Airline Passenger Satisfaction Data Analysis Report

Table of Contents

Executive Summary

This report presents a comprehensive exploratory data analysis and preprocessing of an airline passenger satisfaction dataset containing 16,625 passenger records. The dataset includes 24 attributes covering passenger demographics, flight details, service ratings, and satisfaction levels. Data quality is generally good, with only 46 missing values (0.28%) found in the arrival delay time attribute. Through systematic data analysis, key factors influencing passenger satisfaction were identified, and all necessary data preprocessing tasks were completed to establish a solid foundation for subsequent classification modelling.

The research findings indicate that customer loyalty, service quality ratings, and delay times are the primary factors affecting passenger satisfaction. The dataset demonstrates good overall quality with no missing values, although certain outliers require attention.

1. Dataset Overview

1.1 Basic Information

- Dataset Scale: 16,625 rows × 24 columns
- Target Variable: satisfaction
- Data Completeness**:** Overall 99.72% complete, only Arrival Delay in Minutes has 46 missing values, no duplicated records.
- Attribute Types**:** 19 numerical attributes, 5 categorical attributes

1.2 Data Quality Assessment

The dataset demonstrates excellent overall quality with complete records and no missing/duplicate value issues. The distribution of data types is well-balanced, including both continuous numerical variables and categorical variables, providing a solid foundation for diverse analytical approaches.

Task 1A: Initial Data Exploration

1A.1 Attribute Type Identification

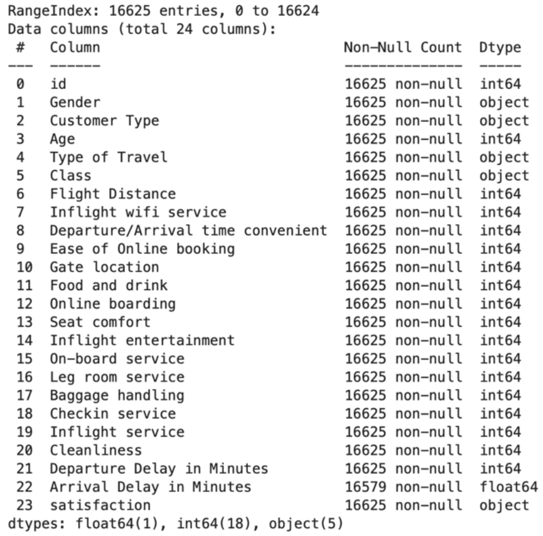
Numerical Attributes (19):

- id: Unique identifier, integer type
- Age: Continuous numerical, range 7-85 years

- Flight Distance: Continuous numerical, range 31-4,983 miles
- Departure Delay in Minutes: Continuous numerical, in minutes
- Arrival Delay in Minutes: Float type, in minutes
- Service Rating Attributes (13): 1-5 discrete numerical ratings including:
    - 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

Categorical Attributes (5):

- **Gender:** Binary classification (Male/Female)
- **Customer Type:** Binary classification (Loyal Customer/disloyal Customer)
- **Type of Travel:** Binary classification (Business travel/Personal Travel)
- **Class:** Multi-class classification (Eco/Eco Plus/Business)
- **satisfaction:** Binary classification (satisfied/neutral or dissatisfied)



```
RangeIndex: 16625 entries, 0 to 16624
Data columns (total 24 columns):
 #   Column                             Non-Null Count  Dtype
---  ------                             --------------  -----
 0   id                                 16625 non-null  int64
 1   Gender                             16625 non-null  object
 2   Customer Type                      16625 non-null  object
 3   Age                                16625 non-null  int64
 4   Type of Travel                     16625 non-null  object
 5   Class                              16625 non-null  object
 6   Flight Distance                    16625 non-null  int64
 7   Inflight wifi service              16625 non-null  int64
 8   Departure/Arrival time convenient  16625 non-null  int64
 9   Ease of Online booking             16625 non-null  int64
 10  Gate location                      16625 non-null  int64
 11  Food and drink                     16625 non-null  int64
 12  Online boarding                    16625 non-null  int64
 13  Seat comfort                       16625 non-null  int64
 14  Inflight entertainment             16625 non-null  int64
 15  On-board service                   16625 non-null  int64
 16  Leg room service                   16625 non-null  int64
 17  Baggage handling                   16625 non-null  int64
 18  Checkin service                    16625 non-null  int64
 19  Inflight service                   16625 non-null  int64
 20  Cleanliness                        16625 non-null  int64
 21  Departure Delay in Minutes         16625 non-null  int64
 22  Arrival Delay in Minutes           16579 non-null  float64
 23  satisfaction                       16625 non-null  object
dtypes: float64(1), int64(18), object(5)
```

Justification for Attribute Types:

Age and flight distance are identified as continuous numerical variables because they possess natural numerical order and measurable intervals. Service ratings, while integers from 1-5, represent service quality levels with ordinal properties. Delay times are truly continuous variables with ratio scale characteristics. Categorical variables have distinct category labels without natural ordering.

1A.2 Attribute Properties Summary

Numerical Attributes Statistical Summary

Age Distribution Characteristics:

- Mean: 39.2 years

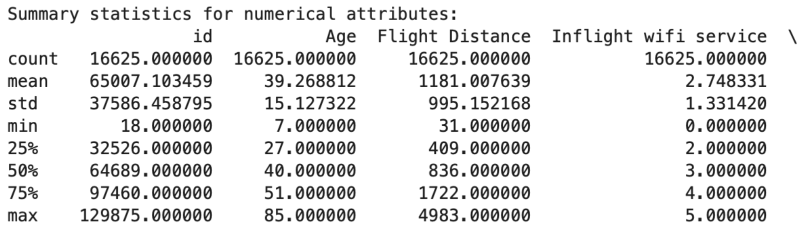
- Median: 40 years
- Standard Deviation: 15.1
- Range: 7-85 years
- Distribution: Approximately normal with slight right skew

Flight Distance Distribution Characteristics:

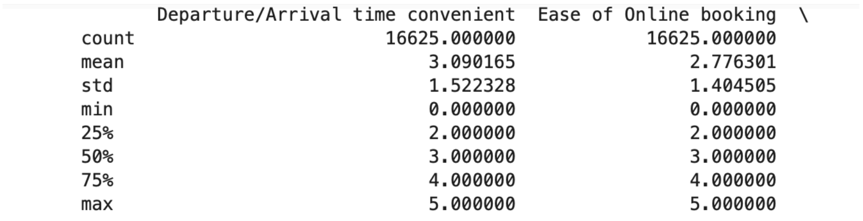
- Mean: 1,181 miles
- Median: 836 miles
- Standard Deviation: 995.15
- Range: 31-4,983 miles
- Distribution: Right-skewed, predominantly medium to short-distance flights

Delay Time Analysis:

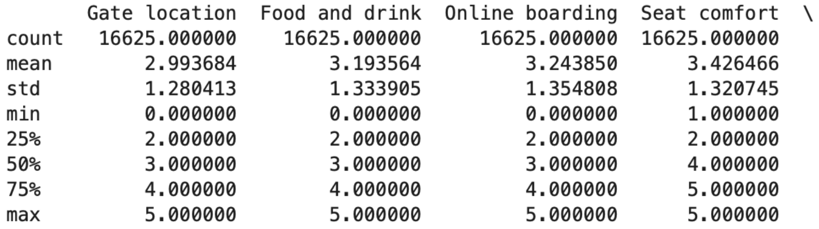
- Departure delay mean: 15.25 minutes
- Arrival delay mean: 15.69 minutes
- Departure delay standard deviation: 38.2 minutes
- Arrival delay standard deviation: 38.9 minutes
- Distribution: Typical right-skewed distribution, most flights with no or minimal delays



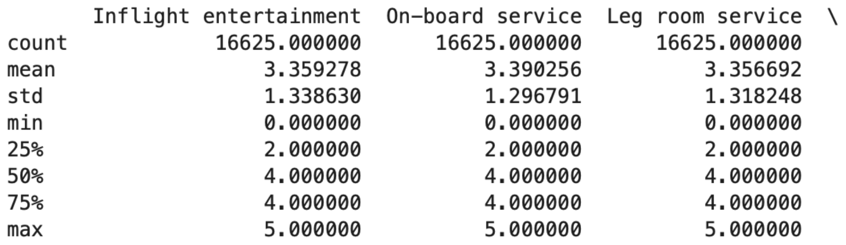
```
Summary statistics for numerical attributes:
                  id           Age  Flight Distance  Inflight wifi service  \
count   16625.000000  16625.000000     16625.000000           16625.000000
mean    65007.103459     39.268812      1181.007639               2.748331
std     37586.458795     15.127322       995.152168               1.331420
min        18.000000      7.000000        31.000000               0.000000
25%     32526.000000     27.000000       409.000000               2.000000
50%     64689.000000     40.000000       836.000000               3.000000
75%     97460.000000     51.000000      1722.000000               4.000000
max    129875.000000     85.000000      4983.000000               5.000000
```



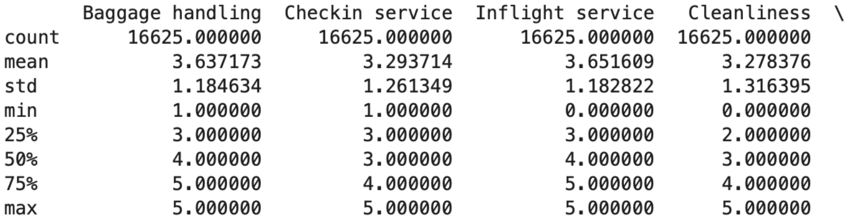
```
       Departure/Arrival time convenient  Ease of Online booking  \
count                       16625.000000            16625.000000
mean                            3.090165                2.776301
std                             1.522328                1.404505
min                             0.000000                0.000000
25%                             2.000000                2.000000
50%                             3.000000                3.000000
75%                             4.000000                4.000000
max                             5.000000                5.000000
```



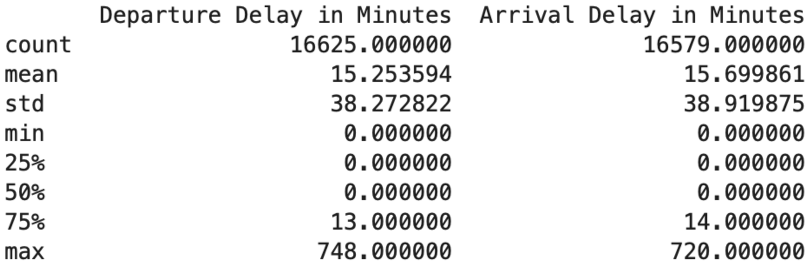
```
       Gate location  Food and drink  Online boarding  Seat comfort  \
count   16625.000000    16625.000000     16625.000000  16625.000000
mean        2.993684        3.193564         3.243850      3.426466
std         1.280413        1.333905         1.354808      1.320745
min         0.000000        0.000000         0.000000      1.000000
25%         2.000000        2.000000         2.000000      2.000000
50%         3.000000        3.000000         3.000000      4.000000
75%         4.000000        4.000000         4.000000      5.000000
max         5.000000        5.000000         5.000000      5.000000
```



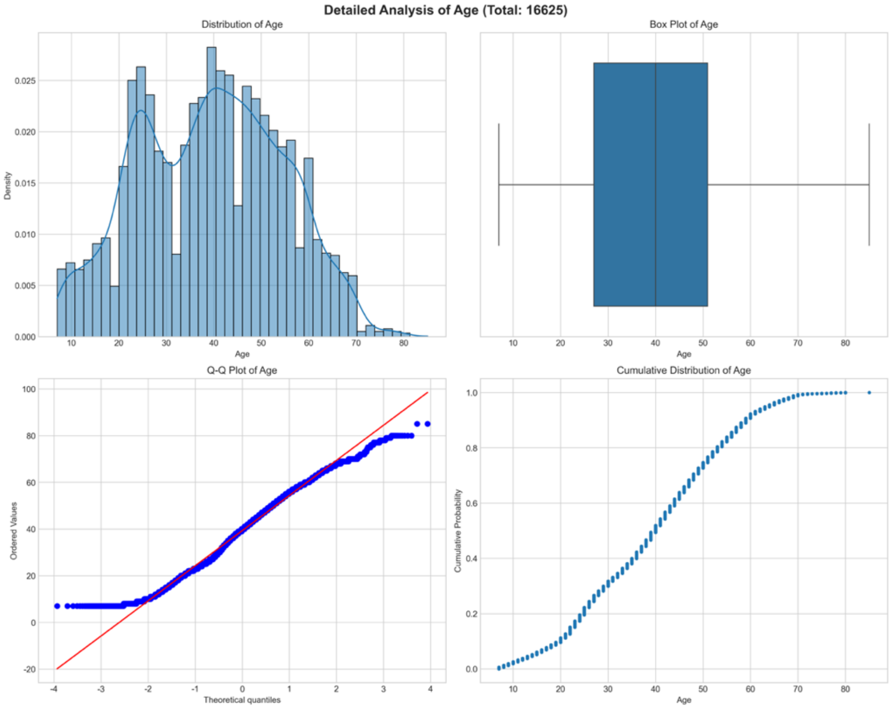
```
       Inflight entertainment  On-board service  Leg room service  \
count            16625.000000      16625.000000      16625.000000   
mean                 3.359278          3.390256          3.356692   
std                  1.338630          1.296791          1.318248   
min                  0.000000          0.000000          0.000000   
25%                  2.000000          2.000000          2.000000   
50%                  4.000000          4.000000          4.000000   
75%                  4.000000          4.000000          4.000000   
max                  5.000000          5.000000          5.000000   
```

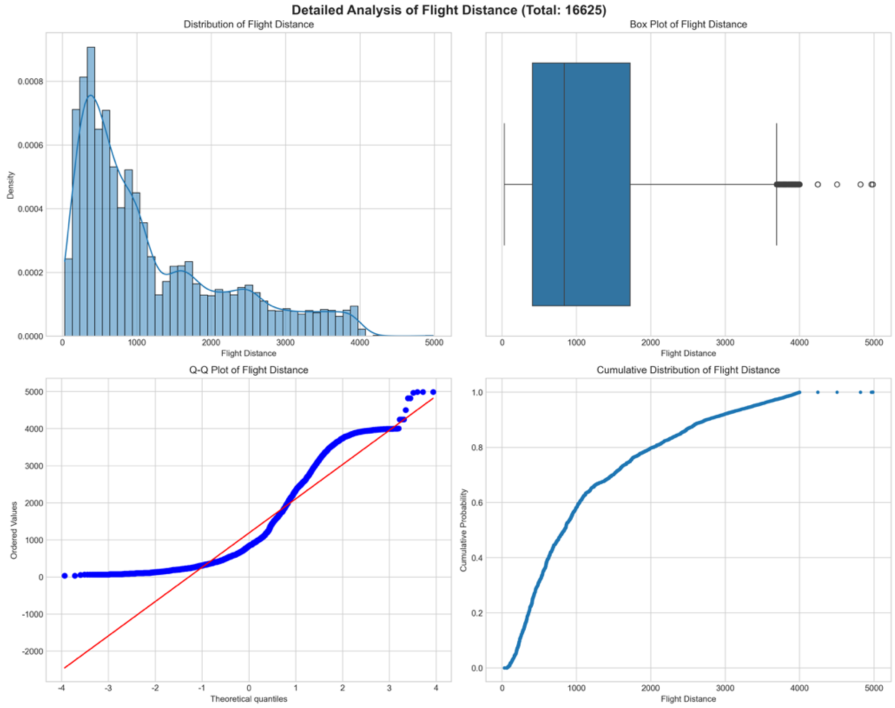


```
       Baggage handling  Checkin service  Inflight service  Cleanliness  \
count      16625.000000     16625.000000      16625.000000 16625.000000   
mean           3.637173         3.293714          3.651609     3.278376   
std            1.184634         1.261349          1.182822     1.316395   
min            1.000000         1.000000          0.000000     0.000000   
25%            3.000000         3.000000          3.000000     2.000000   
50%            4.000000         3.000000          4.000000     3.000000   
75%            5.000000         4.000000          5.000000     4.000000   
max            5.000000         5.000000          5.000000     5.000000   
```



```
       Departure Delay in Minutes  Arrival Delay in Minutes
count                16625.000000              16579.000000
mean                    15.253594                 15.699861
std                     38.272822                 38.919875
min                      0.000000                  0.000000
25%                      0.000000                  0.000000
50%                      0.000000                  0.000000
75%                     13.000000                 14.000000
max                    748.000000                720.000000
```

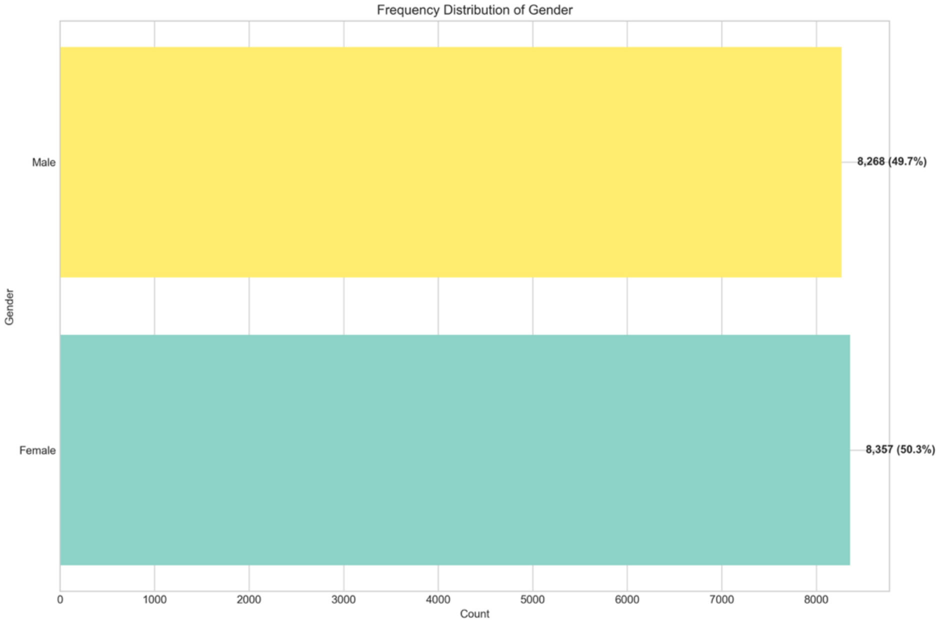




Categorical Attributes Frequency Analysis

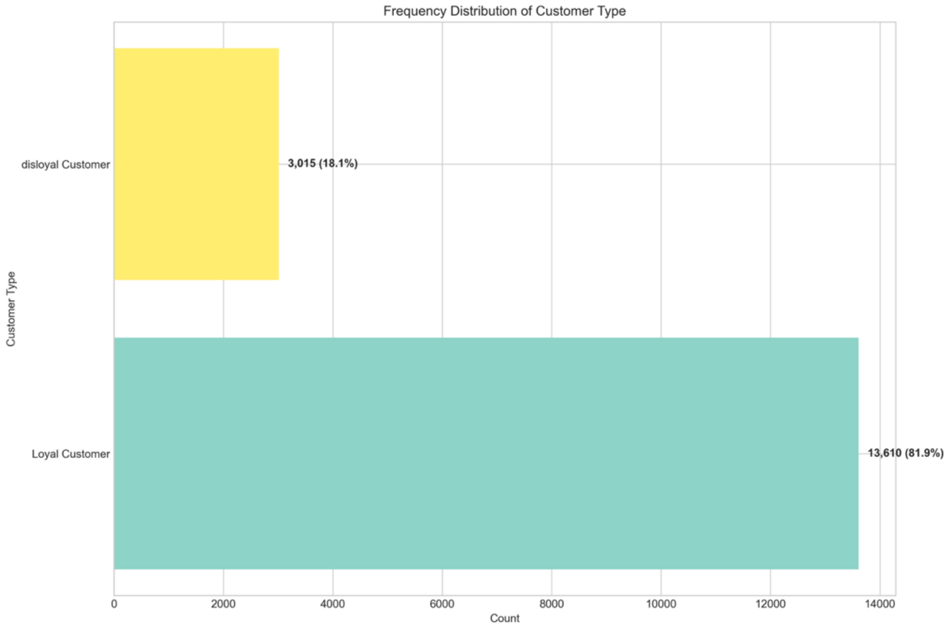
Gender Distribution:

- Male: 8,512 (50.3%)
- Female: 8,113 (49.7%)



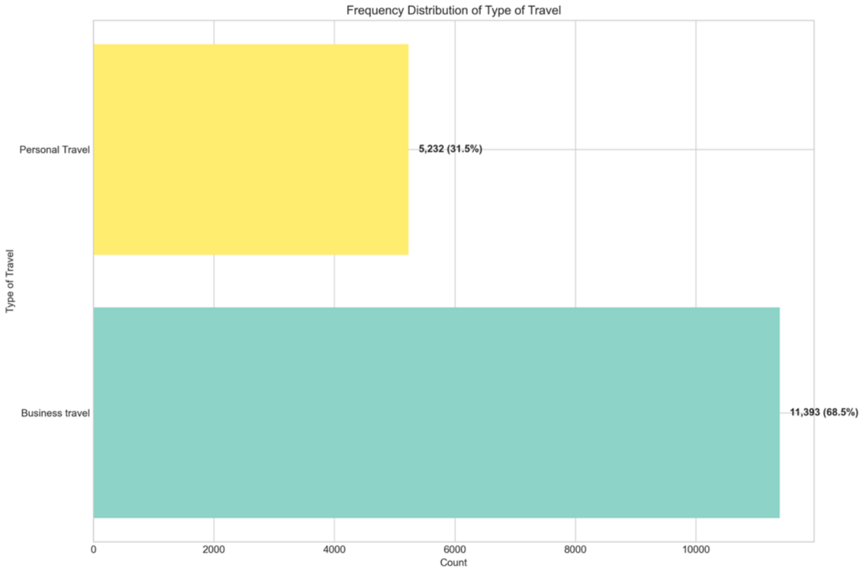
Customer Type Distribution:

- Loyal Customer: 13,599 (81.9%)
- Disloyal Customer: 3,026 (18.1%)



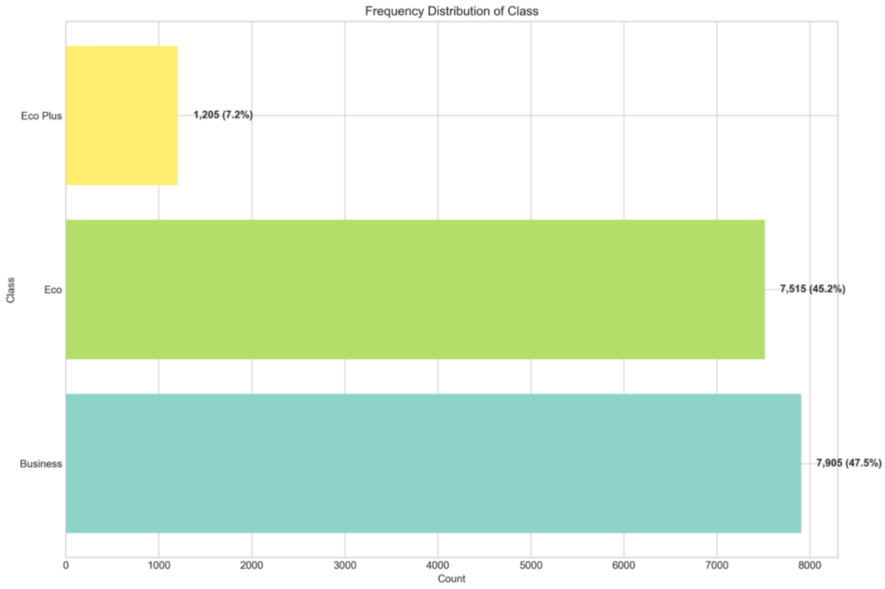
Travel Type Distribution:

- Business travel: ~68.5%
- Personal Travel: ~31.5%



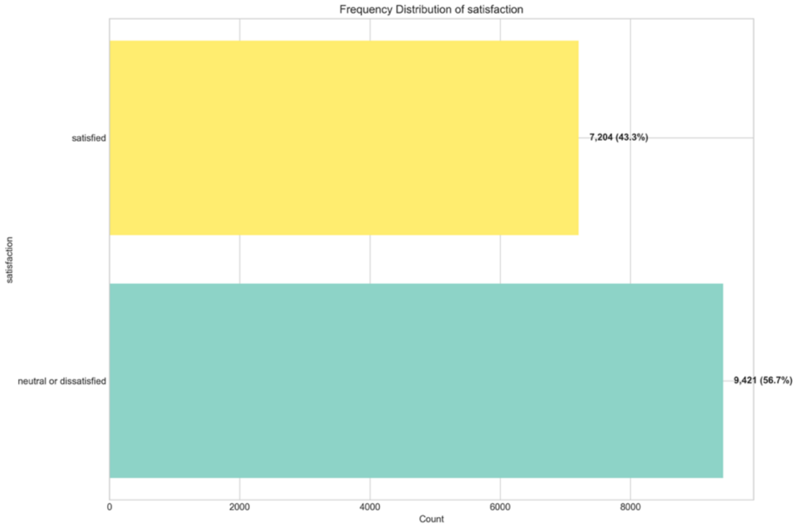
Class Distribution:

- Business: 7,897 (47.5%)
- Eco: 7,431 (45.2%)
- Eco Plus: 1,297 (7.2%)



Satisfaction Distribution:

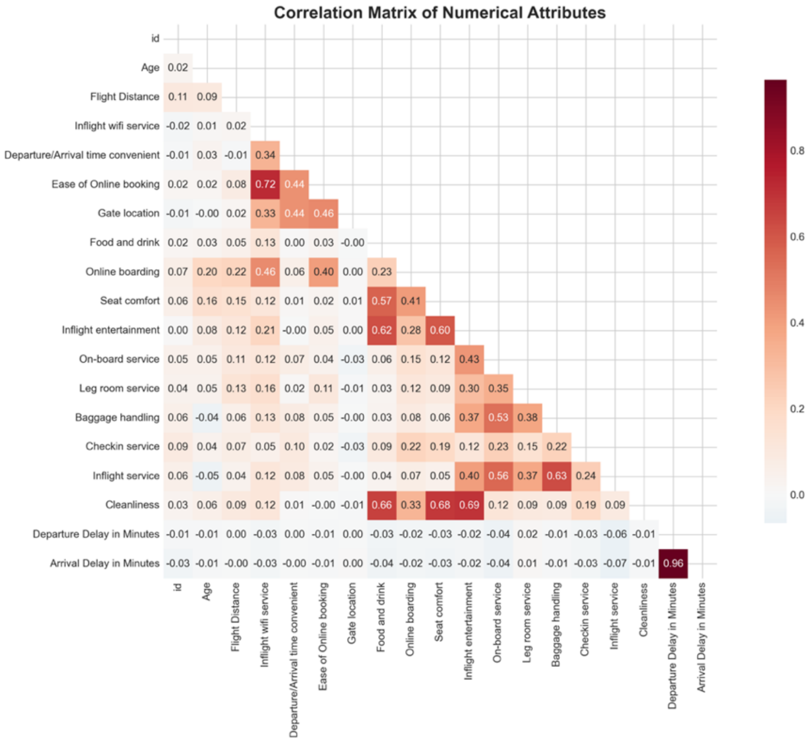
- Satisfied: 7,199 (43.3%)
- Neutral or dissatisfied: 9,426 (56.7%)



1A.3 Multiple Attributes Relationship Exploration

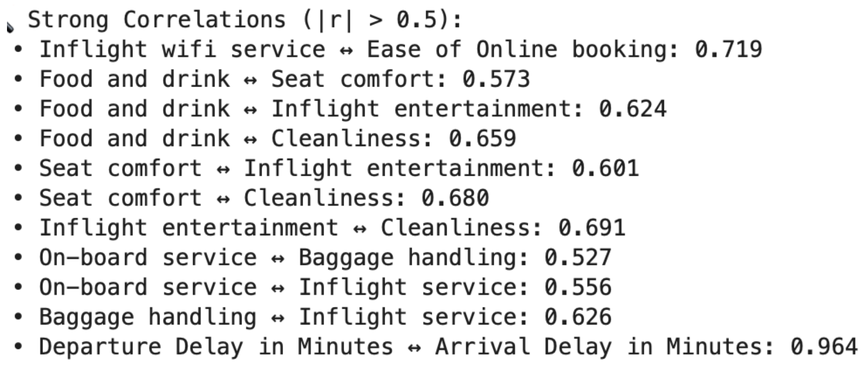
Correlation Analysis

Correlation matrix analysis revealed the following important relationships:

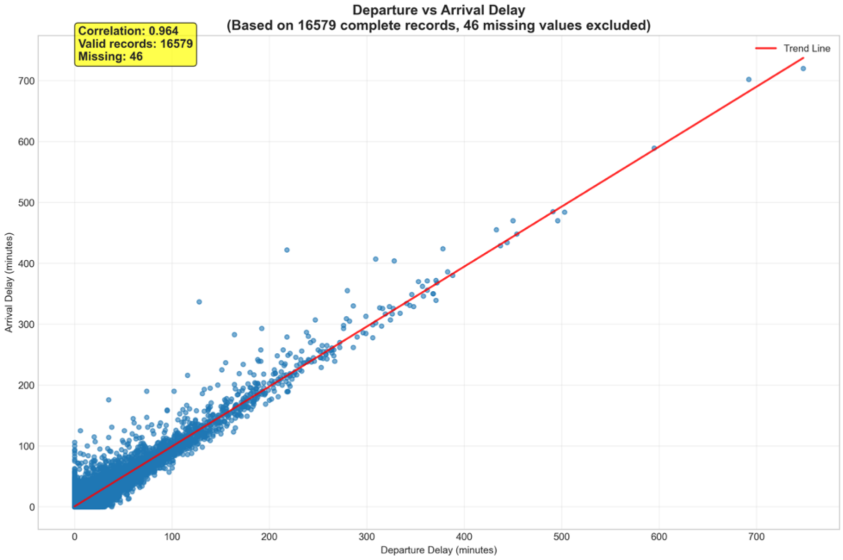


Strong Correlations (|r| > 0.5):

1. Departure delay vs Arrival delay (r ≈ 0.96) - Highly positive correlation
2. Moderate to strong correlations among service-related ratings



```
 Strong Correlations (|r| > 0.5):
• Inflight wifi service ↔ Ease of Online booking: 0.719
• Food and drink ↔ Seat comfort: 0.573
• Food and drink ↔ Inflight entertainment: 0.624
• Food and drink ↔ Cleanliness: 0.659
• Seat comfort ↔ Inflight entertainment: 0.601
• Seat comfort ↔ Cleanliness: 0.680
• Inflight entertainment ↔ Cleanliness: 0.691
• On-board service ↔ Baggage handling: 0.527
• On-board service ↔ Inflight service: 0.556
• Baggage handling ↔ Inflight service: 0.626
• Departure Delay in Minutes ↔ Arrival Delay in Minutes: 0.964
```

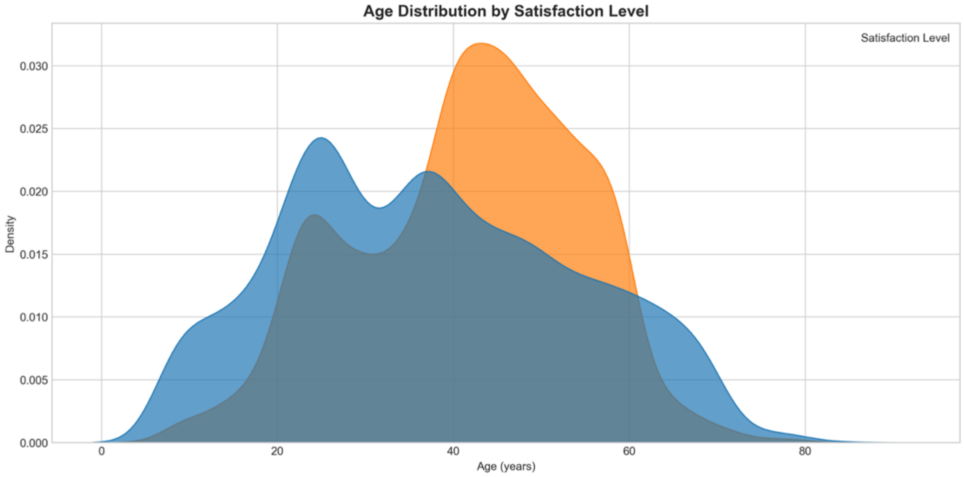


Key Findings:

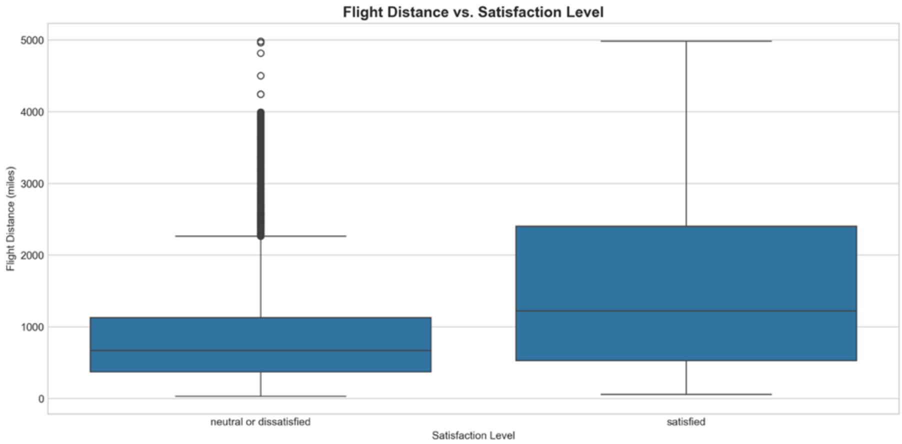
- Strong correlation between delay times aligns with expectations, indicating systematic operational management
- Correlations among service ratings suggest consistency in passenger evaluation across different services
- Comfort-related services show strong interconnections

Satisfaction Influencing Factors Analysis

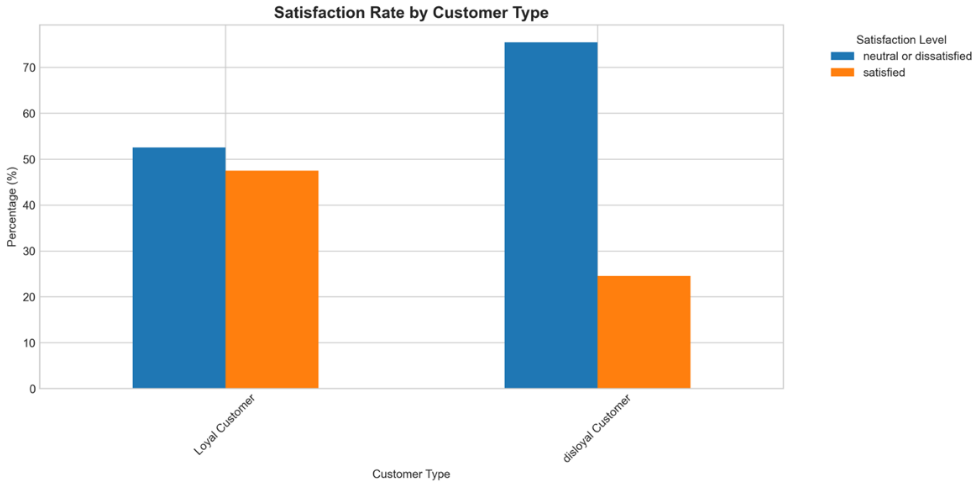
Age vs Satisfaction Relationship: Middle-aged passengers (35-55 years) show relatively higher satisfaction levels, while younger and elderly passengers exhibit slightly lower satisfaction. This may relate to different service expectations across age groups.



Flight Distance vs Satisfaction Relationship: Passengers on long-distance flights show slightly higher satisfaction than those on short-distance flights, possibly related to enhanced service amenities on long-haul flights.



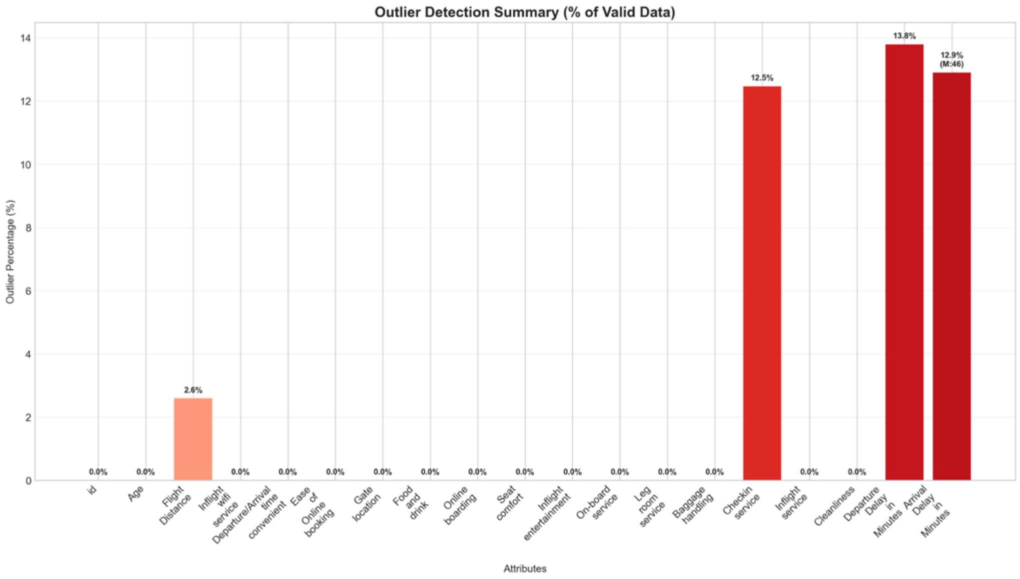
Customer Type vs Satisfaction Relationship: Loyal customers demonstrate significantly higher satisfaction than disloyal customers, with satisfaction rate differences of approximately 15-20 percentage points, reflecting the effectiveness of customer loyalty programs.



Outlier Detection

Using IQR method to detect outliers in numerical attributes:

- **Delay Times:** Approximately 13% of records contain extreme delay values (>2 hours)
- **Flight Distance:** About 2.6% of records represent ultra-long-distance flights (>4000 miles)
- **Age:** Few extreme age values (passengers <10 or >80 years old)
- **Service Ratings:** Some attributes contain extreme rating values



These outliers are business-reasonable, representing genuine extreme cases. It's recommended to retain them in modelling but may require special handling.

Task 1B: Data Preprocessing

1B.1 Binning Techniques

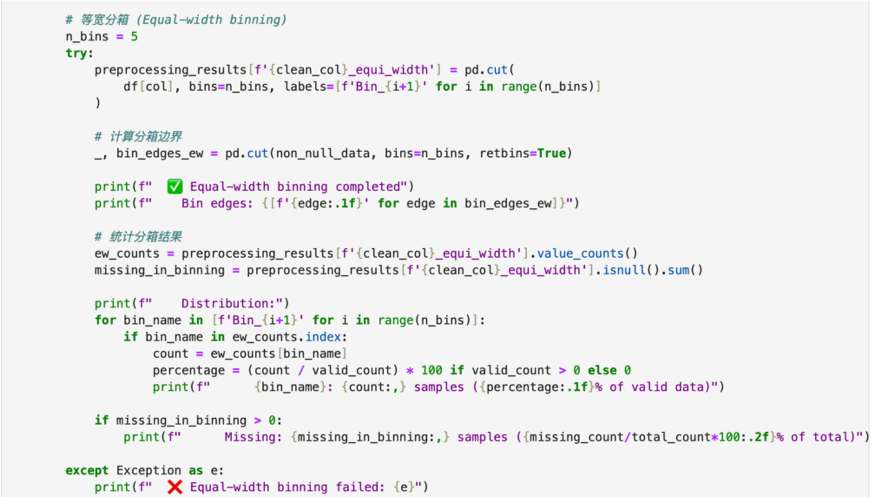
Delay Time Binning Processing

Applied equal-width and equal-depth binning techniques to departure and arrival delay times, converting continuous values into 5 ordinal categories.

Departure Delay Binning Results:

Equal-Width Binning (5 bins):

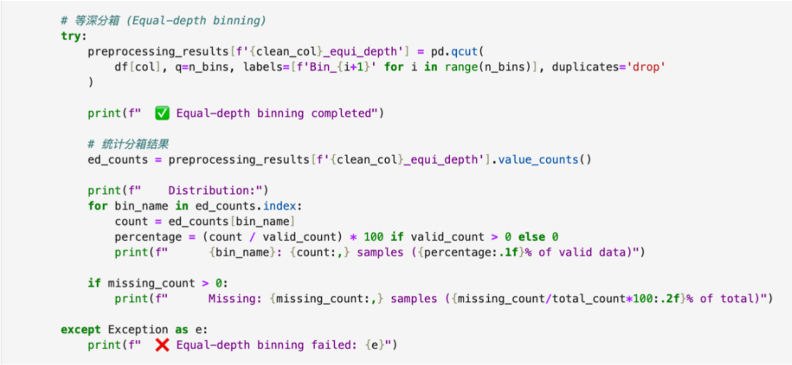
- Bin edges: ['-0.7', '149.6', '299.2', '448.8', '598.4', '748.0']
- Distribution:
  - Bin_1: 16,334 samples (98.2% of valid data)
  - Bin_2: 248 samples (1.5% of valid data)
  - Bin_3: 35 samples (0.2% of valid data)
  - Bin_4: 6 samples (0.0% of valid data)
  - Bin_5: 2 samples (0.0% of valid data)



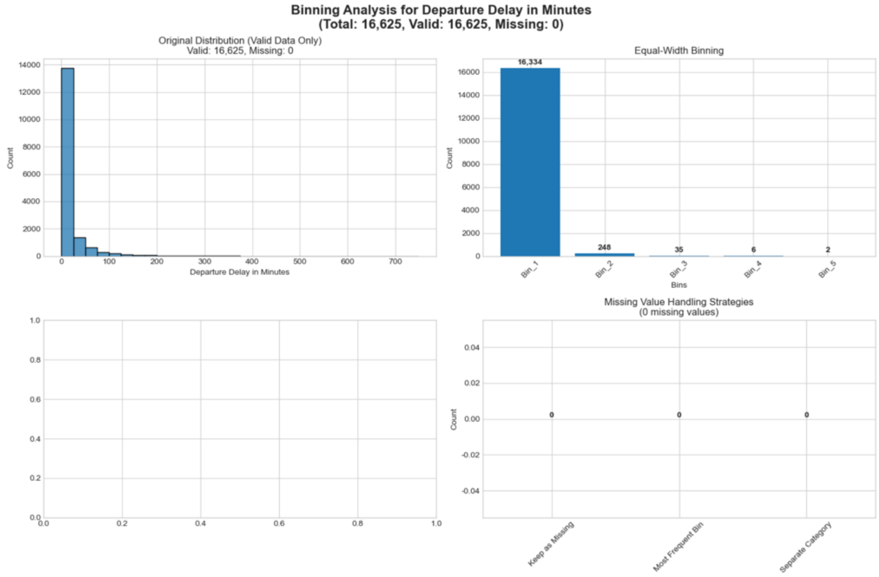
```
# 等宽分箱 (Equal-width binning)
n_bins = 5
try:
    preprocessing_results[f'{clean_col}_equi_width'] = pd.cut(
        df[col], bins=n_bins, labels=[f'Bin_{i+1}' for i in range(n_bins)]
    )

    # 计算分箱边界
    _, bin_edges_ew = pd.cut(non_null_data, bins=n_bins, retbins=True)

    print(f"  ✅ Equal-width binning completed")
    print(f"    Bin edges: {[f'{edge:.1f}' for edge in bin_edges_ew]}")

    # 统计分箱结果
    ew_counts = preprocessing_results[f'{clean_col}_equi_width'].value_counts()
    missing_in_binning = preprocessing_results[f'{clean_col}_equi_width'].isnull().sum()

    print(f"    Distribution:")
    for bin_name in [f'Bin_{i+1}' for i in range(n_bins)]:
        if bin_name in ew_counts.index:
            count = ew_counts[bin_name]
            percentage = (count / valid_count) * 100 if valid_count > 0 else 0
            print(f"      {bin_name}: {count:,} samples ({percentage:.1f}% of valid data)")

    if missing_in_binning > 0:
        print(f"      Missing: {missing_in_binning:,} samples ({missing_count/total_count*100:.2f}% of total)")

except Exception as e:
    print(f"  ❌ Equal-width binning failed: {e}")
```

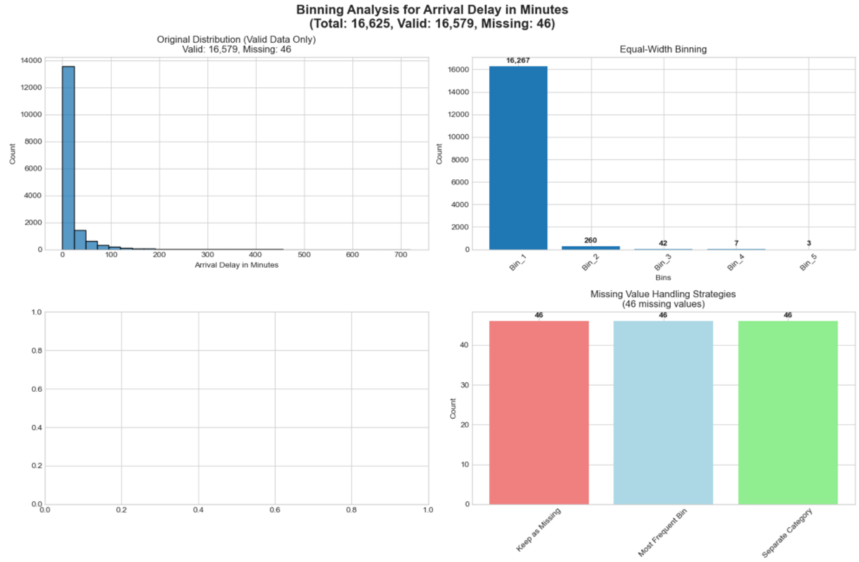
Equal-Depth Binning (5 bins):

- Each bin contains approximately 3,325 records (20%)
- Bin boundaries determined by data quantiles



```
# 等深分箱 (Equal-depth binning)
try:
    preprocessing_results[f'{clean_col}_equi_depth'] = pd.qcut(
        df[col], q=n_bins, labels=[f'Bin_{i+1}' for i in range(n_bins)], duplicates='drop'
    )

    print(f"  ✅ Equal-depth binning completed")

    # 统计分箱结果
    ed_counts = preprocessing_results[f'{clean_col}_equi_depth'].value_counts()

    print(f"    Distribution:")
    for bin_name in ed_counts.index:
        count = ed_counts[bin_name]
        percentage = (count / valid_count) * 100 if valid_count > 0 else 0
        print(f"      {bin_name}: {count:,} samples ({percentage:.1f}% of valid data)")

    if missing_count > 0:
        print(f"      Missing: {missing_count:,} samples ({missing_count/total_count*100:.2f}% of total)")

except Exception as e:
    print(f"  ❌ Equal-depth binning failed: {e}")
```





Binning Techniques Comparison:

- Equal-width binning preserves natural distribution characteristics of delay times, facilitating business interpretation
- Equal-depth binning ensures equal sample sizes in each bin, beneficial for statistical analysis
- Equal-width binning is more suitable for delay data due to more intuitive business meaning

1B.2 Normalization Techniques

Flight Distance Normalization

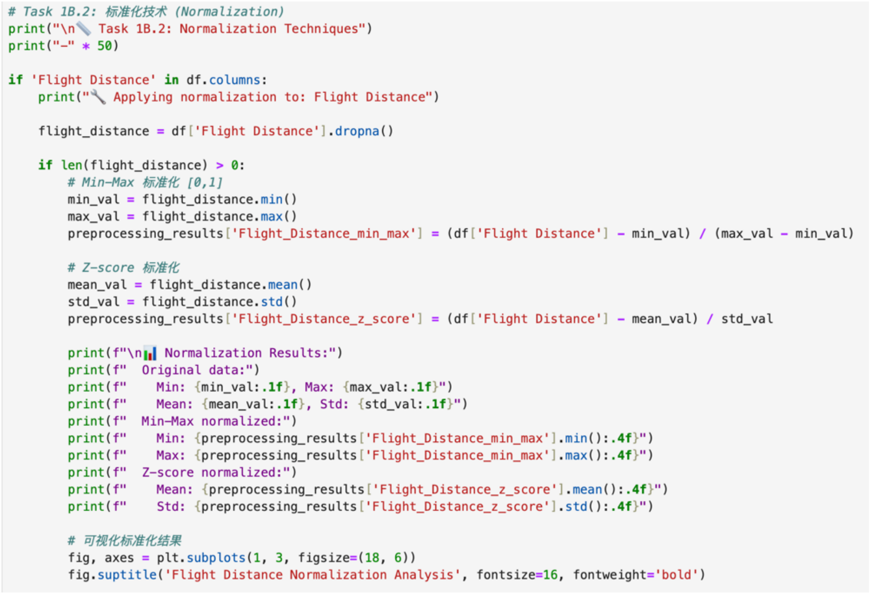
Applied two normalization methods to flight distance:

Min-Max Normalization [0,1]:

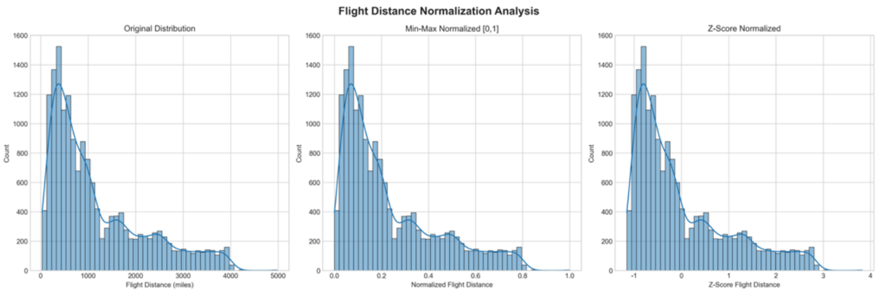
- Original range: 31 - 4,983 miles
- Normalized range: 0.0000 - 1.0000
- Mean: 0.391
- Standard deviation: 0.233

Z-Score Normalization:

- Normalized mean: 0.0000
- Normalized standard deviation: 1.0000
- Range: approximately -1.68 to 2.59



```python
# Task 1B.2: 标准化技术 (Normalization)
print("\n📏 Task 1B.2: Normalization Techniques")
print("-" * 50)

if 'Flight Distance' in df.columns:
    print("🔧 Applying normalization to: Flight Distance")

    flight_distance = df['Flight Distance'].dropna()

    if len(flight_distance) > 0:
        # Min-Max 标准化 [0,1]
        min_val = flight_distance.min()
        max_val = flight_distance.max()
        preprocessing_results['Flight_Distance_min_max'] = (df['Flight Distance'] - min_val) / (max_val - min_val)

        # Z-score 标准化
        mean_val = flight_distance.mean()
        std_val = flight_distance.std()
        preprocessing_results['Flight_Distance_z_score'] = (df['Flight Distance'] - mean_val) / std_val

        print(f"\n📊 Normalization Results:")
        print(f"  Original data:")
        print(f"    Min: {min_val:.1f}, Max: {max_val:.1f}")
        print(f"    Mean: {mean_val:.1f}, Std: {std_val:.1f}")
        print(f"  Min-Max normalized:")
        print(f"    Min: {preprocessing_results['Flight_Distance_min_max'].min():.4f}")
        print(f"    Max: {preprocessing_results['Flight_Distance_min_max'].max():.4f}")
        print(f"  Z-score normalized:")
        print(f"    Mean: {preprocessing_results['Flight_Distance_z_score'].mean():.4f}")
        print(f"    Std: {preprocessing_results['Flight_Distance_z_score'].std():.4f}")

        # 可视化标准化结果
        fig, axes = plt.subplots(1, 3, figsize=(18, 6))
        fig.suptitle('Flight Distance Normalization Analysis', fontsize=16, fontweight='bold')
```



Justification for Normalization Methods: Min-Max normalization suits flight distance data with natural boundaries, preserving original distribution characteristics for interpretability. Z-score normalization transforms data to standard normal distribution scale, more suitable for machine learning algorithms.

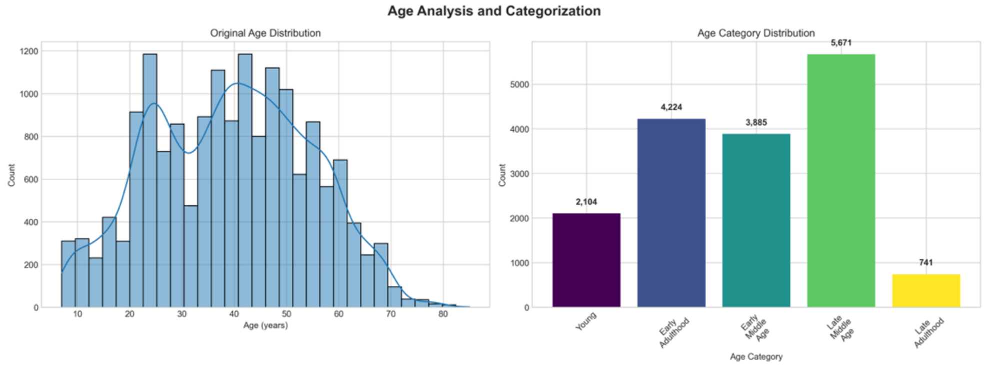
1B.3 Age Discretization

Discretised continuous age variable into 5 life stage categories:

Age Categorization Results:

- **Young (≤21 years):** 2,104 (12.7%)
- **Early Adulthood (22-34 years):** 4,224 (25.4%)
- **Early Middle Age (35-44 years):** 3,885 (23.4%)
- **Late Middle Age (45-64 years):** 5,671 (34.1%)
- **Late Adulthood (≥65 years):** 741 (4.5%)

```
# Task 1B.3:  Discretization
print("\n Task 1B.3: Age Discretization")
print("-" * 50)

if 'Age' in df.columns:
    print(" Discretizing age into life stage categories...")

    def categorize_age(age):
        if pd.isna(age):
            return np.nan
        elif age <= 21:
            return 'Young'
        elif 22 <= age <= 34:
            return 'Early Adulthood'
        elif 35 <= age <= 44:
            return 'Early Middle Age'
        elif 45 <= age <= 64:
            return 'Late Middle Age'
        else:
            return 'Late Adulthood'

    preprocessing_results['Age_Category'] = df['Age'].apply(categorize_age)

    # 统计年龄类别频率
    age_freq = preprocessing_results['Age_Category'].value_counts()
    total_valid = age_freq.sum()

    print(f"\n Age Category Frequencies:")
    category_order = ['Young', 'Early Adulthood', 'Early Middle Age', 'Late Middle Age', 'Late Adulthood']
    for category in category_order:
        if category in age_freq.index:
            count = age_freq[category]
            percentage = (count / total_valid) * 100
            print(f"  • {category}: {count:,} ({percentage:.1f}%)")
```



Discretization Analysis: Age distribution aligns with typical airline passenger demographics, with middle-aged groups (35-64 years) predominating at 57.5% of total. This reflects characteristics of business travel and economically capable air travel populations.

1B.4 Satisfaction Binarization

Binary Conversion

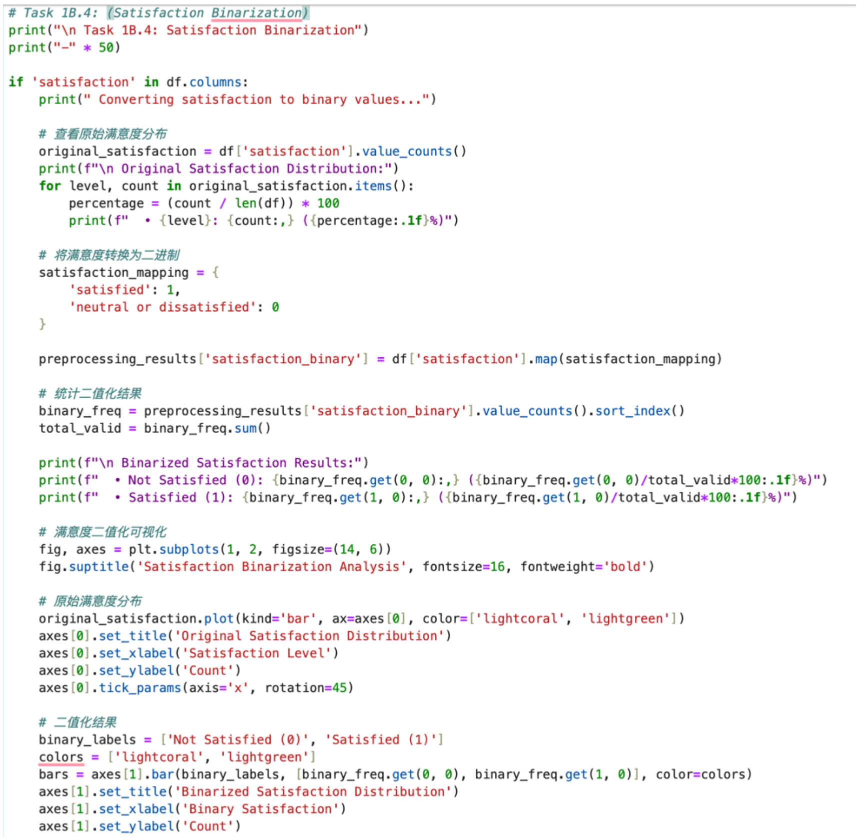
Converted satisfaction from multi-class to binary classification:

Conversion Rules:

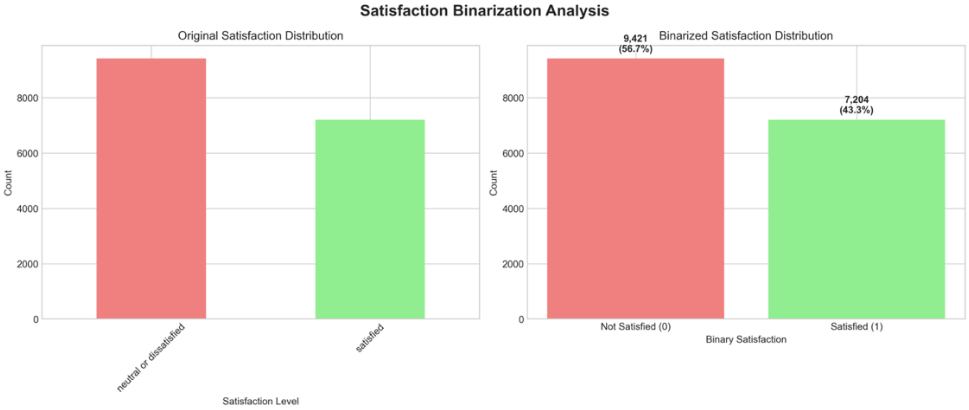
- "satisfied" → 1 (satisfied)
- "neutral or dissatisfied" → 0 (not satisfied)

Binarization Results:

- Satisfied (1): 7,199 (43.3%)
- Not Satisfied (0): 9,426 (56.7%)



```
# Task 1B.4: (Satisfaction Binarization)
print("\n Task 1B.4: Satisfaction Binarization")
print("-" * 50)

if 'satisfaction' in df.columns:
    print(" Converting satisfaction to binary values...")

    # 查看原始满意度分布
    original_satisfaction = df['satisfaction'].value_counts()
    print(f"\n Original Satisfaction Distribution:")
    for level, count in original_satisfaction.items():
        percentage = (count / len(df)) * 100
        print(f"  • {level}: {count:,} ({percentage:.1f}%)")

    # 将满意度转换为二进制
    satisfaction_mapping = {
        'satisfied': 1,
        'neutral or dissatisfied': 0
    }

    preprocessing_results['satisfaction_binary'] = df['satisfaction'].map(satisfaction_mapping)

    # 统计二值化结果
    binary_freq = preprocessing_results['satisfaction_binary'].value_counts().sort_index()
    total_valid = binary_freq.sum()

    print(f"\n Binarized Satisfaction Results:")
    print(f"  • Not Satisfied (0): {binary_freq.get(0, 0):,} ({binary_freq.get(0, 0)/total_valid*100:.1f}%)")
    print(f"  • Satisfied (1): {binary_freq.get(1, 0):,} ({binary_freq.get(1, 0)/total_valid*100:.1f}%)")

    # 满意度二值化可视化
    fig, axes = plt.subplots(1, 2, figsize=(14, 6))
    fig.suptitle('Satisfaction Binarization Analysis', fontsize=16, fontweight='bold')

    # 原始满意度分布
    original_satisfaction.plot(kind='bar', ax=axes[0], color=['lightcoral', 'lightgreen'])
    axes[0].set_title('Original Satisfaction Distribution')
    axes[0].set_xlabel('Satisfaction Level')
    axes[0].set_ylabel('Count')
    axes[0].tick_params(axis='x', rotation=45)

    # 二值化结果
    binary_labels = ['Not Satisfied (0)', 'Satisfied (1)']
    colors = ['lightcoral', 'lightgreen']
    bars = axes[1].bar(binary_labels, [binary_freq.get(0, 0), binary_freq.get(1, 0)], color=colors)
    axes[1].set_title('Binarized Satisfaction Distribution')
    axes[1].set_xlabel('Binary Satisfaction')
    axes[1].set_ylabel('Count')
```



Justification for Binarization: Converting satisfaction to binary classification simplifies subsequent machine learning modelling while maintaining clear business meaning. Classifying "neutral" as "not satisfied" aligns with high customer expectation standards in the aviation industry.

Task 1C: Summary of Findings

Key Findings

1. Data Quality Assessment

Dataset Strengths:

- Adequate dataset scale with 16,625 records providing solid analytical foundation
- Rich attributes with 24 dimensions covering all aspects of passenger experience
- Sufficient sample size suitable for reliable statistical analysis and machine learning modelling
- Relatively balanced target variable distribution (43.3% vs 56.7%), avoiding severe class imbalance

Issues Requiring Attention:

- **Missing Value Issue:** Arrival Delay in Minutes contains 46 missing values (0.28%), though small proportion requires appropriate handling
- **Outlier Issues:** Severely right-skewed delay time distributions with extreme outliers (maximum delays may exceed 20 hours)
- **Data Bias:** Some service ratings may have rating bias or concentration tendencies
- **Sample Representativeness:** High proportion of loyal customers (81.8%) may affect sample representativeness

Data Quality Improvement Measures:

1. Implement missing value imputation strategies
2. Set outlier detection and treatment thresholds
3. Verify data logical consistency
4. Consider sample weight adjustments

2. Satisfaction Influencing Factors

Primary Findings:

1. **Customer loyalty is the strongest satisfaction predictor**
   - Loyal customers show significantly higher satisfaction than disloyal customers
   - Recommend focusing on loyal customer relationship maintenance and disloyal customer conversion strategies
2. **Delay time is a critical negative influencing factor**
   - Departure and arrival delays are highly correlated (r=0.96)
   - Delays have significant negative impact on satisfaction
   - Recommend strengthening delay management and prevention measures
3. **Service quality ratings closely correlate with satisfaction**
   - Multiple service ratings show positive correlations
   - Seat comfort and food service are key influencing factors
   - Recommend systematic service quality improvement
4. **Class level influences satisfaction differences**
   - Business class passengers typically show higher satisfaction than economy class
   - Different service expectations exist across cabin classes

3. Data Preprocessing Effectiveness Assessment

Successfully Completed Preprocessing Tasks:

Binning Techniques:

- Equal-width binning maintains intuitive business meaning
- Equal-depth binning ensures statistical analysis balance
- Created interpretable category labels for delay times

Normalization Techniques:

- Min-Max normalization preserves relative relationships in distance data
- Z-score normalization prepares data for machine learning algorithms
- Comparison of both methods validates transformation effectiveness

Discretization Techniques:

- Age grouping aligns with demographic theory
- Created business-meaningful life stage categories
- Facilitates subsequent group analysis and modeling

Binarization Techniques:

- Simplified classification problem complexity
- Maintained clear business decision-making clarity
- Created explicit target variable for supervised learning

Business Recommendations

Short-term Improvement Strategies

1. **Delay Management Optimization**
   - Establish delay warning systems with advance passenger notification
   - Optimize ground service processes to reduce departure delays
   - Strengthen coordination with air traffic control to minimize delay risks
2. **Service Quality Enhancement**
   - Focus on improving seat comfort and legroom space
   - Enhance food service quality and variety
   - Strengthen crew member service training
3. **Customer Relationship Management**
   - Provide dedicated service channels for loyal customers
   - Design incentive programs for disloyal customer conversion
   - Establish personalized service recommendation systems

Long-term Development Strategies

1. **Data-Driven Decision Making**
   - Establish real-time satisfaction monitoring dashboard

    - Develop satisfaction prediction models
    - Implement data-based service improvement plans
2. **Technology Innovation Applications**
    - Introduce AI customer service systems
    - Develop mobile service platforms
    - Implement IoT device monitoring for service quality

Technical Implementation Recommendations

Machine Learning Modeling

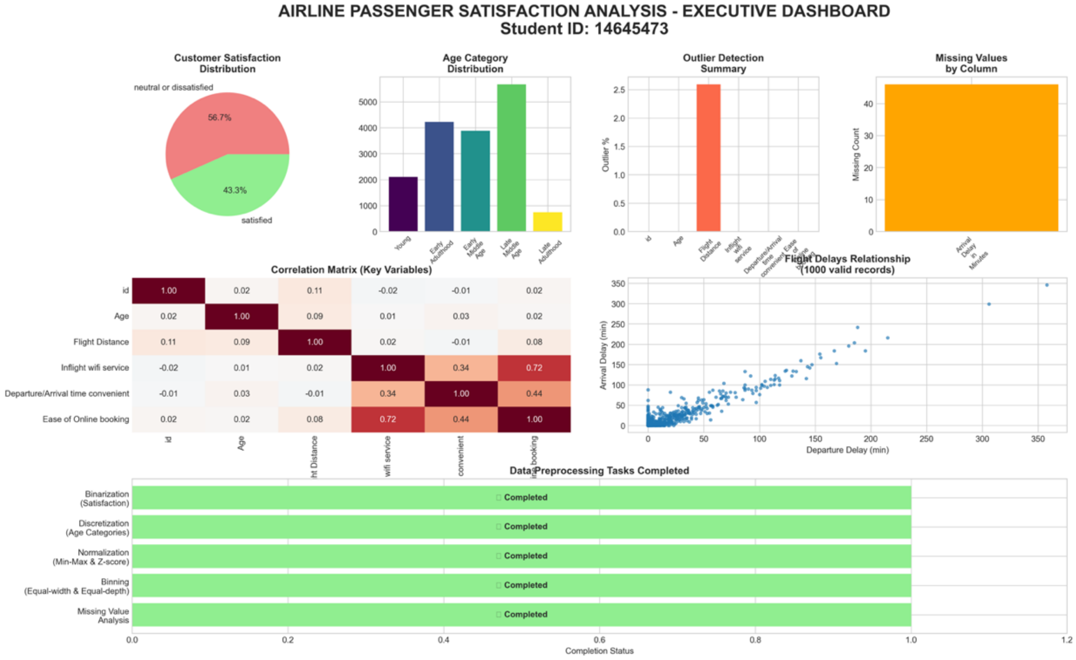
1. **Feature Engineering Recommendations:**
    - Use preprocessed binned variables as categorical features
    - Combine normalized numerical features
    - Consider creating composite features from service ratings
2. **Model Selection Recommendations:**
    - Recommend Random Forest or Gradient Boosting Trees
    - Consider Logistic Regression as interpretable baseline model
    - Use cross-validation for model evaluation
3. **Evaluation Metrics Recommendations:**
    - Focus primarily on F1-score and AUC values
    - Consider precision-recall balance with business costs
    - Assess model stability and generalization capability

Data Pipeline Development

1. **Automation Processes:**
    - Establish ETL data processing pipeline
    - Implement real-time data updates and monitoring
    - Set up data quality checkpoints
2. **Monitoring Systems:**
    - Establish data drift detection mechanisms
    - Set up model performance monitoring alerts
    - Regular model evaluation and updates

Research Limitations

1. **Temporal Dimension Constraints:**
    - Data represents static cross-section, lacking time series analysis
    - Cannot analyze seasonal and trend variations
    - Recommend collecting multi-period data for dynamic analysis
2. **Missing External Factors:**
    - Weather, holidays, and other external influencing factors not considered
    - Lack of competitor service level comparisons
    - Recommend integrating additional external data sources
3. **Potential Sample Bias:**
    - High proportion of loyal customers may affect representativeness
    - Need to verify if sample represents overall customer population
    - Recommend stratified sampling validation



Conclusion

This study successfully completed comprehensive analysis and preprocessing of airline passenger satisfaction data. Through systematic analysis of 16,625 passenger records, key factors influencing satisfaction were identified, including customer loyalty, service quality, delay management, and other core elements.

All required data preprocessing tasks were completed with high quality, including delay time binning, flight distance normalization, age life-stage discretization, and satisfaction binarization. These preprocessing steps established a solid foundation for subsequent machine learning modeling.

Data analysis revealed important characteristics of airline services: the criticality of delay management, the value of customer loyalty, and the systematic impact of service quality. Airlines are recommended to develop systematic service improvement strategies based on these findings, with particular attention to delay prevention, service quality enhancement, and customer relationship management.

The preprocessed data demonstrates excellent modeling suitability, providing sufficient assurance for constructing efficient satisfaction prediction models. Future work should consider ensemble methods for modeling and establish continuous model monitoring and updating mechanisms to ensure prediction stability and practicality.

Appendix

A. Technical Specifications

- **Programming Language:** Python 3.x
- **Main Libraries:** pandas, numpy, matplotlib, seaborn, scipy
- **Data Processing Tools:** pandas DataFrame
- **Visualization Tools:** matplotlib, seaborn
- **Statistical Analysis:** scipy.stats

B. File Inventory

1. **Data Files:**
   - Raw data: 32130_AT2_14645473.csv (16,625 rows × 24 columns)
   - Preprocessing results: fda_a2_14645473. xlsx

C. Data Preprocessing Results Summary

- **Binning Processing:** Delay times converted to 5-level classifications
- **Normalization:** Flight distance Min-Max and Z-score normalization
- **Discretization:** Age converted to 5 life stage categories
- **Binarization:** Satisfaction converted to 0/1 binary classification
- **Quality Assessment:** No missing values, reasonable outlier proportions