

# SC1015 Mini Project – Airline Passenger Satisfaction

Yaxin (U2220733G)

Nivedha (U2220745J)

Kay (U2220423G)



# Table of contents

01

Introduction

02

Problem Formulation

03

Data Preparation

04

Exploratory Data Analysis

05

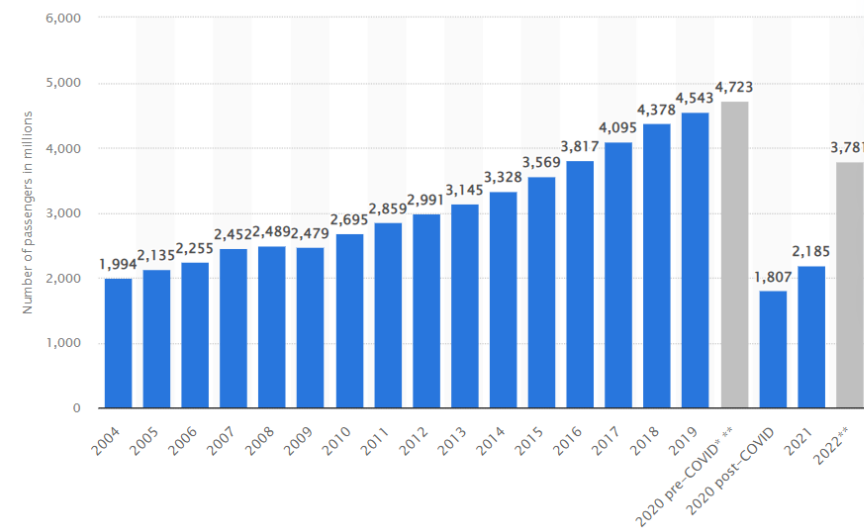
Model Building &  
Machine Learning

06

Conclusion

# Introduction

## How to deal with High Demand in Aviation Industry: Post Pandemic Travel Trends



[Additional Information](#)

© Statista 2023

[Show source](#)





## Motivation - Problem Formulation

**How can airline improve services to increase passenger satisfaction?**

**How different variables affect the overall *satisfaction* in different generation groups?**

# What are the variables included in data set?

Data columns (total 24 columns):

#	Column	Non-Null Count	Dtype
0	ID	129880 non-null	int64
1	Gender	129880 non-null	object
2	Age	129880 non-null	int64
3	Customer Type	129880 non-null	object
4	Type of Travel	129880 non-null	object
5	Class	129880 non-null	object
6	Flight Distance	129880 non-null	int64
7	Departure Delay	129880 non-null	int64
8	Arrival Delay	129487 non-null	float64
9	Departure and Arrival Time Convenience	129880 non-null	int64
10	Ease of Online Booking	129880 non-null	int64
11	Check-in Service	129880 non-null	int64
12	Online Boarding	129880 non-null	int64
13	Gate Location	129880 non-null	int64
14	On-board Service	129880 non-null	int64
15	Seat Comfort	129880 non-null	int64
16	Leg Room Service	129880 non-null	int64
17	Cleanliness	129880 non-null	int64
18	Food and Drink	129880 non-null	int64
19	In-flight Service	129880 non-null	int64
20	In-flight Wifi Service	129880 non-null	int64
21	In-flight Entertainment	129880 non-null	int64
22	Baggage Handling	129880 non-null	int64
23	Satisfaction	129880 non-null	object



# Data Preparation for EDA



01

Removing unnecessary column ("ID")

02

Dealing with Missing value (Arrival Delay)

03

Handling outliers ("Departure Delay" & "Arrival Delay" & "Flight Distance")

04

Checking balance

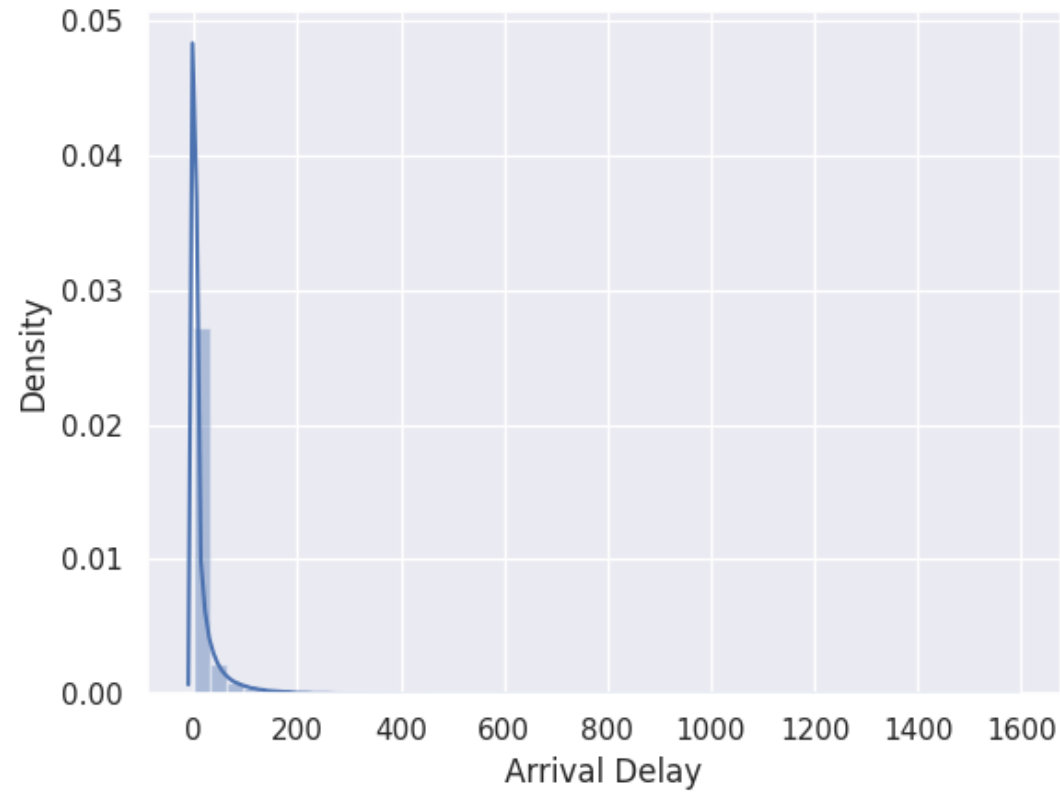
05

Replacing rating '0' with MEAN ('0' means not applicable)

06

Split the dataset according to generations

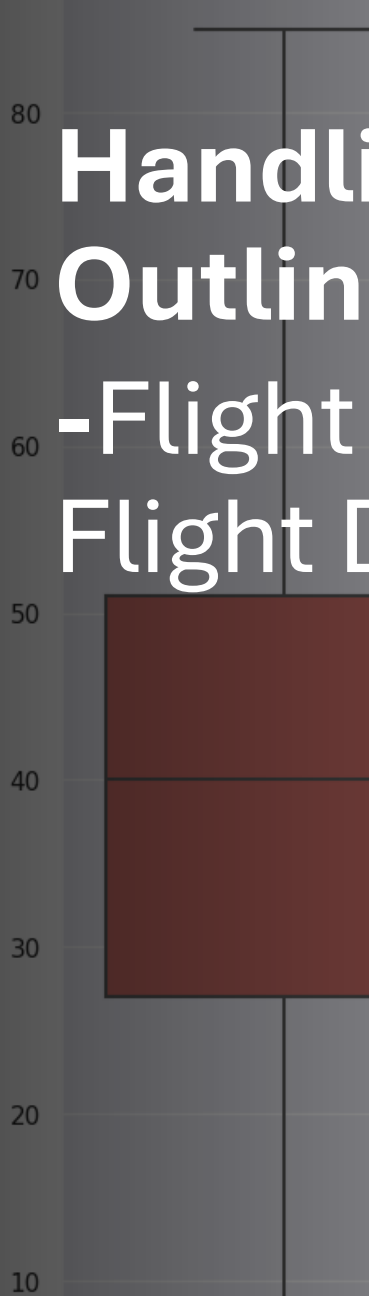
### Missing value for "Arrival Delay" - Median



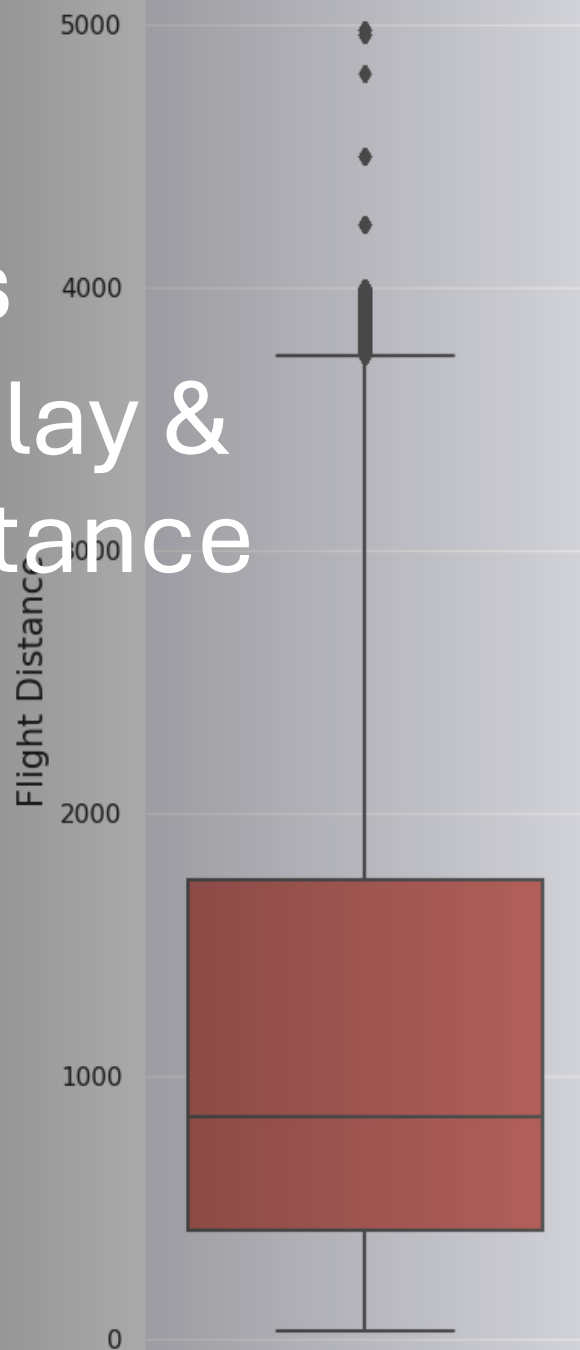
- Skewed arrival delay graph requires median instead of mean for analysis
- Missing value for arrival delay is filled with median value.

# Handling Outliers -Flight Delay & Flight Distance

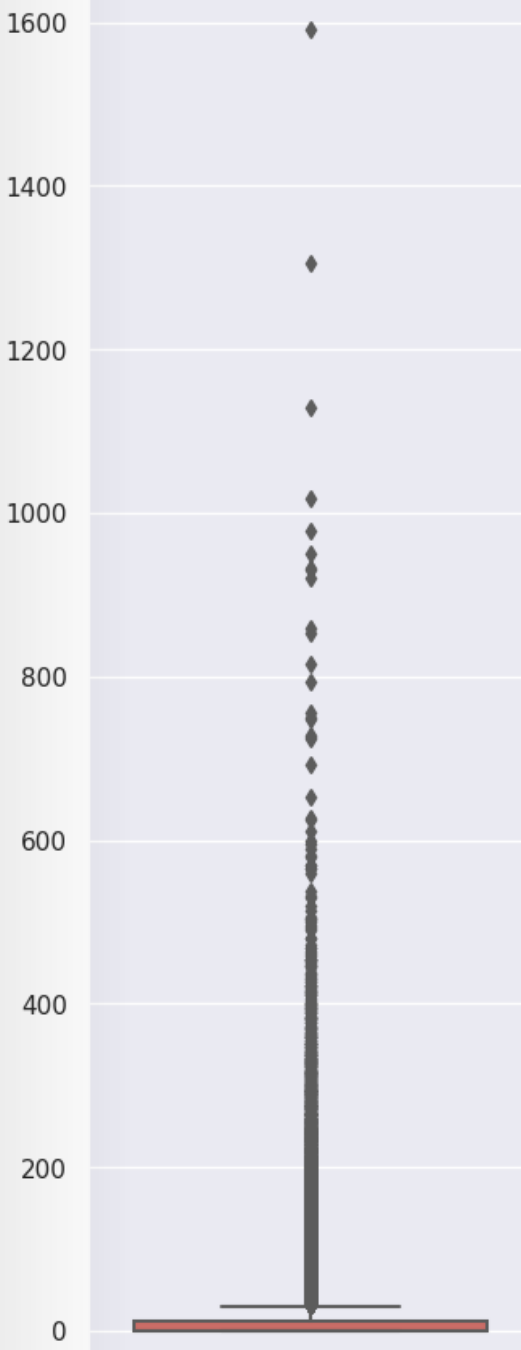
Age



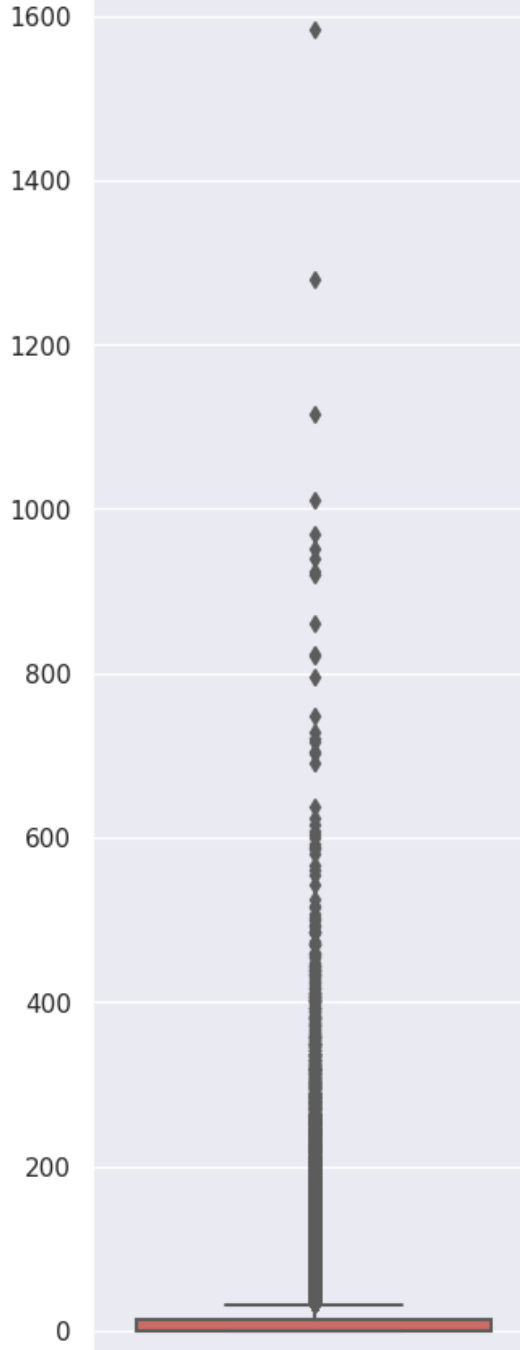
Flight Distance



Departure Delay



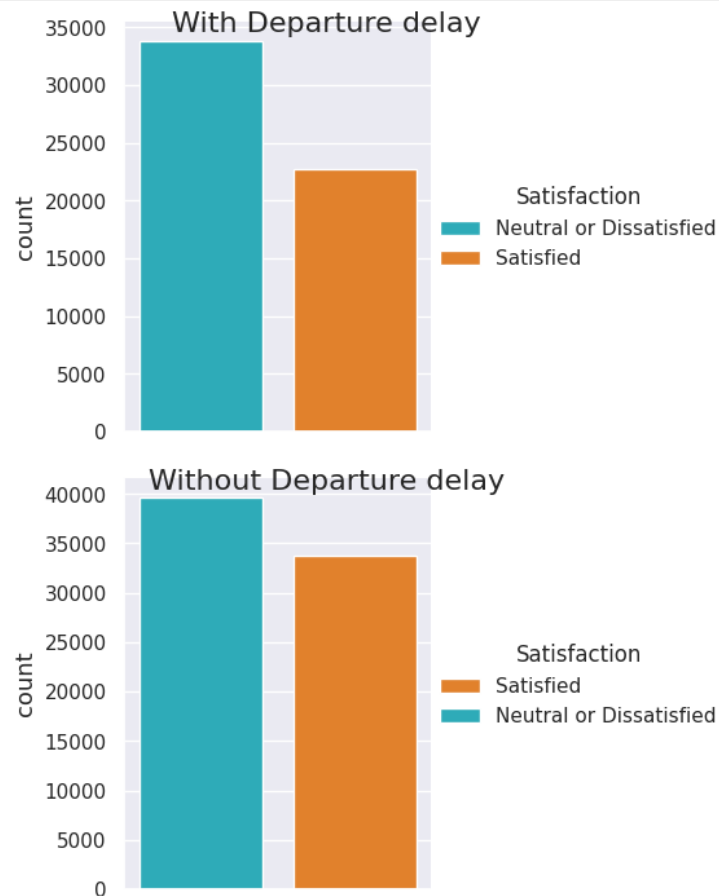
Arrival Delay



1



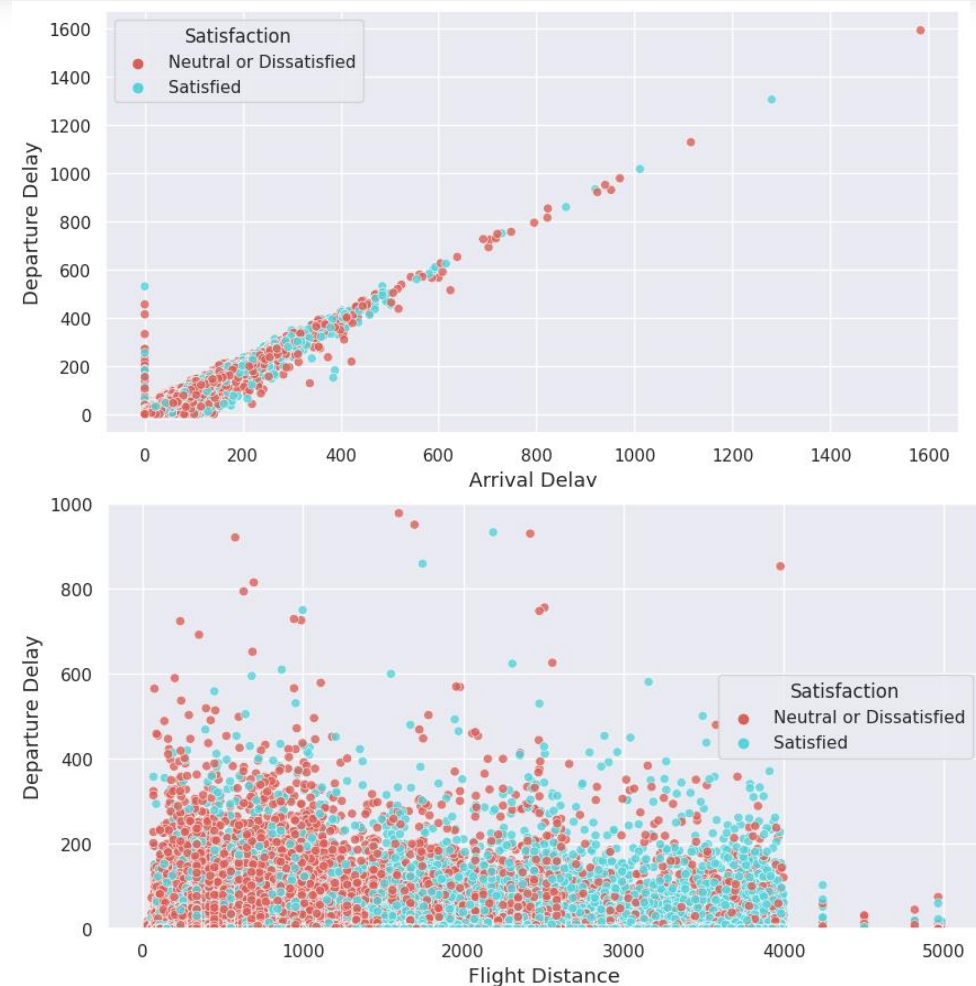
# Satisfaction VS Flight delays ?



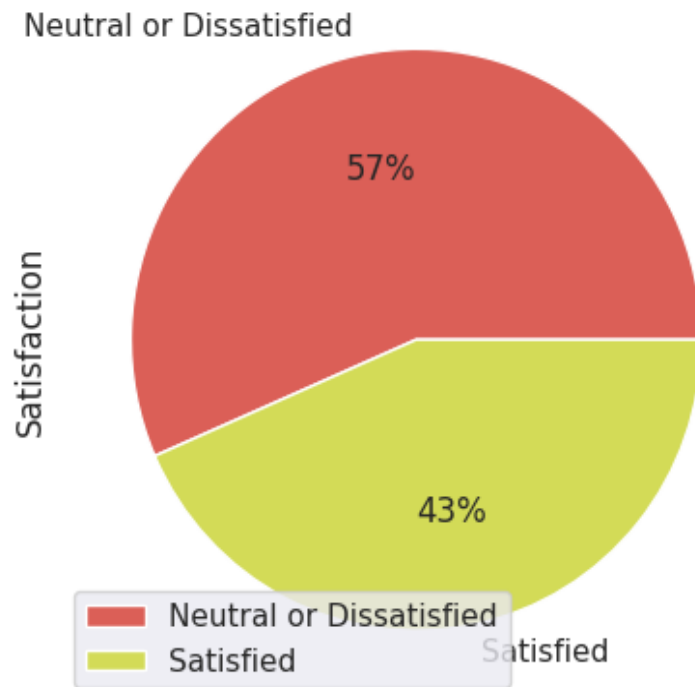
Analysis before handling outliers

# Relationship between Arrival Delay and Departure Delay

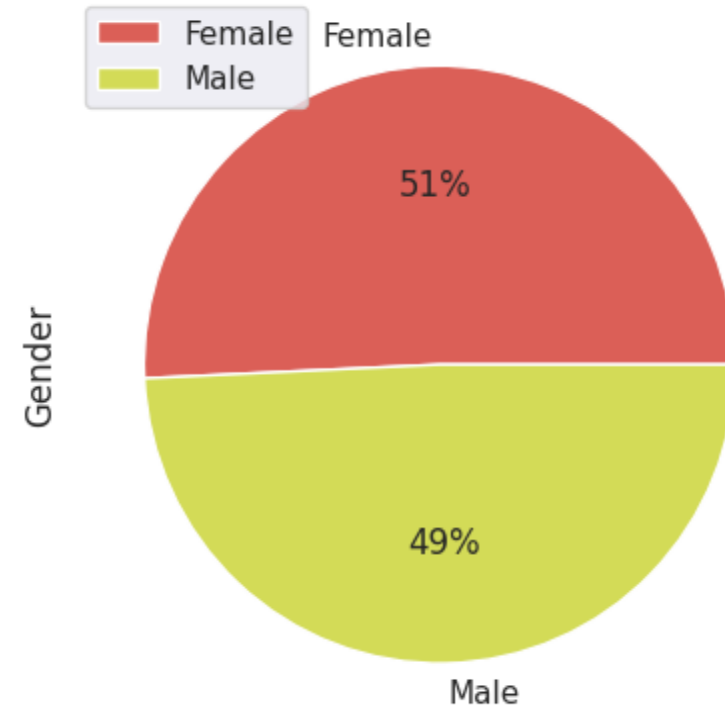
- Departure and arrival delays have a linear relationship.
- Majority of passengers are fine with slight departure delay for longer flights.
- Travelers with short flight distances are unhappy with delayed departure.



## Overall Satisfaction Level :



## By Gender:



- Survey is balanced and unbiased with 43% and 57% numbers being tolerable.
- No modification required for the data.
- Gender has negligible impact on satisfaction.





# Exploratory Data Analysis





# **Let's Explore Satisfaction Ratings Across Different Categories**

Gender

Customer Type

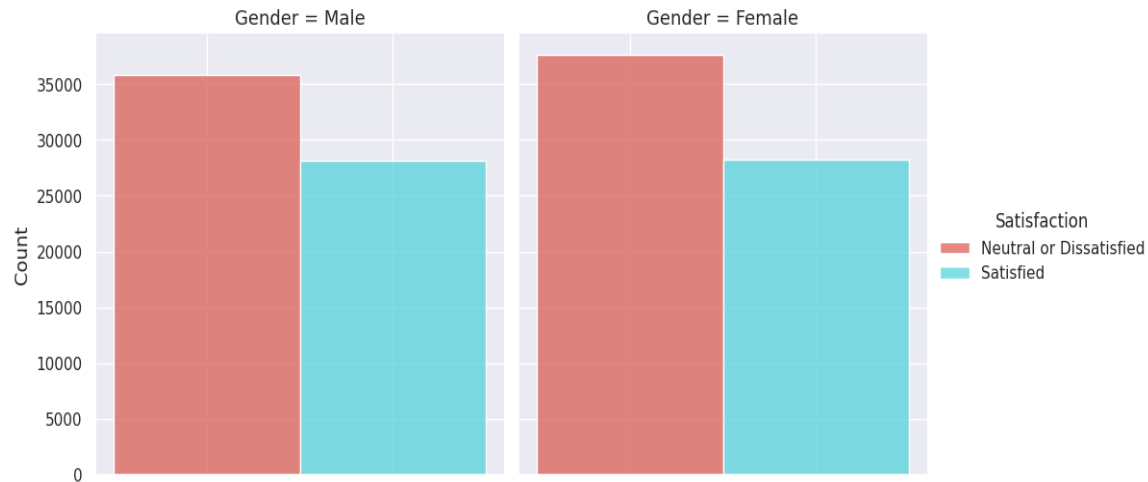
Pre-boarding and onboarding

Age

Class Type



## Gender – Male & Female

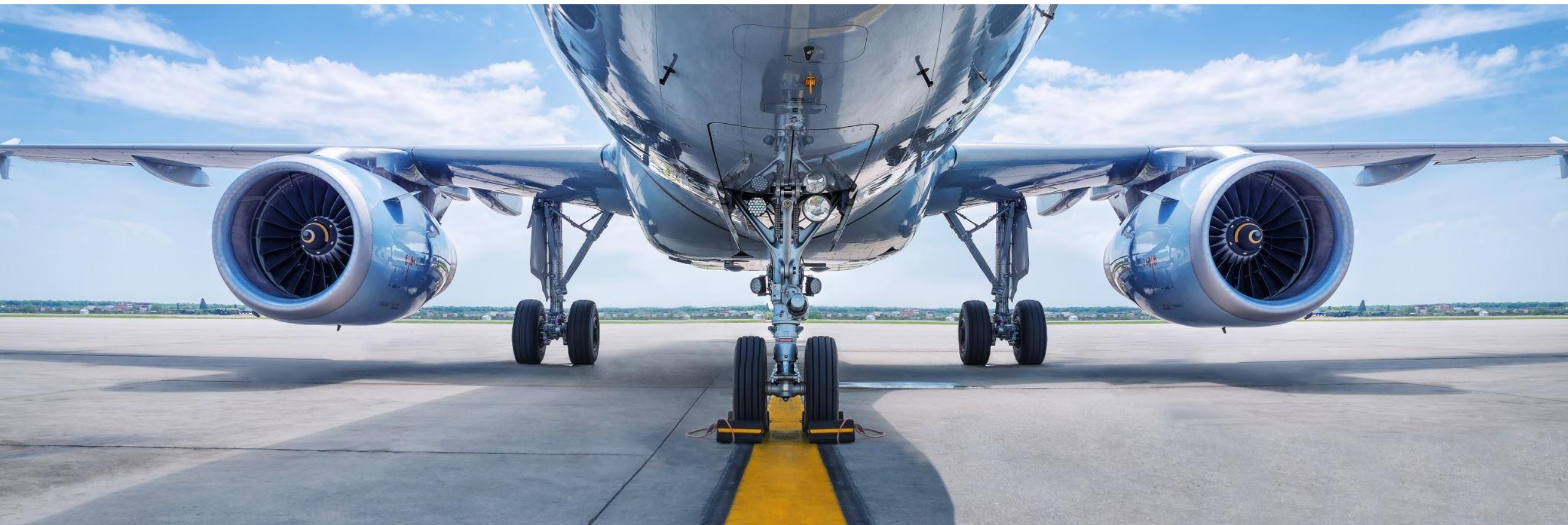


## Customer Type – First Timer & Existing customer



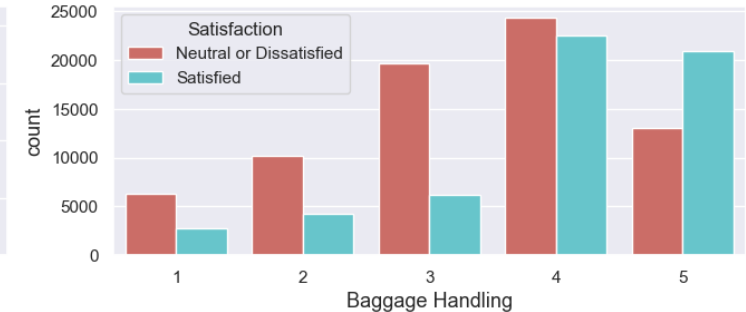
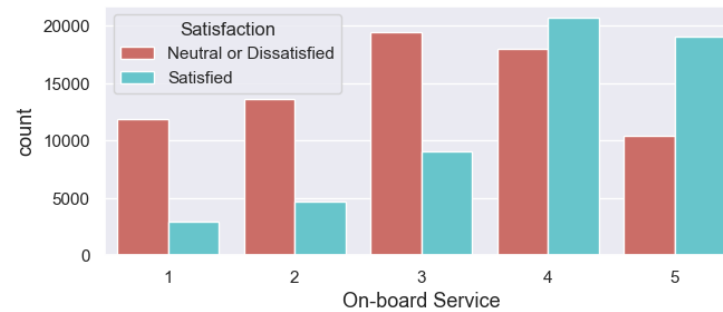
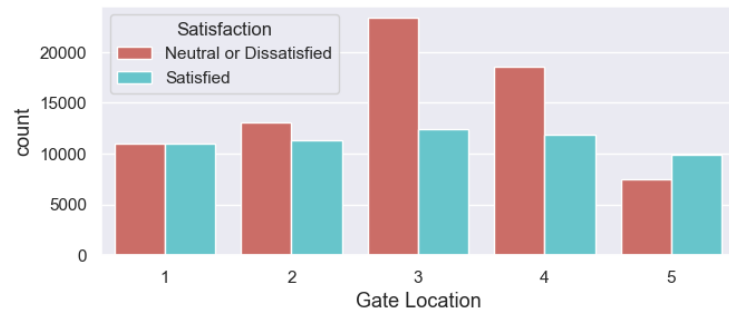
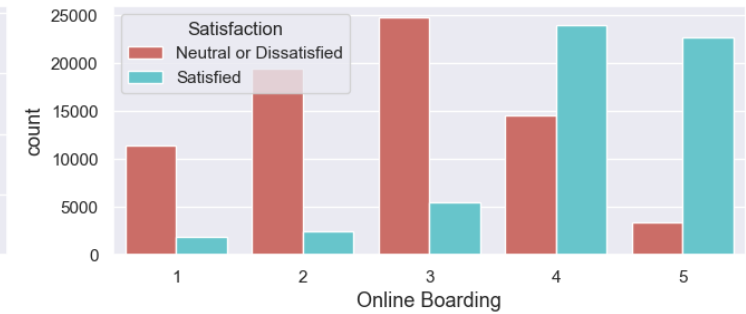
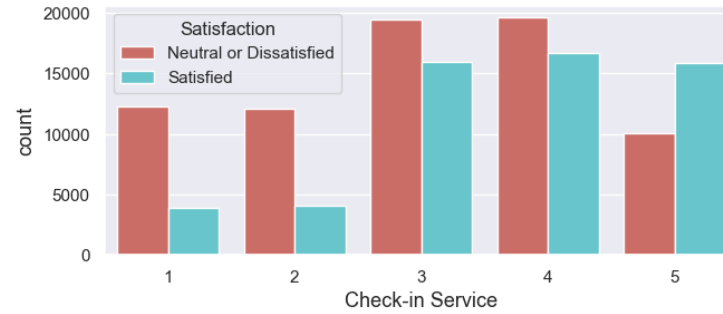
- Gender has negligible impact on satisfaction
- First-time passengers are generally dissatisfied with the flight
- 82% of passengers are returning customers
- High level of dissatisfaction among both first-time and loyal customers

# Exploring How Service Ratings for Pre-boarding and Onboard Services Affect Passengers' Flight Experience

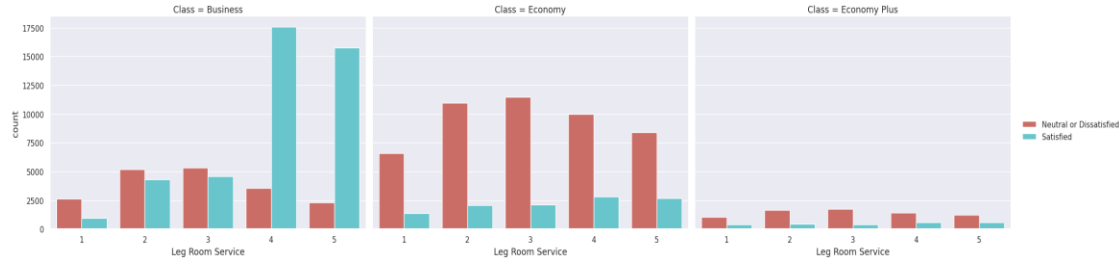


# Analyzing on Pre-Boarding services

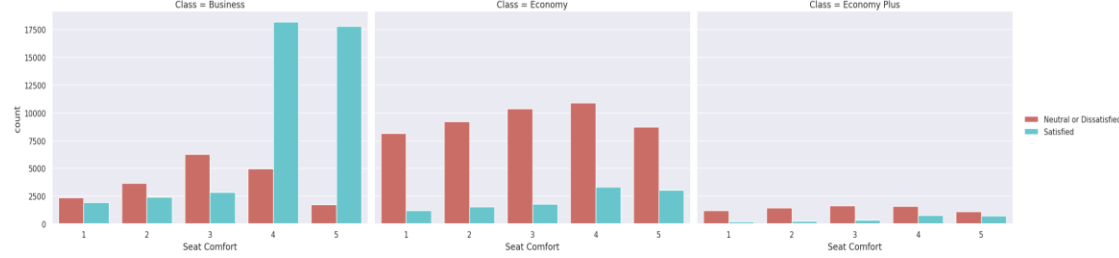
- Baggage handling, check-in, and gate location received low satisfaction ratings of 3-4.
- Online booking and on-board services received higher satisfaction ratings.
- Suggestions to improve the pre-boarding experience by addressing baggage handling, check-in, and gate location services.



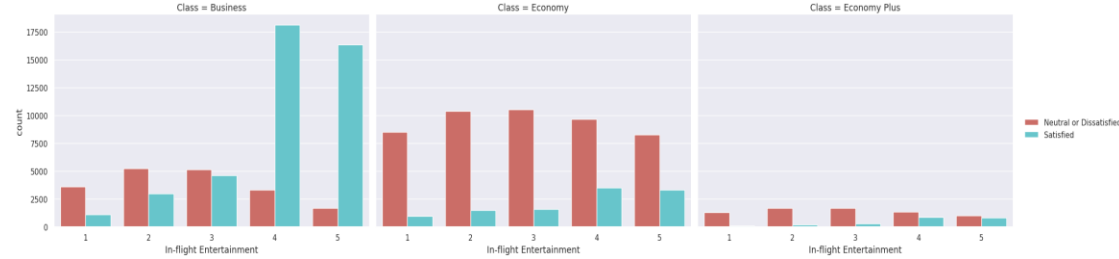
## Economy Plus




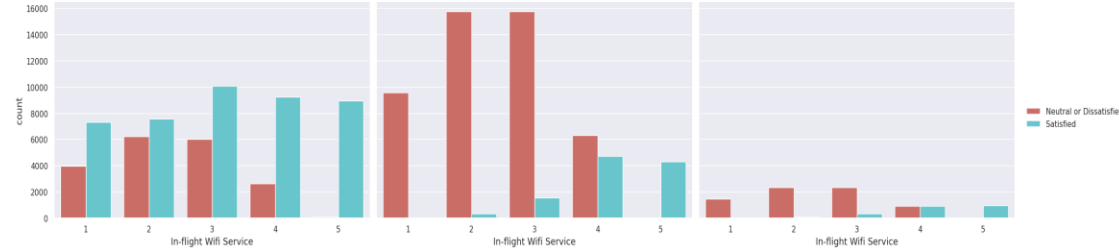
## Economy



Class = Economy

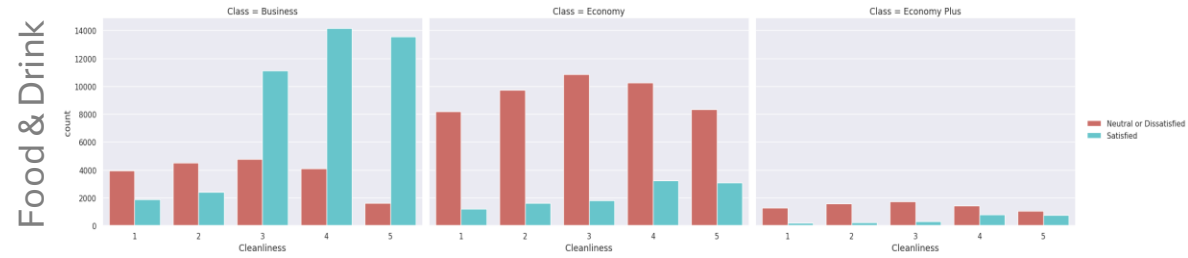


Class = Economy

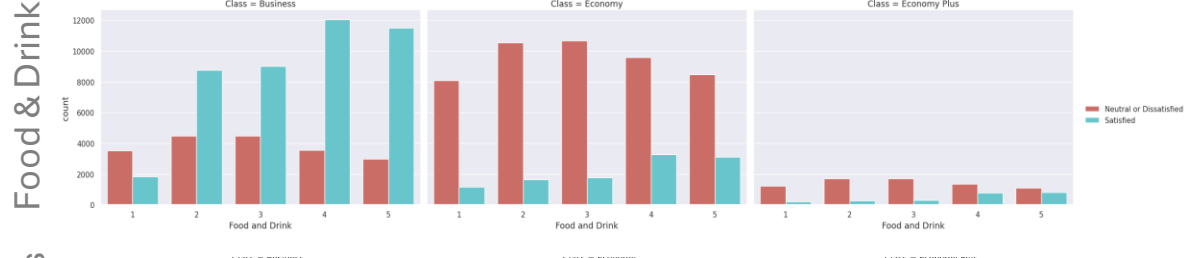


## Economy Plus

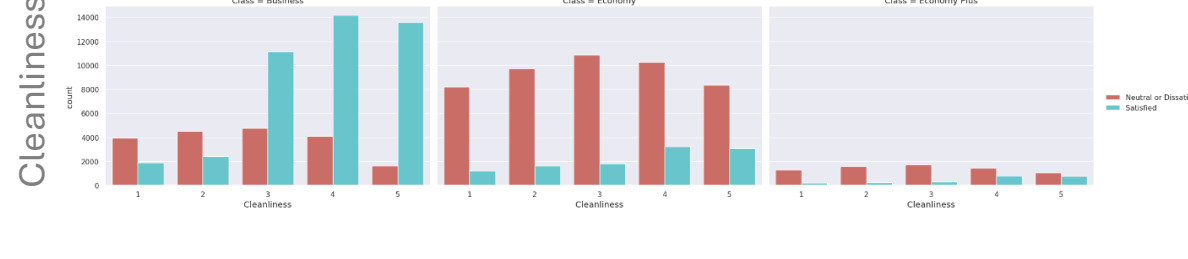
## Business



## Economy

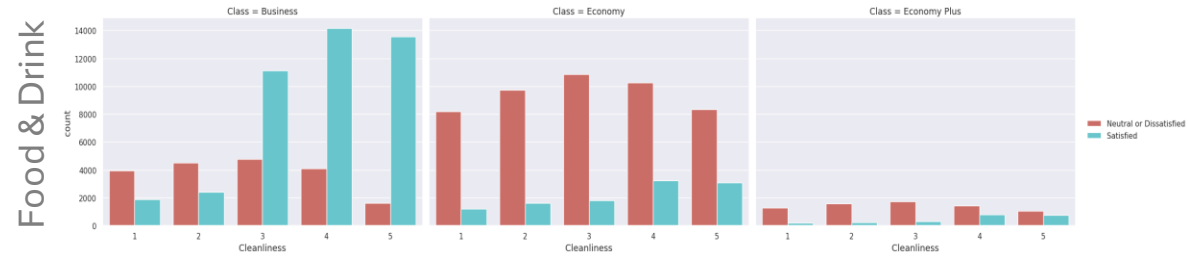


Class = Economy

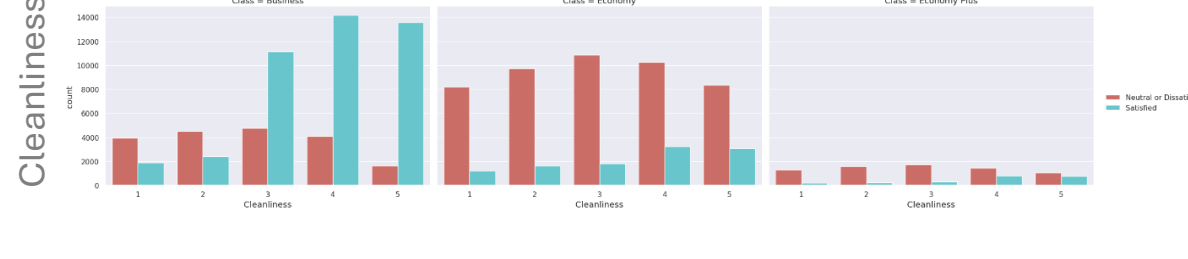


Class = Economy

## Economy Plus



Class = Economy Plus



Class = Economy Plus

Food &amp; Drink

## Food & Drink

## Cleanliness

Leg room

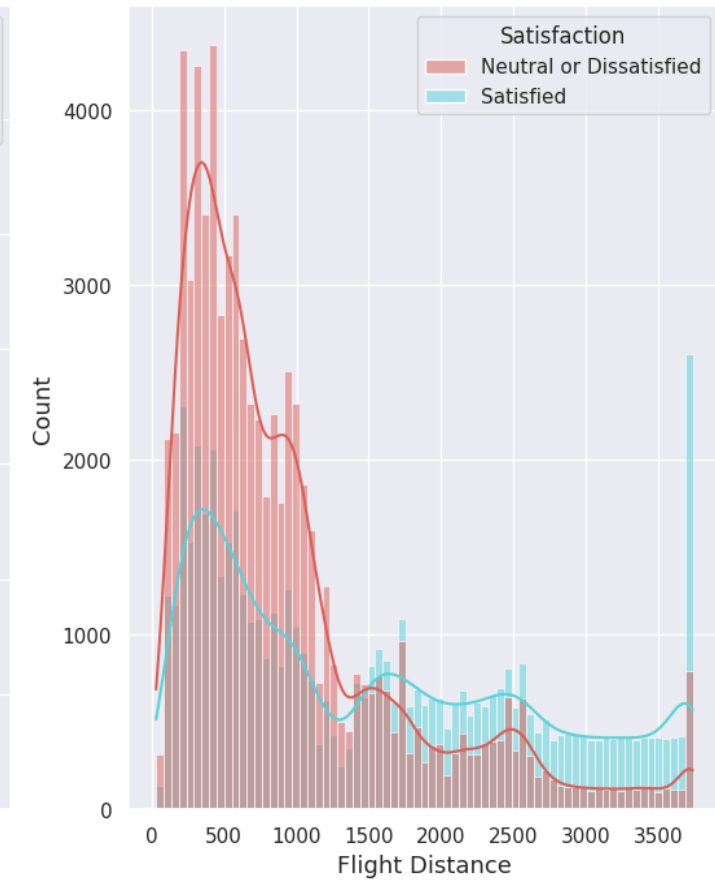
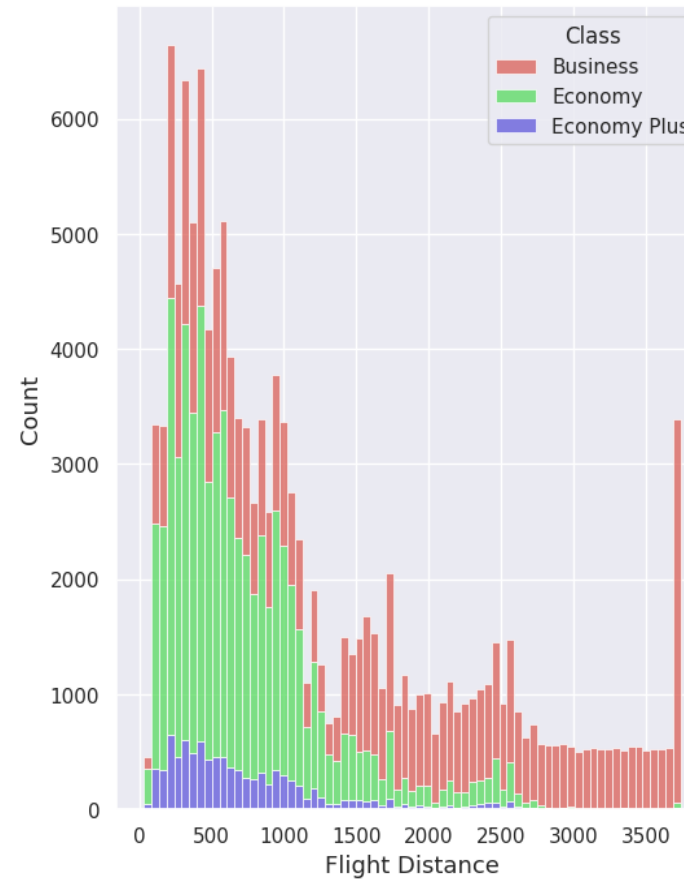
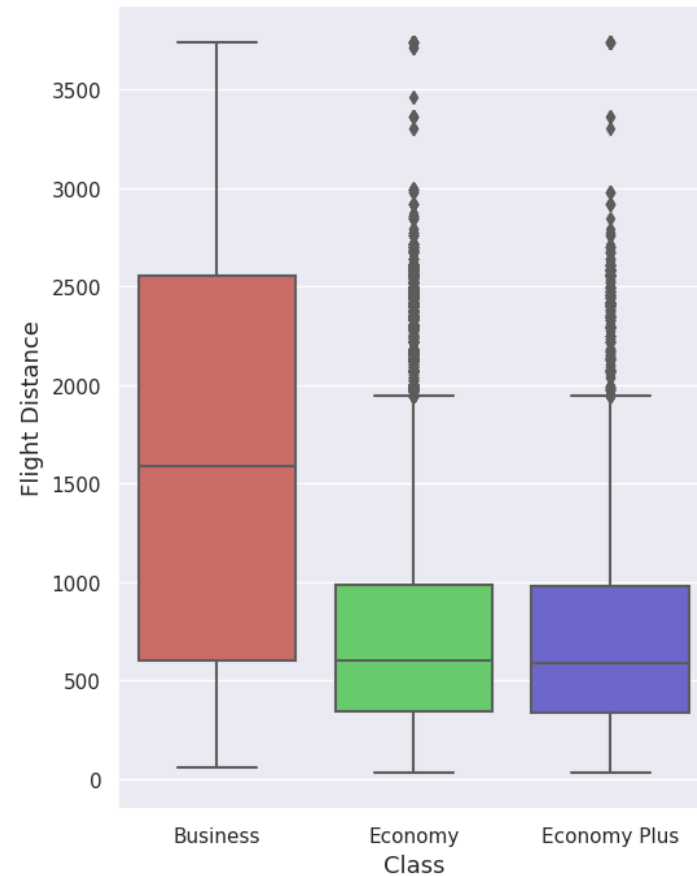
## Seat Comfort

Entertainment

Wifi



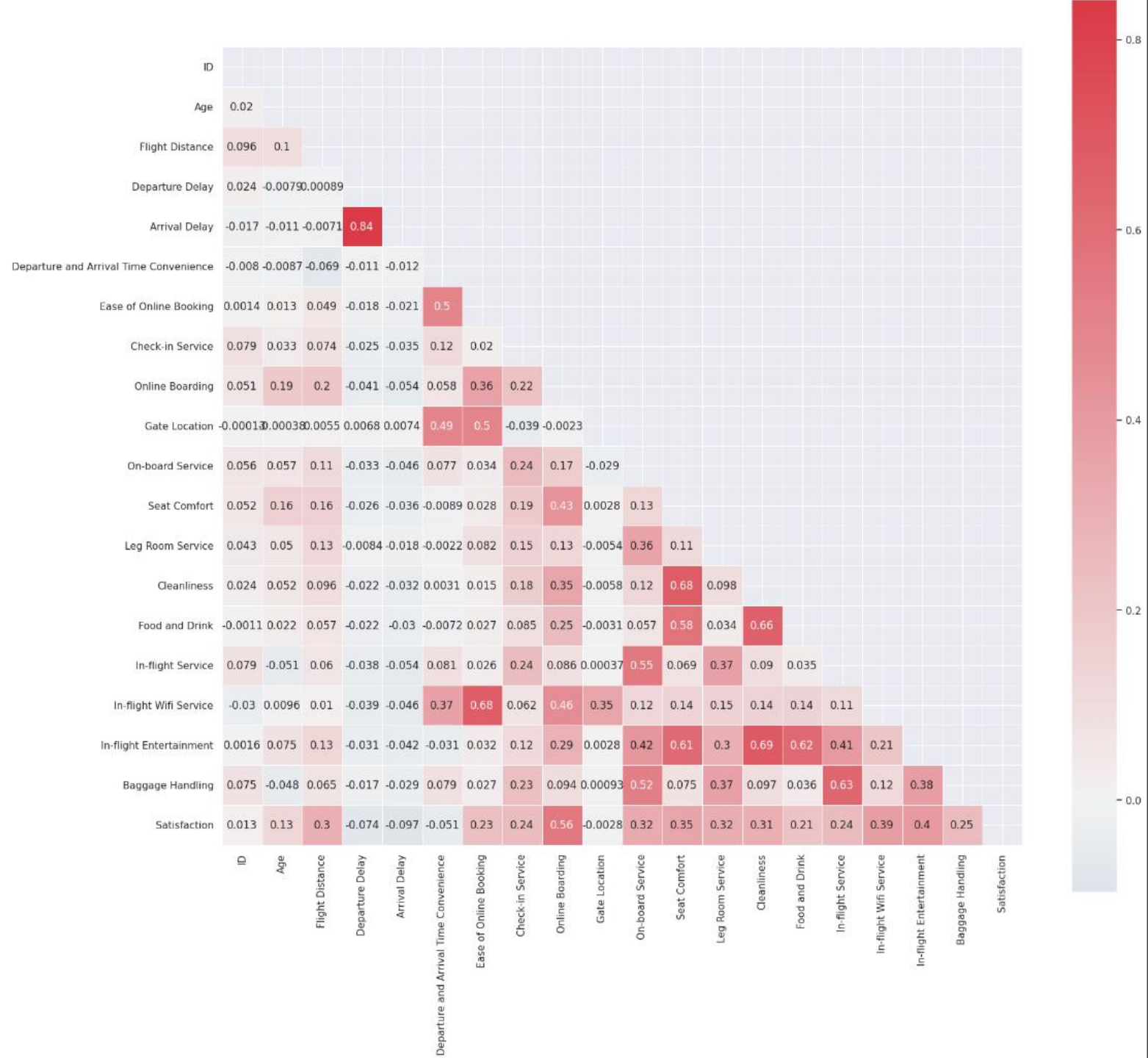
# Examining Passenger Travel Class Preferences and Satisfaction Levels Across Varying Flight Distances



- Business Class preferred for long-distance and business travel
- Longer flight distances associated with higher satisfaction levels
- Business Class amenities may contribute to increased satisfaction on longer flights

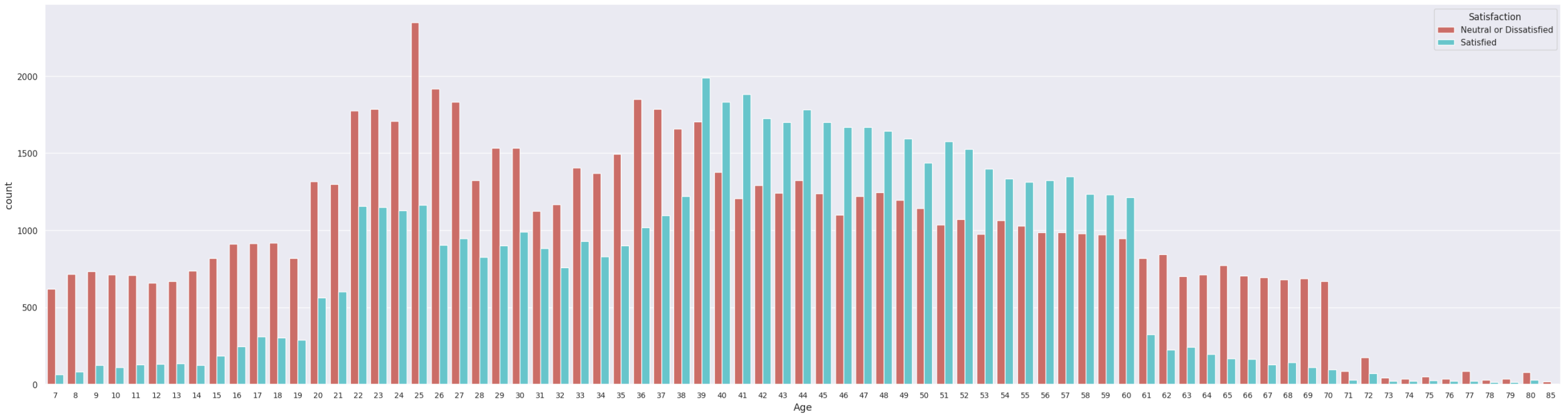
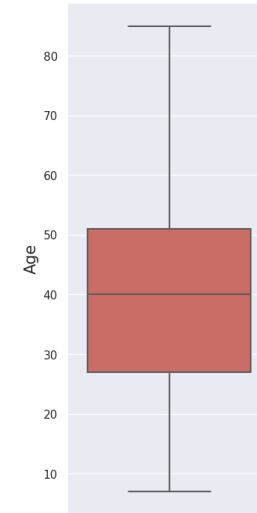
# Variable Correlation Analysis using Heatmap

It is estimated that the length of delay in arrival depends on the length of delay in departure which is linear. If the airline passenger's flight time is delayed for a certain time at departure, then the flight time will also be delayed by the same amount of time at landing.



## Comparing the satisfaction level by Age

- Age group 39-60: Higher satisfaction levels, indicating higher expectations.
- Age group 20-39: Mostly unsatisfied with the services.
- Children and elderly: More likely to feel not satisfied or neutral.



# Split the dataset according to generations

Generations defined by name, birth year, and ages in 2023

Generations	Born	Current Ages
Gen Z	1997 – 2012	11 – 26
Millennials	1981 – 1996	27 – 42
Gen X	1965 – 1980	43 – 58
Boomers II (a/k/a Generation Jones)*	1955 – 1964	59 – 68
Boomers I*	1946 – 1954	69 – 77
Post War	1928 – 1945	78 – 95
WWII	1922 – 1927	96 – 101

Age 7-85

group\_0\_9

group\_10\_25

group\_26\_41

group\_42\_57

group\_58\_67

group\_68\_76

group\_77\_94



# Comparing features that may vary across different age groups



**Food and Drink**

**Leg Room Service**

**Seat Comfort**

**In-flight Wifi Service**

**In-flight Entertainment**

Gen Z

Millennials

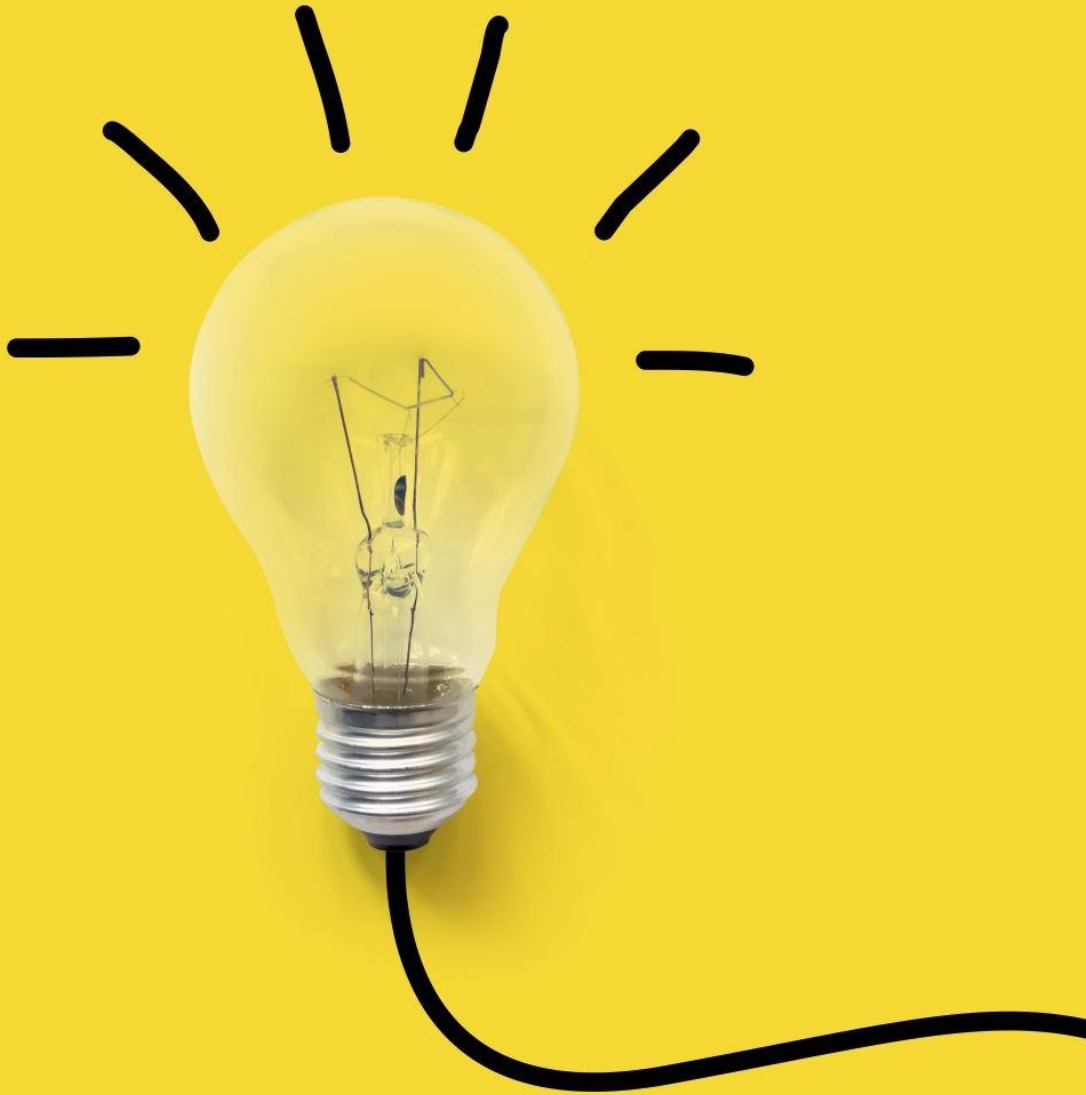
Gen X

Boomer I

Boomer II

Post War

WWII



## Predicting Key Factors for Airline Customer Satisfaction through Machine Learning

- Due to the scattered nature of customer satisfaction data, it can be difficult to identify which areas require improvement or attention.
- To address this challenge, we will use machine learning to predict the important features that contribute to customer satisfaction.
- By providing airlines with these predictive insights, they can focus on improving specific areas and catering to different generations of customers to enhance overall satisfaction and loyalty

# Feature Selection: Top 10 features for each age group



- better visualization
- Reduces training time
- Avoids over-fitting
- Improves accuracy of the model



Chi-Square – filter out top 15 features



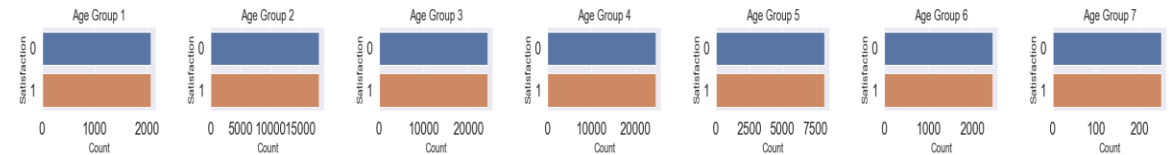
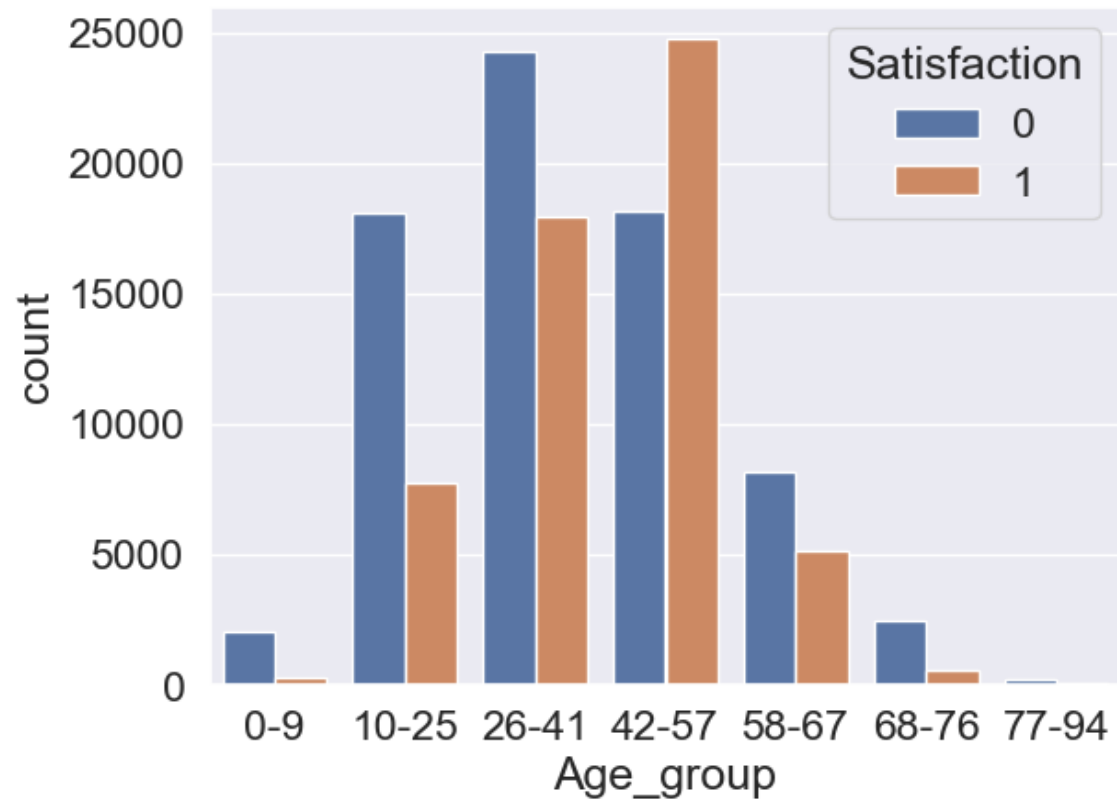
Wrapper Method – features are inherently important in contributing towards the passenger satisfaction

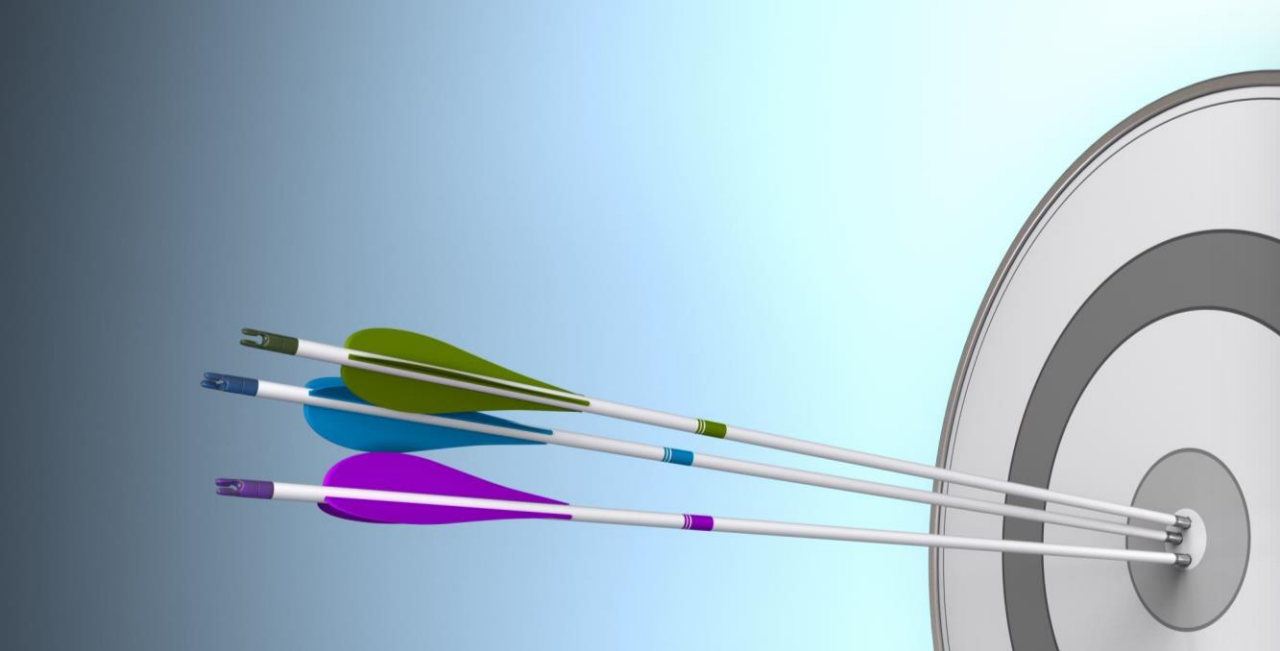
# Top 10 important features + target: ['Satisfaction'] -->Classification

Datasets name	
group_0_9	'In-flight Wifi Service', 'Online Boarding', 'Ease of Online Booking', 'Arrival Delay', 'Type of Travel', 'Class', 'Food and Drink', 'In-flight Entertainment', 'Cleanliness', 'Flight Distance'
group_10_25	'Online Boarding', 'Type of Travel', 'In-flight Wifi Service', 'Ease of Online Booking', 'Class', 'Food and Drink', 'Arrival Delay', 'In-flight Entertainment', 'Seat Comfort', 'Flight Distance'
group_26_41	'Type of Travel', 'Online Boarding', 'Class', 'In-flight Entertainment', 'In-flight Wifi Service', 'Seat Comfort', 'Customer Type', 'Cleanliness', 'Flight Distance', 'Food and Drink'
group_42_57	'Type of Travel', 'Class', 'Leg Room Service', 'In-flight Entertainment', 'Online Boarding', 'On-board Service', 'In-flight Wifi Service', 'Flight Distance', 'Baggage Handling', 'Seat Comfort'
group_58_67	'Type of Travel', 'Class', 'Leg Room Service', 'In-flight Entertainment', 'In-flight Wifi Service', 'Online Boarding', 'On-board Service', 'Flight Distance', 'Baggage Handling', 'In-flight Service'
group_68_76	'Type of Travel', 'In-flight Wifi Service', 'Leg Room Service', 'Arrival Delay', 'In-flight Entertainment', 'Ease of Online Booking', 'Class', 'On-board Service', 'Online Boarding', 'Baggage Handling', 'Flight Distance'
group_77_94	'In-flight Entertainment', 'On-board Service', 'In-flight Service', 'Arrival Delay', 'Leg Room Service', 'Cleanliness', 'In-flight Wifi Service', 'Check-in Service', 'Seat Comfort', 'Flight Distance'



# Data Balancing - using Random Over-Sampling





- Choosing the right algorithm for machine learning
- Quick performance comparison with Lazy Predict
- Simplified evaluation of multiple algorithms
- Time-saving approach for algorithm selection
- Informed decision for model performance

# Lazy Predict library

## Focusing on one group: group\_26\_41

	Accuracy	Balanced Accuracy	ROC AUC	F1 Score	Time Taken
Model					
LGBMClassifier	0.93	0.92	0.92	0.93	0.24
XGBClassifier	0.92	0.92	0.92	0.92	0.65
SVC	0.92	0.91	0.91	0.92	12.15
LabelSpreading	0.92	0.91	0.91	0.92	61.11
LabelPropagation	0.91	0.91	0.91	0.91	40.52
KNeighborsClassifier	0.91	0.91	0.91	0.91	0.47
RandomForestClassifier	0.91	0.90	0.90	0.91	1.52
BaggingClassifier	0.91	0.90	0.90	0.91	0.35
ExtraTreesClassifier	0.90	0.90	0.90	0.90	1.33
DecisionTreeClassifier	0.89	0.89	0.89	0.89	0.07
ExtraTreeClassifier	0.89	0.89	0.89	0.89	0.04
AdaBoostClassifier	0.88	0.88	0.88	0.88	0.59
LinearSVC	0.87	0.87	0.87	0.87	0.63
LogisticRegression	0.87	0.87	0.87	0.87	0.08
CalibratedClassifierCV	0.87	0.87	0.87	0.87	1.76
LinearDiscriminantAnalysis	0.87	0.86	0.86	0.87	0.09
RidgeClassifier	0.87	0.86	0.86	0.87	0.03
RidgeClassifierCV	0.86	0.86	0.86	0.86	0.05
NuSVC	0.87	0.85	0.85	0.86	35.33
SGDClassifier	0.86	0.85	0.85	0.86	0.06
GaussianNB	0.85	0.85	0.85	0.85	0.04
BernoulliNB	0.83	0.83	0.83	0.83	0.03
QuadraticDiscriminantAnalysis	0.83	0.83	0.83	0.83	0.04
PassiveAggressiveClassifier	0.83	0.81	0.81	0.82	0.05
NearestCentroid	0.81	0.81	0.81	0.81	0.02
Perceptron	0.77	0.74	0.74	0.76	0.04
DummyClassifier	0.58	0.50	0.50	0.42	0.02

# Machine Learning



**DECISION  
TREE**



**RANDOM  
FOREST**



**GRID-  
SEARCH**

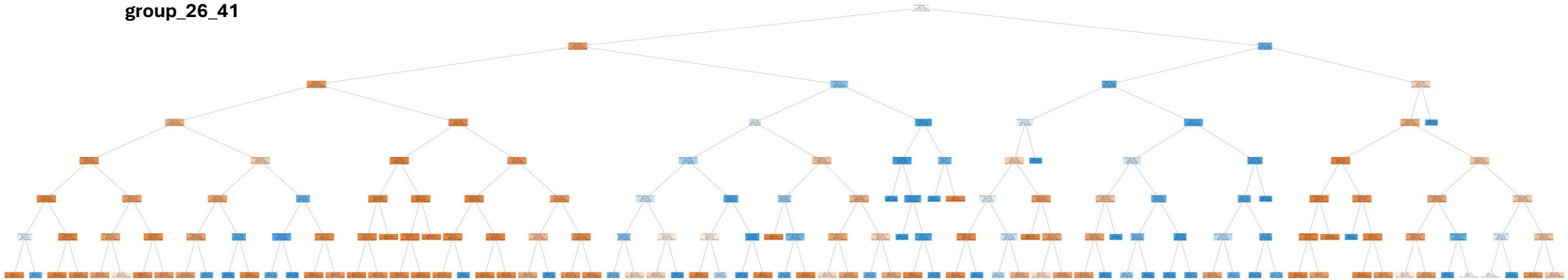


**LGBM  
CLASSIFIER**



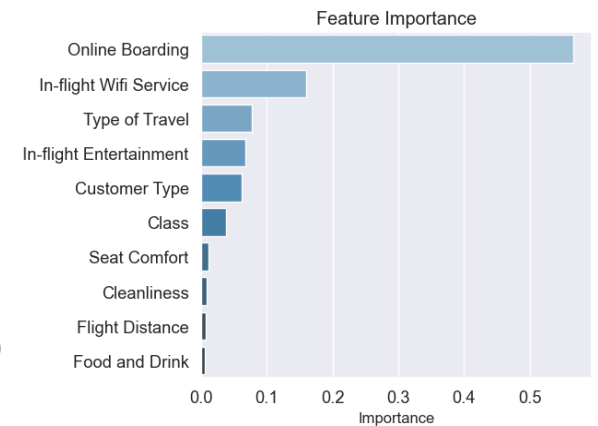
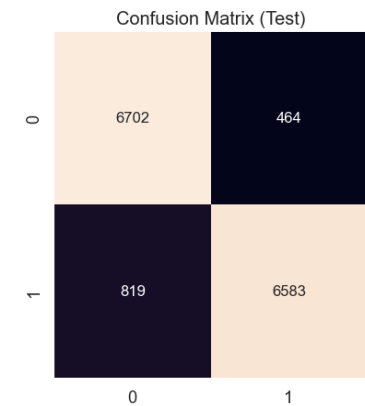
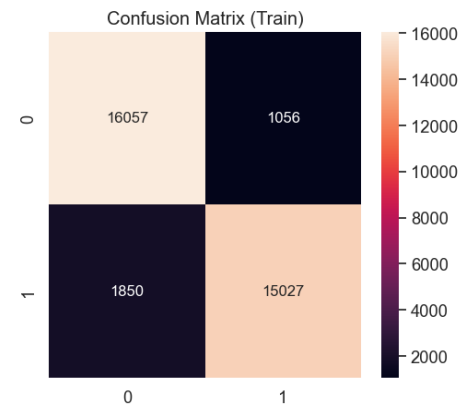
# Decision Tree Classifier

group\_26\_41



The classification accuracy was around 90% for all age groups.

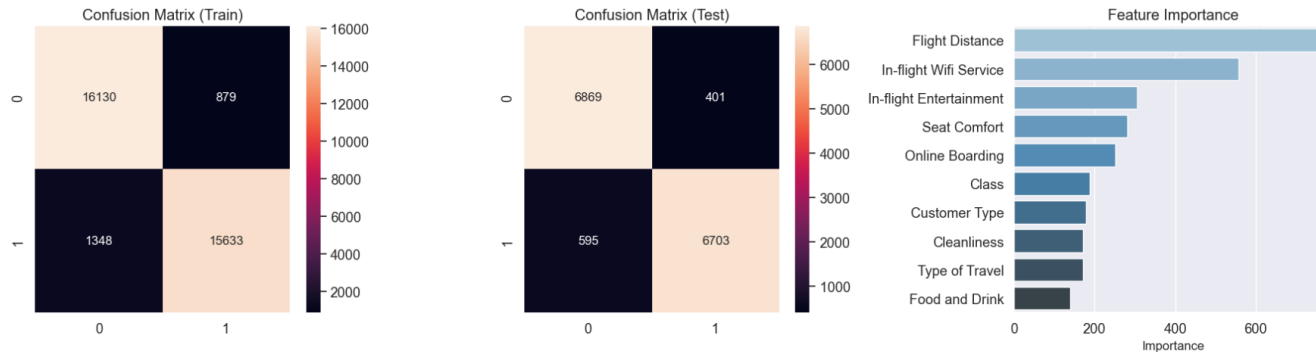
The highest accuracy is for group\_77\_94 (Train Accuracy: 0.96581 | Test Accuracy: 0.93377)



# LightGBM Classifier

## group\_26\_41

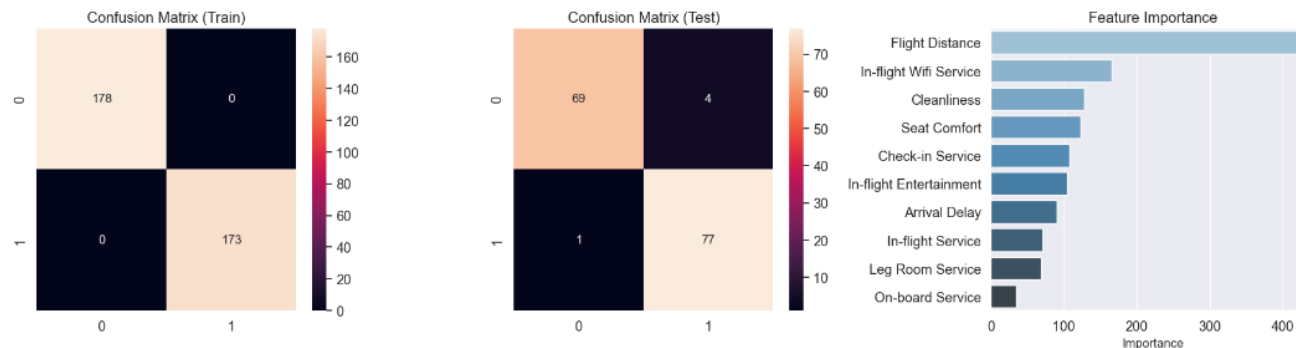
Train Accuracy: 0.93448 | Test Accuracy: 0.93163  
TPR Train: 0.92062 | TPR Test: 0.91847  
TNR Train: 0.94832 | TNR Test: 0.94484  
FPR Train: 0.05168 | FPR Test: 0.05516  
FNR Train: 0.07938 | FNR Test: 0.08153



- Faster training speed and higher efficiency.
- Lower memory usage.
- Better accuracy.

## group\_77\_94

Train Accuracy: 1.00000 | Test Accuracy: 0.96689  
TPR Train: 1.00000 | TPR Test: 0.98718  
TNR Train: 1.00000 | TNR Test: 0.94521  
FPR Train: 0.00000 | FPR Test: 0.05479  
FNR Train: 0.00000 | FNR Test: 0.01282



## Overfitting

reasons: training data size is too small

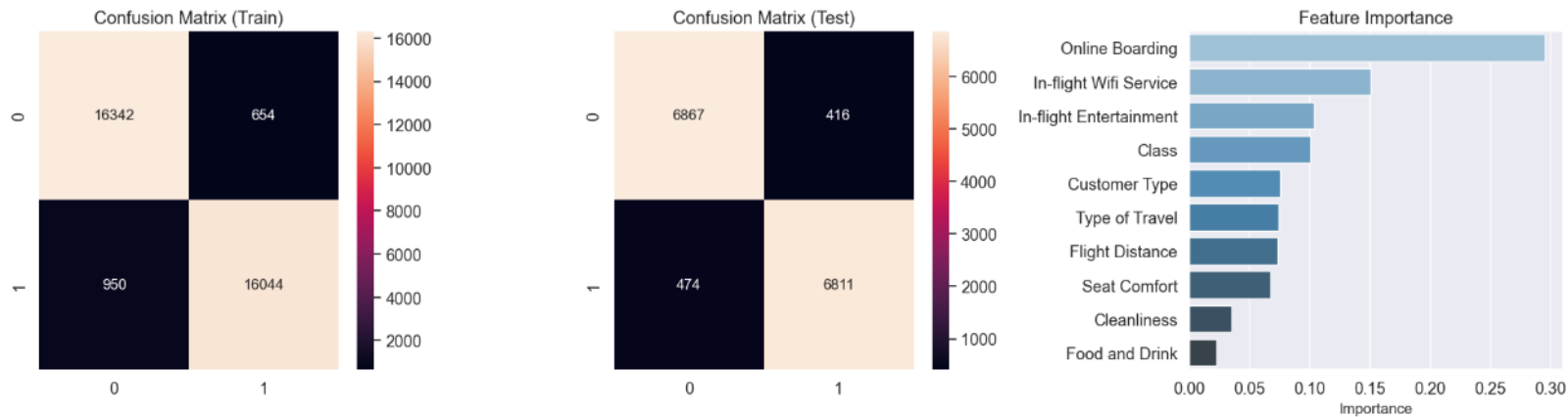


# After Cross-Validation Grid Search - Random Forest with best hyperparameter

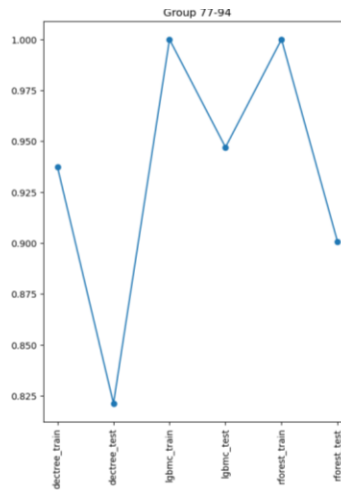
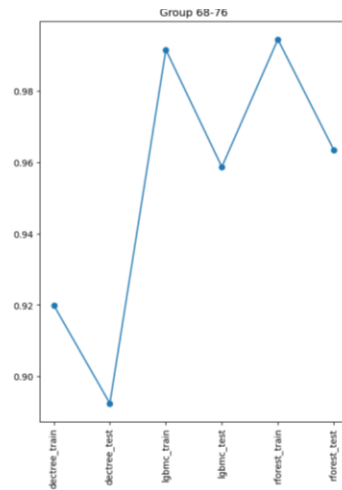
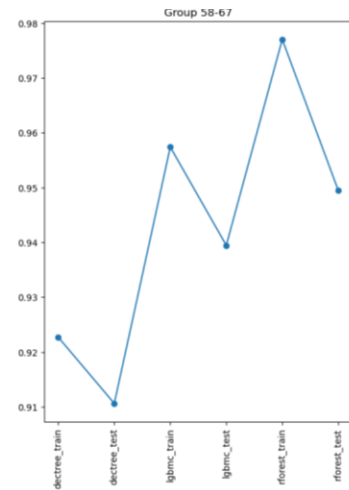
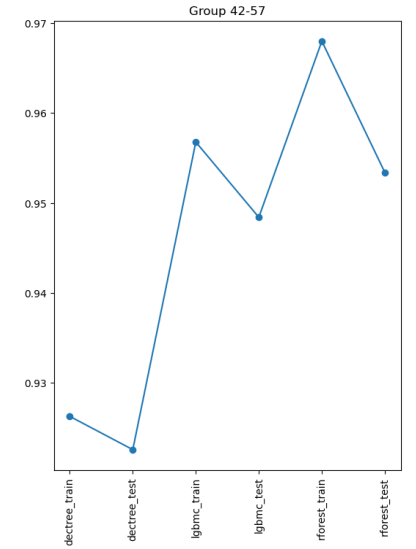
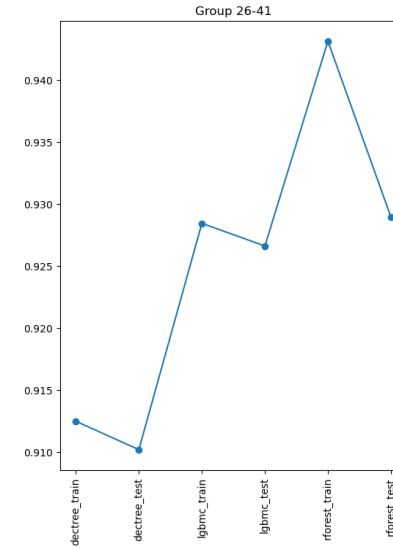
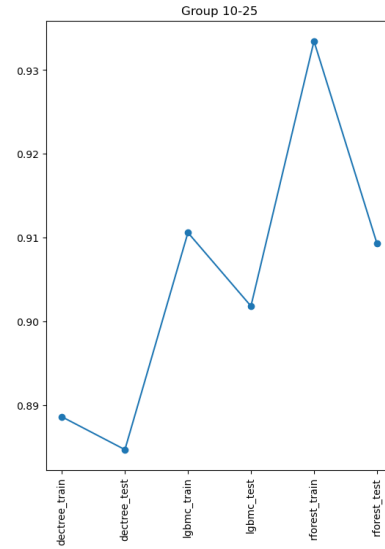
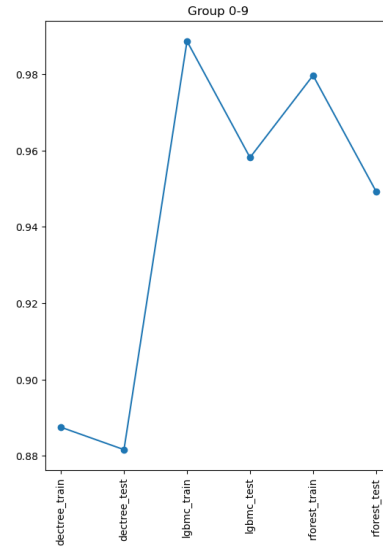
	group_0_9	group_10_25	group_26_41	group_42_57	group_58_67	group_68_76	group_77_94
n_estimators	500	900	600	600	1000	1000	600
max_depth	14	14	14	14	14	14	13
best_score	0.95480	0.91838	0.93663	0.95649	0.96116	0.96315	0.94298

group\_26\_41:

Train Accuracy: 0.95281 | Test Accuracy: 0.93891  
TPR Train: 0.94410 | TPR Test: 0.93493  
TNR Train: 0.96152 | TNR Test: 0.94288  
FPR Train: 0.03848 | FPR Test: 0.05712  
FNR Train: 0.05590 | FNR Test: 0.06507



# Accuracy comparison



Among these 3 models, we can see from this graph that we can consider random forest with grid search as the best model for our prediction.

# Feature importance

group\_0\_9

group\_10\_25

group\_26\_41

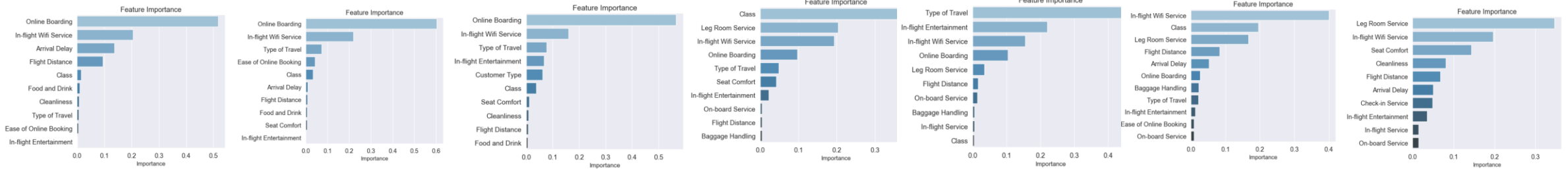
group\_42\_57

group\_58\_67

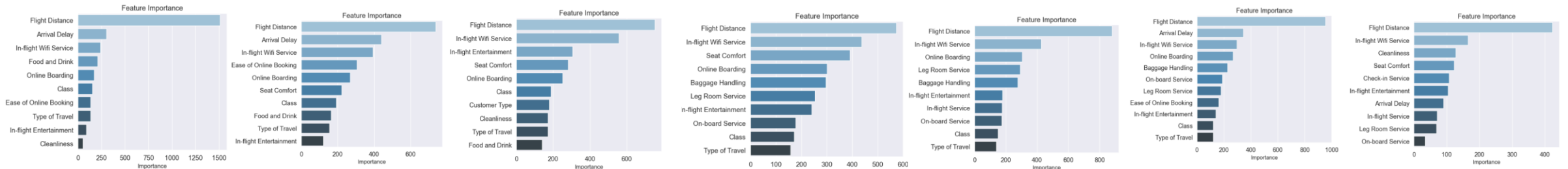
group\_68\_76

group\_77\_94

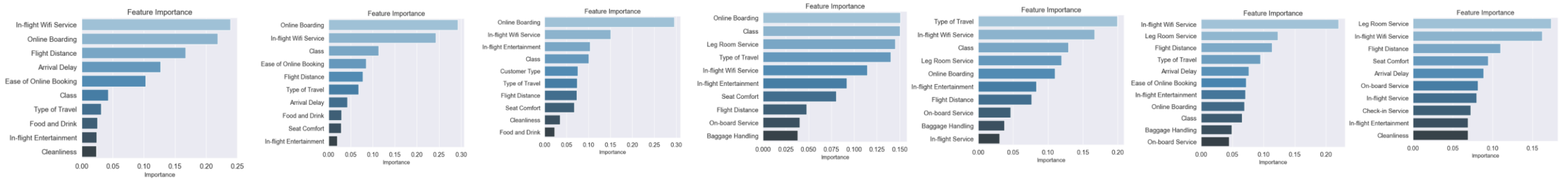
## Decision Tree Classifier



## LightGBM Classifier



## Random Forest with GridSearch



# Conclusion

Generation	Key Services
Gen Z	Online Boarding, In-flight Wifi Service (Flight Distance)
Millennials	Online Boarding, In-flight Wifi Service, Arrival Delay (Flight Distance)
Gen X	Online Boarding, In-flight Wifi Service,(Flight Distance)
Boomer I	Class, Leg Room Service, In-flight Wifi Service, Online Boarding (Flight Distance)
Boomer II	Type of Travel, In-flight Wifi Service, Online Boarding (Flight Distance)
Post War	In-flight Wifi Service, Leg Room Service (Flight Distance)
WWII	Leg Room Service, In-flight Wifi Service (Flight Distance)

## Recommendations

Focusing on providing efficient **online boarding** and **WiFi services**

Other factors:

Class

- Leg Room Service
- Type of Travel