

**AIB503**

**End-of-Course Assessment - January Semester 2025**

**Foundation to Python for AI**

Name: Wang YiMing

PI Number: E2510862

Date of Submission: 19 February 2025

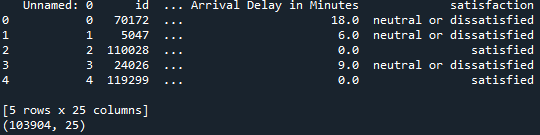
# 

# **1. Introduction**

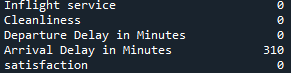
This report aims to apply clustering techniques to the Airline Passenger Satisfaction dataset to identify distinct passenger segments. Through data cleaning, exploratory analysis, K-means clustering, and visualization, we seek to derive actionable insights that the airline can leverage to improve customer satisfaction and operational efficiency.

# **2. Data Preprocessing**

An initial inspection of the data structure (by using df.info() and df.head()) provided an overview of the columns, data types, and any obvious inconsistencies or anomalies. Columns that are not needed for the analysis were dropped to reduce noise in the data. In this case, the columns named ‘Unnamed: 0’ and ‘id’ were removed.

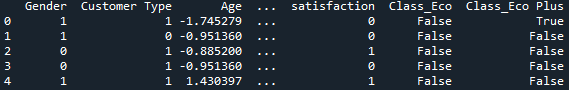


A check for missing values showed that the ‘Arrival Delay in Minutes’ column had 310 entries missing. As this is a numerical variable, I can use the median to fill in these gaps. The median is a solid choice because it minimizes the influence of outliers compared to the mean. After filling in the missing values, I rechecked the dataset to ensure that there were no remaining missing entries.



Binary mapping was applied to features such as ‘Gender (Female → 0, Male → 1)’, ‘Customer Type (disloyal Customer → 0, Loyal Customer → 1)’, ‘Type of Travel (Personal Travel → 0, Business travel → 1)’, and ‘satisfaction (neutral or dissatisfied → 0, satisfied → 1)’. One-hot encoding was used for the ‘Class’ column, creating dummy variables for Economy (Class\_Eco) and Economy Plus (Class\_Eco Plus). The original ‘Class’ value for Business was dropped as a baseline to avoid the dummy variable trap.

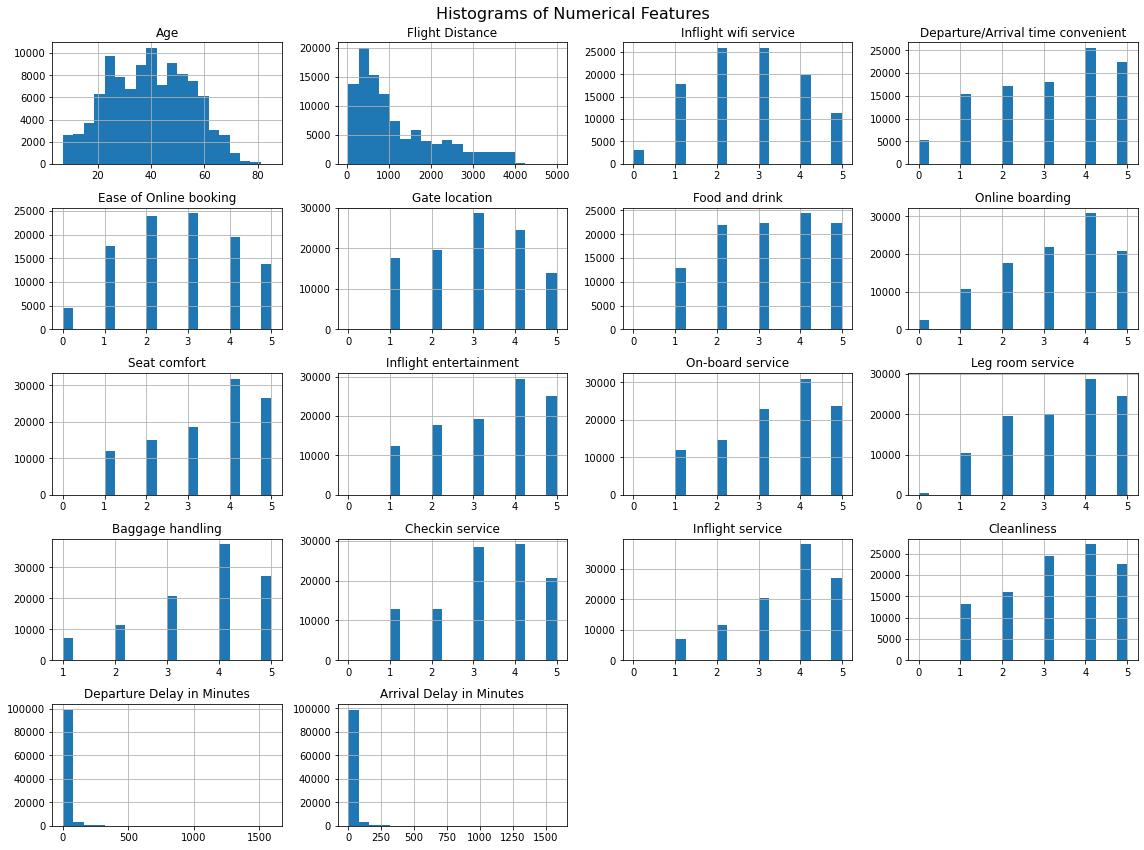
To prevent variables with large numeric ranges from skewing the distance calculations in K-means, we used StandardScaler from sklearn.preprocessing. This tool adjusts each numerical feature to have a mean of 0 and a standard deviation of 1. This step is particularly important for clustering algorithms, as they can be quite sensitive to variations in scale.



# **3. Exploratory Data Analysis**

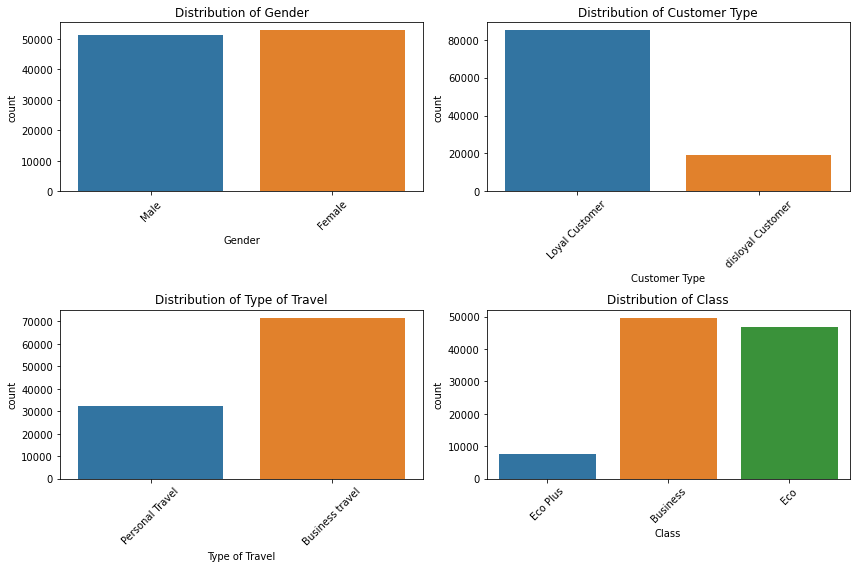
After preprocessing and cleaning the data, we created several plots to explore the distribution and relationships among the features in the Airline Passenger Satisfaction dataset. In this section, we highlight the main findings from the histograms of numerical variables, bar charts of categorical variables, a correlation heatmap, and a pair plot.

**1. Histograms of Numerical Features**



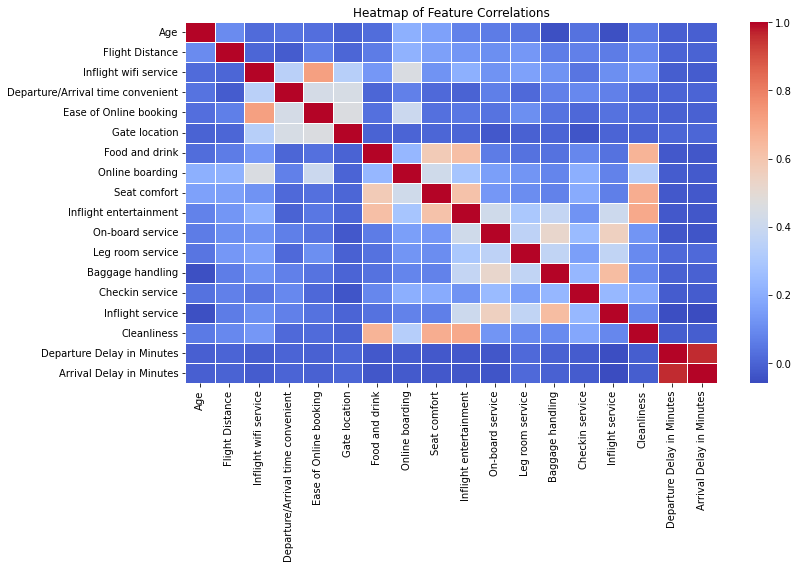
* **Age:** The age distribution of passengers varies widely, spanning from teenagers to those in their eighties, with a significant number of travelers falling between the ages of 20 and 50.
* **Flight Distance:** This characteristic shows a right-skewed pattern, which means that while the majority of flights are relatively short to medium in time, there are also a few that are considerably longer.
* **Rating-Based Features:** These ratings are measured on a scale from 0 to 5, where 0 typically means “Not Applicable.” Most ratings tend to cluster between 3 and 5, indicating a generally high level of satisfaction across various service aspects.
* **Departure/Arrival Delay in Minutes:** Both of these metrics are heavily right-skewed, suggesting that while most flights have minimal delays, there are a few that experience extremely long delays, potentially exceeding 1,000 minutes.

**2. Bar Chart of Categorical Variables**

****

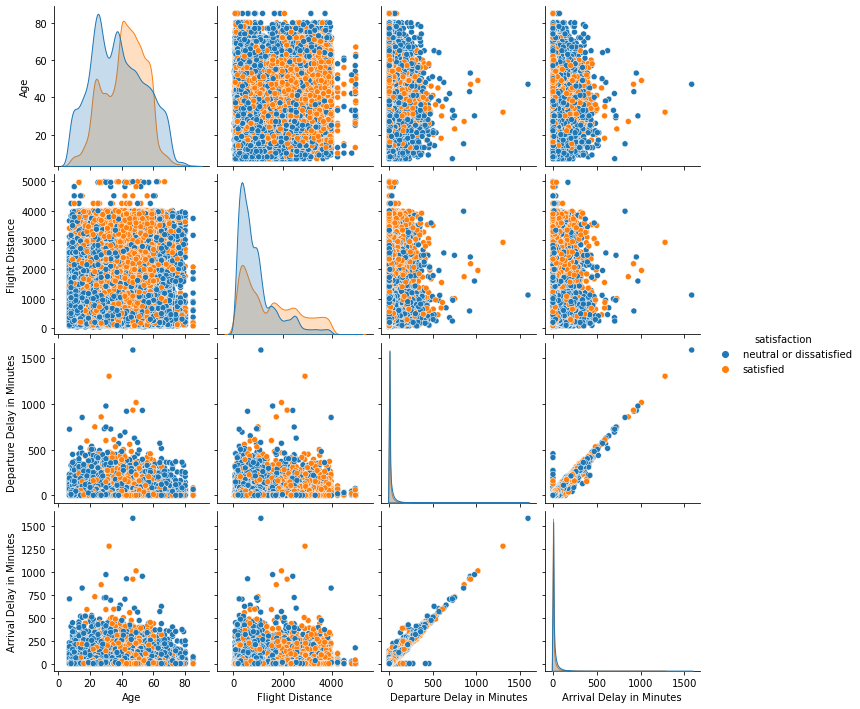
* **Gender:** The dataset shows a fairly balanced gender distribution, although there is a slightly higher number of male passengers compared to female passengers.
* **Customer Type:** Most individuals are categorized as Loyal Customers, while a smaller segment is identified as disloyal.
* **Type of Travel:** A greater proportion of passengers traveled for Business rather than Personal reasons, indicating a focus on business-related travel.
* **Class:** The majority of travelers are in Business or Eco class, with Eco Plus representing the smallest category.

**3. Correlation Heatmap**

****

* The heatmap indicates a strong positive correlation between 'Departure Delay' and 'Arrival Delay,' which makes sense. When departures are delayed, arrivals tend to be delayed as well

**4. Pair Plot**

****

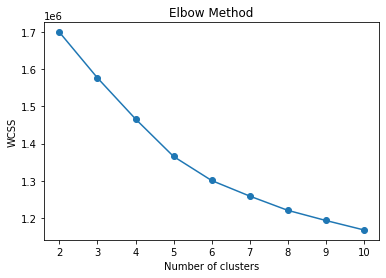
* **Departure vs. Arrival Delay:** The diagonal pattern highlights a strong positive correlation. Interestingly, passengers identified as satisfied (orange points) are more common on flights with minimal departure and arrival delays. In contrast, neutral or dissatisfied passengers (blue points) are scattered across various delay ranges, although dissatisfaction tends to rise with longer delays.

# **4. K-Means Clustering**

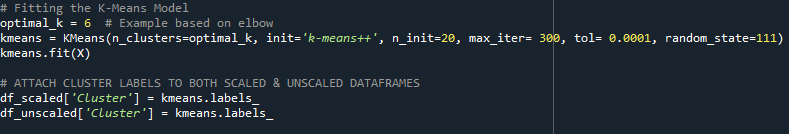
**1. Determining the Optimal Number of Clusters**

To cluster the airline passenger dataset, we employed K-Means, a widely used unsupervised machine learning algorithm that partitions data points into K distinct clusters. However, selecting an appropriate K is crucial. We used the **Elbow Method.**

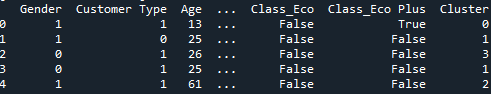
* We trained K-Means models for a range of K values (K=2 to K=10)
* For each model, we calculated the Within-Cluster Sum of Squares (WCSS), which measures the variance within each cluster.
* We then plotted K on the x-axis and WCSS on the y-axis.



Based on the results from the elbow method, I chose **K=6**. With **Silhouette Score for k = 6: 0.177**, indicates a moderate separation of data points into their assigned clusters. I then fitted the final model.



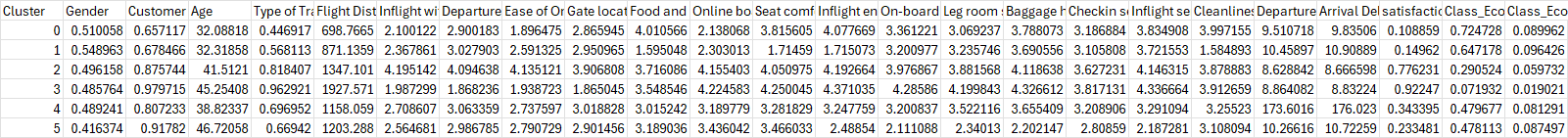
Below are the top few data with their respective clusters.



# 

# **5. Interpretation of Clusters and Recommendations**

After applying K-Means (K=6), we obtained the following mean feature values for each cluster. These results illuminate how passengers differ by demographics (age, loyalty), travel patterns (flight distance, travel type), service ratings (seat comfort, inflight entertainment, etc.), and overall satisfaction.



### 

**Cluster 0**

* Age: ~32.09 (one of the younger segments)
* Flight Distance: ~699 miles (short-haul)
* Business vs. Personal Travel: ~45% business (Type of Travel: 0.4469)
* Inflight Wi-Fi: 2.10 (below mid-range on a 0–5 scale)
* Food & Drink: 4.01 (relatively high)
* Seat Comfort: 3.82 (good)
* Inflight Entertainment: 4.08 (good)
* Ease of Online Booking: 1.90 (low)
* Departure/Arrival Delays: ~9.51 / 9.83 minutes (moderate)
* Class: ~72% Economy (High)
* Satisfaction: 0.1089 (extremely low overall)

While certain inflight services like food, seat comfort, and entertainment receive relatively high ratings, overall satisfaction is the lowest among all groups, sitting at around 11%. It's worth noting that online booking convenience and Wi-Fi are perceived as weaker areas. Additionally, a significant portion of travelers are economy-class flyers, making up about 72%, which may reflect a focus on budget-friendly travel.

1. Enhance Digital Experience: Improve the airline’s app/website to create a more seamless booking experience, particularly for short-haul travelers who tend to book more often.
2. Boost Wi-Fi: Reliable Wi-Fi can enhance even short flights, which is especially important for younger passengers
3. Target Quick Upgrades: Provide affordable Wi-Fi options or other upgrades to improve satisfaction without incurring significant costs.

**Cluster 1**

* Age: ~32.32 (also young)
* Flight Distance: ~871 miles (short-to-medium haul)
* Business vs. Personal Travel: ~57% business (Type of Travel: 0.5681)
* Food & Drink: 1.59 (very low)
* Seat Comfort: 1.71 (quite low)
* Inflight Entertainment: 1.72 (also low)
* Cleanliness: 1.58 (low)
* Ease of Online Booking: 2.59 (better than Cluster 0)
* Inflight Wi-Fi: 2.37 (slightly higher than Cluster 0)
* Delays: ~10.46 / 10.90 min
* Satisfaction: 0.1496 (low, though higher than Cluster 0)

Cluster 1 is far more dissatisfied with onboard basics: seat comfort, food quality, entertainment, and even cleanliness. Despite these issues, they find the airline’s digital/booking experience more tolerable than Cluster 0.

1. Improve Cabin Experience: Prioritize improved seat padding, upgraded meal options, and comprehensive cabin cleaning.
2. Mid-Flight Entertainment: Even moderate-distance flights require reliable inflight entertainment.

**Cluster 2**

* Age: ~41.51
* Customer Type: ~88% Loyal (High)
* Flight Distance: ~1347 miles (medium-haul)
* Business vs. Personal Travel: ~82% business (Type of Travel: 0.8184)
* Inflight Wi-Fi: 1.20 (low)
* Ease of Online Booking: 4.14 (very high)
* Seat Comfort: 4.05, Inflight Entertainment: 4.19 (both excellent)
* Baggage Handling: 4.12 (high)
* Departure/Arrival Time Convenience: ~4.09 (high)
* Delays: ~8.63 / 8.67 min (fairly low)
* Satisfaction: 0.7762 (high)

This group, primarily made up of loyal business travelers (about 88% loyal), expresses high levels of satisfaction across various service aspects, including seat comfort, punctuality, and particularly the simplicity of the booking process. Interestingly, the inflight Wi-Fi rating is only around 1.20, yet this doesn't seem to affect their overall satisfaction, likely because other services, like entertainment and baggage handling, are outstanding.

1. Maintain High Standards: They are already quite satisfied, it's important to keep providing priority boarding, comfortable seating, and minimal delays for these frequent flyers.
2. Consider Wi-Fi Upgrades: Although the Wi-Fi rating is low, improving this service could enhance overall satisfaction even further.

**Cluster 3**

* Age: ~45.25
* Flight Distance: ~1927.57 (long-haul)
* Type of Travel: ~96% business (very high)
* Departure/Arrival time convenient: ~1.86 (very low)
* Customer Type: 0.9797 (~98% loyal)
* Seat Comfort: 4.25, Inflight Entertainment: 4.37 (very high)
* Food & Drink: 3.55 (above average)
* Delays: ~8.86 / 8.83 min (moderate-low)
* Satisfaction: 0.9225 (the highest among all clusters)

This group mainly consists of older business travelers on long-haul flights, who value excellent seat comfort, entertainment options, and strong loyalty programs. Even though they rate the convenience of departure and arrival times at only around 1.86, their overall satisfaction remains very high, indicating that onboard services and reliability outweigh some scheduling issues.

1. Retain Premium Experience: Keep investing in long-haul cabin comfort, including legroom, advanced entertainment systems, and high-quality meals.
2. Optimize Flight Schedules: By improving departure and arrival times or providing more flight options, the overall experience could be nearly flawless.

**Cluster 4**

* Age: ~38.82
* Flight Distance: ~1158.06
* Type of Travel: ~97% business
* Inflight Wi-Fi: 2.71 (relatively high vs. other clusters)
* Seat Comfort: 3.28 (moderate)
* Delays: 173.60 / 176.02 min (extremely high)
* Satisfaction: 0.3434 (low, but not as low as Clusters 0–1)

Even with reliable Wi-Fi and solid service ratings, extensive delays significantly impact overall satisfaction. Business travelers, who depend on punctuality, find that frequent major delays overshadow any other benefits.

1. Priority: Reduce Delays: Look into route scheduling, ground handling, and aircraft availability to tackle the significant delay problem in this area.
2. Compensation & Communication: Provide lounge access, real-time updates, or flight vouchers during major delays to help reduce dissatisfaction.

**Cluster 5**

* Age: ~46.72 (the oldest cluster)
* Flight Distance: ~1203.29
* Type of Travel: ~67% business, ~33% personal
* Customer Type: ~92% Loyal (very high)
* Seat Comfort: 3.47 (moderate)
* Inflight Entertainment: 2.49 (fairly low)
* On-board Service: 2.11 (lowest among all clusters)
* Baggage Handling: 2.20 (also low)
* Delays: ~10.27 / 10.72 min (moderate)
* Satisfaction: 0.2335 (low)

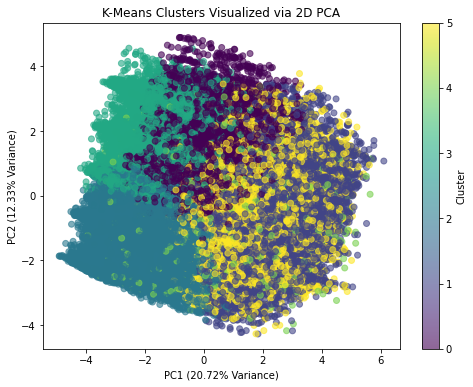
This cluster tend to be loyal (around 92%) and are generally older travelers, their satisfaction levels are quite low (approximately 23%). The primary issues seem to stem from aspects of the travel experience, such as onboard service (rated around 2.11) and subpar inflight entertainment (rated about 2.49). Although delays are not particularly frequent, the overall quality of their travel experience leaves much to be desired.

1. Improve Service Basics: These frequent flyers, who are mostly older, require consistent and dependable experiences, from baggage handling to inflight service.
2. Mixed Travel Motivation: Since one-third of these travelers are on personal trips, consider offering promotions or loyalty benefits that address both business and leisure travel needs.

# **6. Visualizing Clusters and Discussing Key Findings, Challenges, and Next Steps**

**PCA-Based Cluster Visualization**

To visualize our six clusters in a two-dimensional space, we applied Principal Component Analysis (PCA) to simplify the dataset by reducing its dimensionality from numerous features (both numeric and encoded categorical) to just two principal components (PC1 and PC2).

****

1. **Partial Overlap:** The clusters show considerable overlap in this 2D representation, implying that the true separation might exist in a higher-dimensional space.
2. **Some Regional Tendencies:** A few clusters seem to be more densely packed in the upper left area (like cluster 3 or cluster 2), while others are more spread out towards the center and right (such as cluster 1 and cluster 5).
3. **Moderate Differentiation:** While the clusters aren't distinctly separated, there are subtle variations in color and areas where one cluster is more prominent, suggesting that K-Means has identified some structure that PCA can't fully capture in just two dimensions.

**Key Findings from the Overall Analysis**

* **Cluster Profiles Show Significant Variation**
  + Clusters vary in terms of average age, travel distance, service ratings (such as seat comfort, entertainment, and cleanliness), and operational metrics (like delays). While certain clusters report high satisfaction levels due to minimal delays and quality onboard service, others face challenges like poor Wi-Fi or inadequate cabin conditions.
* **High-Dimensional Separation**
  + Even though the 2D PCA plot indicates a fair amount of overlap, the K-Means algorithm utilizes all features in an n-dimensional space. As a result, the clusters are more clearly defined in that higher-dimensional context.
* **Actionable Segmentation**
  + Each segment has its own specific needs, some prioritize strong Wi-Fi, while others focus on seat comfort or punctuality. Customized enhancements can lead to more effective resource distribution and improved overall customer satisfaction.

**Challenges Faced**

* **Data Imbalance and Missing Values**
  + A considerable number of missing entries for ‘Arrival Delay in Minutes’ necessitated imputation. While using the median for filling is a reliable method, it could still influence cluster formation if there are significant outliers present.
  + Certain categories (like Economy versus Economy Plus) have uneven sample sizes, which might skew the cluster centers towards the more frequently represented categories.
* **High-Dimensional and Correlated Features**
  + The dataset includes numerous service ratings and delay metrics, complicating the interpretation of cluster boundaries.
* **Difficulties Selecting K**
  + The Elbow Method did not reveal a distinct bend, and the silhouette score is not particularly close to 1. Overlapping clusters in the 2D PCA plot suggest that the “optimal” K might still lead to some uncertainty.
* **Interpretation of Categorical vs. Numeric**
  + Combining both numeric (scaled) and binary/categorical variables (unscaled) in K-Means can complicate the interpretation of the distance metric, necessitating careful data preparation.

**Potential Next Steps**

1. **Enrich the Dataset**
   * **Add More Features**: Include route-specific details (such as distinctions between short-haul and long-haul flights), loyalty program tiers, or booking lead times to identify more nuanced passenger segments.
   * **Address Outliers**: Particularly for delay times, employing advanced outlier detection or winsorizing could enhance the quality of clusters.
2. **Refine the Clustering Approach**
   * Explore Alternative Algorithms: Techniques like Hierarchical Clustering may reveal more intricate cluster shapes or automatically filter out outliers.
3. **Conduct Robust Validation**
   * Stability Analysis: Execute K-Means multiple times with varying random seeds or sample subsets to verify the consistency of clusters.
4. **Continuous Feedback Loop**
   * Implement Segment-Specific Strategies: For each cluster’s operational or service deficiencies (such as delays, Wi-Fi issues, or seat comfort), initiate pilot improvements.
   * Track Satisfaction Over Time: Regularly re-cluster with updated data to assess whether changes in service or operations lead to more distinct (and more satisfied) passenger segments.
5. **Better Dimensionality Reduction**
   * Consider using 3D PCA for cluster visualization, as it might uncover separations that are not apparent in a standard 2D PCA plot.