BACKGROUND

As someone who has always been interested in travel, I understand that the journey, primarily consisting of inflight time, is an essential part of the traveling experience. Growing up in Singapore, every year I would read the news about the Skytrax World Airline Awards , seeing Singapore Airlines in one of the top spots, often first place.
But, what exactly makes an airline the best airline in the world?

The problem questions:

- How can airlines know which factors are most important for/the strongest contributors to customer satisfaction?
- Which factors contributing to satisfaction correlate with one another?
- Are a random set of customers more likely to be satisfied or neutral/dissatisfied with an airline experience?

OBJECTIVE

- Clean, fit a model to predict, and visualize & interpret the passenger satisfaction data to find the relationships between category ratings and overall satisfaction of the passenger, as well as evaluate the results.

DATA

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Ease of Online booking	Gate location	Food and drink
2	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus	460	3	4	3	1	5
3	1	5047	Male	disloyal Customer	25	Business travel	Business	235	3	2	3	3	1
4	2	110028	Female	Loyal Customer	26	Business travel	Business	1142	2	2	2	2	5
5	3	24026	Female	Loyal Customer	25	Business travel	Business	562	2	5	5	5	2
6	4	119299	Male	Loyal Customer	61	Business travel	Business	214	3	3	3	3	4
7	5	111157	Female	Loyal Customer	26	Personal Travel	Eco	1180	3	4	2	1	1
8	6	82113	Male	Loyal Customer	47	Personal Travel	Eco	1276	2	4	2	3	2
9	7	96462	Female	Loyal Customer	52	Business travel	Business	2035	4	3	4	4	5
10	8	79485	Female	Loyal Customer	41	Business travel	Business	853	1	2	2	2	4
11	9	65725	Male	disloyal Customer	20	Business travel	Eco	1061	3	3	3	4	2
12	10	34991	Female	disloyal Customer	24	Business travel	Eco	1182	4	5	5	4	2
13	11	51412	Female	Loyal Customer	12	Personal Travel	Eco Plus	308	2	4	2	2	1
14	12	98628	Male	Loyal Customer	53	Business travel	Eco	834	1	4	4	4	1
15	13	83502	Male	Loyal Customer	33	Personal Travel	Eco	946	4	2	4	3	4
16	14	95789	Female	Loyal Customer	26	Personal Travel	Eco	453	3	2	3	2	2
17	15	100580	Male	disloyal Customer	13	Business travel	Eco	486	2	1	2	3	4
18	16	71142	Female	Loyal Customer	26	Business travel	Business	2123	3	3	3	3	4
19	17	127461	Male	Loyal Customer	41	Business travel	Business	2075	4	4	2	4	4
20	18	70354	Female	Loyal Customer	45	Business travel	Business	2486	4	4	4	4	3
21	19	66246	Male	Loyal Customer	38	Personal Travel	Eco	460	2	3	3	2	5

	O	P	Q	R	S	T	U	V	W	X	
1	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes	satisfa
2	5	5	4	3	4	4	5	5	25	18	neutral
3	1	1	1	5	3	1	4	1	1	6	neutral
4	5	5	4	3	4	4	4	5	0	0	satisfie
5	2	2	2	5	3	1	4	2	11	9	neutral
6	5	3	3	4	4	3	3	3	0	0	satisfie
7	1	1	3	4	4	4	4	1	0	0	neutral
8	2	2	3	3	4	3	5	2	9	23	neutral
9	5	5	5	5	5	4	5	4	4	0	satisfie
10	3	1	1	2	1	4	1	2	0	0	neutral
11	3	2	2	3	4	4	3	2	0	0	neutral
12	2	2	3	3	5	3	5	2	0	0	neutral
13	1	1	1	2	5	5	5	1	0	0	neutral
14	1	1	1	1	3	4	4	1	28	8	neutral
15	4	4	4	5	2	2	2	4	0	0	satisfie
16	2	2	4	3	2	2	1	2	43	35	neutral
17	1	4	2	1	4	1	3	4	1	0	neutral
18	4	4	5	3	4	5	4	4	49	51	satisfie
19	4	5	5	5	5	3	5	5	0	10	satisfie
20	5	5	5	5	5	3	5	4	7	5	satisfie
21	5	5	1	2	4	3	2	5	17	18	neutral

WHAT WE HAVE DONE ACTUALLY?

Table of Contents:

DATA CLEANING

Understanding the Structure of the Dataset

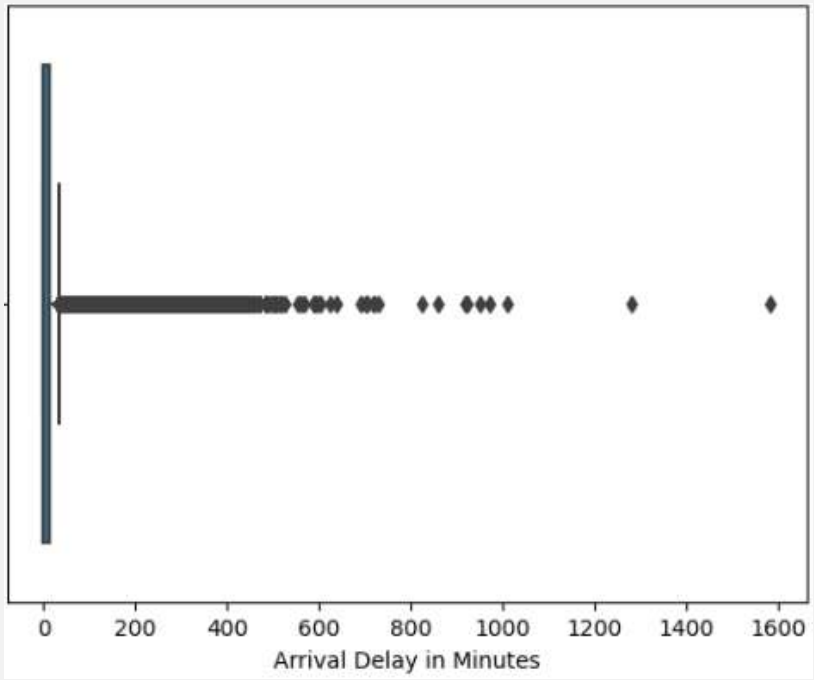
- The data has 103,904 entries of 25 columns.
- We can see that each row represents the data for one customer, and that there is information about their 1-5 ratings (0 if there is no data on what they rated that category) for each category.
- *There are a few things to notice in terms of data cleaning:*

  a)The column "Arrival Delay in Minutes" has 103904-103594 = **310** missing values

  b)The first two columns are unnecessary, as they are not relevant to the analysis

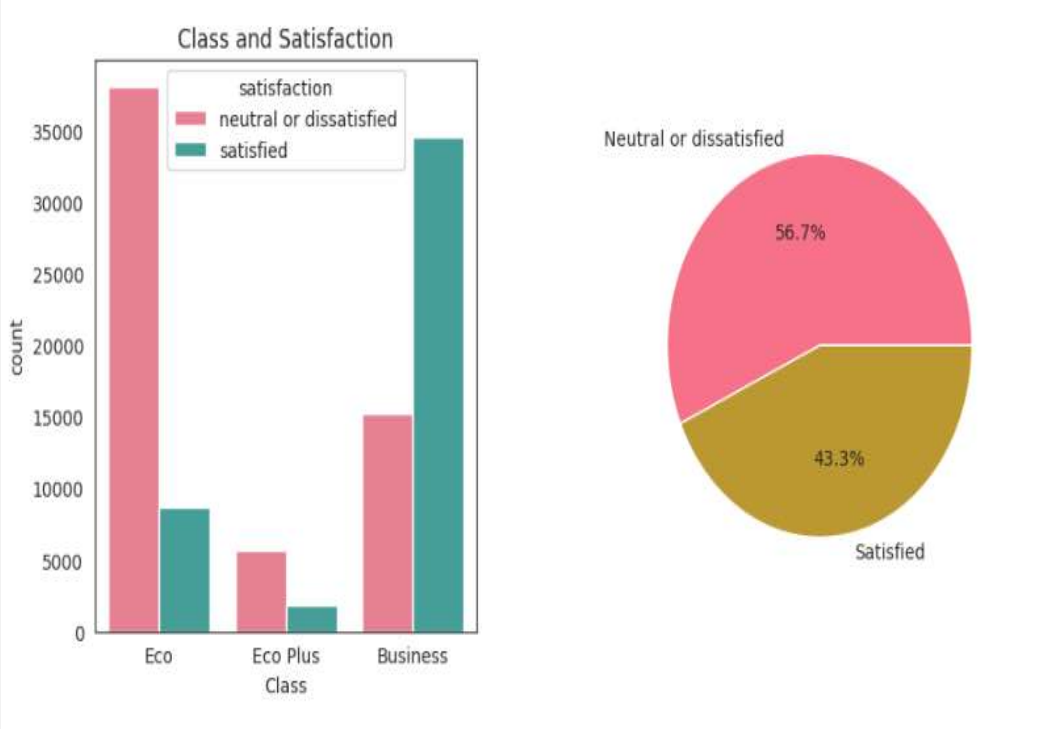
PREPARING DATA FOR ANALYSIS

(THAT MISSING VALUE)

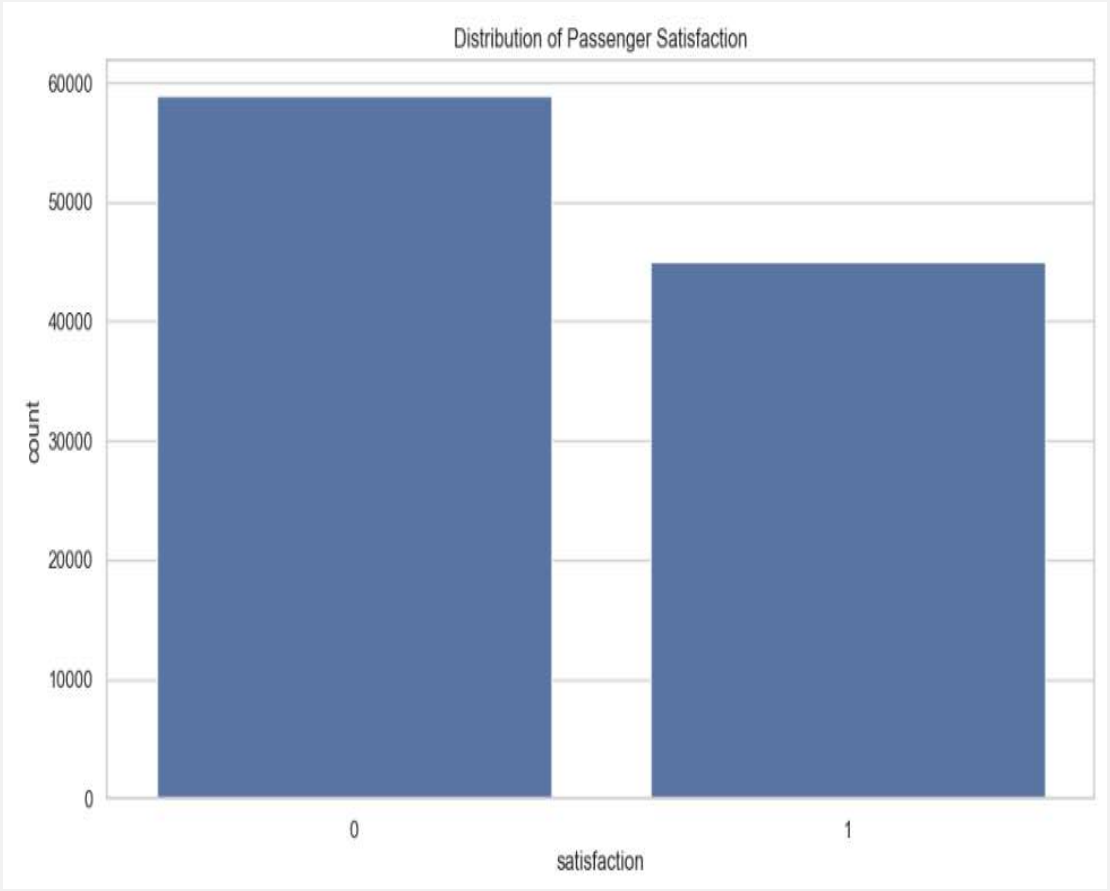


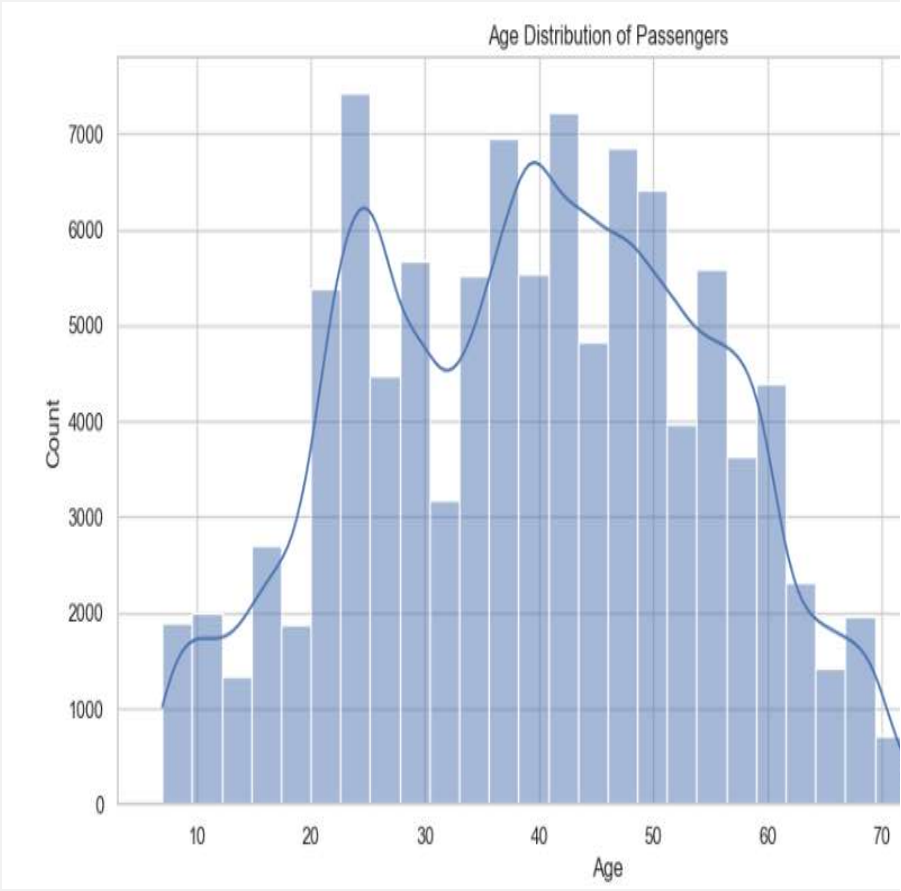
- Despite the mean of the 'Arrival Delay in Minutes' column being 15.13, we can see from the boxplot that almost any value except 0 is an outlier.
- The mode of the column is 0 by a large margin, and the 310 missing values are few compared to the large number of total entries.
- In considering this along with most values being outliers other than 0, I will fill in the Average Delay in Minutes column's null values with 0.

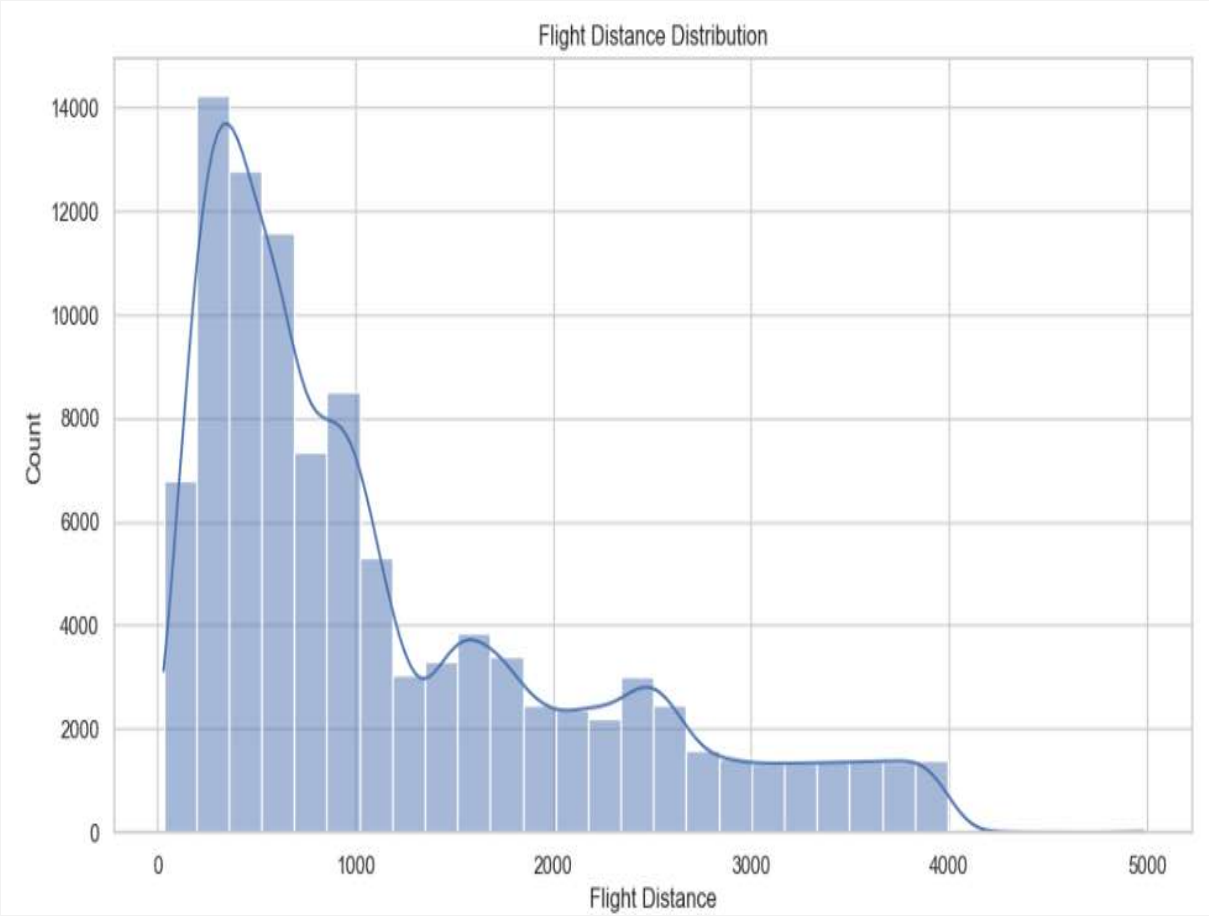
DATA VISUALIZATION AND ANALYSIS

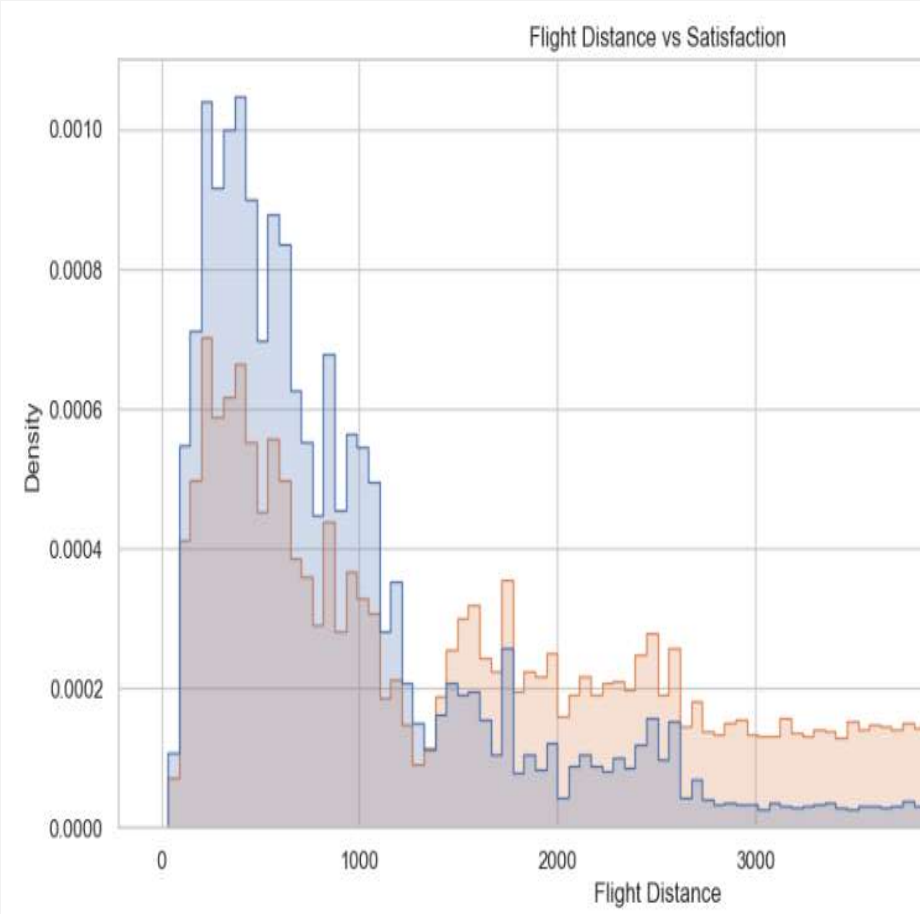


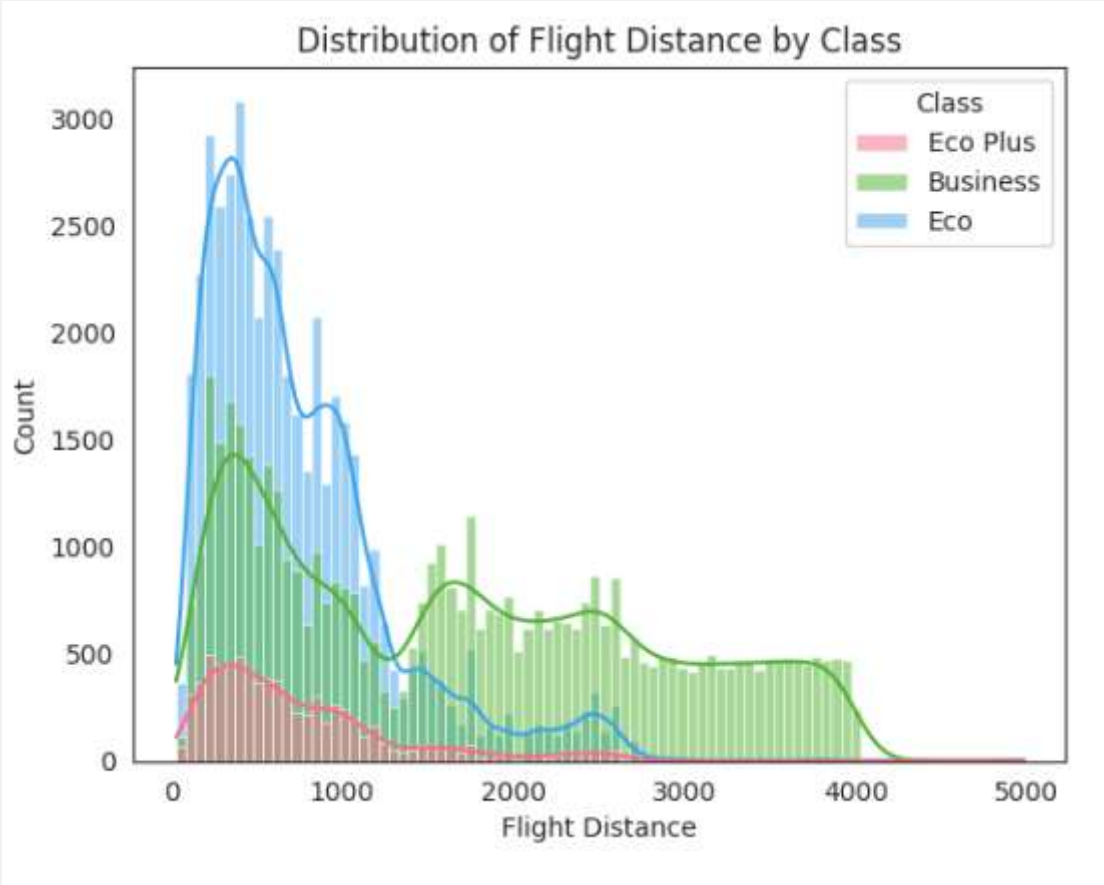
- 56.7% of the passengers in the survey were neutral or dissatisfied, the rest being satisfied with the airline.
- It is reasonable to assume that the class passengers are traveling in affects their satisfaction, which is corroborated by the bar plot on the left.
- Economy and Economy Plus tend to have more neutral or dissatisfied customers than satisfied customers, whereas Business Class has more satisfied customers than dissatisfied customers.
- Notably, Economy Plus has the lowest count, therefore, there is an unequal distribution in the class of customers in the dataset, however, this may also be because Economy Plus is generally flown the least.

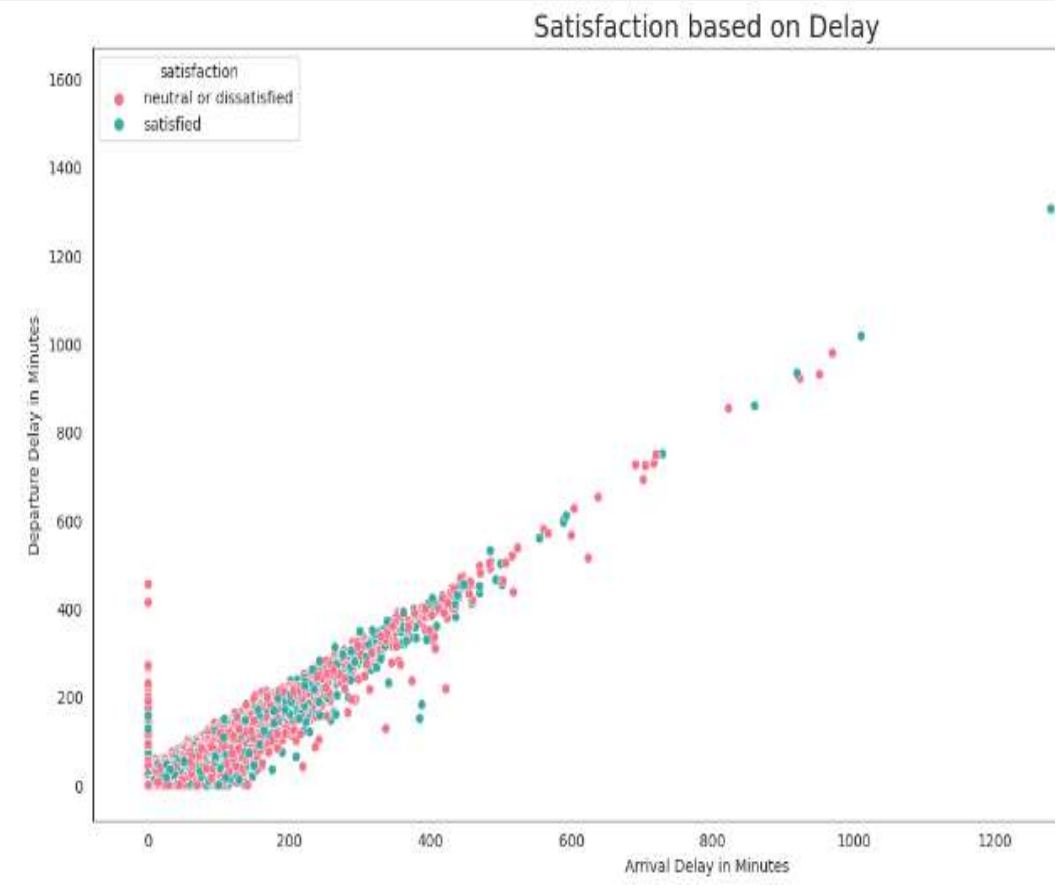

Distribution of Passenger Satisfaction
60000
50000
40000
30000
20000
10000
0
count
0
1
satisfaction

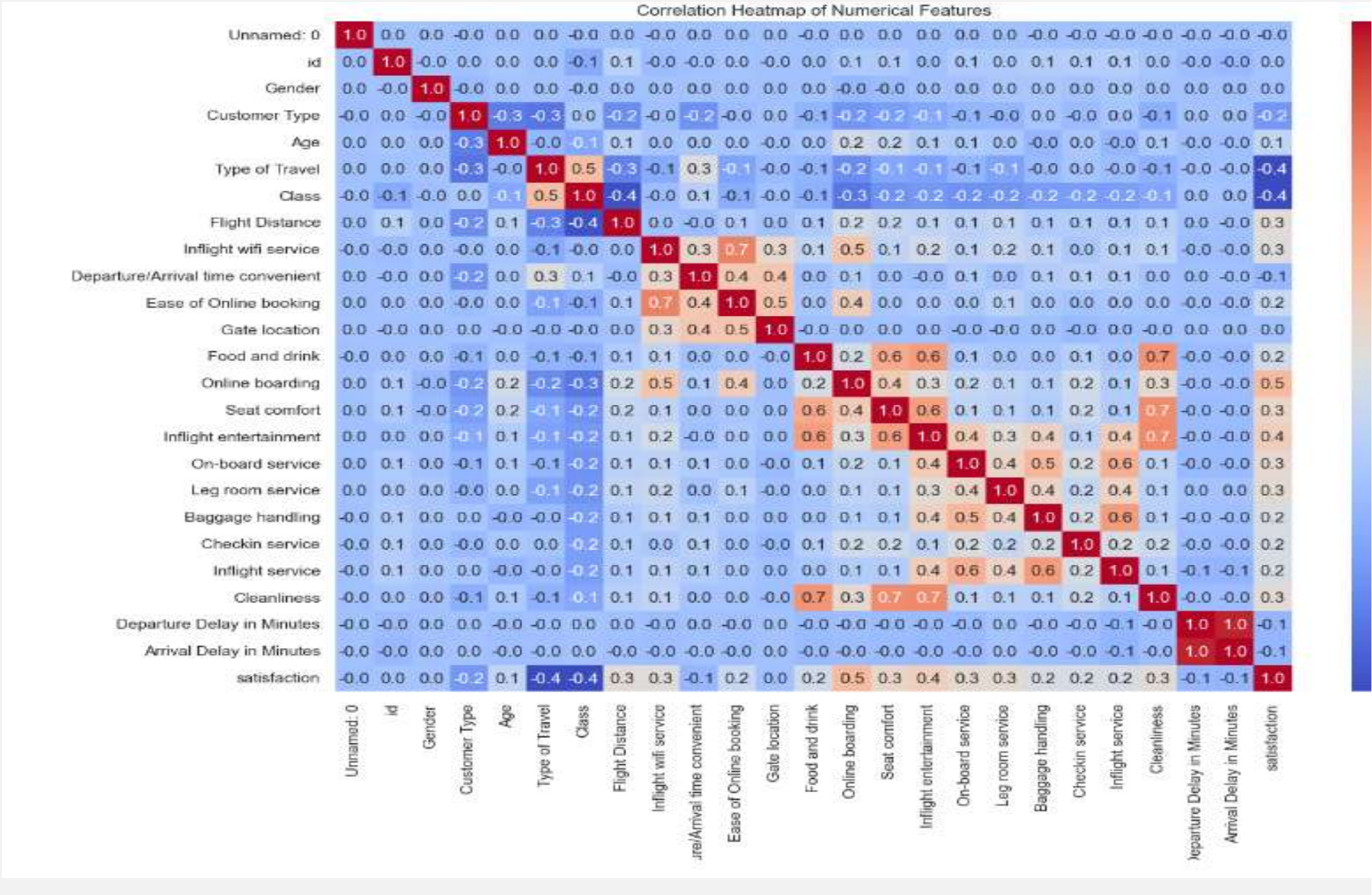

Age Distribution of Passengers
7000
6000
5000
4000
3000
2000
1000
0
Count
10
20
30
40
50
60
70
Age


Flight Distance Distribution
Count
Flight Distance
0
2000
4000
6000
8000
10000
12000
14000
0
1000
2000
3000
4000
5000

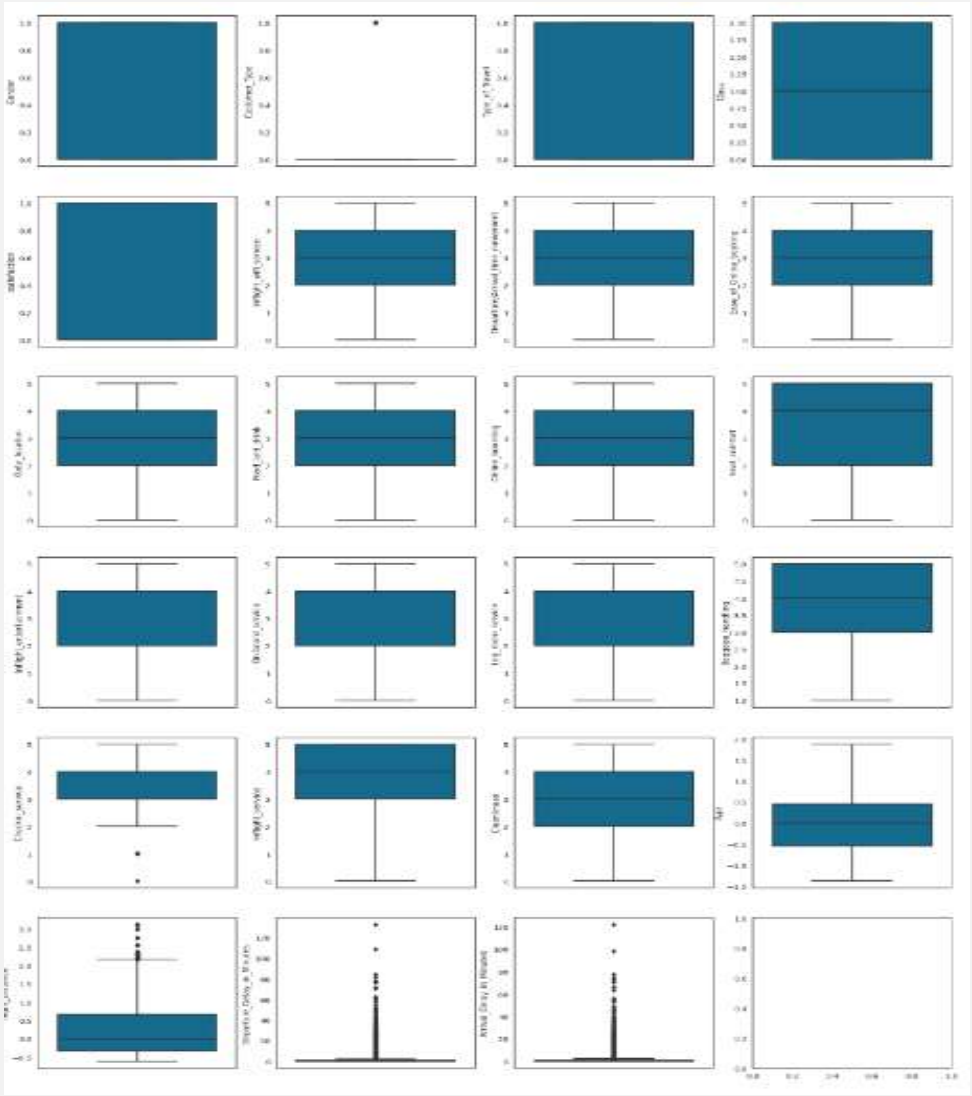

Flight Distance vs Satisfaction
Density
Flight Distance
0.0000
0.0002
0.0004
0.0006
0.0008
0.0010
0
1000
2000
3000


Distribution of Flight Distance by Class
Class
Eco Plus
Business
Eco
Count
Flight Distance
0
500
1000
1500
2000
2500
3000
0
1000
2000
3000
4000
5000


Satisfaction based on Delay
satisfaction
neutral or dissatisfied
satisfied
Departure Delay in Minutes
Arrival Delay in Minutes
0
200
400
600
800
1000
1200
1400
1600


Correlation Heatmap of Numerical Features
Unnamed: 0
id
Gender
Customer Type
Age
Type of Travel
Class
Flight Distance
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
Departure Delay in Minutes
Arrival Delay in Minutes
satisfaction

OUTLIER DETECTION



There are four variables that catch our eye in terms of

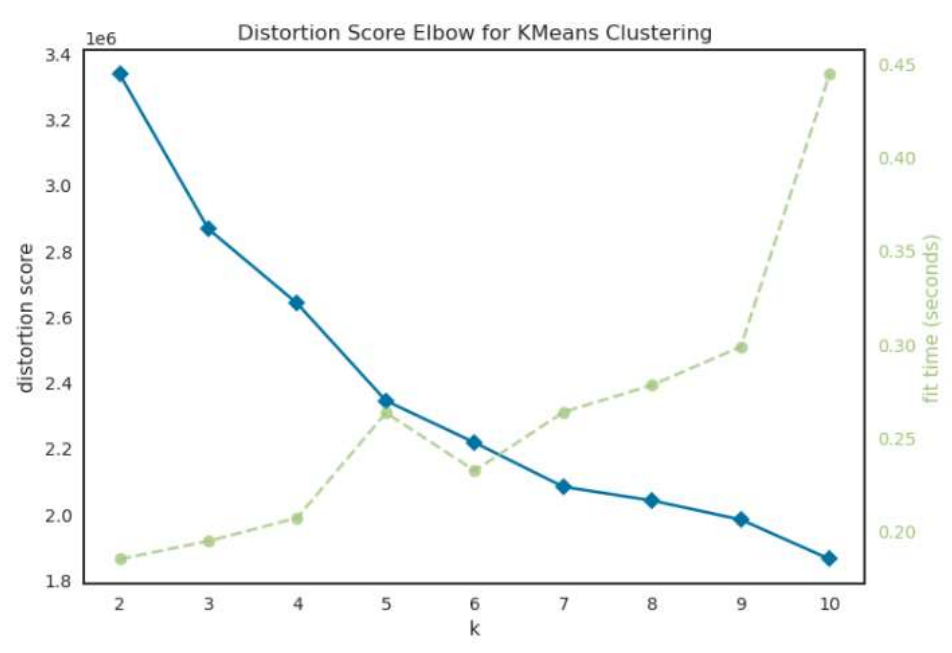
- ‘Flight Distance’>2.5: 58 observations (~0.05%)
- ‘Check-in service’ < 1: 1 observation (~0%)
- ‘Departure Delay in Minutes’ >100: 2 observations (~0
- ‘Arrival Delay in Minutes’ >100: 1 observation (~0%)

DIMENSIONALITY REDUCTION (PCA METHOD)

	0	1	2	3	4	5	6	7	8	9	10	11	12
0	0.643481	-2.838255	1.528182	0.416362	1.525427	0.455497	1.696304	-0.958217	-0.635927	-0.612674	0.803281	1.032057	0.0
1	-1.139223	4.151076	-1.740010	-1.240153	-2.275673	-2.218773	-1.359591	-0.771596	-0.129489	-1.628197	-0.369051	-0.299546	-0.
2	-1.823286	-2.830938	2.882431	1.003769	-0.619484	0.646969	0.459538	0.622320	-0.066356	-0.250733	0.190945	-0.480558	-0.
3	-0.429559	1.984216	-3.671168	-0.678198	0.751446	-1.945542	-2.566615	0.207154	-0.608445	-1.069061	0.279280	0.010339	-0.
4	-1.730046	-1.267699	-0.075760	0.944302	-1.408517	0.170834	-0.730823	0.082743	-0.701223	-0.326915	-0.593180	-0.996033	1.2

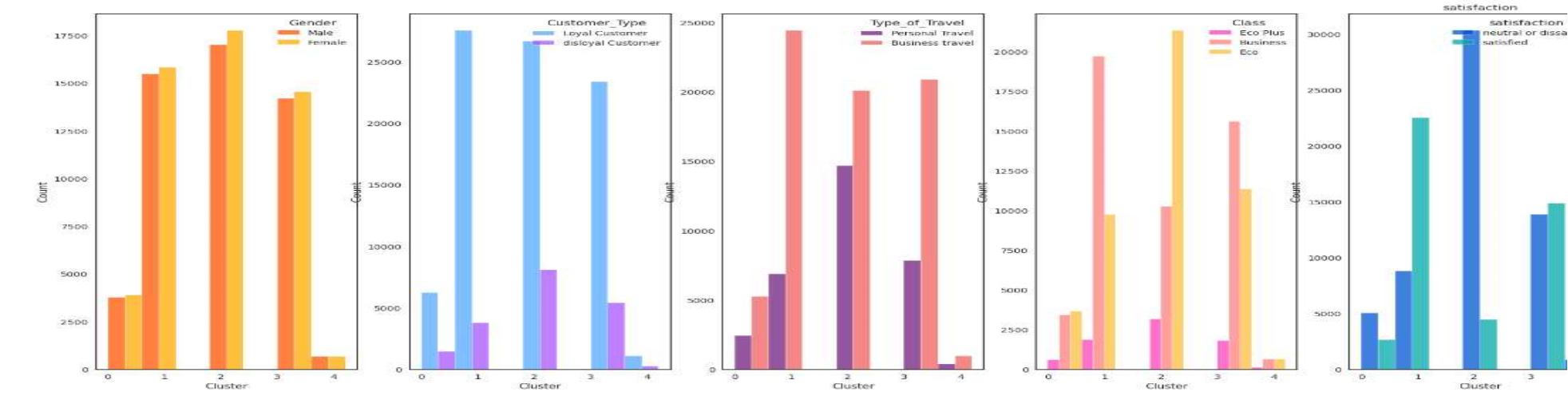
- **Conclusion**: The "dimensionally reduced" data frame above has 13 nameless variables (principle components) which would explain 95% of variance.

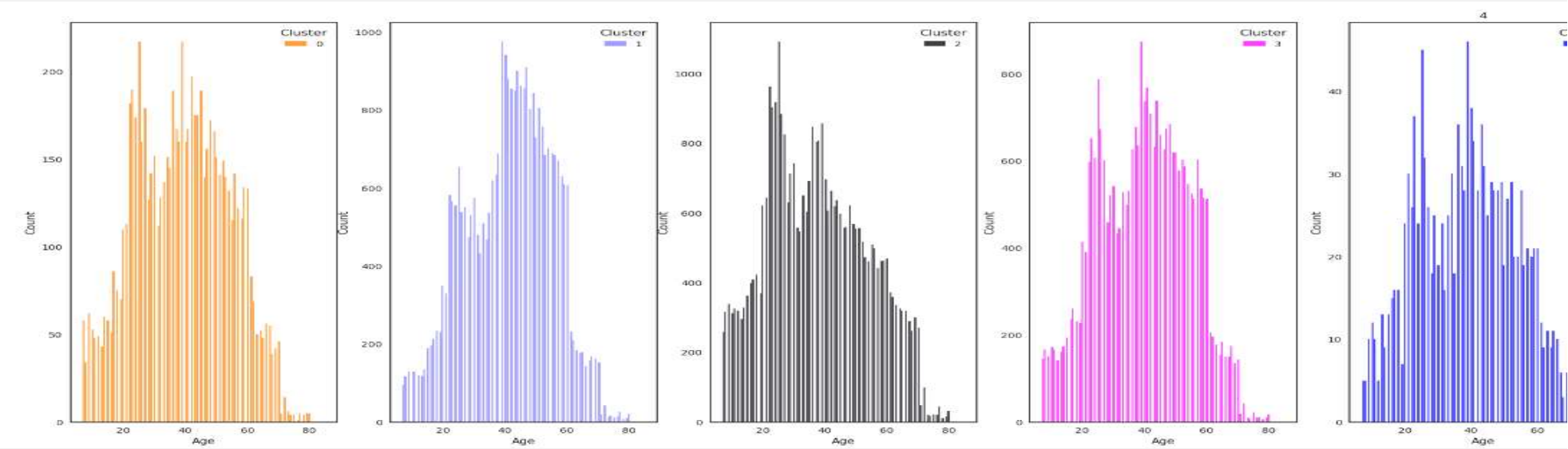
CLUSTER SEGMENTATION

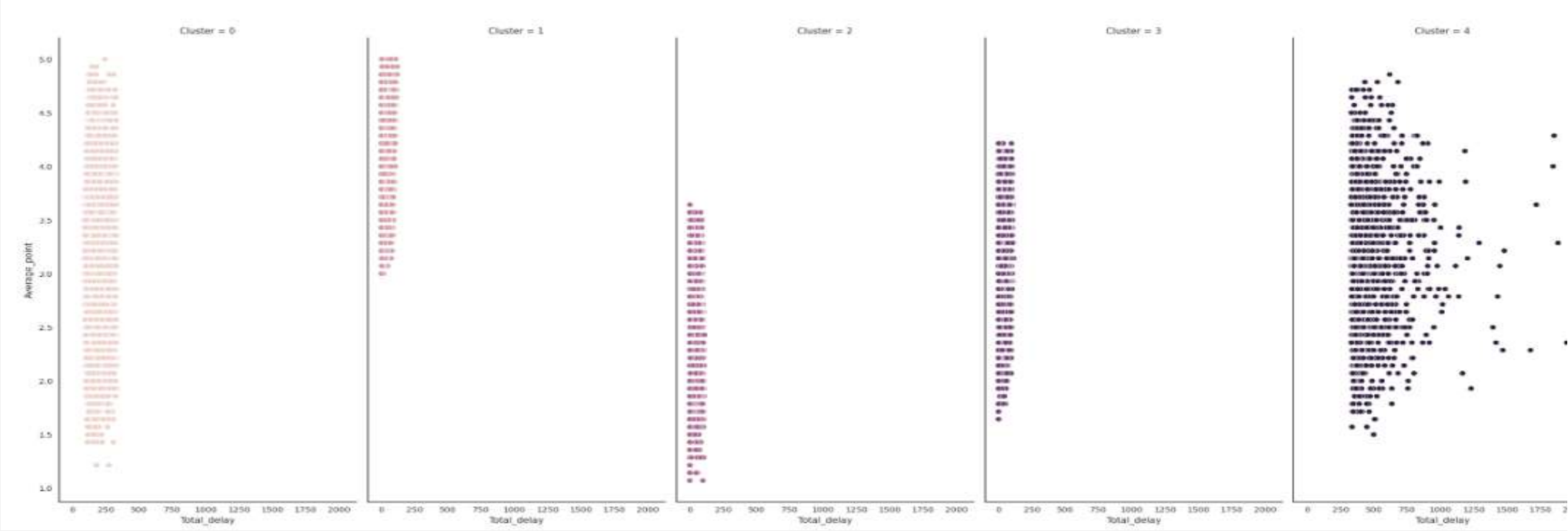


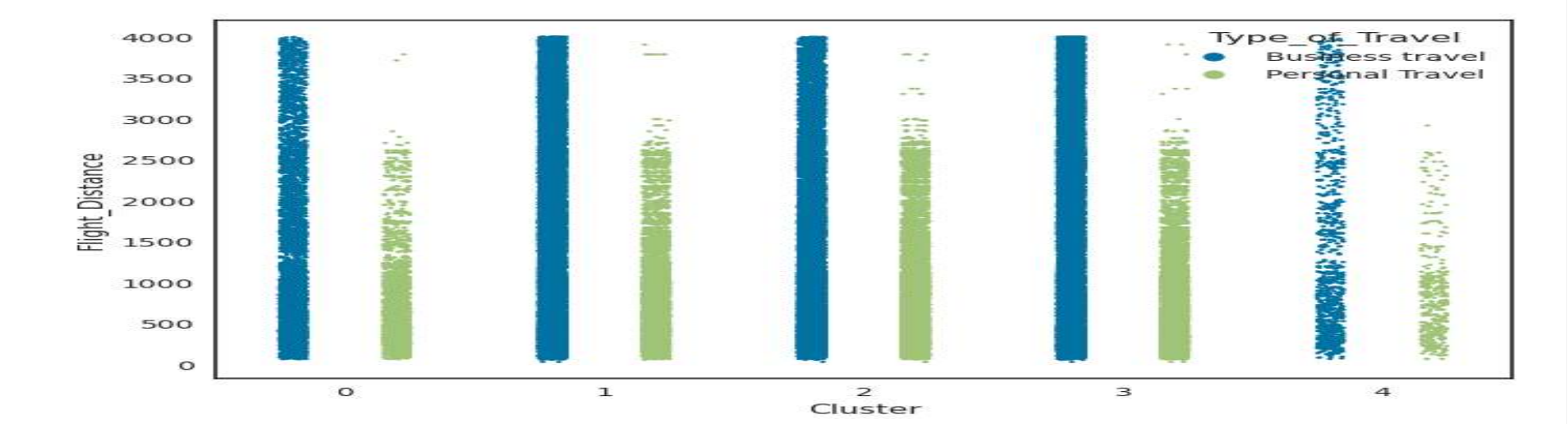
With every new iteration, the curves may be di
the result always boils down to 5 or 6 as the op
number of clusters. **Let's take 5 clusters**.

DIFFERENT CLUSTERS ANALYSIS






Cluster = 0
Cluster = 1
Cluster = 2
Cluster = 3
Cluster = 4
Average_point
Total_delay


Type_of_Travel
Business travel
Personal Travel
Flight_Distance
Cluster

CLUSTER ANALYSIS

Cluster 0

- not numerous cluster
- nearly equal proportion between women and men
- most customers are loyal
- 2 times more "business travels" than "personal travels"
- "Eco" and "Business" seats are booked equally; "Eco Plus" is booked the least
- 1/3 of passengers is "satisfied"
- absolute majority of passengers is of age between 20 and 60 year old
- air flights distances are usually not more than 1000 miles

Cluster 1

- a numerous cluster
- contains slightly more women then men
- most customers are loyal
- 3 times more "business travels" than "personal travels"
- "Business" seats represent 70% of bookings; "Eco Plus" is booked the least
- 72% of passengers are "satisfied" (best result)
- absolute majority of passengers is of age between 20 and 60 y.o.
- passengers tend to give the highest rating points: almost all of them lie between "3" and "5"

Cluster 2

- the most numerous cluster
- contains slightly more women then men
- most customers are loyal
- 30% more "business travels" than "personal travels"
- "Eco" seats are booked the most (65% of bookings); "Eco Plus" is booked the least
- 13% of passengers are "satisfied" (worst result)
- the biggest share of young (~20 y.o.) passengers; the share of older passengers constantly lowers until 65+ age

Cluster 3

- a numerus cluster
- contains slightly more women then men
- most customers are loyal
- 3 times more "business travels" than "personal
- "Business" seats represent 55% of bookings; "E booked the least
- 50% of passengers are "satisfied"
- absolute majority of passengers is of age betwe 60 y.o.
- passengers are conservative in giving ranking p usually, figures lie within "2", "3" and "4"

Cluster 4

- the least numerous cluster
- nearly equal proportion between women and men
- most customers are loyal
- 2 times more "business travels" than "personal travels"
- "Eco" and "Business" seats are booked equally; "Eco Plus" is booked the least
- 1/3 of passengers is "satisfied"
- majority of passengers is of age between 20 and 60 y.o , but children and older passengers (65+ y.o) are also present to some extent
- most travels are up to 1000 miles long, but 2000 miles-long flights occurs as well

MODEL SELECTION

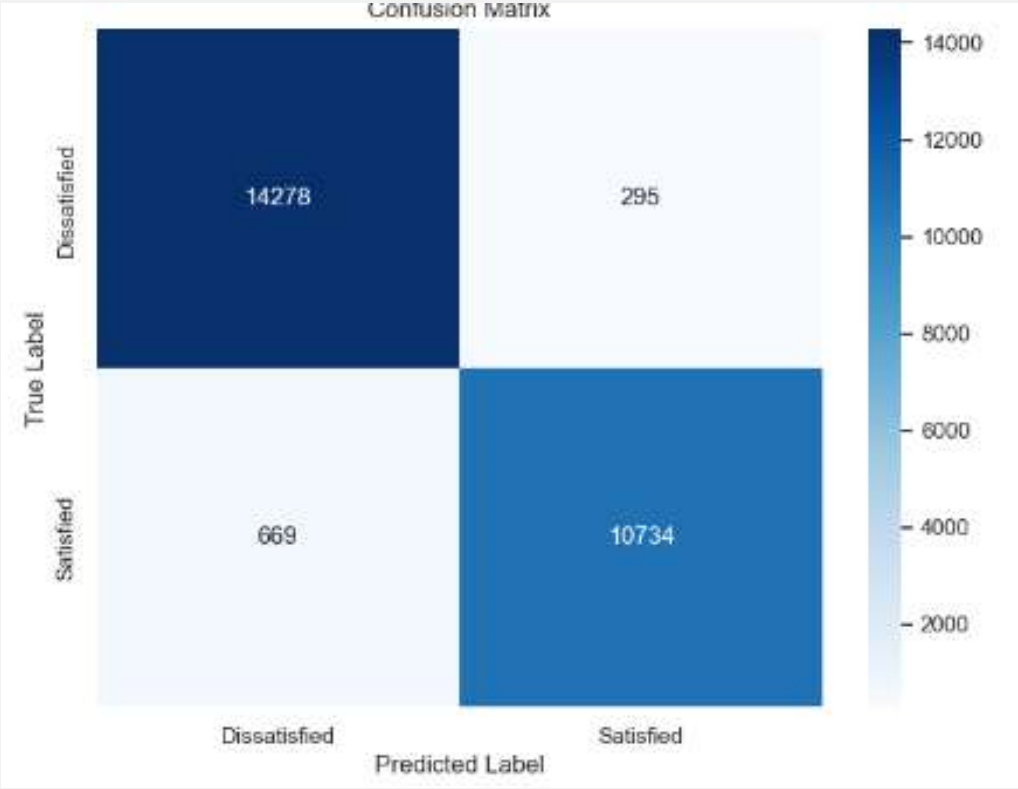
- It is an classification problem, so our basic instinct was to implement decision tree but decision tree is not much robust than input variable so in the end we decided to choose Random forest

- Model Building Random Forest Classifier because it handles both numerical and categorical data well and is robust against overfitting.

- Accuracy: 96.35% Precision, Recall, and F1-Score: Class 0 (Neutral or Dissatisfied): Precision = 95%, Recall = 98%, F1-score = 97% Class 1 (Satisfied): Precision = 98%, Recall = 94%, F1-score96% =

- The metrics indicate that the model performs excellently in predicting passenger satisfaction, with high scores in both precision and recall across the satisfaction categori96%

ACCURACY



The matrix shows a strong predictive performance
true positive and true negative rates, indicating tha
is effective at classifying both satisfied and dissatisf
passengers correctly. True Positives (Satisfied corre
identified): The model has a high number of correc
predictions for satisfied passengers. True Negatives
(Dissatisfied correctly identified): Similarly, it accura
identifies a large number of dissatisfied passengers

RECOMMENDATION

	Feature	Importance
11	Online boarding	0.171534
6	Inflight wifi service	0.151017
4	Class	0.099753
3	Type of Travel	0.097657
13	Inflight entertainment	0.057292
12	Seat comfort	0.045061
5	Flight Distance	0.043389
15	Leg room service	0.039845
8	Ease of Online booking	0.037885
1	Customer Type	0.035961
2	Age	0.035749
14	On-board service	0.030285
18	Inflight service	0.026105
19	Cleanliness	0.025393
17	Checkin service	0.024873
16	Baggage handling	0.024517
9	Gate location	0.018942
7	Departure/Arrival time convenient	0.017735
10	Food and drink	0.012204
0	Gender	0.004804

Feature Importance Analysis Online Boarding (17.15%): The most influential feature, indicating that the online boarding experience plays a crucial role in determining passenger satisfaction.

Inflight Wifi Service (15.10%): Another significant factor, emphasizing the value of connectivity during flights. Class (9.98%): Reflects the impact of travel class on passenger satisfaction, with business class likely offering a more satisfying experience.

Type of Travel (9.77%): Differentiates between personal and business travel, affecting satisfaction levels.

Inflight Entertainment (5.73%): Entertainment options available inflight significantly influence satisfaction, especially on longer Recommendations Based on Analysis Enhance Online Services: Improving the online boarding process and inflight wifi service could lead to significant gains in passenger satisfaction.

Focus on Inflight Experience: Entertainment and comfort in different classes should be areas of focus for service improvements.