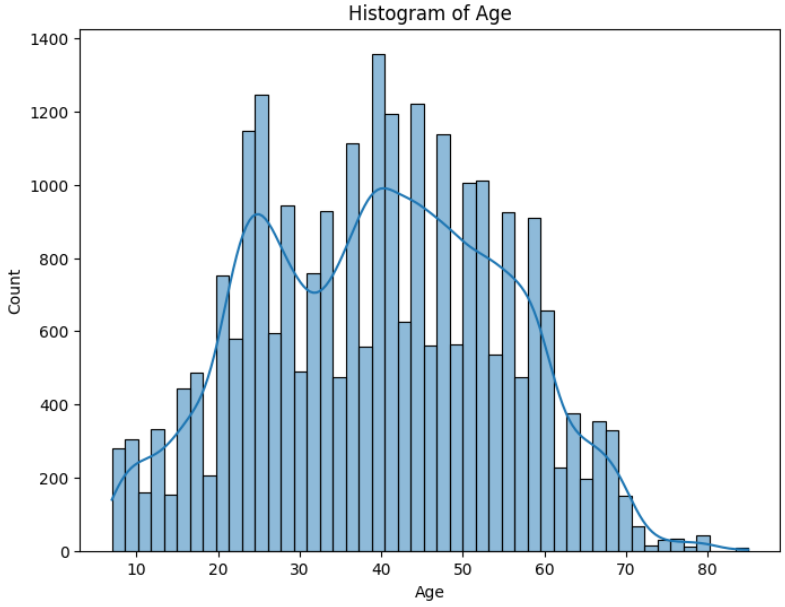
Exploratory Data Analysis of Airline Passenger Characteristics and Flight Metrics

Student Name: J. Padala
Student ID: 23040317
GitHub Repository: https://github.com/Jayendra727/clustering-and-fitting/upload

There are 18 columns and 25976 rows in the dataset. Along with flight parameters like travel type, class, and distance, it includes consumer information like gender, age, and type. Along with evaluations for overall satisfaction and departure/arrival delay times, ratings for other services including Wi-Fi, boarding, and cleanliness are given.

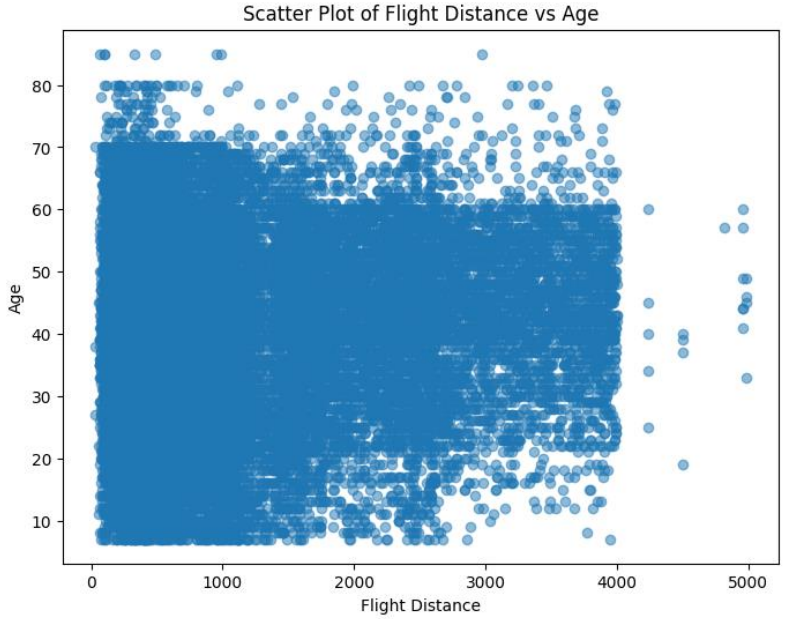
Features	Mean	Median	Standard Deviation	Skewness	Kurtosis
ID	65005.65	65319.5	37611.52	-0.011	-1.206
Age	39.62	40	15.13	-0.000088	-0.71
Flight Distance	1193.78	849	998.68	1.102	0.25
Wi – Fi	2.72	3	1.33	0.0407	-0.858
Departure/ Arrival	3.04	3	1.53	-0.324	-1.05
Online Booking	2.75	3	1.41	-0.0206	-0.92
Gate location	2.97	3	1.28	-0.055	-1.03
Food and Drink	3.21	3	1.331	-0.17	-1.144
Online Boarding	3.26	4	1.355	-0.46	-0.68
Seat Comfort	3.44	4	1.32	-0.498	-0.911
Inflight Entertainment	3.357	4	1.338	-0.371	-1.06
On-board service	3.38	4	1.282	-0.426	-0.874
Leg room service	3.35	4	1.318	-0.3411	-0.99
Baggage Handling	3.63	4	1.176	-0.678	-0.3701
Check-in service	3.31	3	1.269	-0.372	-0.83
Inflight service	3.64	4	1.1806	-0.696	-0.36
Cleanliness	3.28	3	1.319	-0.304	-1.02
Departure Delay	14.306	0	37.42	7.19	102.16
Arrival Delay	14.74	0	37.51	NaN	NaN

The dataset indicates that the average age of the clients is approximately forty years old. With a notable standard deviation of 998 kilometres, flights usually span 1193 kilometres. Services are generally rated between three and four, with the highest mean ratings going to baggage handling and in-flight service, at 3.63 and 3.65, respectively. Arrival and departure delays often take 14–15 minutes on average, with standard deviations of about 37 minutes indicating significant fluctuation. Positively skewed distributions are seen in a few characteristics, including flight distance, arrival delay, and departure delay, which may indicate the existence of outliers. Finding areas for improvement is made easier with the help of the correlation heatmap, which provides insights into the links between features and overall happiness.

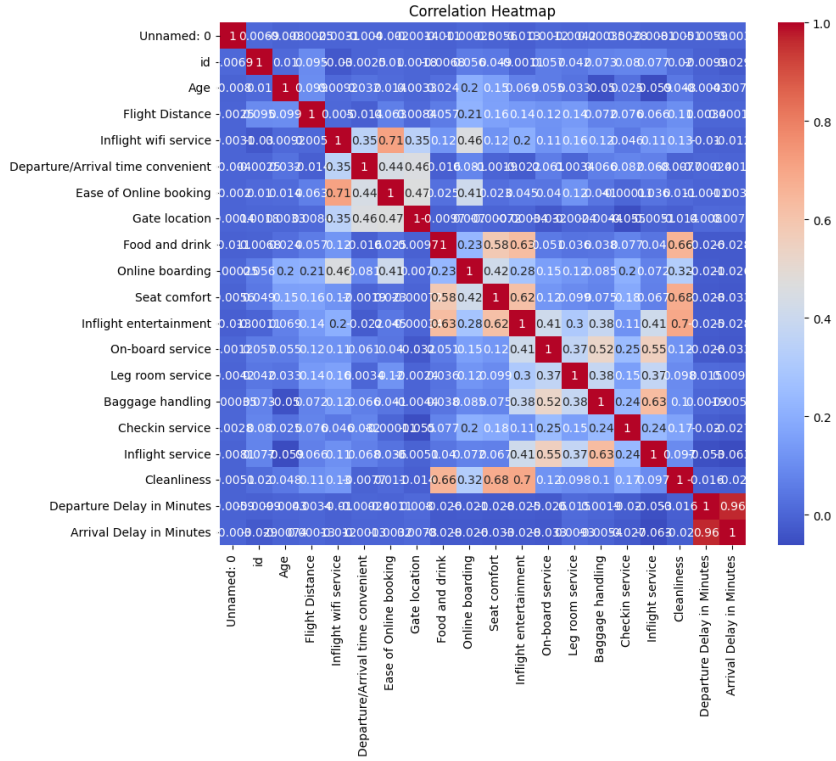


An age distribution histogram is shown on this graph. Age groupings are represented by the x-axis, while the number or frequency of people in each age group is displayed on the y-axis. With the largest

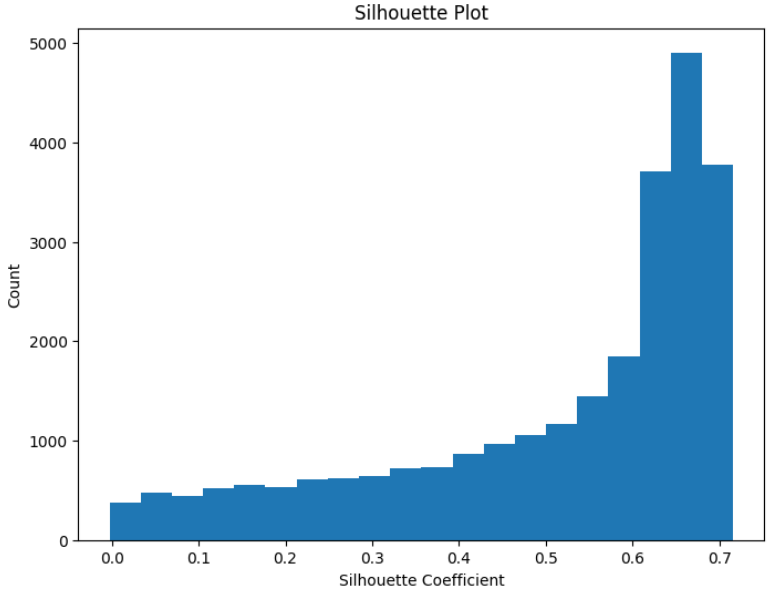
frequency occurring in the 30- to 40-year-old age range, the distribution resembles a bell shape, suggesting that a sizable amount of the data is concentrated in this age range.



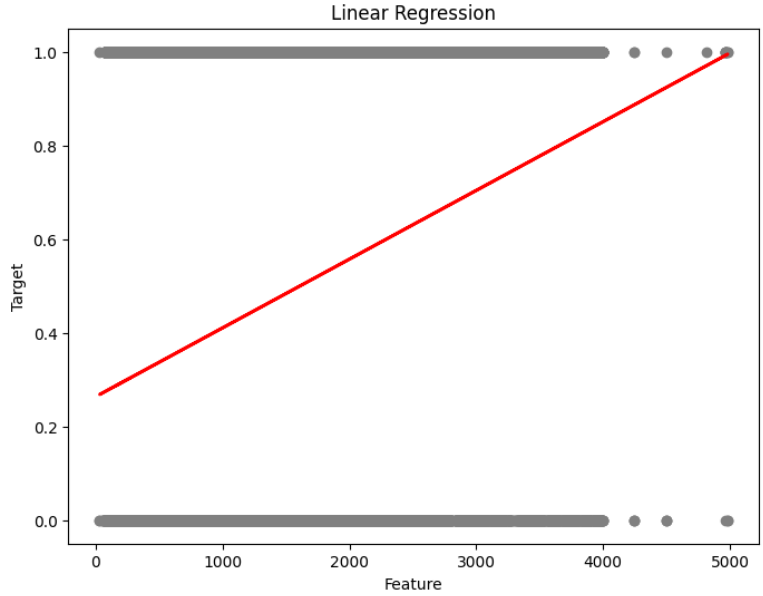
This scatter plot illustrates the correlation between age (y-axis) and flight distance (x-axis). The age is represented by the y-coordinate, and the flight distance is represented by the x-coordinate, for each point on the plot. A dense concentration of points is seen in the lower-left area of the map, indicating that many of the dataset's members are younger and have flown shorter distances. Points are also dispersed among the greater age and flying distance groups, though.



A correlation heatmap, like the one shown above, shows the correlation coefficients between various variables in a dataset. The correlation coefficients are color-coded to indicate the degree and direction of the association, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation). Variables like "Departure Delay in Minutes" and "Arrival Delay in Minutes" appear to have moderate to strong positive relationships with variables like "Ease of Online booking," "Gate location," and "Online boarding" in this heatmap. Conversely, it appears that the variables "Food and drink" and "Seat comfort" have less of an association with the delay variables.



The silhouette coefficients of the data points are shown in this silhouette plot graph. The silhouette coefficient expresses the degree to which each data point, relative to other clusters, fits into the designated cluster. The count or frequency of data points with a specific silhouette coefficient value is displayed on the y-axis, while the x-axis displays the silhouette coefficient values, which range from -1 to 1. A reasonably excellent fit of the data points within their designated clusters is indicated by the plot, which shows a large concentration of data points with silhouette coefficients between 0.6 and 0.7.



The linear regression model is depicted in this graph. Given that the x-axis is labelled "feature," one of the dataset's independent variables or predictors is probably represented by it. "Target," the dependent variable or the variable that the linear regression model is predicting, is indicated on the y-axis. The best-fitting linear relationship between the feature and the target variables is represented by the red line, which is the linear regression line. The black dots, which represent the data points, are dispersed throughout the regression line, suggesting a slight departure from the linear model.

All in all, these visualizations offer a thorough summary of the dataset, making it possible to spot important links, patterns, and possible areas for more research or improvement. Airlines can improve passenger experiences, operational efficiency, and service quality by using these insights to guide their decision-making.