interact to shape perceptions of travel smoothness.

The digital and logistical aspects may have an interesting impact: satisfaction with one component of digital service or logistical effectiveness boosts opinions of others. This impact has two significant consequences. First, adjustments in user-friendly booking platforms and more clear scheduling information can improve perceptions of wifi and boarding. Second, dissatisfaction with one digital stage may influence perceptions of subsequent products, resulting in an especially negative customer experience.

3.4 Onboard Comfort & Cleanliness

This group comprises **Cleanliness, Inflight entertainment, Food and drink,** and **Seat comfort**. These attributes represent tangible, in-cabin experiences with some of the highest inter-attribute associations.

- **Cleanliness vs. Inflight Entertainment** (0.61, Strong): Cleaner cabins coincide with higher ratings for entertainment, potentially reflecting general satisfaction spillover or coordinated upgrades.
- **Cleanliness vs. Food & Drink** (0.60, Strong): High cleanliness ratings correlate with higher food ratings, suggesting hygiene impacts perceptions of catering.
- **Cleanliness vs. Seat Comfort** (0.57, Moderate-Strong): Cleanliness and comfort are naturally linked, as physical ergonomics drive passenger perception.

Cleanliness appears as a keystone variable in this cluster, with a considerable impact on evaluations of other in-cabin aspects. The substantial relationship of cleanliness, comfort, and entertainment demonstrates how travelers evaluate cabin quality holistically, rather than separately. For airlines, this means that expenditures in cleanliness norms might indirectly improve ratings for food, entertainment, and comfort. Similarly, changes in seat comfort and entertainment quality may amplify the benefit of cleanliness, resulting in a broader satisfaction effect.

TASK 1B. DATA PREPROCESSING

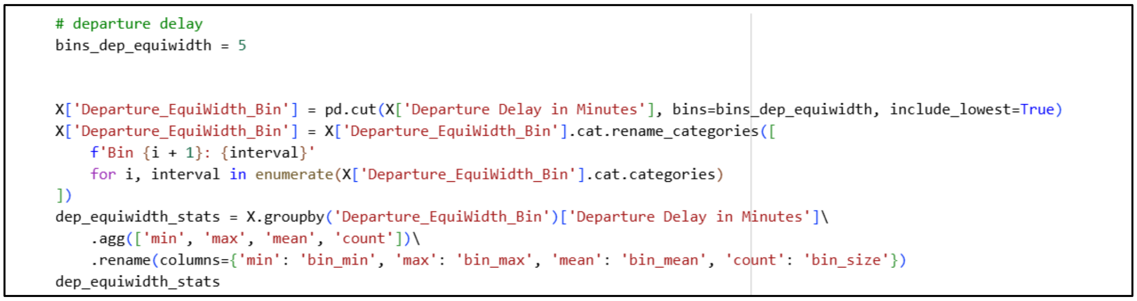
1. SMOOTH THE VALUES OF DEPARTURE DELAY IN MINUTES AND ARRIVAL DELAY IN MINUTES ATTRIBUTES

1.1 EQUI-WIDTH BINNING

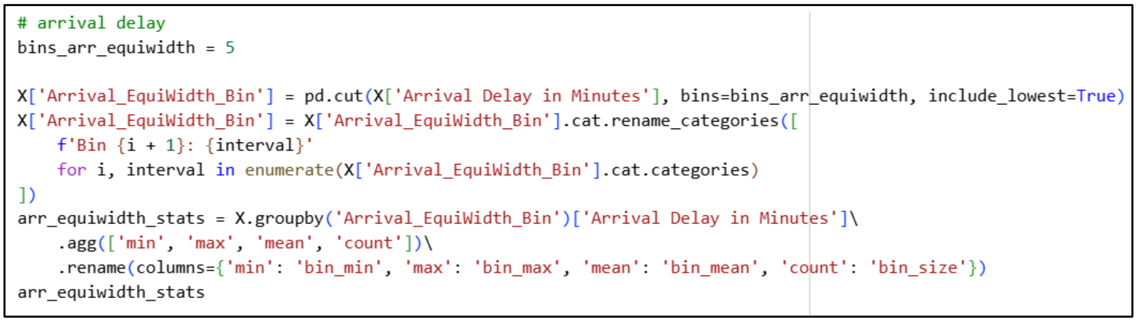
When applying Equi-width binning, both 10 bins and 5 bins have been tested. With 10 bins, the range of 0–1305 minutes of Departure Delay (0-1280 minutes for Arrival Delay) was divided into intervals of about 131 minutes each. Because the data is highly skewed toward small delays, many higher intervals were empty or had very few records, making the results sparse and less meaningful. In contrast, 5 bins produced broader intervals of about 261 minutes each, which reduced the number of empty bins and provided a clearer overall distribution. Therefore, **5 bins** is more suitable for Equi-width in this dataset.

a) Steps for Equi-width binning:

- *Step 1.* Define the number of bins: 05 for both Departure Delay in Minutes and Arrival Delay in Minutes.
- *Step 2.* To assign each observation to its proper bins, use the pandas.cut() function on each attribute together with the bins option.
- *Step 3.* Rename bins to include bin number and the interval.
- *Step 4.* Group the DataFrame by bins and calculate their statistics.



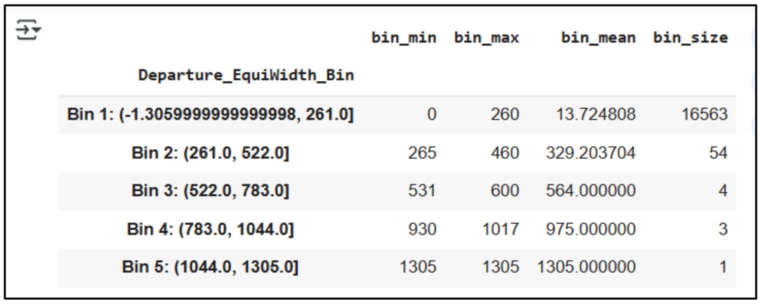
```
# departure delay
bins_dep_equiwidth = 5


X['Departure_EquiWidth_Bin'] = pd.cut(X['Departure Delay in Minutes'], bins=bins_dep_equiwidth, include_lowest=True)
X['Departure_EquiWidth_Bin'] = X['Departure_EquiWidth_Bin'].cat.rename_categories([
    f'Bin {i + 1}: {interval}'
    for i, interval in enumerate(X['Departure_EquiWidth_Bin'].cat.categories)
])
dep_equiwidth_stats = X.groupby('Departure_EquiWidth_Bin')['Departure Delay in Minutes']\
    .agg(['min', 'max', 'mean', 'count'])\
    .rename(columns={'min': 'bin_min', 'max': 'bin_max', 'mean': 'bin_mean', 'count': 'bin_size'})
dep_equiwidth_stats
```



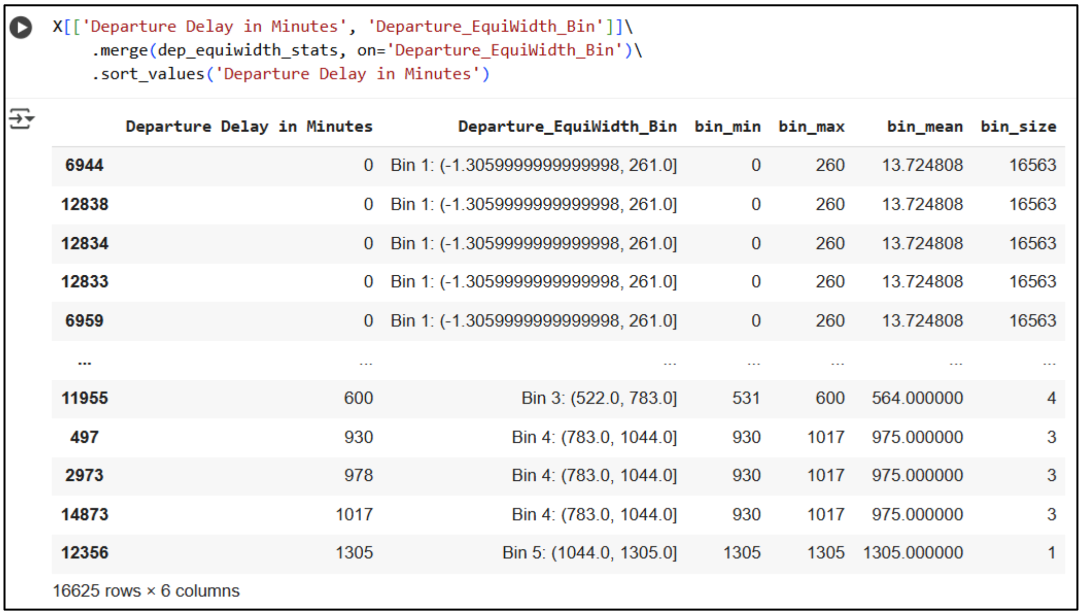
```
# arrival delay
bins_arr_equiwidth = 5

X['Arrival_EquiWidth_Bin'] = pd.cut(X['Arrival Delay in Minutes'], bins=bins_arr_equiwidth, include_lowest=True)
X['Arrival_EquiWidth_Bin'] = X['Arrival_EquiWidth_Bin'].cat.rename_categories([
    f'Bin {i + 1}: {interval}'
    for i, interval in enumerate(X['Arrival_EquiWidth_Bin'].cat.categories)
])
arr_equiwidth_stats = X.groupby('Arrival_EquiWidth_Bin')['Arrival Delay in Minutes']\
    .agg(['min', 'max', 'mean', 'count'])\
    .rename(columns={'min': 'bin_min', 'max': 'bin_max', 'mean': 'bin_mean', 'count': 'bin_size'})
arr_equiwidth_stats
```

b) Departure Delay in Minutes

For the **Departure Delay** attribute, Equi-width binning with 05 bins shows that nearly all flights are concentrated in the first interval between 0 and 260 minutes. This bin alone contains 16,563 of the 16,625 records, with an average delay of about 14 minutes. In contrast, the remaining bins capture only a very small portion of the data. Bins 2 to 4 contain just 61 records in total, while Bin 5 contains only one flight with the maximum recorded delay of 1305 minutes.



Departure_EquiWidth_Bin	bin_min	bin_max	bin_mean	bin_size
Bin 1: (-1.3059999999999998, 261.0]	0	260	13.724808	16563
Bin 2: (261.0, 522.0]	265	460	329.203704	54
Bin 3: (522.0, 783.0]	531	600	564.000000	4
Bin 4: (783.0, 1044.0]	930	1017	975.000000	3
Bin 5: (1044.0, 1305.0]	1305	1305	1305.000000	1



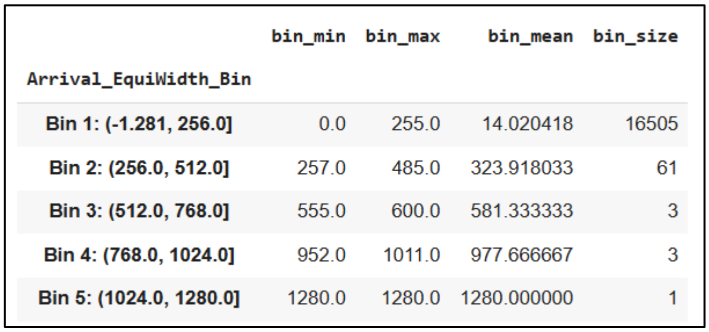
```
X[['Departure Delay in Minutes', 'Departure_EquiWidth_Bin']]\
    .merge(dep_equiwidth_stats, on='Departure_EquiWidth_Bin')\
    .sort_values('Departure Delay in Minutes')
```

	Departure Delay in Minutes	Departure_EquiWidth_Bin	bin_min	bin_max	bin_mean	bin_size
6944	0	Bin 1: (-1.3059999999999998, 261.0]	0	260	13.724808	16563
12838	0	Bin 1: (-1.3059999999999998, 261.0]	0	260	13.724808	16563
12834	0	Bin 1: (-1.3059999999999998, 261.0]	0	260	13.724808	16563
12833	0	Bin 1: (-1.3059999999999998, 261.0]	0	260	13.724808	16563
6959	0	Bin 1: (-1.3059999999999998, 261.0]	0	260	13.724808	16563
...	...	...	...	...	...	...
11955	600	Bin 3: (522.0, 783.0]	531	600	564.000000	4
497	930	Bin 4: (783.0, 1044.0]	930	1017	975.000000	3
2973	978	Bin 4: (783.0, 1044.0]	930	1017	975.000000	3
14873	1017	Bin 4: (783.0, 1044.0]	930	1017	975.000000	3
12356	1305	Bin 5: (1044.0, 1305.0]	1305	1305	1305.000000	1

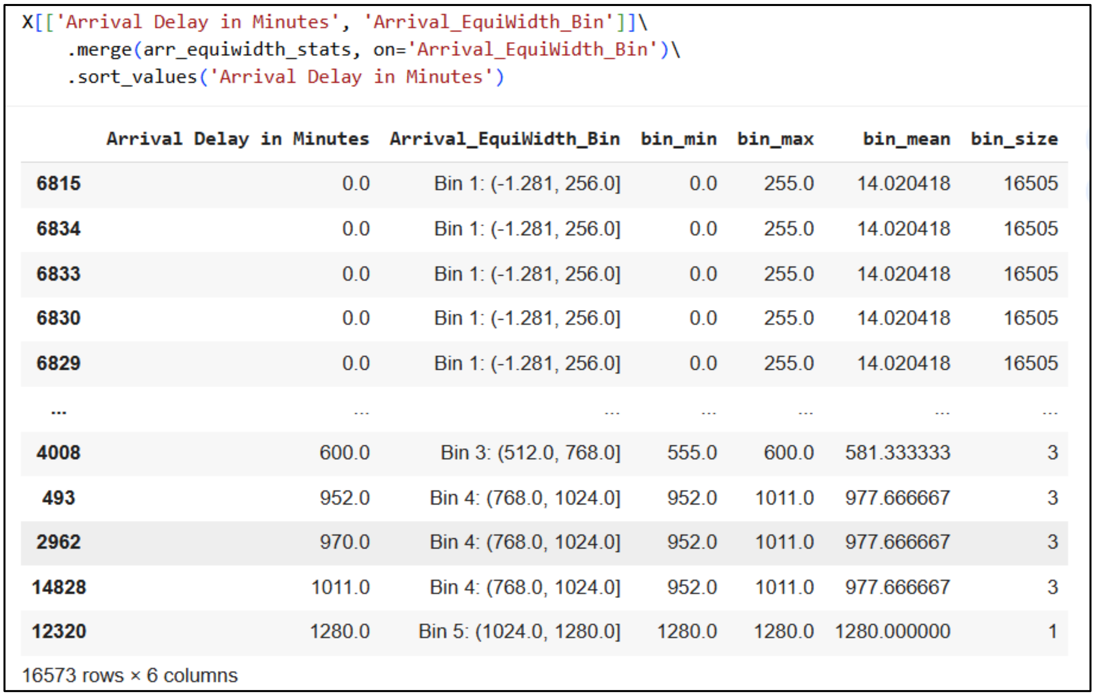
16625 rows × 6 columns

c) Arrival Delay in Minutes

For the **Arrival Delay** attribute, the pattern is almost identical. The first bin, covering 0 to 255 minutes, holds 16,505 of the 16,573 records with an average delay of about 14 minutes. The higher bins account for only 68 records, with Bins 2 to 4 containing 67 records while Bin 5 contains only a single record with a maximum delay of 1280 minutes.



Arrival_EquiWidth_Bin	bin_min	bin_max	bin_mean	bin_size
Bin 1: (-1.281, 256.0]	0.0	255.0	14.020418	16505
Bin 2: (256.0, 512.0]	257.0	485.0	323.918033	61
Bin 3: (512.0, 768.0]	555.0	600.0	581.333333	3
Bin 4: (768.0, 1024.0]	952.0	1011.0	977.666667	3
Bin 5: (1024.0, 1280.0]	1280.0	1280.0	1280.000000	1



```
X[['Arrival Delay in Minutes', 'Arrival_EquiWidth_Bin']]\
    .merge(arr_equiwidth_stats, on='Arrival_EquiWidth_Bin')\
    .sort_values('Arrival Delay in Minutes')
```

	Arrival Delay in Minutes	Arrival_EquiWidth_Bin	bin_min	bin_max	bin_mean	bin_size
6815	0.0	Bin 1: (-1.281, 256.0]	0.0	255.0	14.020418	16505
6834	0.0	Bin 1: (-1.281, 256.0]	0.0	255.0	14.020418	16505
6833	0.0	Bin 1: (-1.281, 256.0]	0.0	255.0	14.020418	16505
6830	0.0	Bin 1: (-1.281, 256.0]	0.0	255.0	14.020418	16505
6829	0.0	Bin 1: (-1.281, 256.0]	0.0	255.0	14.020418	16505
...	...	...	...	...	...	...
4008	600.0	Bin 3: (512.0, 768.0]	555.0	600.0	581.333333	3
493	952.0	Bin 4: (768.0, 1024.0]	952.0	1011.0	977.666667	3
2962	970.0	Bin 4: (768.0, 1024.0]	952.0	1011.0	977.666667	3
14828	1011.0	Bin 4: (768.0, 1024.0]	952.0	1011.0	977.666667	3
12320	1280.0	Bin 5: (1024.0, 1280.0]	1280.0	1280.0	1280.000000	1

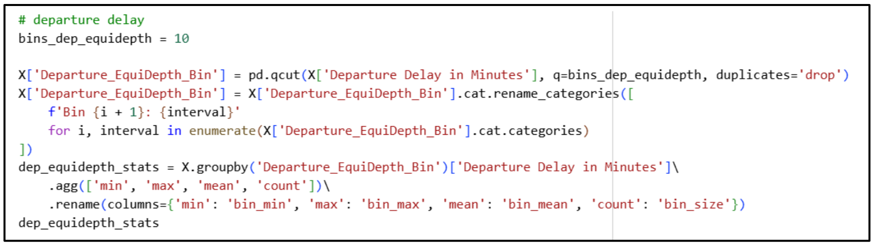
16573 rows × 6 columns

1.2 EQUI-DEPTH BINNING

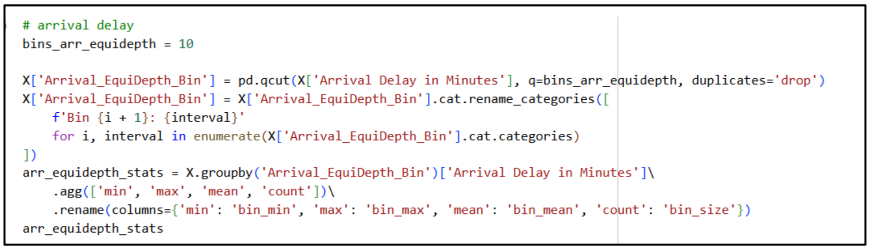
For the Equi-depth binning, **5 bins** have been decided to use. This choice is based on the distribution of data, where the first bin always contains the majority of observations due to many repeated small delay values. Regardless of how many bins that attempted to create, the first bin remains dominant while only the later bins are split further. Increasing the number of bins beyond 5 does not meaningfully improve interpretation but instead adds unnecessary complexity.

a) Steps for Equi-depth binning

- *Step 1.* Define the number of bins: 05 for both Departure Delay in Minutes and Arrival Delay in Minutes.
  (We initially attempted to code Equi-depth binning directly with 5 bins; however, the code did not run as intended because the high frequency of identical values at small delays caused overlapping quantile edges. This made it impossible for pandas to generate five distinct bins. As a solution, we specified 10 bins instead. With duplicates='drop', the overlapping edges were removed, and the result consistently collapsed into exactly 5 meaningful bins.)
- *Step 2.* To divide observations so that each bin has the same size, use the pandas.cut() function on each attribute with the choice of bins in the **q** parameter.
- *Step 3*. Rename bins to include bin number and the boundary.
- *Step 4.* Group the DataFrame by bins and calculate their statistics.



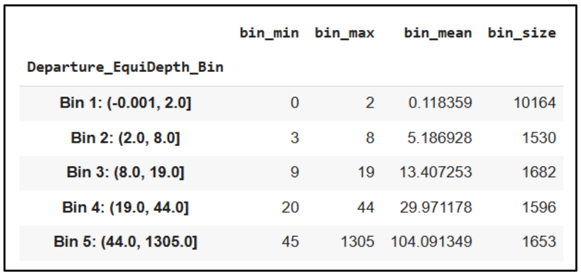
```
# departure delay
bins_dep_equidepth = 10

X['Departure_EquiDepth_Bin'] = pd.qcut(X['Departure Delay in Minutes'], q=bins_dep_equidepth, duplicates='drop')
X['Departure_EquiDepth_Bin'] = X['Departure_EquiDepth_Bin'].cat.rename_categories([
    f'Bin {i + 1}: {interval}'
    for i, interval in enumerate(X['Departure_EquiDepth_Bin'].cat.categories)
])
dep_equidepth_stats = X.groupby('Departure_EquiDepth_Bin')['Departure Delay in Minutes']\
    .agg(['min', 'max', 'mean', 'count'])\
    .rename(columns={'min': 'bin_min', 'max': 'bin_max', 'mean': 'bin_mean', 'count': 'bin_size'})
dep_equidepth_stats
```



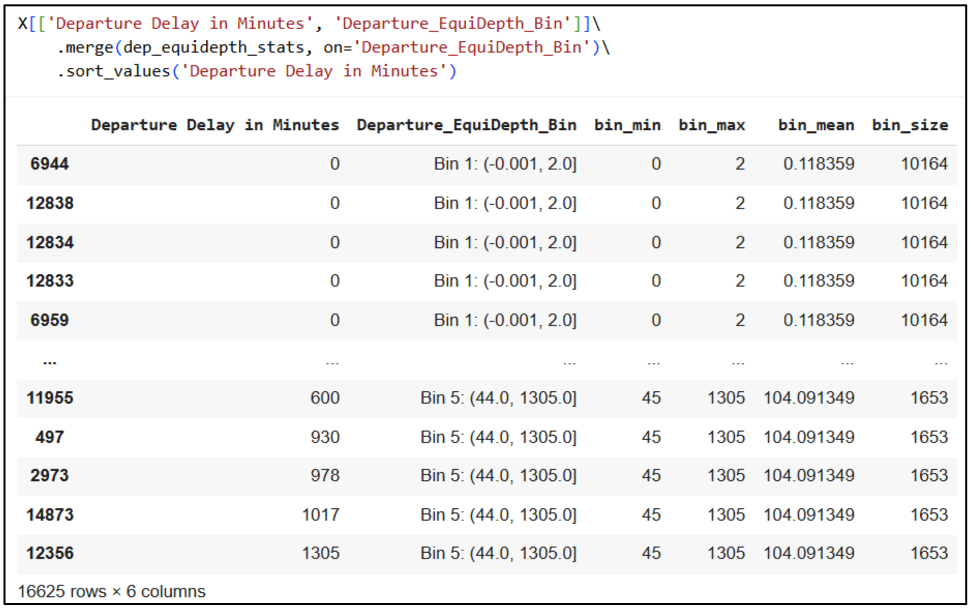
```
# arrival delay
bins_arr_equidepth = 10

X['Arrival_EquiDepth_Bin'] = pd.qcut(X['Arrival Delay in Minutes'], q=bins_arr_equidepth, duplicates='drop')
X['Arrival_EquiDepth_Bin'] = X['Arrival_EquiDepth_Bin'].cat.rename_categories([
    f'Bin {i + 1}: {interval}'
    for i, interval in enumerate(X['Arrival_EquiDepth_Bin'].cat.categories)
])
arr_equidepth_stats = X.groupby('Arrival_EquiDepth_Bin')['Arrival Delay in Minutes']\
    .agg(['min', 'max', 'mean', 'count'])\
    .rename(columns={'min': 'bin_min', 'max': 'bin_max', 'mean': 'bin_mean', 'count': 'bin_size'})
arr_equidepth_stats
```

b) Departure Delay in Minutes

The Equi-depth binning for departure delay shows a clear concentration of records in the lowest delay range. The first bin (0 - 2 minutes) dominates with 10,164 flights and a mean of 0.12 minutes, showing most departures are punctual. Bins covering 3 - 19 minutes (3,212 flights combined) highlight that minor delays are frequent but short. Larger delays occur less often. The 20 - 44 minutes bin contains 1,596 flights with a mean of 29.97 minutes. The final bin (45 - 1305 minutes) has 1,653 flights averaging 104 minutes, reflecting outliers with severe delays. Overall, departure delays are concentrated at the low end, with long delays being uncommon but impactful.



	bin_min	bin_max	bin_mean	bin_size
Departure_EquiDepth_Bin				
Bin 1: (-0.001, 2.0]	0	2	0.118359	10164
Bin 2: (2.0, 8.0]	3	8	5.186928	1530
Bin 3: (8.0, 19.0]	9	19	13.407253	1682
Bin 4: (19.0, 44.0]	20	44	29.971178	1596
Bin 5: (44.0, 1305.0]	45	1305	104.091349	1653



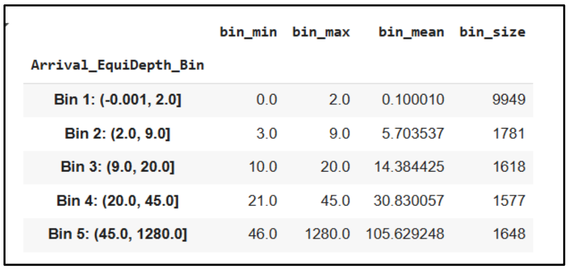
```
X[['Departure Delay in Minutes', 'Departure_EquiDepth_Bin']]\
    .merge(dep_equidepth_stats, on='Departure_EquiDepth_Bin')\
    .sort_values('Departure Delay in Minutes')
```

	Departure Delay in Minutes	Departure_EquiDepth_Bin	bin_min	bin_max	bin_mean	bin_size
6944	0	Bin 1: (-0.001, 2.0]	0	2	0.118359	10164
12838	0	Bin 1: (-0.001, 2.0]	0	2	0.118359	10164
12834	0	Bin 1: (-0.001, 2.0]	0	2	0.118359	10164
12833	0	Bin 1: (-0.001, 2.0]	0	2	0.118359	10164
6959	0	Bin 1: (-0.001, 2.0]	0	2	0.118359	10164
...	...	...	...	...	...	...
11955	600	Bin 5: (44.0, 1305.0]	45	1305	104.091349	1653
497	930	Bin 5: (44.0, 1305.0]	45	1305	104.091349	1653
2973	978	Bin 5: (44.0, 1305.0]	45	1305	104.091349	1653
14873	1017	Bin 5: (44.0, 1305.0]	45	1305	104.091349	1653
12356	1305	Bin 5: (44.0, 1305.0]	45	1305	104.091349	1653

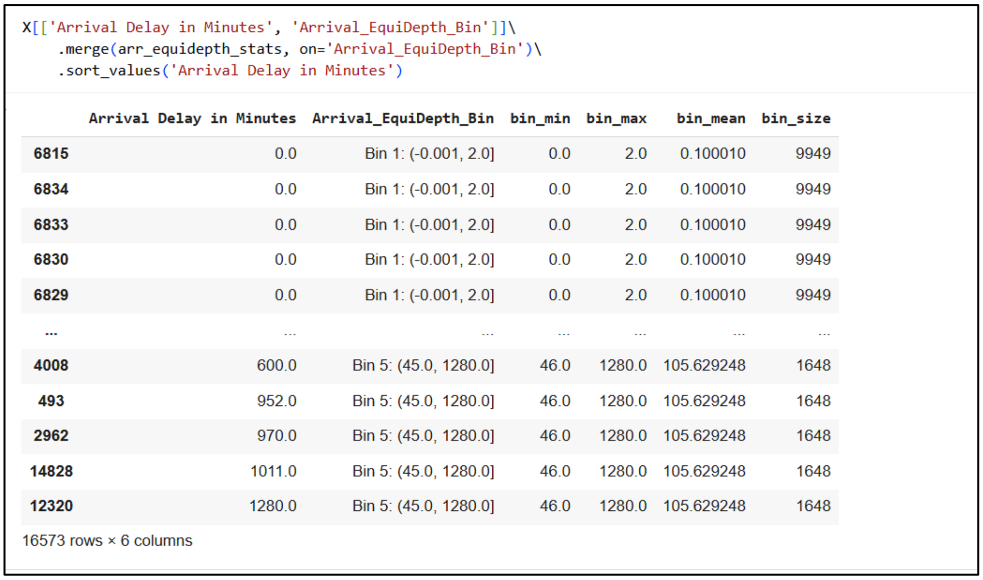
16625 rows × 6 columns

c) Arrival Delay in Minutes

Equi-depth binning for arrival delay shows most flights experience minimal delays. The first bin (0–2 minutes) holds 9,949 flights with a mean of 0.10 minutes, indicating many flights arrive on time. The next two bins (3–20 minutes) also contain high counts, showing short delays are common. Higher bins capture fewer but larger delays. The 21– 45 minutes bin includes 1,577 flights with a mean of 30.83 minutes. The final bin (46–1280 minutes) has 1,648 flights averaging 105.63 minutes, reflecting extreme cases. Overall, arrival delays are skewed toward small values, with rare but significant outliers.



	bin_min	bin_max	bin_mean	bin_size
Arrival_EquiDepth_Bin				
Bin 1: (-0.001, 2.0]	0.0	2.0	0.100010	9949
Bin 2: (2.0, 9.0]	3.0	9.0	5.703537	1781
Bin 3: (9.0, 20.0]	10.0	20.0	14.384425	1618
Bin 4: (20.0, 45.0]	21.0	45.0	30.830057	1577
Bin 5: (45.0, 1280.0]	46.0	1280.0	105.629248	1648

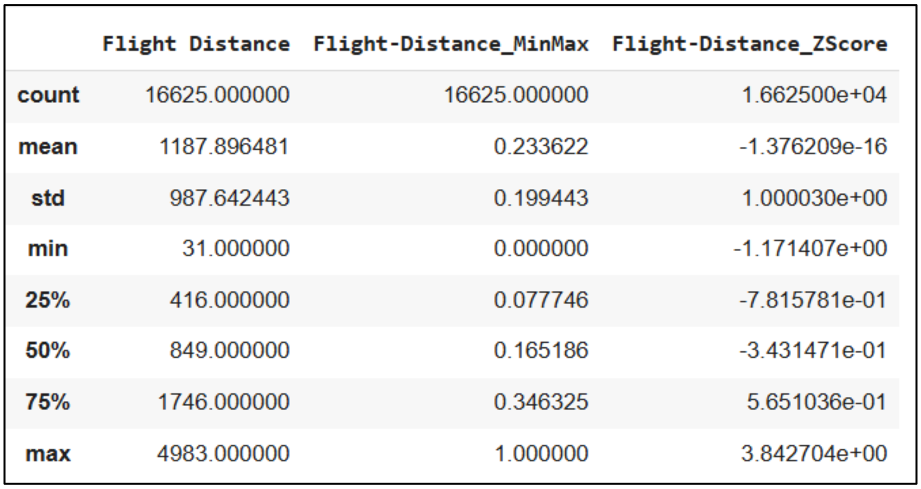


```
X[['Arrival Delay in Minutes', 'Arrival_EquiDepth_Bin']]\
    .merge(arr_equidepth_stats, on='Arrival_EquiDepth_Bin')\
    .sort_values('Arrival Delay in Minutes')
```

	Arrival Delay in Minutes	Arrival_EquiDepth_Bin	bin_min	bin_max	bin_mean	bin_size
6815	0.0	Bin 1: (-0.001, 2.0]	0.0	2.0	0.100010	9949
6834	0.0	Bin 1: (-0.001, 2.0]	0.0	2.0	0.100010	9949
6833	0.0	Bin 1: (-0.001, 2.0]	0.0	2.0	0.100010	9949
6830	0.0	Bin 1: (-0.001, 2.0]	0.0	2.0	0.100010	9949
6829	0.0	Bin 1: (-0.001, 2.0]	0.0	2.0	0.100010	9949
...	...	...	...	...	...	...
4008	600.0	Bin 5: (45.0, 1280.0]	46.0	1280.0	105.629248	1648
493	952.0	Bin 5: (45.0, 1280.0]	46.0	1280.0	105.629248	1648
2962	970.0	Bin 5: (45.0, 1280.0]	46.0	1280.0	105.629248	1648
14828	1011.0	Bin 5: (45.0, 1280.0]	46.0	1280.0	105.629248	1648
12320	1280.0	Bin 5: (45.0, 1280.0]	46.0	1280.0	105.629248	1648

16573 rows × 6 columns

2. NORMALIZE THE ATTRIBUTE FLIGHT DISTANCE



	Flight Distance	Flight-Distance_MinMax	Flight-Distance_ZScore
count	16625.000000	16625.000000	1.662500e+04
mean	1187.896481	0.233622	-1.376209e-16
std	987.642443	0.199443	1.000030e+00
min	31.000000	0.000000	-1.171407e+00
25%	416.000000	0.077746	-7.815781e-01
50%	849.000000	0.165186	-3.431471e-01
75%	1746.000000	0.346325	5.651036e-01
max	4983.000000	1.000000	3.842704e+00

2.1 MIN-MAX NORMALIZATION

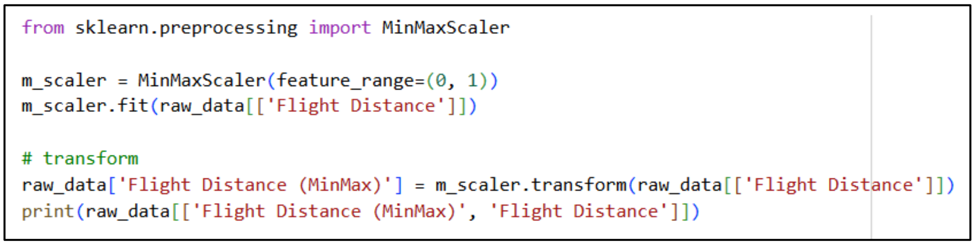
Min-Max normalization rescales values to a fixed range [0.0, 1.0]. The minimum and maximum values are transformed to 0 and 1, respectively. It works well for bounded features but is sensitive to outliers.

Steps for Min–Max Normalization (Flight Distance)

- *Step 1.* Identify the minimum (X_{min}) and maximum (X_{max}) values of the *Flight Distance* column.
- *Step 2.* Apply the Min–Max formula:

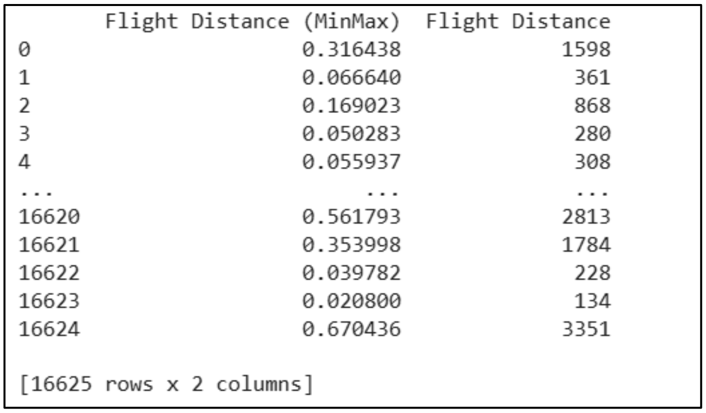
$$X' = \frac{X - X_{min}}{X_{max} - X_{min}}$$

or use sklearn.preprocessing.MinMaxScaler to rescale values into the [0,1] range.



```
from sklearn.preprocessing import MinMaxScaler

m_scaler = MinMaxScaler(feature_range=(0, 1))
m_scaler.fit(raw_data[['Flight Distance']])

# transform
raw_data['Flight Distance (MinMax)'] = m_scaler.transform(raw_data[['Flight Distance']])
print(raw_data[['Flight Distance (MinMax)', 'Flight Distance']])
```

After applying Min–Max normalization to the **Flight Distance** attribute, the transformed values were rescaled into the range 0 to 1, with the minimum flight distance mapped to 0 and the maximum mapped to 1. This preserves the relative spacing between observations while bounding the scale.



```
       Flight Distance (MinMax)  Flight Distance
0                      0.316438             1598
1                      0.066640              361
2                      0.169023              868
3                      0.050283              280
4                      0.055937              308
...                         ...              ...
16620                  0.561793             2813
16621                  0.353998             1784
16622                  0.039782              228
16623                  0.020800              134
16624                  0.670436             3351

[16625 rows x 2 columns]
```

2.2 Z-SCORE NORMALIZATION

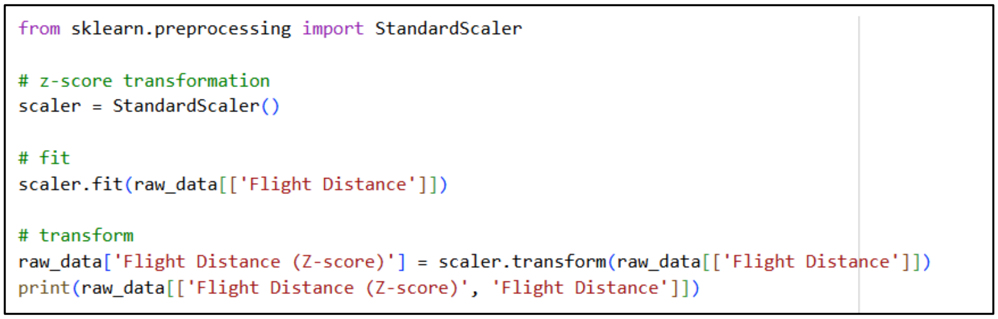
Z-score normalization (Standardization) centers the data at 0 and scales to unit variance (mean ≈ 0, standard deviation ≈ 1). It works well for features with comparable variance in models that assume centered data (such as linear models).

Steps for Z-score Normalization (Flight Distance)

- *Step 1*. Calculate the mean (μ) and standard deviation (σ) of the *Flight Distance* column.
- *Step 2.* Apply the Z-score formula:

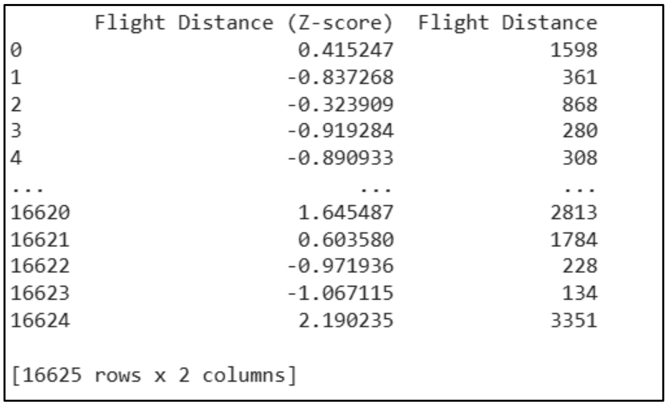
$$Z = X' = \frac{X - \mu}{\sigma}$$

or use sklearn.preprocessing.StandardScaler to standardize values with mean ≈ 0 and std ≈ 1.



```
from sklearn.preprocessing import StandardScaler

# z-score transformation
scaler = StandardScaler()

# fit
scaler.fit(raw_data[['Flight Distance']])

# transform
raw_data['Flight Distance (Z-score)'] = scaler.transform(raw_data[['Flight Distance']])
print(raw_data[['Flight Distance (Z-score)', 'Flight Distance']])
```

After applying Z-score normalization, the **Flight Distance** attribute was standardized so that the transformed values have a mean close to 0 and a standard deviation close to 1. This centers the distribution and makes flight distances directly comparable with other standardized variables.



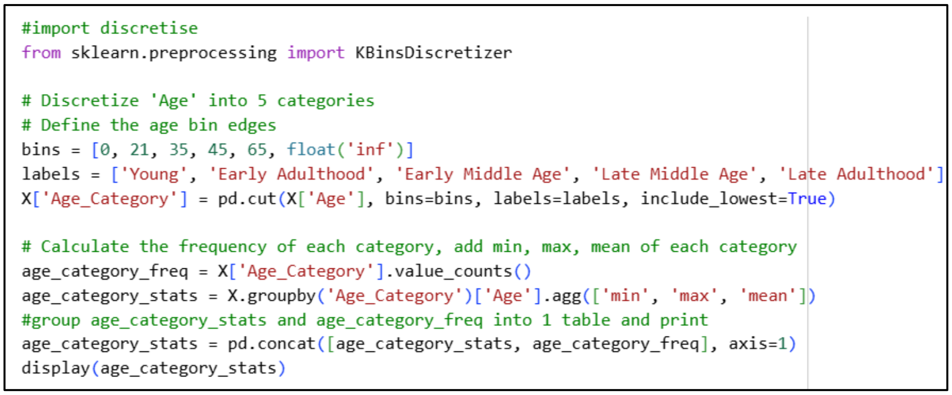
```
       Flight Distance (Z-score)  Flight Distance
0                       0.415247             1598
1                      -0.837268              361
2                      -0.323909              868
3                      -0.919284              280
4                      -0.890933              308
...                          ...              ...
16620                   1.645487             2813
16621                   0.603580             1784
16622                  -0.971936              228
16623                  -1.067115              134
16624                   2.190235             3351

[16625 rows x 2 columns]
```

3. DISCRETISE THE AGE ATTRIBUTE INTO 5 CATEGORIES

Discretization converts continuous data into discrete bins. This technique simplifies complex data, making it easier to analyze and prepare for category ML algorithms. Discretising age groups enhances understanding and enables the capture of non-linear correlations with satisfaction.

Steps for Discretise (the Age attribute)

- *Step 1*. Define Categories: Young (ages ≤ 21), Early Adulthood (22–34), Early Middle Age (35–44), Late Middle Age (45–64), Late Adulthood (≥ 65).
- *Step 2*. Assign Values: Map each Age value to its category using the bin edges [0, 21, 35, 45, 65, inf] (e.g., with pd.cut(..., include_lowest=True)), producing a new Age_Category column.
- *Step 3*. Statistics: Group by Age_Category and compute counts and summary stats (min, max, mean) to understand distribution by life stage in the dataset.



```
#import discretise
from sklearn.preprocessing import KBinsDiscretizer

# Discretize 'Age' into 5 categories
# Define the age bin edges
bins = [0, 21, 35, 45, 65, float('inf')]
labels = ['Young', 'Early Adulthood', 'Early Middle Age', 'Late Middle Age', 'Late Adulthood']
X['Age_Category'] = pd.cut(X['Age'], bins=bins, labels=labels, include_lowest=True)

# Calculate the frequency of each category, add min, max, mean of each category
age_category_freq = X['Age_Category'].value_counts()
age_category_stats = X.groupby('Age_Category')['Age'].agg(['min', 'max', 'mean'])
#group age_category_stats and age_category_freq into 1 table and print
age_category_stats = pd.concat([age_category_stats, age_category_freq], axis=1)
display(age_category_stats)
```

Age is discretized into 5 categories with the following frequencies:

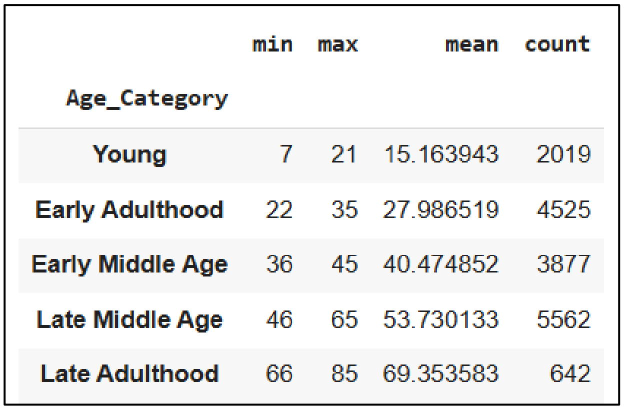
- ***Late Middle Age*** (5,562 passengers, largest group): Represents the dominant segment of the dataset, showing that most travelers are between 46–65 years old, a demographic likely reflecting established working professionals or frequent flyers.

- ***Early Adulthood*** (4,525 passengers): A substantial group of younger adults (22–34), potentially including students, young professionals, and early-career travelers.

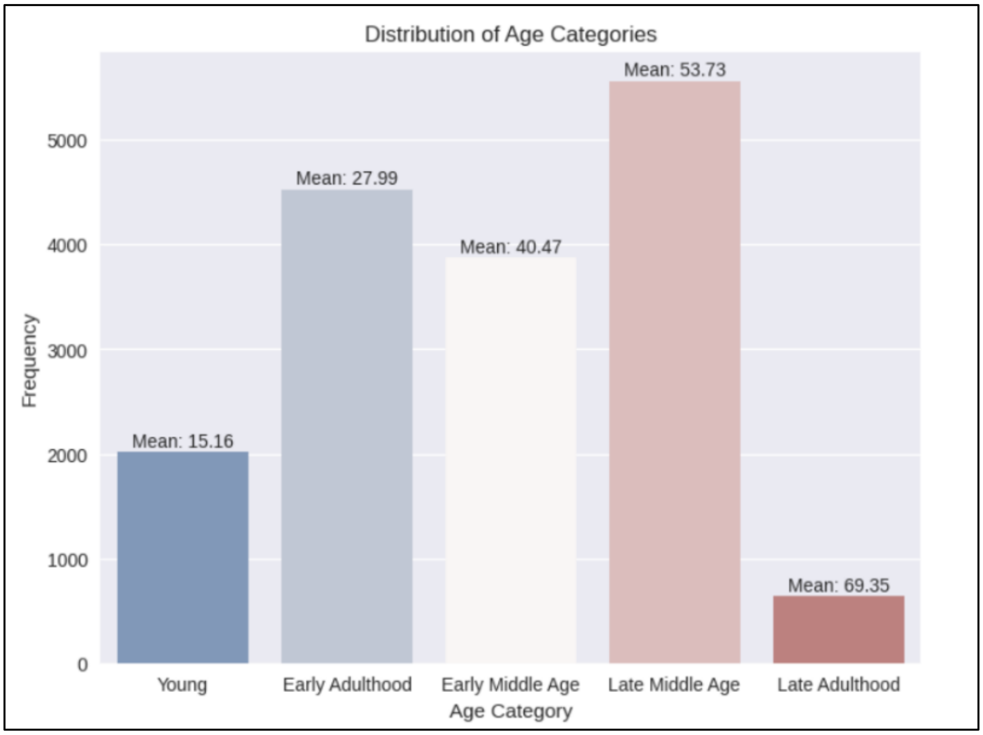
- ***Early Middle Age*** (3,877 passengers): Midlife passengers (35–44), forming another sizeable segment, consistent with the distribution of working-age travelers.

- ***Young*** (2,019 passengers): The smallest of the main age groups, representing passengers aged 21 and under. This suggests younger travelers form a minority in the dataset.

- ***Late Adulthood*** (642 passengers, Fewest overall): A relatively rare segment of older adults (65+), which is expected given lower air travel frequency among senior passengers.



	min	max	mean	count
Age_Category				
Young	7	21	15.163943	2019
Early Adulthood	22	35	27.986519	4525
Early Middle Age	36	45	40.474852	3877
Late Middle Age	46	65	53.730133	5562
Late Adulthood	66	85	69.353583	642

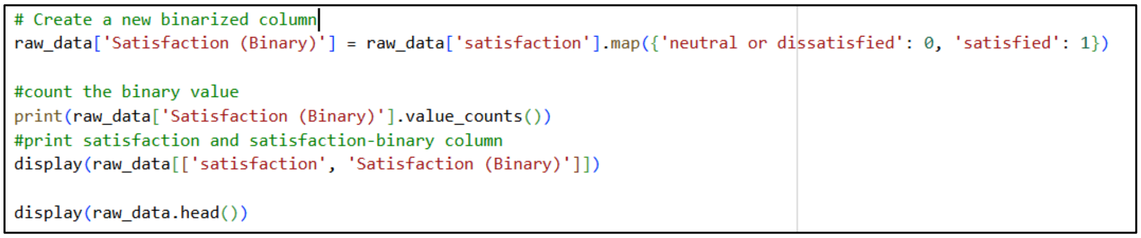


4. BINARISE THE SATISFACTION VARIABLE

Binarisation turns the target into a clear classification problem (satisfied = 1 vs neutral/dissatisfied = 0), simplifying model selection and interpretation. It improves actionable insights (who is satisfied vs not).

Steps for Binarise the Satisfaction attribute

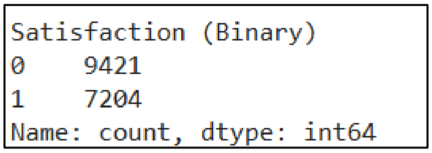
- *Step 1*. Create a new binary column: Map "neutral or dissatisfied" → 0 and "satisfied" → 1 using .map().
- *Step 2.* Count values: Use value_counts() to check the frequency of 0 and 1 in the new column.
- *Step 3.* Verify mapping: Display both the original satisfaction column and the new Satisfaction (Binary) column side by side to ensure correct transformation.



```
# Create a new binarized column
raw_data['Satisfaction (Binary)'] = raw_data['satisfaction'].map({'neutral or dissatisfied': 0, 'satisfied': 1})

#count the binary value
print(raw_data['Satisfaction (Binary)'].value_counts())
#print satisfaction and satisfaction-binary column
display(raw_data[['satisfaction', 'Satisfaction (Binary)']])

display(raw_data.head())
```

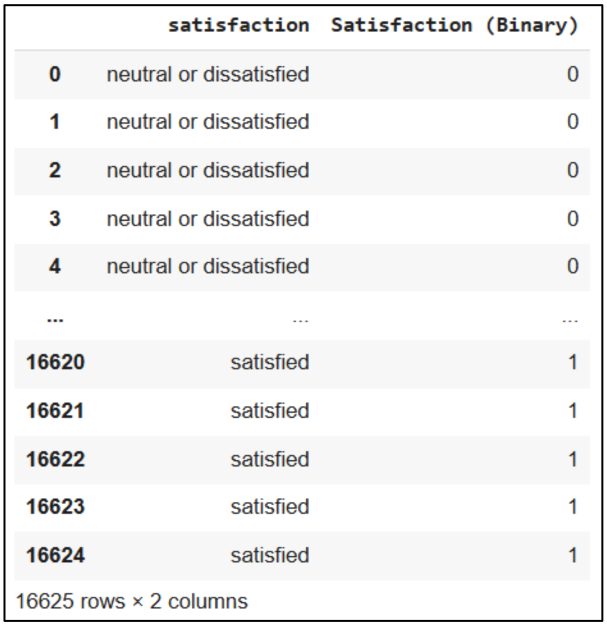
Satisfaction was binarized with the following frequency:

- 0 (Neutral or dissatisfied): 9,421 passengers.
- 1 (Satisfied): 7,204 passengers.



```
Satisfaction (Binary)
0    9421
1    7204
Name: count, dtype: int64
```

The Satisfaction (Binary) column simplifies the target into two clear groups, making it suitable for binary classification tasks (e.g., Logistic Regression, Decision Tree). The counts show a slight class imbalance, with more passengers in the neutral/dissatisfied group, which should be considered in further analysis and modeling.



	satisfaction	Satisfaction (Binary)
0	neutral or dissatisfied	0
1	neutral or dissatisfied	0
2	neutral or dissatisfied	0
3	neutral or dissatisfied	0
4	neutral or dissatisfied	0
...	...	...
16620	satisfied	1
16621	satisfied	1
16622	satisfied	1
16623	satisfied	1
16624	satisfied	1

16625 rows × 2 columns

TASK 1C. SUMMARY

1. ATTRIBUTE FINDING

Most attributes are the ordinal type that is satisfaction level of passenger experience.

The analysis of service attributes shows that overall passenger satisfaction varies considerably across operational and in-flight factors. **Inflight Wi-Fi** service emerges as the weakest attribute, with a mean rating of 2.80 and most passengers assigning scores of 2 or 3. **Ease of Online Booking** also performs poorly, with a mean of 2.88, reflecting customer frustration with the digital interface. These findings suggest that digital services remain a major area for improvement.

Operational aspects such as **Baggage Handling** and **Inflight Service** stand out as the strongest features, with mean ratings of 3.63 and 3.65 respectively. Baggage processes are generally reliable, while cabin service demonstrates consistently high quality, reinforcing them as competitive strengths. **Online Boarding** also shows strong performance (mean = 3.32), highlighting efficiency in pre-flight processes.

Comfort- and cabin-related attributes show mixed results. **Seat Comfort** (3.43) and **Leg Room Service** (3.37) indicate persistent issues with passenger space allocation, while **Food and Drink** (3.20) and **Cleanliness** (3.29) reflect variability in service quality and consistency**. Entertainment** performs relatively well (3.36), reinforcing its role as a positive in-flight experience.

The most critical operational challenges are revealed in the delay variables: **Departure Delay in Minutes** averages 15.1 minutes, with a median of 0 and an interquartile range (IQR) of 13 minutes. While most flights depart on time or with minimal delay, the distribution is heavily right-skewed (skewness = 8.25). A small but significant group of flights experience extreme delays, with the maximum recorded at 1,305 minutes (over 21 hours). This results in a very large variance (1,629.5), demonstrating operational inconsistency. **Arrival Delay in Minutes** shows a nearly identical pattern, with a mean of 15.5 minutes, median of 0, and IQR of 13 minutes. The skewness (8.02) and extreme outliers (up to 1,280 minutes) again indicate that while most flights arrive on time, a small number of extreme cases create severe disruptions. The variance is 1,652.4, confirming high variability. Importantly, 13.7% of values are flagged as outliers, and 0.3% are missing, suggesting data quality considerations.

2. RELATIONSHIP FINDING

The correlation matrix does not show strong/clear linear relationships among most attributes. However, the strongest pattern **Departure Delay** and **Arrival Delay** are almost perfectly correlated (r = 0.97). This demonstrates that delays at departure nearly always propagate to arrival, making departure punctuality a key predictor of perceived reliability. In practice, passengers interpret departure punctuality as a signal of overall airline performance, reinforcing the importance of managing delays at the source.

Digital and logistical convenience emerges as another cluster of interrelated attributes. Strong associations are observed between **Ease of Online Booking** and **Wi-Fi** (r = 0.66), as well as between **Online Booking** and **Departure/Arrival Convenience** (r = 0.53). These findings suggest a “halo effect,” where satisfaction with digital touchpoints influences perceptions of other travel aspects. Improvements in booking platforms and communication about schedules may therefore enhance overall satisfaction beyond individual attributes.

A second cluster centers on **Cleanliness**, where cleanliness shows strong associations with both **Food and Drink** (r = 0.60), **Inflight Entertainment** (r = 0.61), and **Seat Comfort** (r = 0.57). These relationships highlight how passengers evaluate cabin quality holistically: improvements in one area (such as hygiene) can reinforce positive perceptions of catering, comfort, and entertainment. This interrelationship emphasizes the importance of integrated cabin management methods, in which seemingly disparate service changes are planned together to maximize overall passenger experience.

3. PREPROCESSING INSIGHTS

Smoothing and binning of **delay attributes** confirm the skewed distribution. Equi-width binning shows that over 99% of flights fall into the first bin (0–260 minutes departure; 0–255 minutes arrival), with average delays of ~14 minutes. Outliers being above 1,200 minutes is rare but critical in shaping dissatisfaction.

Equi-depth binning provides further clarity: over 10,000 flights depart within 2 minutes of schedule, and around 9,900 arrive similarly on time. Minor delays (3–19 minutes) are common, while severe delays (over 45 minutes) are infrequent but impactful, averaging 100+ minutes. This highlights a two-tiered reality: most flights are punctual, but a small cluster of extreme delays creates disproportionate damage to customer experience and satisfaction metrics.

Discretizing **Age** reveals five distinct passenger groups. The dominant segment is Late Middle Age (46–65 years, 5,562 passengers), followed by Early Adulthood (22–34 years, 4,525 passengers) and Early Middle Age (35–44 years, 3,877 passengers). These groups together make up most travelers, emphasizing the importance of meeting the expectations of working-age passengers. Younger passengers (≤21, 2,019) and older adults (65+, 642) are smaller segments, suggesting they are less central to the airline's customer base.

4. FURTHER EXAMINATION

Visual patterns indicate that delay attributes form two clear clusters: one of mostly punctual or minimally delayed flights, and a smaller but distinct cluster of extreme delays (over 100 minutes). These clusters should be examined further to identify the routes, airports, or operational conditions driving severe disruptions.

Service attributes also group naturally into clusters. Digital services (online booking, Wi-Fi, boarding, schedule convenience) form one cluster, while in-cabin experiences (cleanliness, comfort, food, entertainment) form another. Dissatisfaction within one attribute often coincides with dissatisfaction in others, suggesting interdependencies worth deeper statistical testing.

Several associations merit further investigation. The strong dependency between departure and arrival delays (0.97) could be tested causally to confirm whether improvements in turnaround efficiency reduce both simultaneously. The moderate link between passenger class and type of travel (0.56) also deserves closer examination across different age groups. Finally, cleanliness appears to act as a keystone factor for in-cabin satisfaction and should be analyzed more rigorously using regression or structural modelling to assess its direct and indirect effects.