Cairo University - Faculty of Engineering
Computer Engineering Department
CMP4040 – Machine Intelligence
Spring 2024

Big Data Project Report

Team #12

Airline Passenger Satisfaction Prediction

Team members:

Name	Section	BN
Donia Gameel	1	24
Shaza Mohammed	1	32
Heba Ashraf Raslan	2	32

Supervised by:

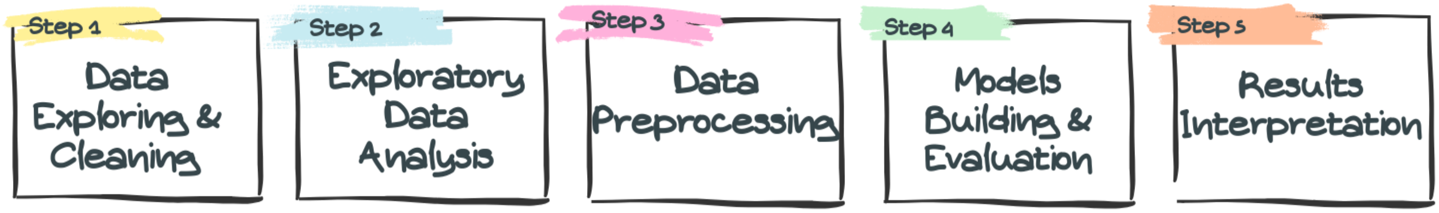
Dr. Lydia Waheed

Eng. Omar Samir

Problem description:

The project aims to predict airline passenger satisfaction based on various factors such as flight distance, in-flight service, ease of online booking, and departure/arrival time convenience. Understanding the factors that contribute to passenger satisfaction is crucial for airlines to improve their services, enhance customer experience, and increase customer loyalty. By analyzing this dataset, we aim to provide valuable insights into what drives passenger satisfaction and how airlines can better meet customer expectations.

Project Pipeline:


Step 1
Data Exploring & Cleaning
Step 2
Exploratory Data Analysis
Step 3
Data Preprocessing
Step 4
Models Building & Evaluation
Step 5
Results Interpretation

(1) Data Exploring & Cleaning:

- Explore the features and the unique values for each

Column name	# unique values	Example of the values
Unnamed: 0	103904	[0 1 2 ... 103901 103902 103903]
id	103904	[70172 5047 110028 ... 54173 62567]
Gender	2	['Male' 'Female']
Customer Type	2	['Loyal Customer' 'disloyal Customer']
Age	75	[9 12..20 24..37 40 ..60 66 64..72 79]
Type of travel	2	['Personal Travel' 'Business travel']
class	3	['Eco Plus' 'Business' 'Eco']
Flight Diastance	3802	[460 235 1142 ... 974 1479 400]
Inflight wifi service	6	[0 1 3 2 5 4]
Departure/Arrival time convenient	6	[0 1 3 2 5 4]
Ease of Online booking	6	[0 1 3 2 5 4]
Gate location	6	[0 1 3 2 5 4]
Food and drink	6	[0 1 3 2 5 4]
Online boarding	6	[0 1 3 2 5 4]
Seat comfort	6	[0 1 3 2 5 4]
Inflight entertainment	6	[0 1 3 2 5 4]
On-board service	6	[0 1 3 2 5 4]
Leg room service	6	[0 1 3 2 5 4]
Baggage handling	5	[0 1 3 2 5 4]
Checkin service	6	[0 1 3 2 5 4]
Inflight service	6	[0 1 3 2 5 4]
Cleanliness	6	[0 1 3 2 5 4]

Departure Delay in Minutes	446	[25 1 0 49 109 435 1592 1305 652 726..]
Arrival Delay in Minutes	455	[1.800e+01 6.000e+00 2.800e+02 1.410e+02..]
Satisfaction	2	['neutral or dissatisfied' 'satisfied']

- Explore dataset size:
  - Training data: 103904 rows × 25 columns
  - Test data: 25976 rows × 25 columns
- Remove nulls from training & test dataset

Training data	Arrival Delay in Minutes column	310 null values
Test data	Arrival Delay in Minutes column	83 null values

- Check if there are duplicate rows

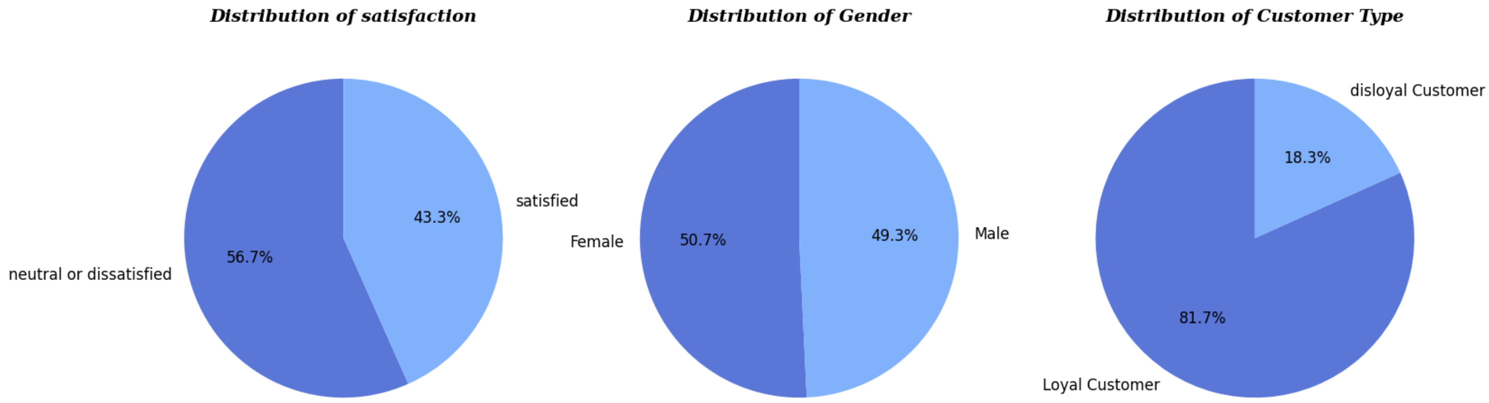
- Remove unnecessary columns

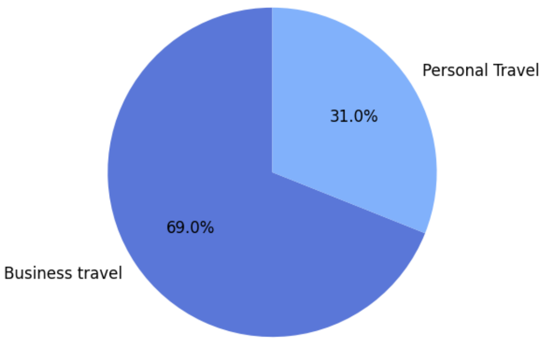
id
Unnamed:0

(2) Exploratory Data Analysis

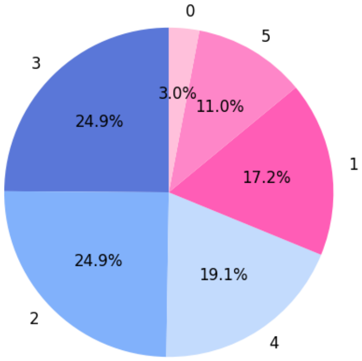
- *Univariate Analysis*

1- Charts for categorical variables

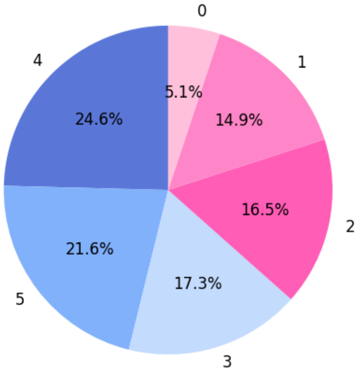




Distribution of Type of Travel

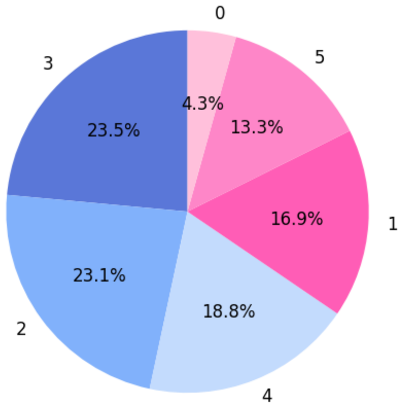


Distribution of Inflight wifi service

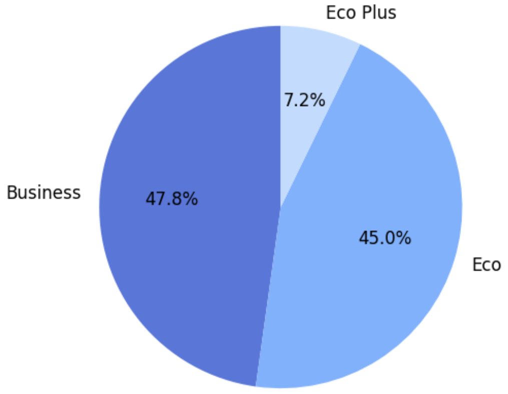


Distribution of Departure/Arrival time convenient

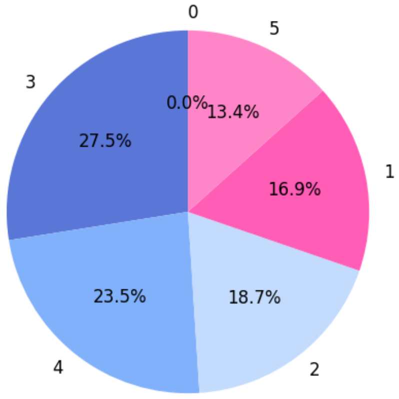
Distribution of Ease of Online booking



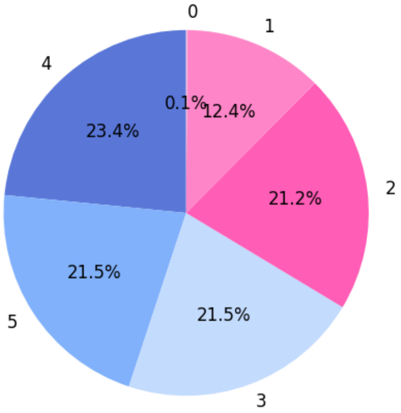
Distribution of Class



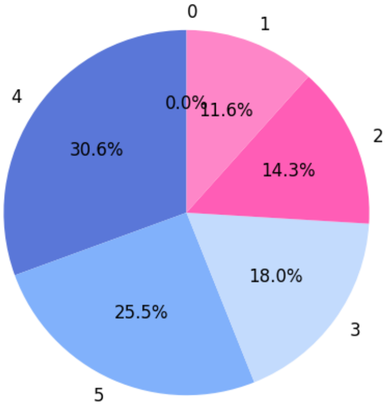
Distribution of Gate location



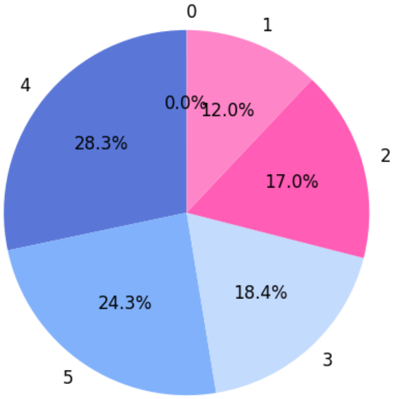
Distribution of Food and drink



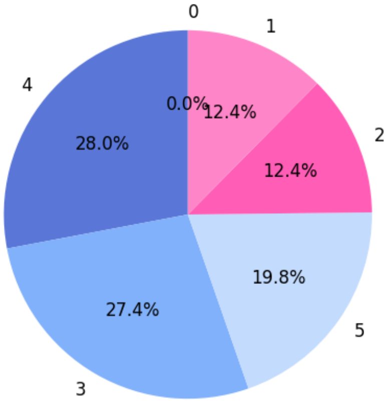
Distribution of Seat comfort



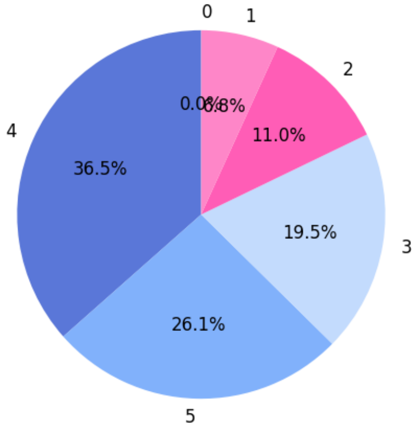
Distribution of Inflight entertainment



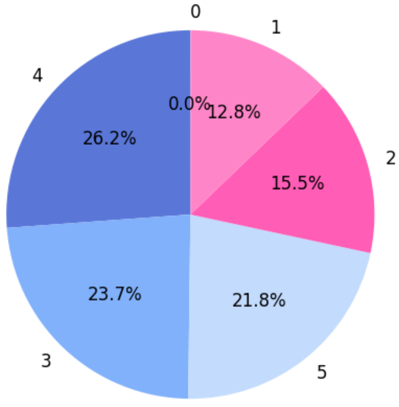
Distribution of Checkin service

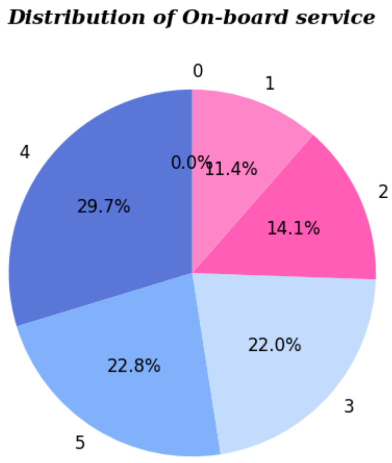


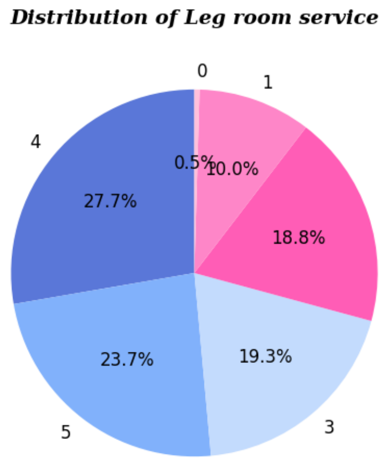
Distribution of Inflight service

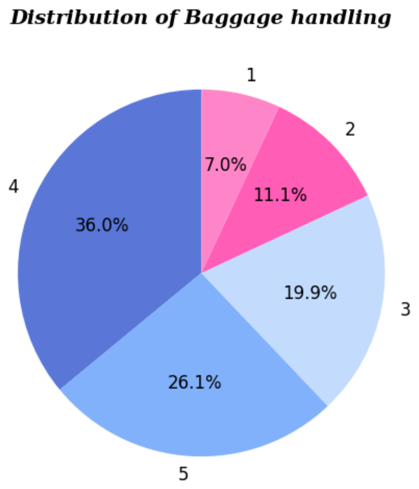


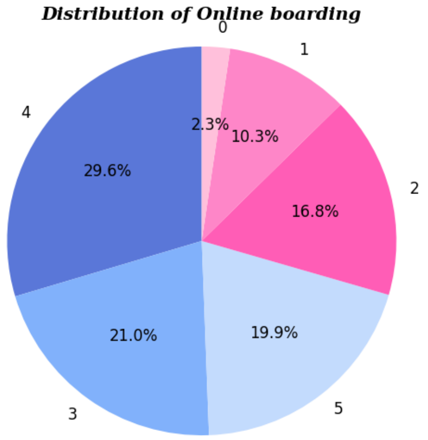
Distribution of Cleanliness







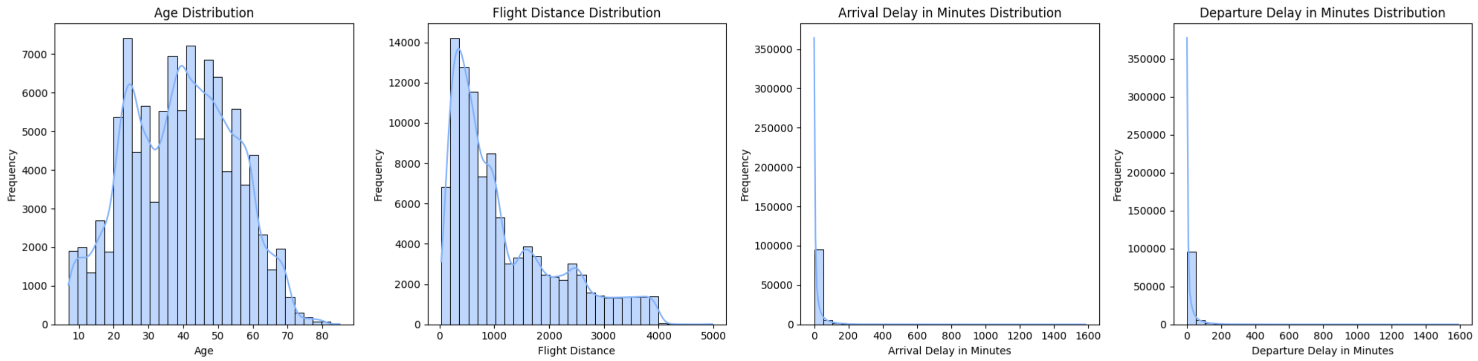




Insights:

- There is an almost equal number of male and female participants in the survey.
- Most of passengers are neutral or dissatisfied = 56.7% ==> we need to analysis the reasons and try to find business solutions to make them more satisfied
- We have more loyal customer data (81.7%)
- Most of travels are for Business travel (69%)
- Very few people fly in the economy plus class. They usually prefer Economy or Business.

2- Histogram for numerical variables



- Proportion of rows with 0 values in column 'Departure Delay in Minutes' to total rows: 0.5646365876193409
- Proportion of rows with 0 values in column 'Arrival Delay in Minutes' to total rows: 0.5597378349245458

Insights:

- Most of the delays are 0, which is a good indicator.

- The variables Flight Distance and Departure Delay and Arrival Delay are all heavily right-skewed.

Investigate problem of outliers:

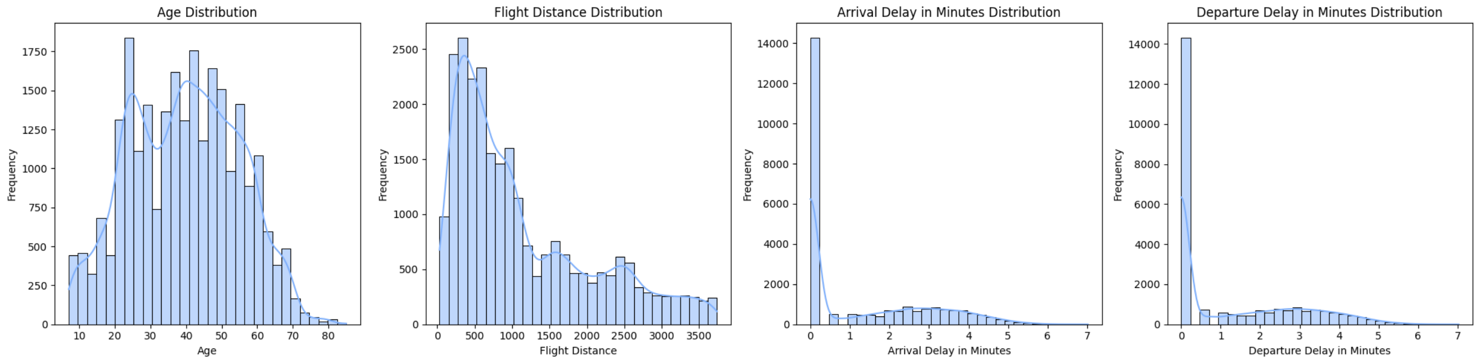
Number of outliers in each column:

Departure Delay in Minutes	14529
Arrival Delay in Minutes	13954
Flight Distance	2291

Portion of outliers in each column:

Departure Delay in Minutes	0.139831
Arrival Delay in Minutes	0.134699
Flight Distance	0.022049

- Since the portion of rows having the outliers in the "Flight Distance" is very small so we will remove it.
- We will normalize "Departure Delay in Minutes","Arrival Delay in Minutes"

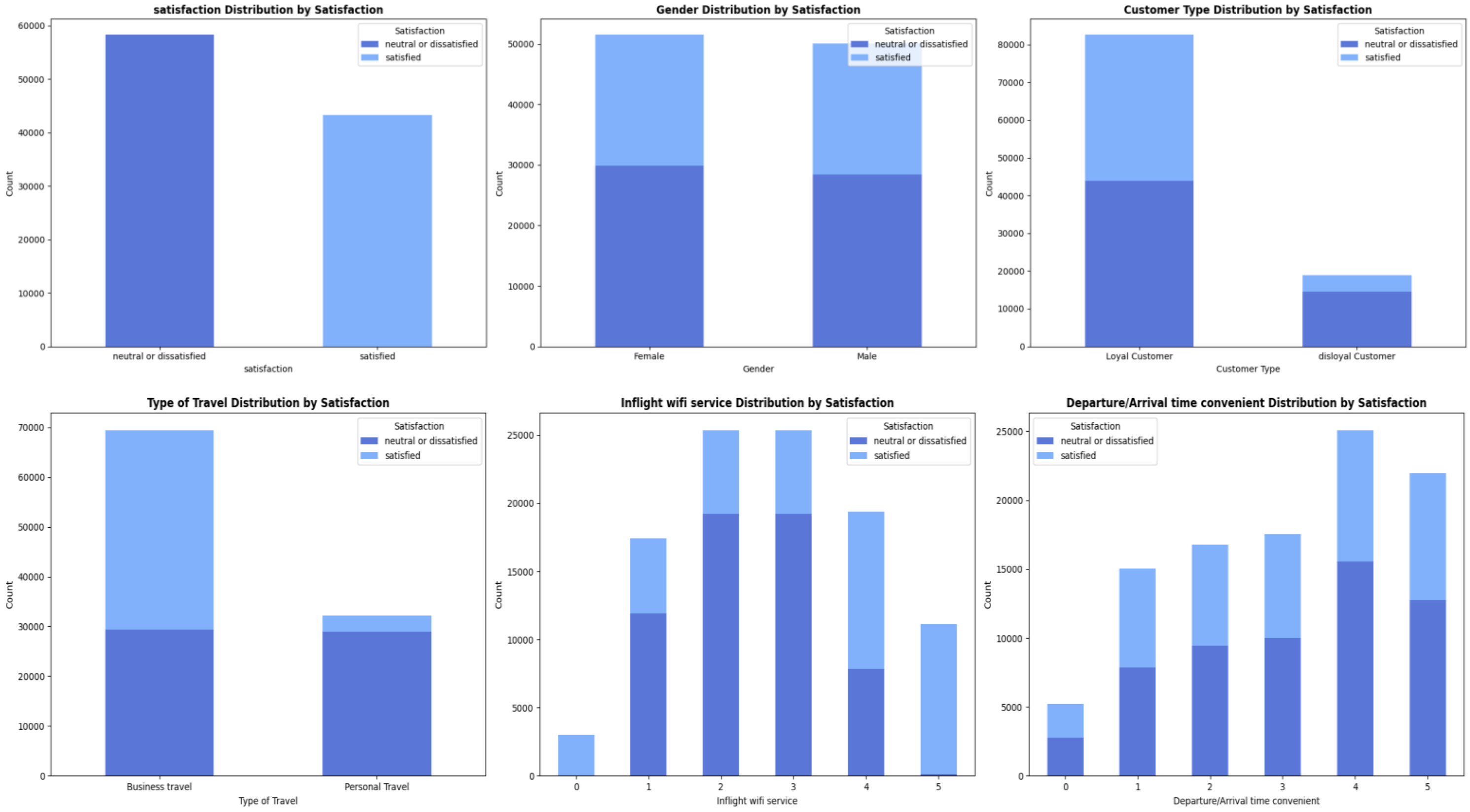


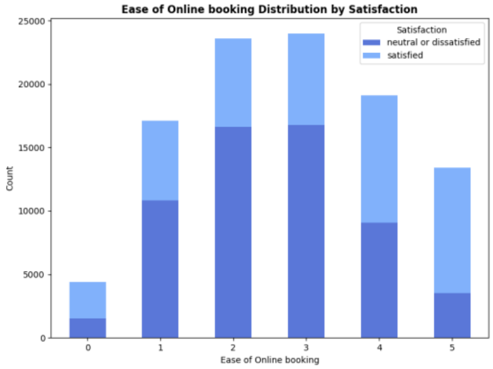
(figure) After solving the outliers problem

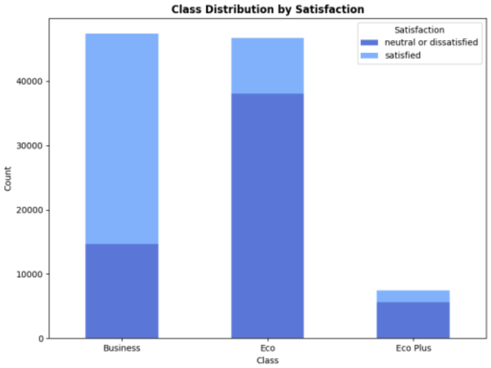
- The variables Departure Delay and Arrival Delay are still heavily right-skewed which is expected as most of the values are 0

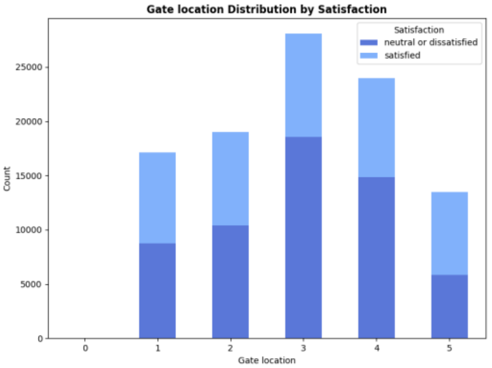
- *Bivariate Analysis*

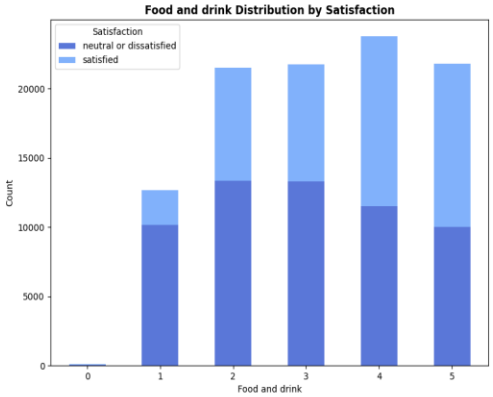
Bar charts & Pie charts for categorical features and histograms for numerical columns showing distribution of satisfaction

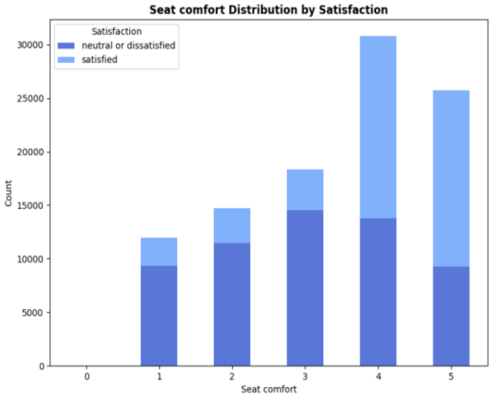


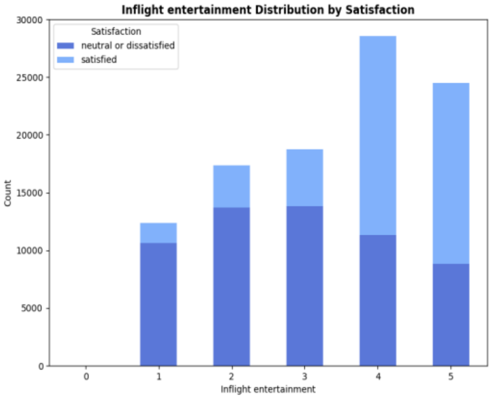

Ease of Online booking Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Ease of Online booking

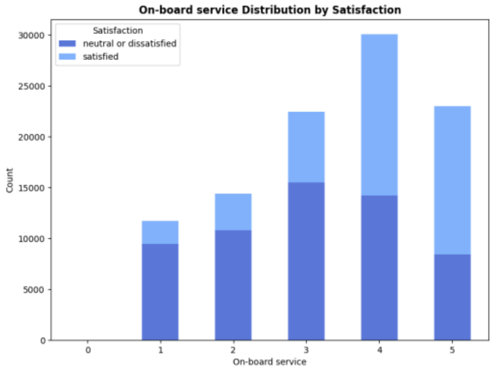

Class Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
40000
30000
20000
10000
0
Business
Eco
Eco Plus
Class

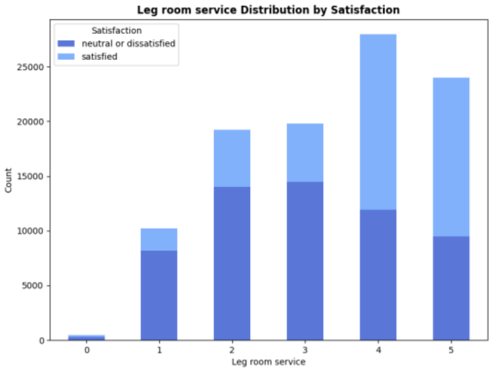

Gate location Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Gate location

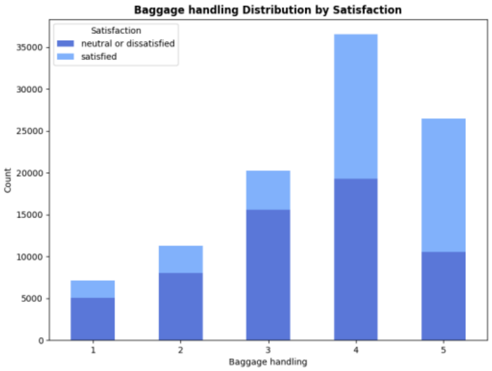

Food and drink Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
20000
15000
10000
5000
0
0
1
2
3
4
5
Food and drink

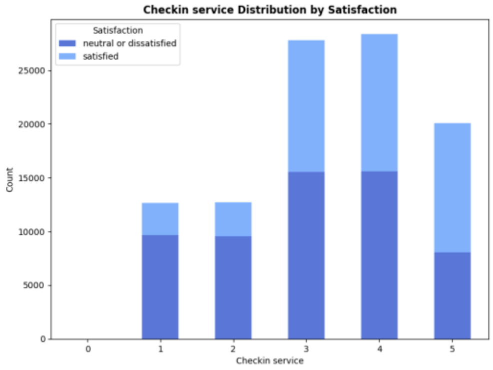

Seat comfort Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Seat comfort

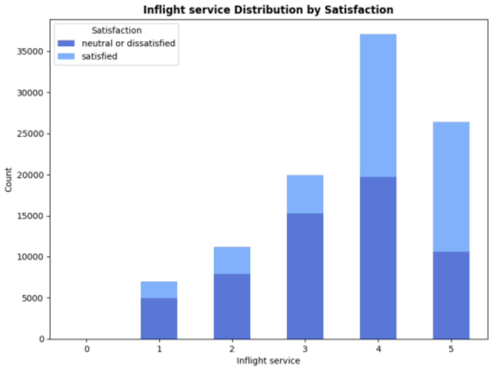

Inflight entertainment Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Inflight entertainment

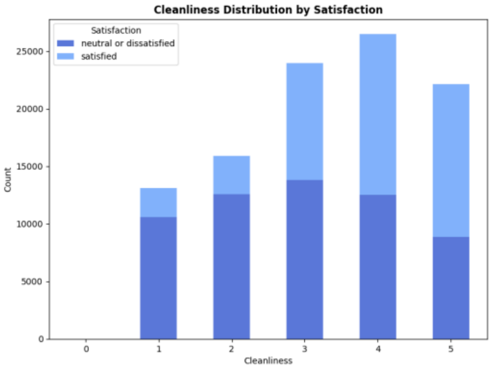

On-board service Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
On-board service

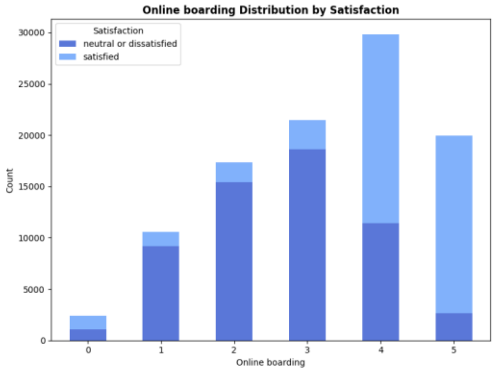

Leg room service Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Leg room service


Baggage handling Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
35000
30000
25000
20000
15000
10000
5000
0
1
2
3
4
5
Baggage handling

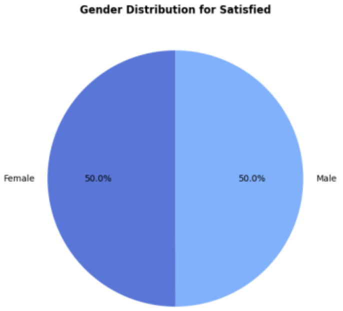

Checkin service Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Checkin service

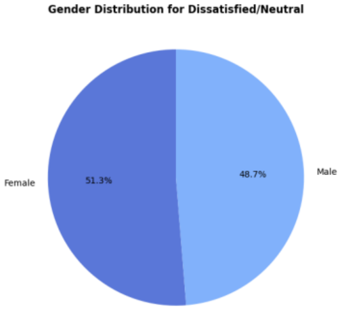

Inflight service Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
35000
30000
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Inflight service

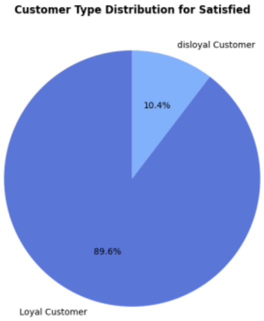

Cleanliness Distribution by Satisfaction
Satisfaction
neutral or dissatisfied
satisfied
Count
25000
20000
15000
10000
5000
0
0
1
2
3
4
5
Cleanliness

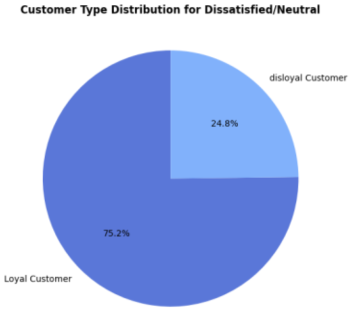


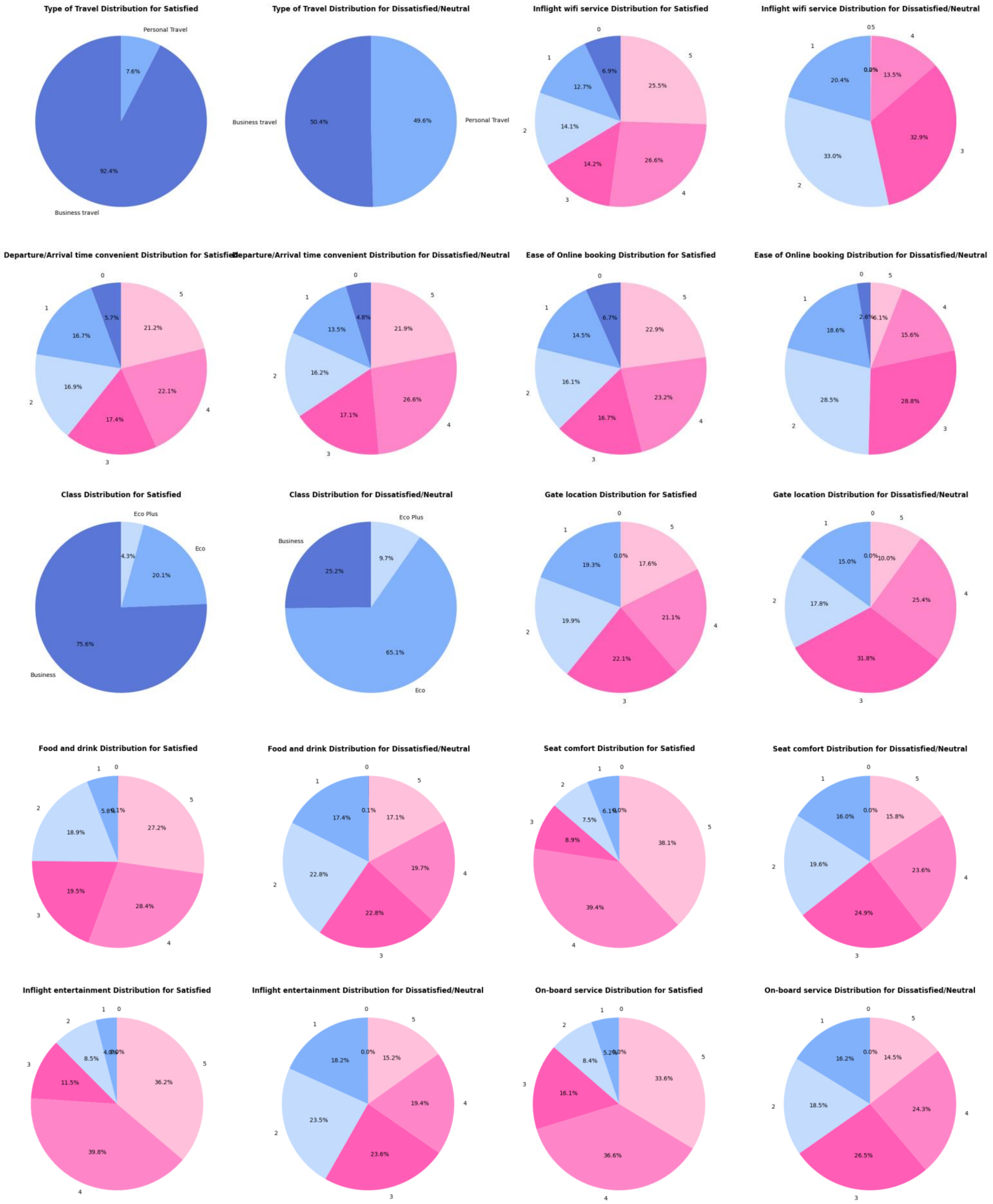
Then we focused on plotting the distribution for satisfied & dissatisfied for each feature:

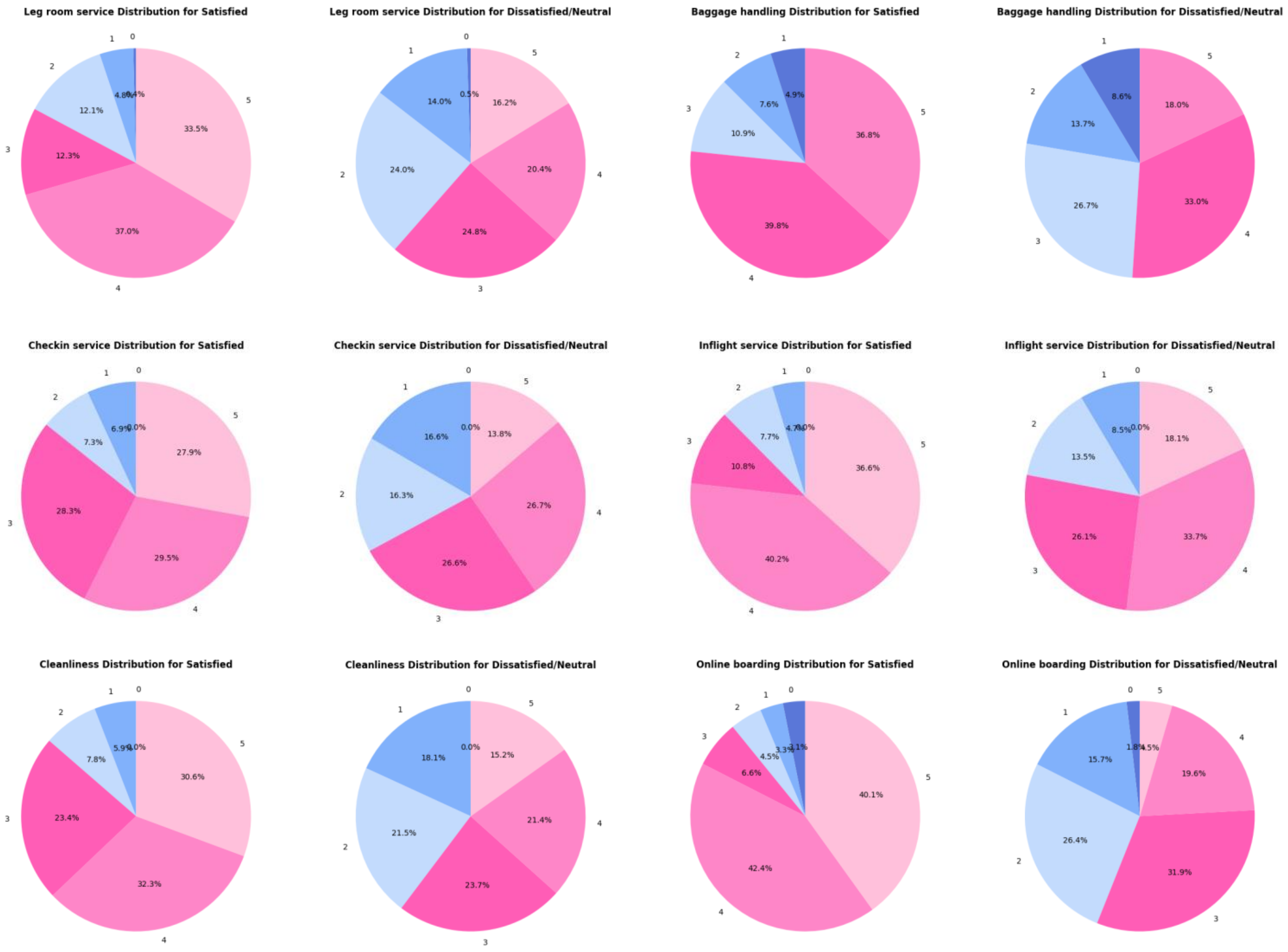


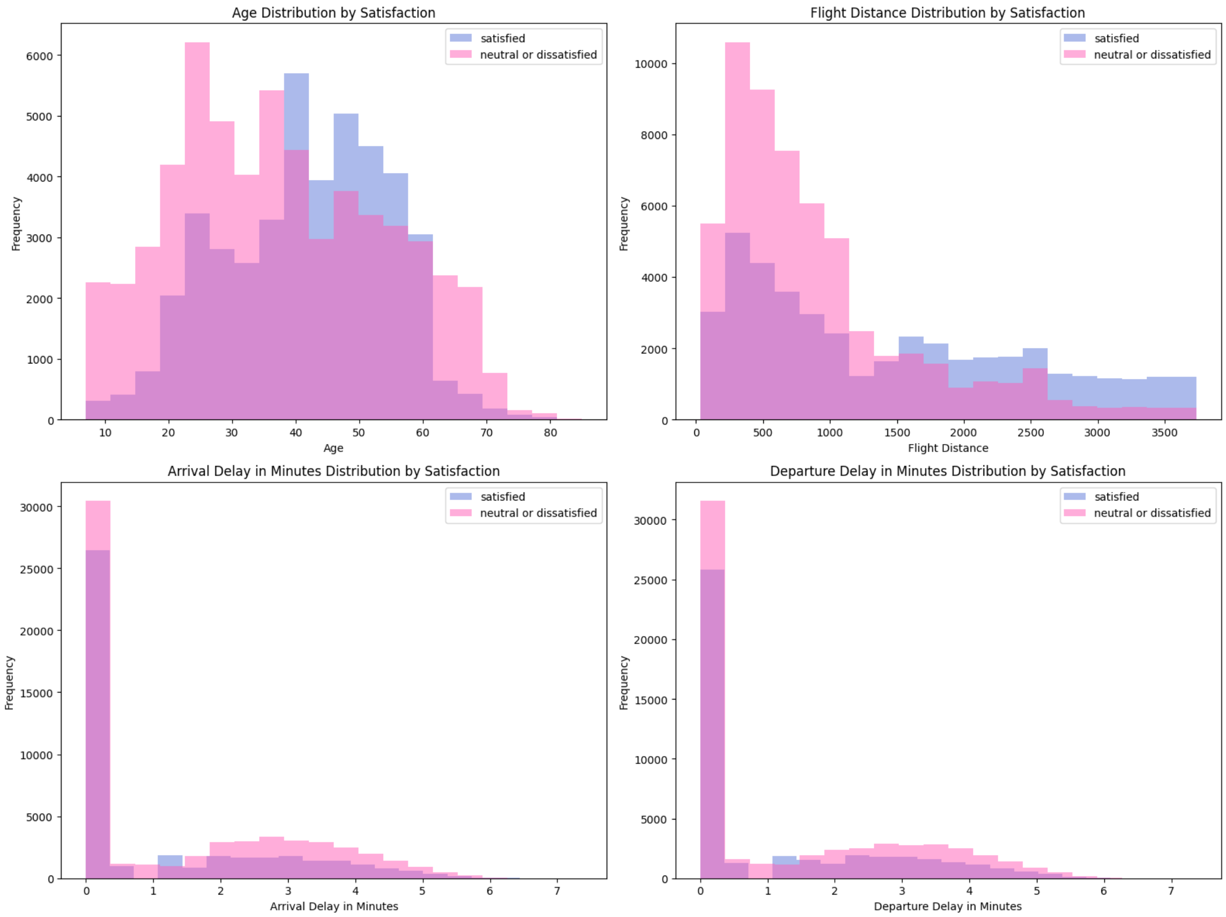




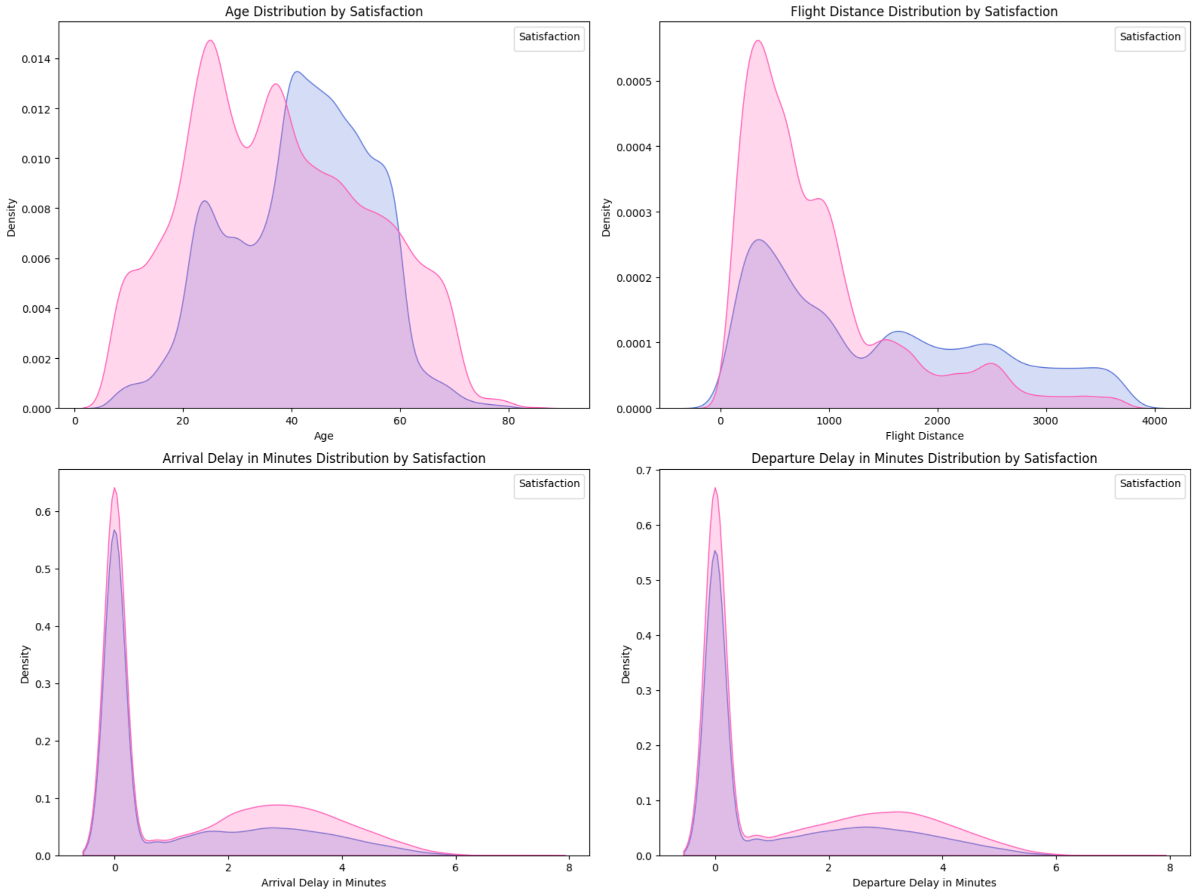


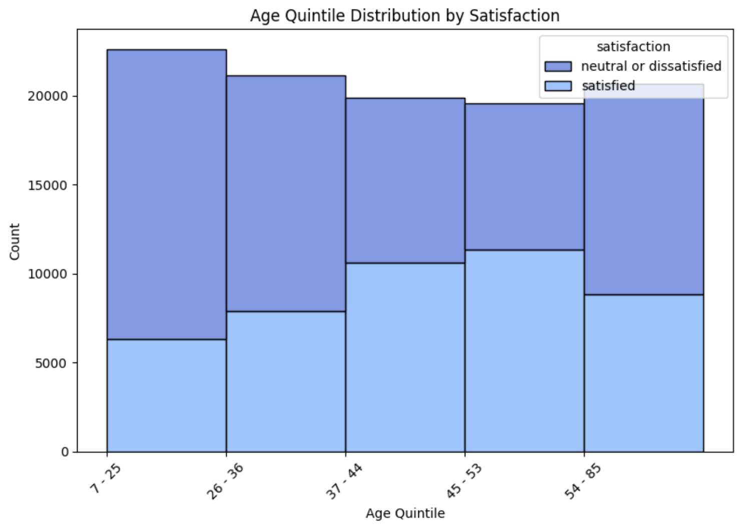

Type of Travel Distribution for Satisfied
Personal Travel
7.6%
92.4%
Business travel
Type of Travel Distribution for Dissatisfied/Neutral
Business travel
50.4%
49.6%
Personal Travel
Inflight wifi service Distribution for Satisfied
0
1
6.9%
12.7%
25.5%
5
2
14.1%
14.2%
26.6%
3
4
Inflight wifi service Distribution for Dissatisfied/Neutral
05
4
1
0.0%
13.5%
20.4%
32.9%
33.0%
3
2
Departure/Arrival time convenient Distribution for Satisfied
0
1
5
5.7%
16.7%
21.2%
16.9%
22.1%
2
4
17.4%
3
Departure/Arrival time convenient Distribution for Dissatisfied/Neutral
0
1
5
13.5%
4.8%
21.9%
2
16.2%
26.6%
17.1%
4
3
Ease of Online booking Distribution for Satisfied
0
5
1
6.7%
14.5%
22.9%
2
16.1%
23.2%
16.7%
4
3
Ease of Online booking Distribution for Dissatisfied/Neutral
0
5
1
4
2.6%
6.1%
18.6%
15.6%
28.5%
28.8%
2
3
Class Distribution for Satisfied
Eco Plus
Eco
4.3%
20.1%
75.6%
Business
Class Distribution for Dissatisfied/Neutral
Eco Plus
Business
9.7%
25.2%
65.1%
Eco
Gate location Distribution for Satisfied
0
1
5
19.3%
0.0%
17.6%
2
19.9%
21.1%
4
22.1%
3
Gate location Distribution for Dissatisfied/Neutral
0
1
5
15.0%
0.0%
10.0%
2
17.8%
25.4%
4
31.8%
3
Food and drink Distribution for Satisfied
1
0
2
5
5.8%0.1%
18.9%
27.2%
19.5%
28.4%
3
4
Food and drink Distribution for Dissatisfied/Neutral
0
1
5
17.4%
0.1%
17.1%
2
22.8%
19.7%
4
22.8%
3
Seat comfort Distribution for Satisfied
1
0
2
3
5
6.1%0.0%
7.5%
8.9%
38.1%
39.4%
4
Seat comfort Distribution for Dissatisfied/Neutral
0
1
5
16.0%
0.0%
15.8%
2
19.6%
23.6%
4
24.9%
3
Inflight entertainment Distribution for Satisfied
1
0
2
3
5
4.0%0.0%
8.5%
11.5%
36.2%
39.8%
4
Inflight entertainment Distribution for Dissatisfied/Neutral
0
1
5
18.2%
0.0%
15.2%
19.4%
4
2
23.5%
23.6%
3
On-board service Distribution for Satisfied
1
0
2
5
3
5.2%0.0%
8.4%
16.1%
33.6%
36.6%
4
On-board service Distribution for Dissatisfied/Neutral
0
1
5
16.2%
0.0%
14.5%
2
18.5%
24.3%
4
26.5%
3


Leg room service Distribution for Satisfied
0
1
2
3
4
5
0.4%
4.8%
12.1%
12.3%
37.0%
33.5%
Leg room service Distribution for Dissatisfied/Neutral
0
1
2
3
4
5
0.5%
14.0%
24.0%
24.8%
20.4%
16.2%
Baggage handling Distribution for Satisfied
1
2
3
4
5
4.9%
7.6%
10.9%
39.8%
36.8%
Baggage handling Distribution for Dissatisfied/Neutral
1
2
3
4
5
8.6%
13.7%
26.7%
33.0%
18.0%
Checkin service Distribution for Satisfied
0
1
2
3
4
5
0.0%
6.9%
7.3%
28.3%
29.5%
27.9%
Checkin service Distribution for Dissatisfied/Neutral
0
1
2
3
4
5
0.0%
16.6%
16.3%
26.6%
26.7%
13.8%
Inflight service Distribution for Satisfied
0
1
2
3
4
5
0.0%
4.7%
7.7%
10.8%
40.2%
36.6%
Inflight service Distribution for Dissatisfied/Neutral
0
1
2
3
4
5
0.0%
8.5%
13.5%
26.1%
33.7%
18.1%
Cleanliness Distribution for Satisfied
0
1
2
3
4
5
0.0%
5.9%
7.8%
23.4%
32.3%
30.6%
Cleanliness Distribution for Dissatisfied/Neutral
0
1
2
3
4
5
0.0%
18.1%
21.5%
23.7%
21.4%
15.2%
Online boarding Distribution for Satisfied
0
1
2
3
4
5
3.1%
3.3%
4.5%
6.6%
42.4%
40.1%
Online boarding Distribution for Dissatisfied/Neutral
0
5
4
3
2
1
1.8%
4.5%
19.6%
31.9%
26.4%
15.7%

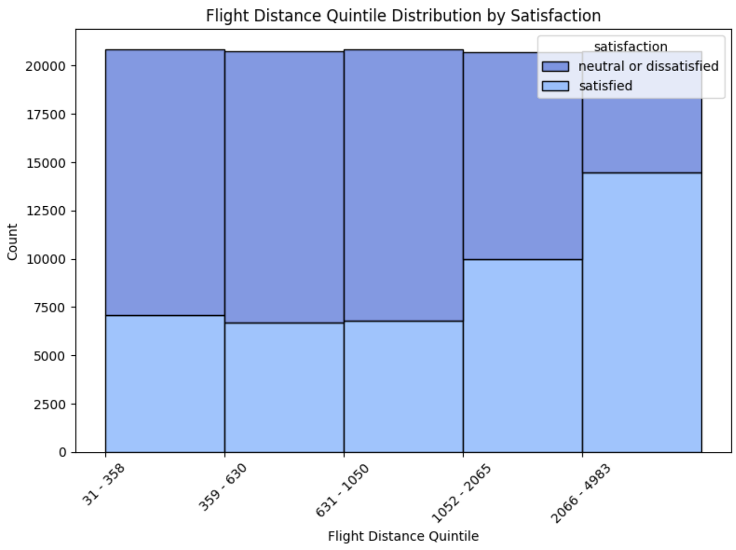


Clearer representation:

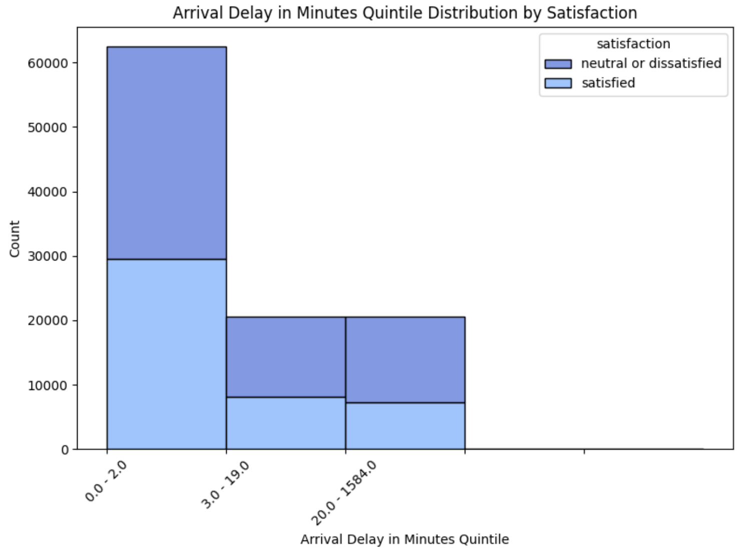




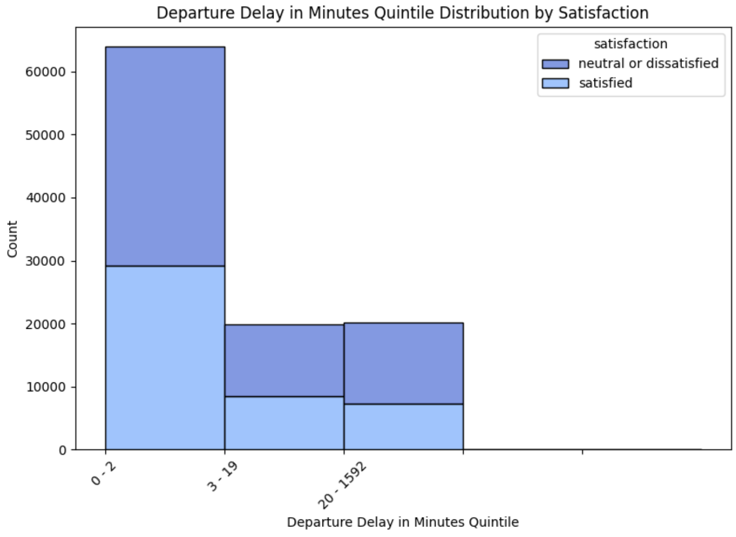
Age Quintile Distribution by Satisfaction



Flight Distance Quintile Distribution by Satisfaction



Arrival Delay in Minutes Quintile Distribution by Satisfaction



Departure Delay in Minutes Quintile Distribution by Satisfaction

Insights:

- Gender nearly doesn't affect satisfaction.
- Loyal passengers have higher satisfaction percentages than Disloyal ones.
- Satisfied Passengers usually go for Business travel.
- Most people of Passengers going for Personal Travel are not satisfied.
- Satisfied Passengers use Business Class while travelling.
- Passengers using Eco travelling are the least Satisfied Passengers
- More than 80% of passengers flying in economy are either Neutral or Dissatisfied. That shows us that it needs some improvement.
- Most Satisfied Passengers are in range [37-53] year & Most Unsatisfied are in range [7-36] year.
- Satisfied Passengers have more long-distance flights than the dissatisfied.
- The more the delay the less the satisfied passenger's portion.
- The most frequency in the levels of satisfaction is 4 for all except: [Inflight Wi-Fi service, Ease of Online booking, Gate location] is 3

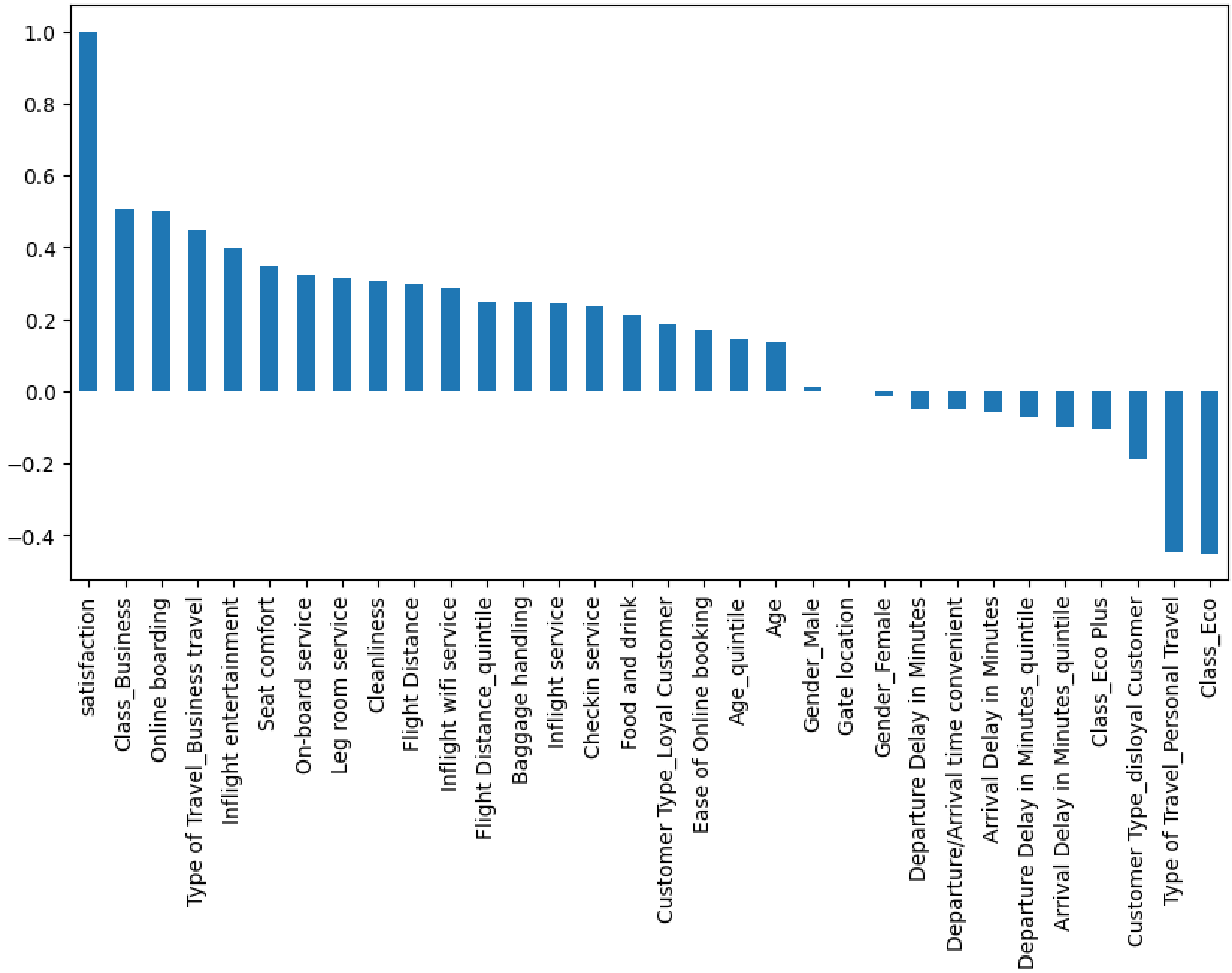
- Rate 3 is the most frequent between unsatisfied passengers in services
- Rate 4 is the most frequent between satisfied passengers in services
- The ratings are almost evenly distributed between 1 and 5.

With that in mind, the positive thing is that there are more positive or neutral ratings (3 through 5)
than negative ones (0 through 2).

- Our passengers have mixed opinions about the Departure and Arrival Time Convenience.

We concluded that there is not that much correlation between total Satisfaction
and Departure and Arrival Time Convenience.

Showing Correlation between satisfaction and other columns



Insights:

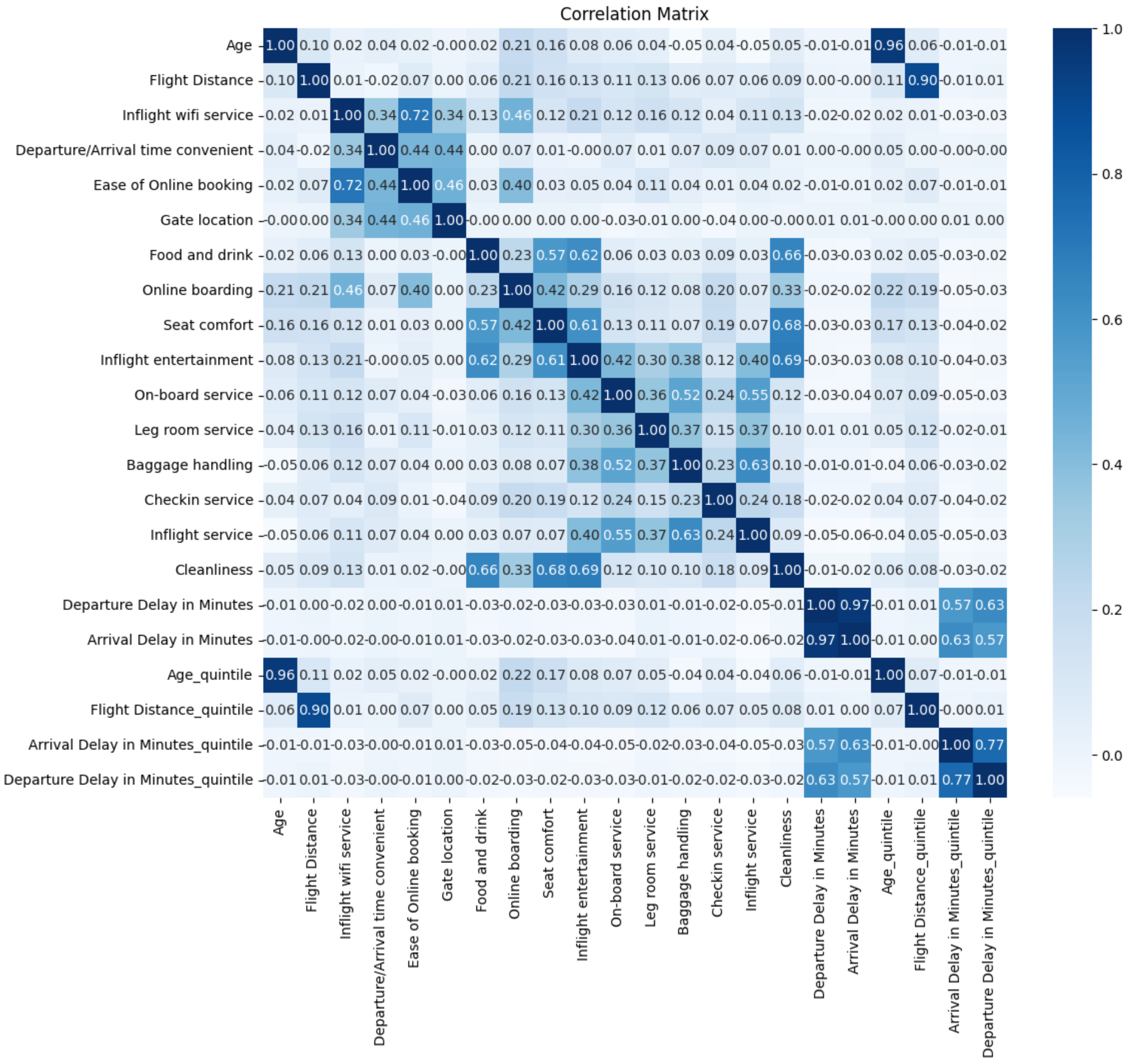
Positively Correlated:

- Business Class ,online boarding, inflight entertainment, seat comfort, on-board service, Legroom service, cleanliness, Flight distance, and Business travels are strong reasons for people satisfaction.

Negatively Correlated:

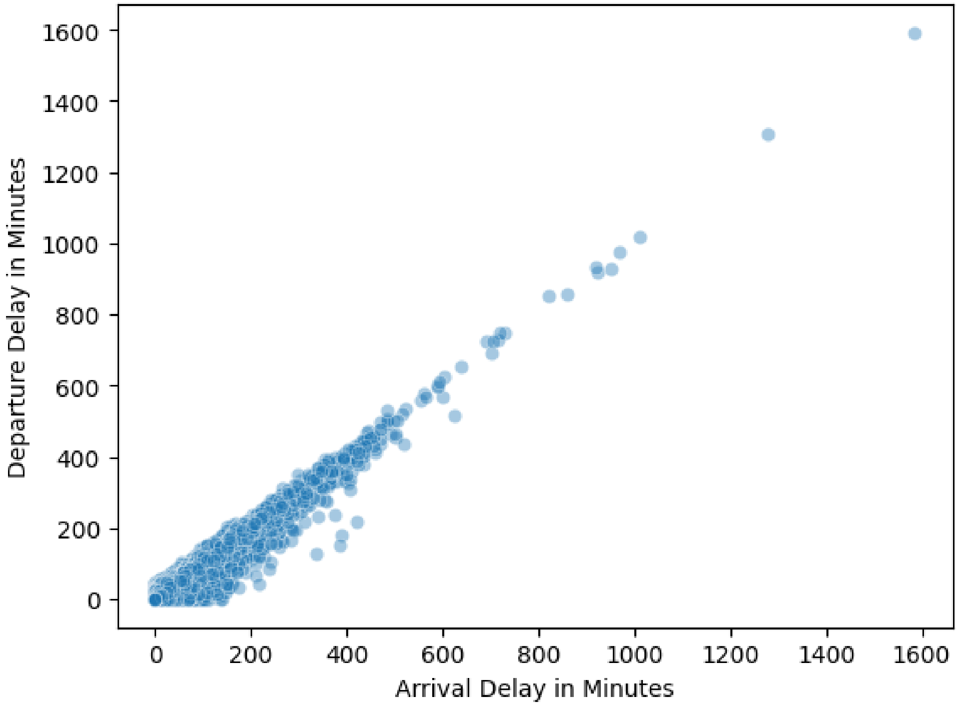
- Personal Travels, Economy Class, Eco plus Class or being Disloyal Customer results in Unsatisfaction.

- *Multivariate Analysis:*



Insights:

- Departure Delay is highly correlated with Arrival Delay. [Will deal with this in feature engineering].
- Inflight WiFi service and Ease of online booking are + correlated.
- Inflight entertainment, Food and Drink, Seat comfort and cleanliness are + correlated .
- Baggage handling is + correlated with Inflight service.



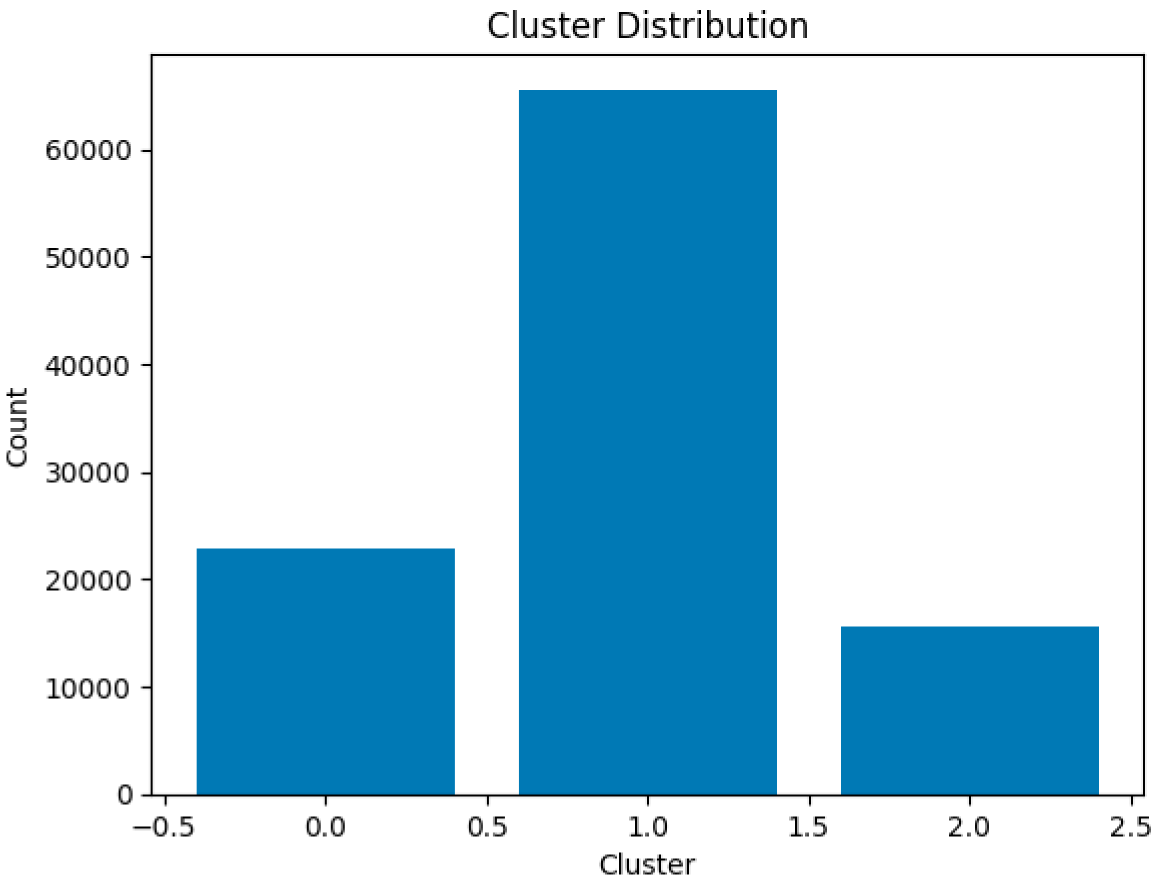
Insights:

- There is a strong correlation between the two columns
we can drop one of the two columns and as Arrival Delay in Minutes column has some null, we can drop it.

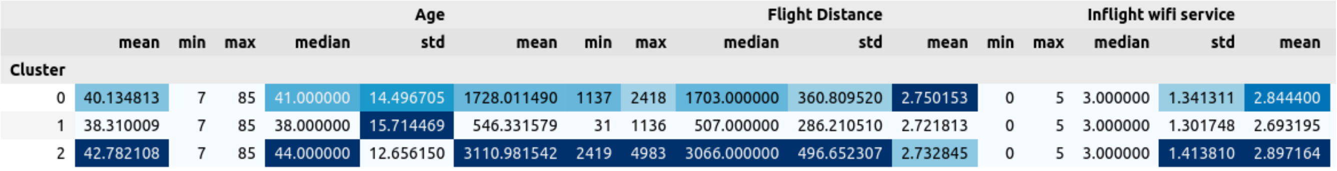
- Remove quintile columns ['Age_quintile', 'Flight Distance_quintile', 'Arrival Delay in Minutes_quintile', 'Departure Delay in Minutes_quintile']

- *<u>Clustering</u>*

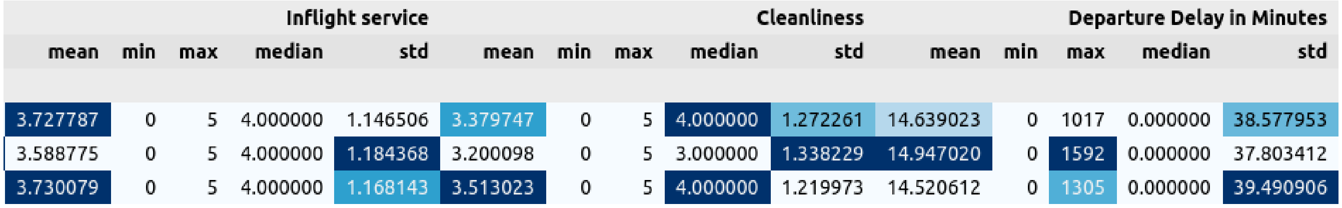
Applying Kmeans on the data with 3 clusters



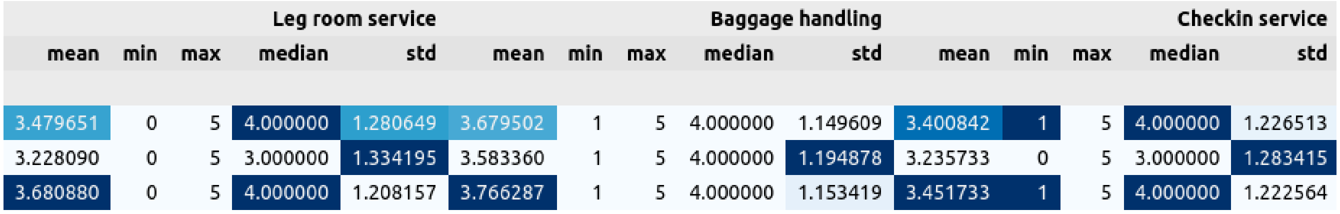
- The portion of each class:
    - Class 1 ==> 0.630900
    - Class 0 ==> 0.219453
    - Class 2 ==> 0.149648
- Most of the data (63%) is in one cluster (cluster 1)
- showing aggregates of each numeric column grouped by the cluster



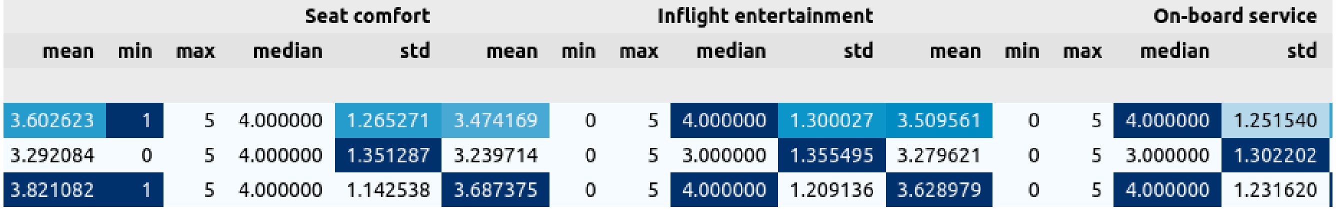
	Age					Flight Distance					Inflight wifi service					
	mean	min	max	median	std	mean	min	max	median	std	mean	min	max	median	std	mean
Cluster																
0	40.134813	7	85	41.000000	14.496705	1728.011490	1137	2418	1703.000000	360.809520	2.750153	0	5	3.000000	1.341311	2.844400
1	38.310009	7	85	38.000000	15.714469	546.331579	31	1136	507.000000	286.210510	2.721813	0	5	3.000000	1.301748	2.693195
2	42.782108	7	85	44.000000	12.656150	3110.981542	2419	4983	3066.000000	496.652307	2.732845	0	5	3.000000	1.413810	2.897164



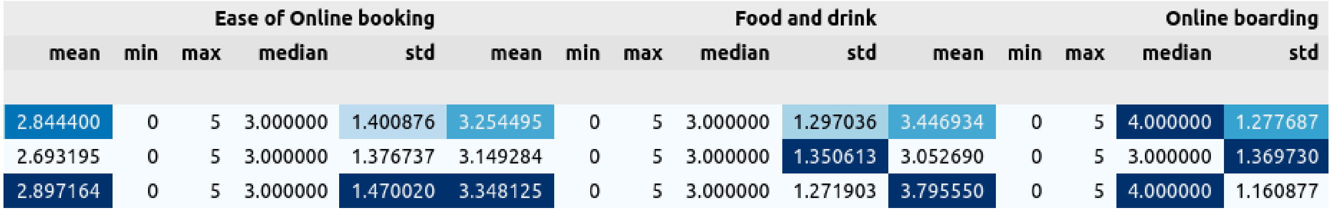
Inflight service					Cleanliness					Departure Delay in Minutes				
mean	min	max	median	std	mean	min	max	median	std	mean	min	max	median	std
3.727787	0	5	4.000000	1.146506	3.379747	0	5	4.000000	1.272261	14.639023	0	1017	0.000000	38.577953
3.588775	0	5	4.000000	1.184368	3.200098	0	5	3.000000	1.338229	14.947020	0	1592	0.000000	37.803412
3.730079	0	5	4.000000	1.168143	3.513023	0	5	4.000000	1.219973	14.520612	0	1305	0.000000	39.490906



Leg room service					Baggage handling					Checkin service				
mean	min	max	median	std	mean	min	max	median	std	mean	min	max	median	std
3.479651	0	5	4.000000	1.280649	3.679502	1	5	4.000000	1.149609	3.400842	1	5	4.000000	1.226513
3.228090	0	5	3.000000	1.334195	3.583360	1	5	4.000000	1.194878	3.235733	0	5	3.000000	1.283415
3.680880	0	5	4.000000	1.208157	3.766287	1	5	4.000000	1.153419	3.451733	1	5	4.000000	1.222564

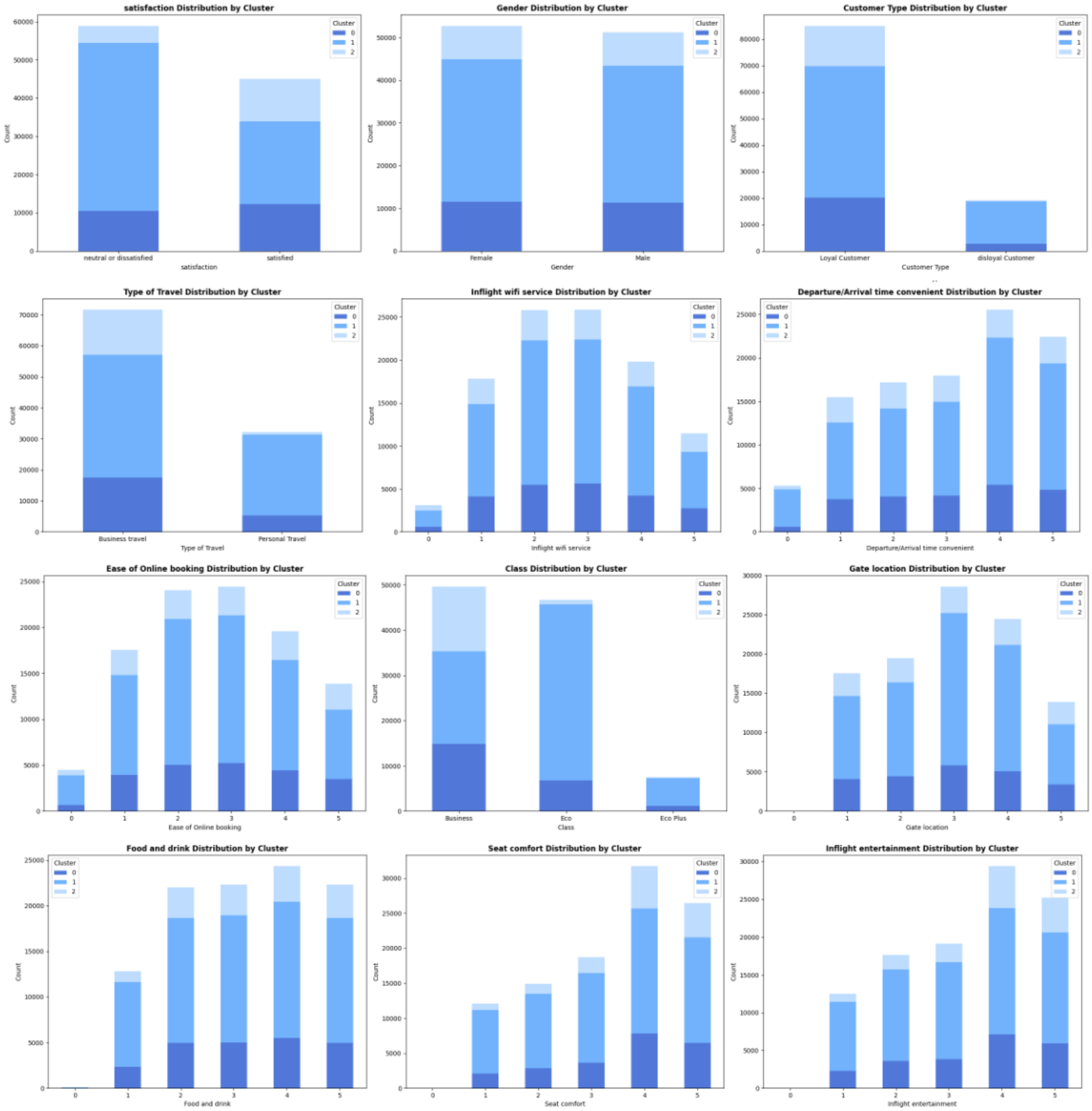


Seat comfort					Inflight entertainment					On-board service				
mean	min	max	median	std	mean	min	max	median	std	mean	min	max	median	std
3.602623	1	5	4.000000	1.265271	3.474169	0	5	4.000000	1.300027	3.509561	0	5	4.000000	1.251540
3.292084	0	5	4.000000	1.351287	3.239714	0	5	3.000000	1.355495	3.279621	0	5	3.000000	1.302202
3.821082	1	5	4.000000	1.142538	3.687375	0	5	4.000000	1.209136	3.628979	0	5	4.000000	1.231620



Ease of Online booking					Food and drink					Online boarding				
mean	min	max	median	std	mean	min	max	median	std	mean	min	max	median	std
2.844400	0	5	3.000000	1.400876	3.254495	0	5	3.000000	1.297036	3.446934	0	5	4.000000	1.277687
2.693195	0	5	3.000000	1.376737	3.149284	0	5	3.000000	1.350613	3.052690	0	5	3.000000	1.369730
2.897164	0	5	3.000000	1.470020	3.348125	0	5	3.000000	1.271903	3.795550	0	5	4.000000	1.160877

- **Insights:**
  - All age values are in the 3 clusters [7-85]
  - Flight Distances is distributed on all clusters without intersection between them:
  - Cluster 1 contains flight distance in the range [1137:2418]
  - Cluster 0 contains flight distance in the range [31:1136]
  - Cluster 2 contains flight distance in the range [2419:4983]
  - Value 0 for seat comfort column is only in cluster 1
  - Value 0 for Checkin service column is only in cluster 1
  - Values in range [1017:1592] don't exist in class 0
  - Values in range [1305:1592] don't exist in class 2
- Show the cluster distribution over the categories of categorical features



satisfaction Distribution by Cluster

Gender Distribution by Cluster

Customer Type Distribution by Cluster

Type of Travel Distribution by Cluster

Inflight wifi service Distribution by Cluster

Departure/Arrival time convenient Distribution by Cluster

Ease of Online booking Distribution by Cluster

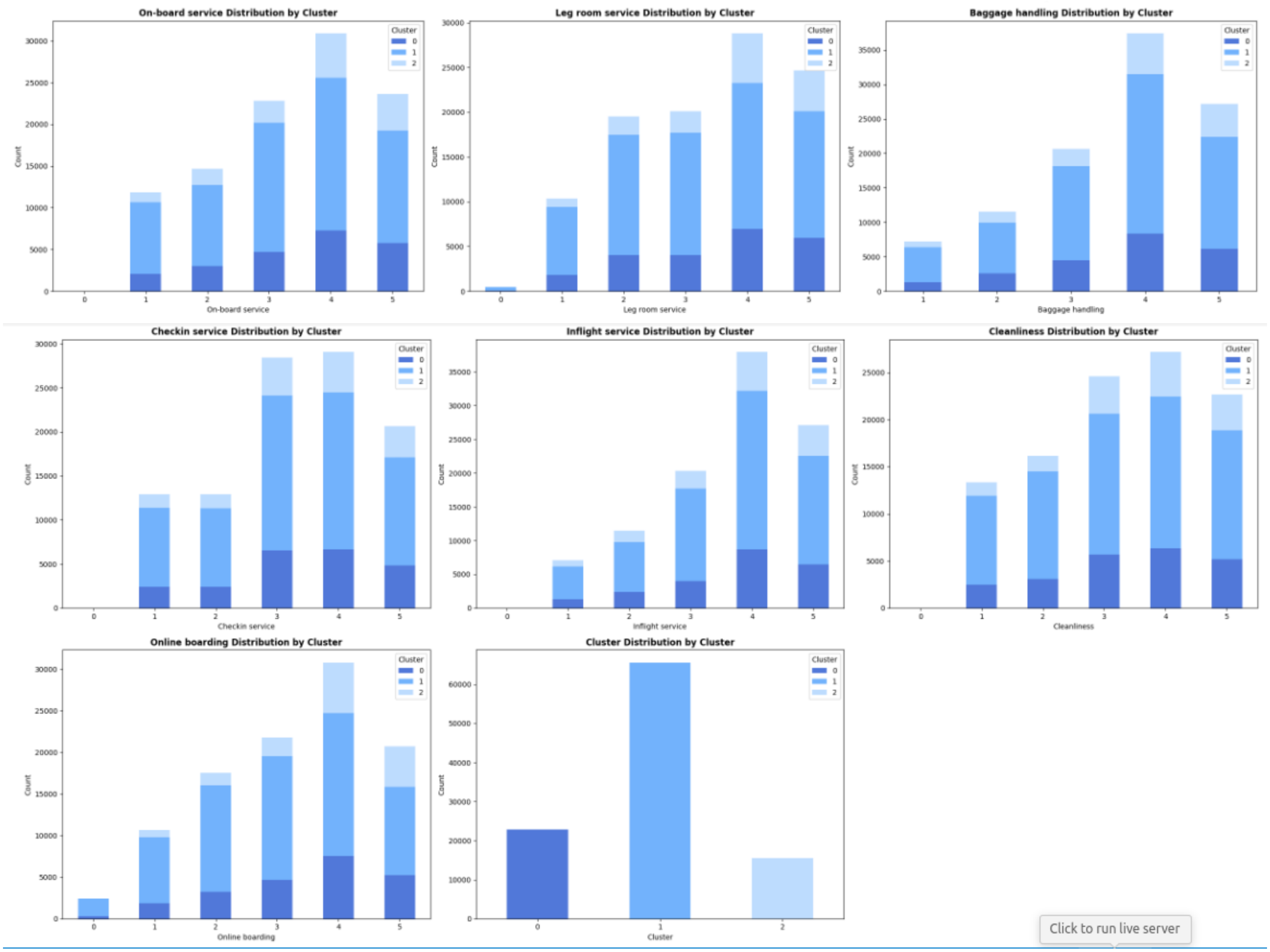
Class Distribution by Cluster

Gate location Distribution by Cluster

Food and drink Distribution by Cluster

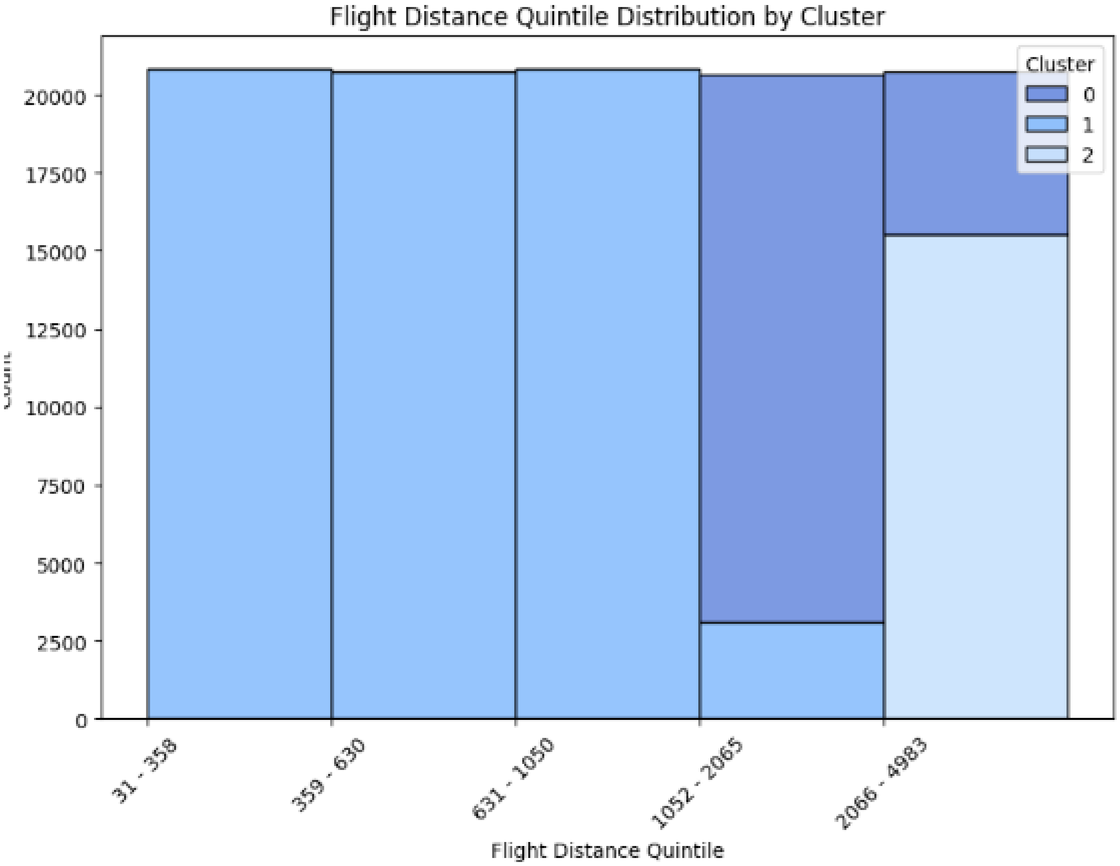
Seat comfort Distribution by Cluster

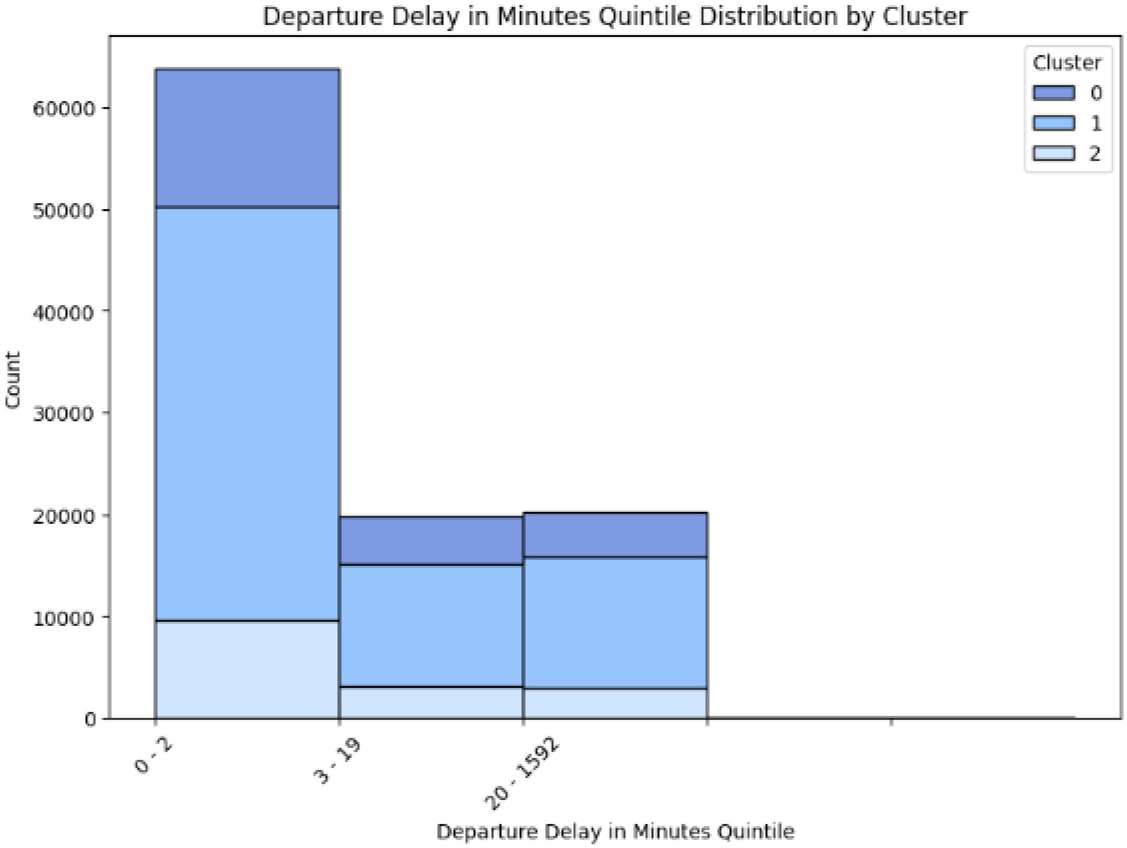
Inflight entertainment Distribution by Cluster

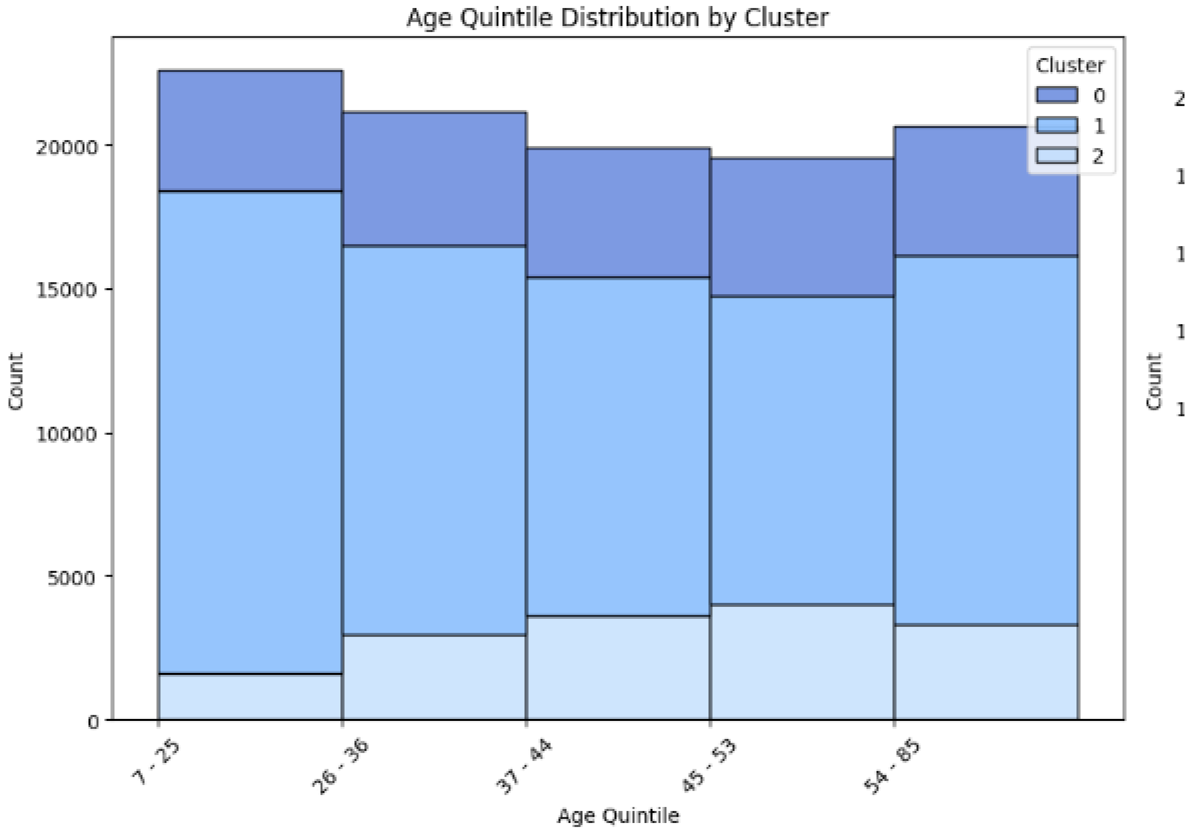


- **Insights:**
  - Cluster 1 is the major class in all values in all columns
  - Satisfied customers are distributed over all cluster
  - Dissatisfied or neutral customers are distributed over all cluster
  - Each gender is distributed over all clusters
  - Loyal customer is distributed over all clusters
  - Cluster 2 doesn't contain Disloyal customers
  - The portion of customers with type of travel is personal in cluster 2 is very small
  - All values of [Inflight WiFi service, departure arrival time convenient, Ease of online booking, Gate Location, Food and drink,
  - seat comfort, inflight entertainment, on-board service, Baggage handling, Checkin service, inflight service & cleanliness] are distributed over all clusters
  - The portion of departure arrival time convenient with value 0 in cluster 2 is very small
  - Cluster 2 doesn't contain customers of Eco plus class
  - The portion of of customers of Eco class in cluster 2 is very small
  - Cluster 2 doesn't contain values 0 of Online boarding

- Values 0 of Leg room service are all in cluster 1


Flight Distance Quintile Distribution by Cluster
Count
0
2500
5000
7500
10000
12500
15000
17500
20000
31 - 358
359 - 630
631 - 1050
1052 - 2065
2066 - 4983
Flight Distance Quintile
Cluster
0
1
2


Departure Delay in Minutes Quintile Distribution by Cluster
Count
0
10000
20000
30000
40000
50000
60000
0 - 2
3 - 19
20 - 1592
Departure Delay in Minutes Quintile
Cluster
0
1
2

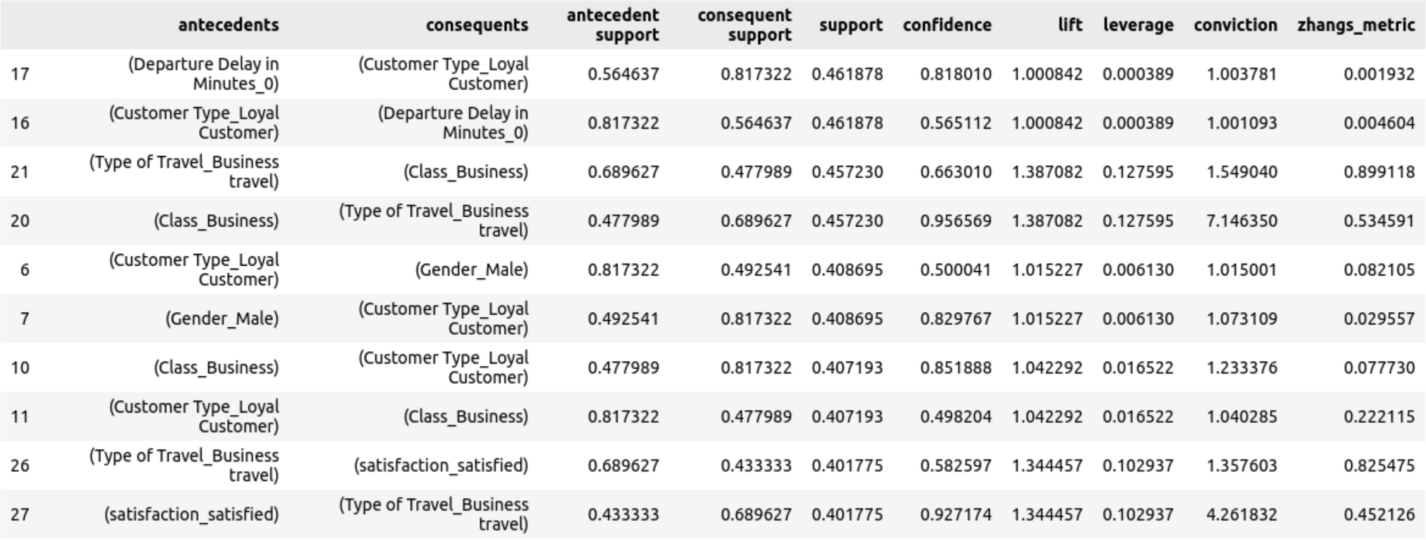


- **Insights:**
  - Cluster 1 is the major class in all values in all columns
  - Satisfied customers are distributed over all cluster
  - Dissatisfied or neutral customers are distributed over all cluster
  - Each gender is distributed over all clusters
  - Loyal customer is distributed over all clusters
  - Cluster 2 doesn't contain Disloyal customers
  - The portion of customers with type of travel is personal in cluster 2 is very small
  - All values of [Inflight WiFi service, departure arrival time convenient, Ease of online booking, Gate Location, Food and drink,
  - seat comfort, inflight entertainment, on-board service, Baggage handling, Checkin service, inflight service & cleanliness] are distributed over all clusters
  - The portion of departure arrival time convenient with value 0 in cluster 2 is very small
  - Cluster 2 doesn't contain customers of Eco plus class
  - The portion of of customers of Eco class in cluster 2 is very small
  - Cluster 2 doesn't contain values 0 of Online boarding
  - Values 0 of Leg room service are all in cluster 1

- *Association Rules*

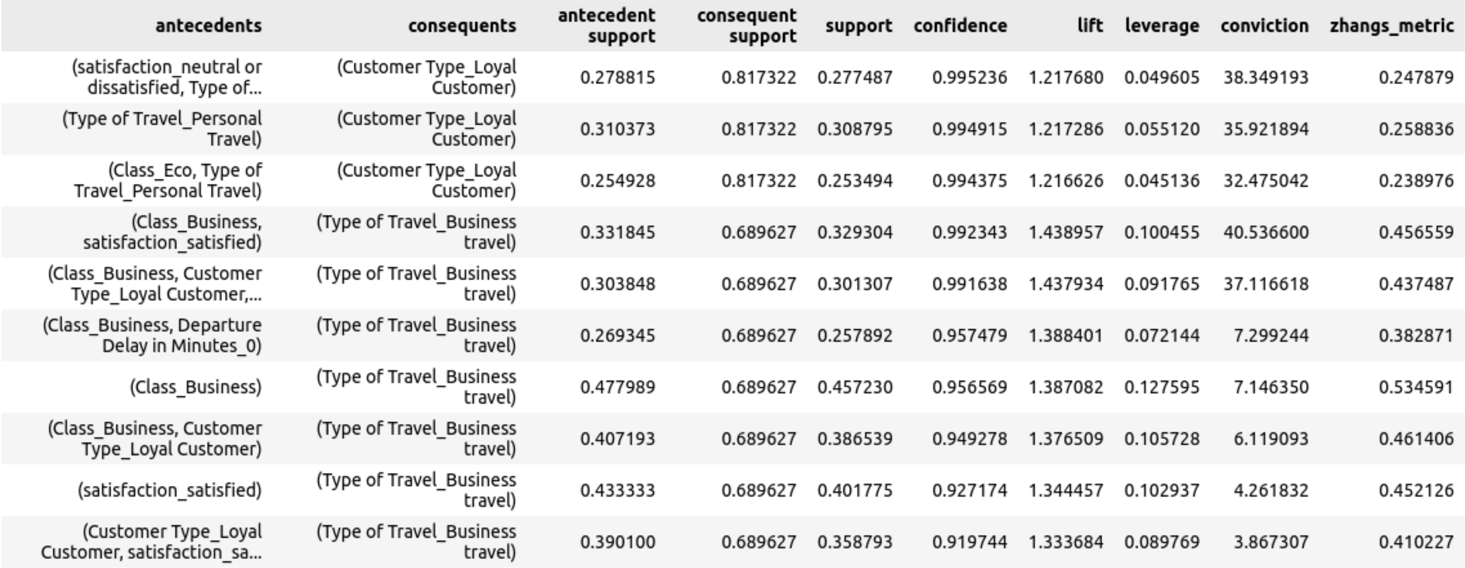
Applying Apriori algorithm with minimum support = 0.25 and minimum cardinality = 2

- Top 10 rules sorted by support:



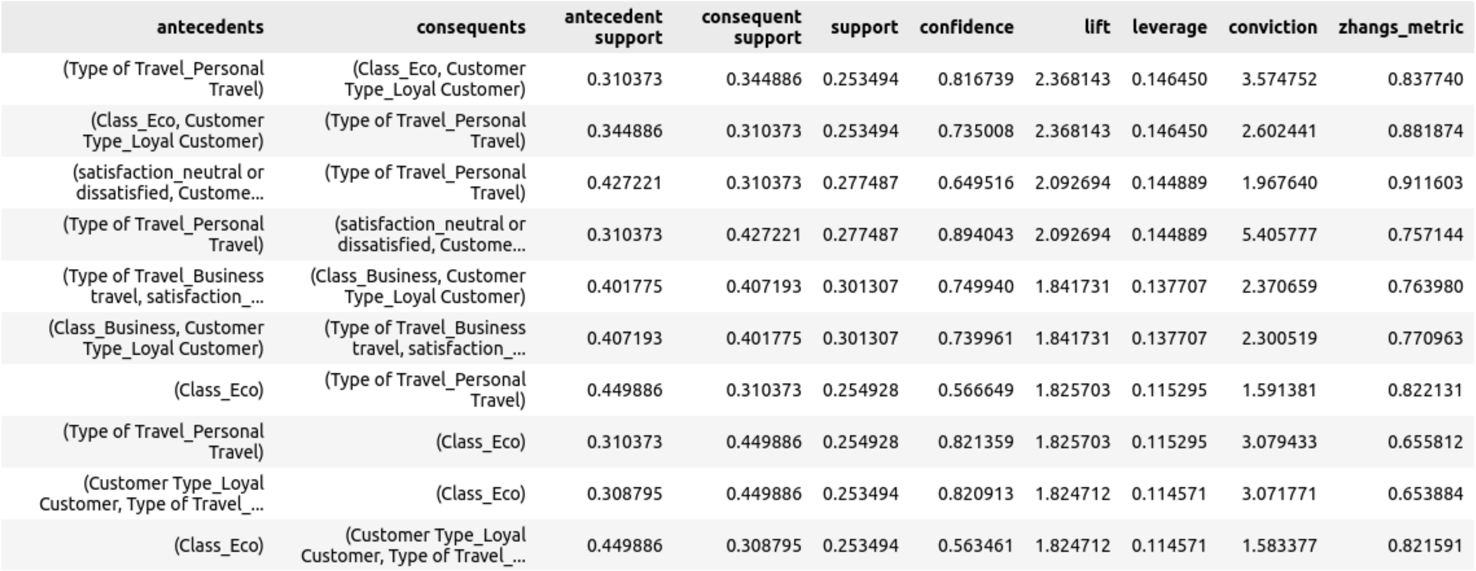
	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
17	(Departure Delay in Minutes_0)	(Customer Type_Loyal Customer)	0.564637	0.817322	0.461878	0.818010	1.000842	0.000389	1.003781	0.001932
16	(Customer Type_Loyal Customer)	(Departure Delay in Minutes_0)	0.817322	0.564637	0.461878	0.565112	1.000842	0.000389	1.001093	0.004604
21	(Type of Travel_Business travel)	(Class_Business)	0.689627	0.477989	0.457230	0.663010	1.387082	0.127595	1.549040	0.899118
20	(Class_Business)	(Type of Travel_Business travel)	0.477989	0.689627	0.457230	0.956569	1.387082	0.127595	7.146350	0.534591
6	(Customer Type_Loyal Customer)	(Gender_Male)	0.817322	0.492541	0.408695	0.500041	1.015227	0.006130	1.015001	0.082105
7	(Gender_Male)	(Customer Type_Loyal Customer)	0.492541	0.817322	0.408695	0.829767	1.015227	0.006130	1.073109	0.029557
10	(Class_Business)	(Customer Type_Loyal Customer)	0.477989	0.817322	0.407193	0.851888	1.042292	0.016522	1.233376	0.077730
11	(Customer Type_Loyal Customer)	(Class_Business)	0.817322	0.477989	0.407193	0.498204	1.042292	0.016522	1.040285	0.222115
26	(Type of Travel_Business travel)	(satisfaction_satisfied)	0.689627	0.433333	0.401775	0.582597	1.344457	0.102937	1.357603	0.825475
27	(satisfaction_satisfied)	(Type of Travel_Business travel)	0.433333	0.689627	0.401775	0.927174	1.344457	0.102937	4.261832	0.452126

- Top10 rules sorted by confidence:



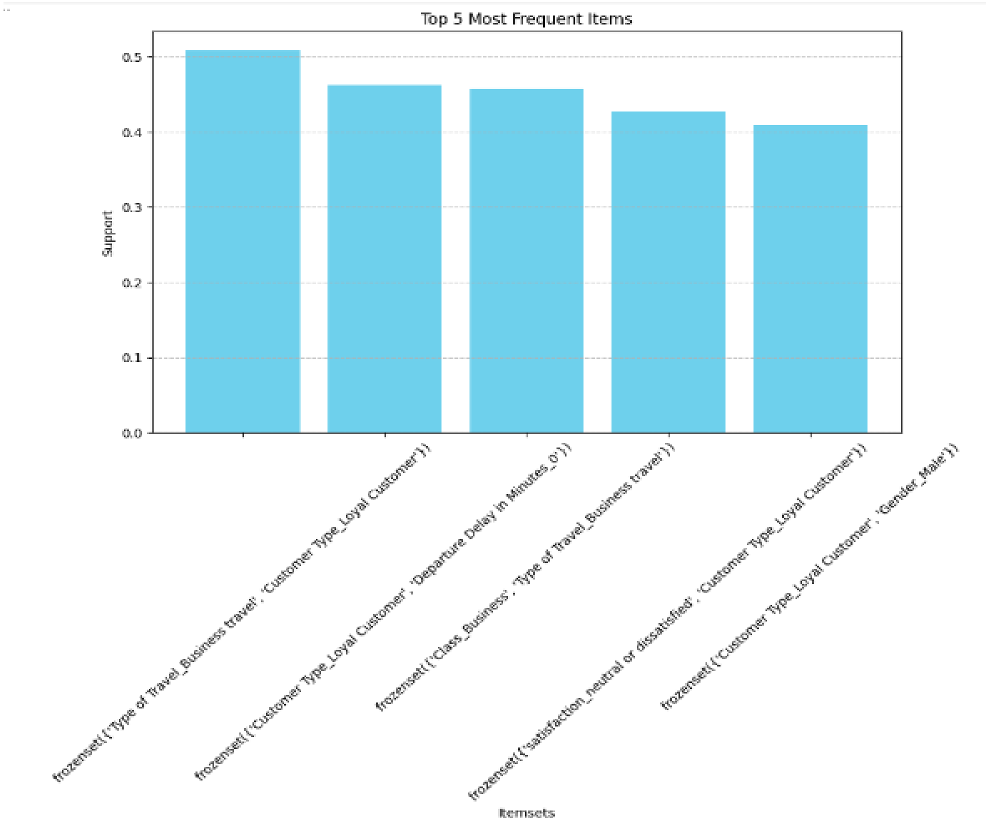
antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
(satisfaction_neutral or dissatisfied, Type of...	(Customer Type_Loyal Customer)	0.278815	0.817322	0.277487	0.995236	1.217680	0.049605	38.349193	0.247879
(Type of Travel_Personal Travel)	(Customer Type_Loyal Customer)	0.310373	0.817322	0.308795	0.994915	1.217286	0.055120	35.921894	0.258836
(Class_Eco, Type of Travel_Personal Travel)	(Customer Type_Loyal Customer)	0.254928	0.817322	0.253494	0.994375	1.216626	0.045136	32.475042	0.238976
(Class_Business, satisfaction_satisfied)	(Type of Travel_Business travel)	0.331845	0.689627	0.329304	0.992343	1.438957	0.100455	40.536600	0.456559
(Class_Business, Customer Type_Loyal Customer,...	(Type of Travel_Business travel)	0.303848	0.689627	0.301307	0.991638	1.437934	0.091765	37.116618	0.437487
(Class_Business, Departure Delay in Minutes_0)	(Type of Travel_Business travel)	0.269345	0.689627	0.257892	0.957479	1.388401	0.072144	7.299244	0.382871
(Class_Business)	(Type of Travel_Business travel)	0.477989	0.689627	0.457230	0.956569	1.387082	0.127595	7.146350	0.534591
(Class_Business, Customer Type_Loyal Customer)	(Type of Travel_Business travel)	0.407193	0.689627	0.386539	0.949278	1.376509	0.105728	6.119093	0.461406
(satisfaction_satisfied)	(Type of Travel_Business travel)	0.433333	0.689627	0.401775	0.927174	1.344457	0.102937	4.261832	0.452126
(Customer Type_Loyal Customer, satisfaction_sa...	(Type of Travel_Business travel)	0.390100	0.689627	0.358793	0.919744	1.333684	0.089769	3.867307	0.410227

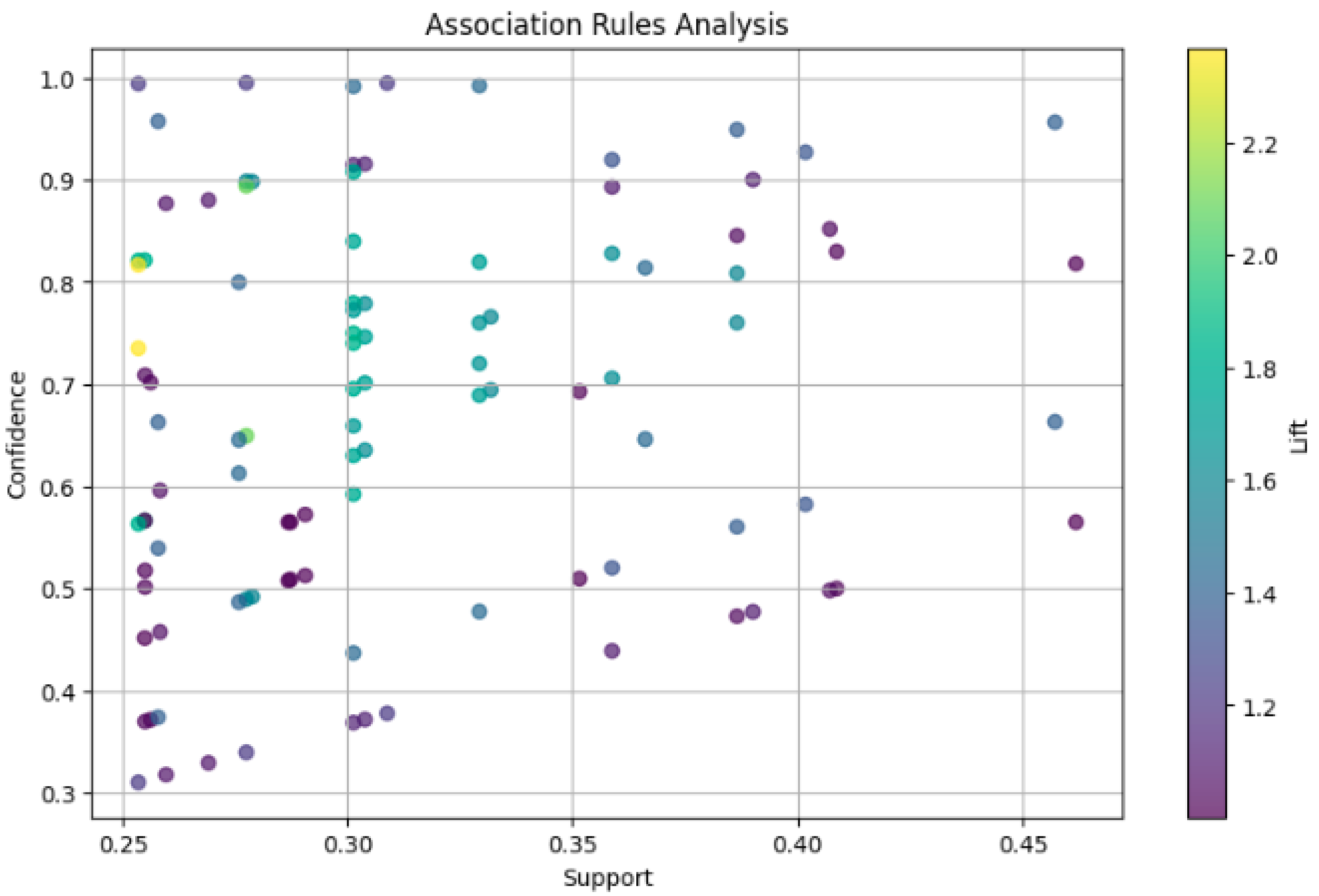
- Top 10 rules sorted by lift:



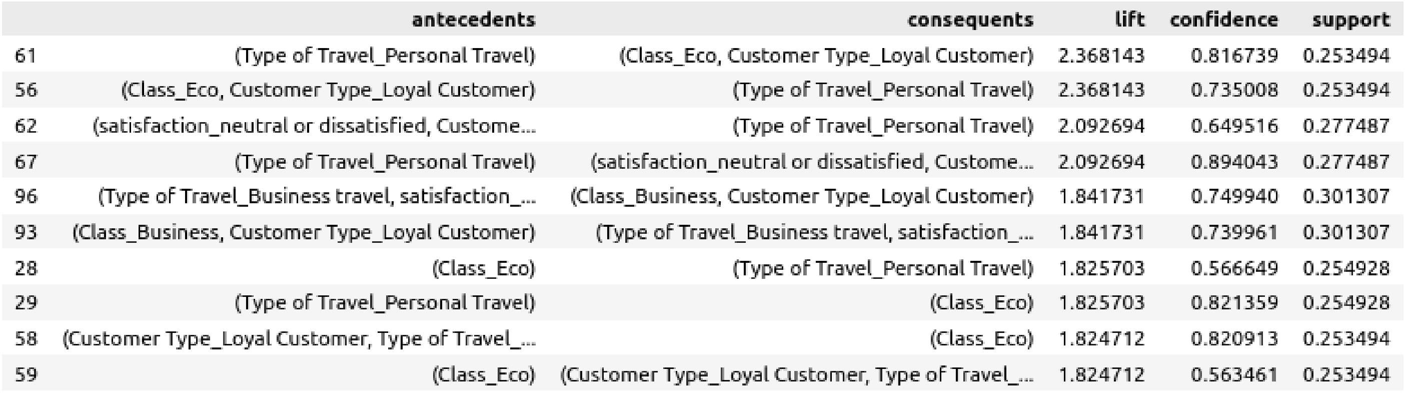
antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
(Type of Travel_Personal Travel)	(Class_Eco, Customer Type_Loyal Customer)	0.310373	0.344886	0.253494	0.816739	2.368143	0.146450	3.574752	0.837740
(Class_Eco, Customer Type_Loyal Customer)	(Type of Travel_Personal Travel)	0.344886	0.310373	0.253494	0.735008	2.368143	0.146450	2.602441	0.881874
(satisfaction_neutral or dissatisfied, Custome...	(Type of Travel_Personal Travel)	0.427221	0.310373	0.277487	0.649516	2.092694	0.144889	1.967640	0.911603
(Type of Travel_Personal Travel)	(satisfaction_neutral or dissatisfied, Custome...	0.310373	0.427221	0.277487	0.894043	2.092694	0.144889	5.405777	0.757144
(Type of Travel_Business travel, satisfaction_...	(Class_Business, Customer Type_Loyal Customer)	0.401775	0.407193	0.301307	0.749940	1.841731	0.137707	2.370659	0.763980
(Class_Business, Customer Type_Loyal Customer)	(Type of Travel_Business travel, satisfaction_...	0.407193	0.401775	0.301307	0.739961	1.841731	0.137707	2.300519	0.770963
(Class_Eco)	(Type of Travel_Personal Travel)	0.449886	0.310373	0.254928	0.566649	1.825703	0.115295	1.591381	0.822131
(Type of Travel_Personal Travel)	(Class_Eco)	0.310373	0.449886	0.254928	0.821359	1.825703	0.115295	3.079433	0.655812
(Customer Type_Loyal Customer, Type of Travel_...	(Class_Eco)	0.308795	0.449886	0.253494	0.820913	1.824712	0.114571	3.071771	0.653884
(Class_Eco)	(Customer Type_Loyal Customer, Type of Travel_...	0.449886	0.308795	0.253494	0.563461	1.824712	0.114571	1.583377	0.821591

- Most frequent 2-itemset: {'Type of Travel_Business travel', 'Customer Type_Loyal Customer'}
  - Frequency: 0.5085
- Top 5 frequent items:





- Top 10 rules sorted by lift:



	antecedents	consequents	lift	confidence	support
61	(Type of Travel_Personal Travel)	(Class_Eco, Customer Type_Loyal Customer)	2.368143	0.816739	0.253494
56	(Class_Eco, Customer Type_Loyal Customer)	(Type of Travel_Personal Travel)	2.368143	0.735008	0.253494
62	(satisfaction_neutral or dissatisfied, Custome...	(Type of Travel_Personal Travel)	2.092694	0.649516	0.277487
67	(Type of Travel_Personal Travel)	(satisfaction_neutral or dissatisfied, Custome...	2.092694	0.894043	0.277487
96	(Type of Travel_Business travel, satisfaction_...	(Class_Business, Customer Type_Loyal Customer)	1.841731	0.749940	0.301307
93	(Class_Business, Customer Type_Loyal Customer)	(Type of Travel_Business travel, satisfaction_...	1.841731	0.739961	0.301307
28	(Class_Eco)	(Type of Travel_Personal Travel)	1.825703	0.566649	0.254928
29	(Type of Travel_Personal Travel)	(Class_Eco)	1.825703	0.821359	0.254928
58	(Customer Type_Loyal Customer, Type of Travel_...	(Class_Eco)	1.824712	0.820913	0.253494
59	(Class_Eco)	(Customer Type_Loyal Customer, Type of Travel_...	1.824712	0.563461	0.253494

- The most interesting rules that are likely to provide real business value and insights are those with high lift values.
- Lift measures how much more likely the consequent (rhs) is, given the antecedent (lhs), compared to if the two were independent.

- **# Looking at the rules sorted by lift:**

- {Type of Travel_Personal Travel} ==>
  {Class_Eco, Customer Type_Loyal Customer}
  with Lift: 2.3681

confidence: 0.8167
support: 0.2535

- {Class_Eco, Customer Type_Loyal Customer}
  ==> {Type of Travel_Personal Travel}
  with lift = 2.3681
  confidence: 0.7350
  support: 0.2535

- {satisfaction_neutral or dissatisfied, Customer Type_Loyal Customer}
  ==> {Type of Travel_Personal Travel}
  with lift = 2.0927
  confidence: 0.6495
  support: 0.2775

- {Type of Travel_Personal Travel}
  ==> {satisfaction_neutral or dissatisfied, Customer Type_Loyal Customer}
  with lift = 2.0927
  confidence: 0.8940
  support: 0.2775

- {Type of Travel_Business travel, satisfaction_satisfied}
  ==> {Class_Business, Customer Type_Loyal Customer}
  with lift = 1.8417
  confidence: 0.7499
  support: 0.3013

- {Class_Business, Customer Type_Loyal Customer}
  ==> {Type of Travel_Business travel, satisfaction_satisfied}
  with lift = 1.8417
  confidence: 0.7400
  support: 0.3013

- {Class_Eco}
  ==> {Type of Travel_Personal Travel}
  with lift = 1.8257
  confidence: 0.5666
  support: 0.2549

- {Type of Travel_Personal Travel}
  ==> {Class_Eco}
  with lift = 1.8257

confidence: 0.8214
support: 0.2549

- {Customer Type_Loyal Customer, Type of Travel_Personal Travel}
  ==> {Class_Eco}
  with lift = 1.8247
  confidence: 0.8209
  support: 0.2535

- {Class_Eco}
  ==> {Customer Type_Loyal Customer, Type of Travel_Personal Travel}
  with lift = 1.8247
  confidence: 0.5635
  support: 0.2535

- **Insights:**
- If a customer's type of travel is "Personal Travel", then there is a strong association with the customer being classified as "Eco" class and a "Loyal Customer".
  - The lift value of 2.3681 indicates that the occurrence of the antecedent and consequent together is 2.3681 times more likely than if they were statistically independent.
  - This means that customers who travel for personal reasons are 2.3681 times more likely to be classified as "Eco" class and "Loyal Customers" compared to what would be expected if these attributes were unrelated.
  - The confidence value of 0.8167 indicates that 81.67% of the transactions that contain "Personal Travel" also contain "Eco" class and "Loyal Customer".
  - The support value of 0.2535 indicates that 25.35% of the transactions contain both "Personal Travel" and "Eco" class and "Loyal Customer".
- If a customer is classified as "Eco" class and is a "Loyal Customer", then there is a strong association with their type of travel being "Personal Travel".
  - The lift value of 2.3681 indicates that the occurrence of the consequent given the antecedent is 2.3681 times more likely than if they were statistically independent.
  - This means that customers who are classified as "Eco" class and "Loyal Customers" are 2.3681 times more likely to travel for personal reasons compared to what would be expected if these attributes were unrelated.

    - The confidence value of 0.7350 indicates that 73.50% of the transactions that contain "Eco" class and "Loyal Customer" also contain "Personal Travel".
    - The support value of 0.2535 indicates that 25.35% of the transactions contain both "Eco" class and "Loyal Customer", and "Personal Travel".
- If a customer is classified as a "Loyal Customer" and their satisfaction level is "neutral or dissatisfied", then there is a strong association with their type of travel being "Personal Travel".
    - The lift value of 2.0927 indicates that the occurrence of the consequent given the antecedent is 2.0927 times more likely than if they were statistically independent.
    - This means that if a customer is classified as a "Loyal Customer" and their satisfaction level is "neutral or dissatisfied", there is 2.0927 times more likely that their type of travel will be "Personal Travel" compared to what would be expected if these attributes were unrelated.
    - The confidence value of 0.6495 indicates that 64.95% of the transactions that contain "Loyal Customer" with a satisfaction level of "neutral or dissatisfied" also contain "Personal Travel".
    - The support value of 0.2775 indicates that 27.75% of the transactions contain both "Loyal Customer" with a satisfaction level of "neutral or dissatisfied", and "Personal Travel".
- If a customer's type of travel is "Personal Travel", then there is a strong association with the customer being classified as a "Loyal Customer" and having a satisfaction level of "neutral or dissatisfied".
    - The lift value of 2.0927 indicates that the occurrence of the consequent given the antecedent is 2.0927 times more likely than if they were statistically independent.
    - This means that if a customer's type of travel is "Personal Travel", there is a higher likelihood that the customer will be classified as a "Loyal Customer" and have a satisfaction level of "neutral or dissatisfied" compared to what would be expected if these attributes were unrelated.
    - The confidence value of 0.8940 indicates that 89.40% of the transactions that contain "Personal Travel" also contain "Loyal Customer" with a satisfaction level of "neutral or dissatisfied".
    - The support value of 0.2775 indicates that 27.75% of the transactions contain both "Personal Travel" and "Loyal Customer" with a satisfaction level of "neutral or dissatisfied".

- If a customer's type of travel is "Business travel" and their satisfaction level is "satisfied", then there is a moderate association with the customer being classified as "Business" class and a "Loyal Customer".
    - The lift value of 1.8417 indicates that the occurrence of the consequent given the antecedent is 1.8417 times more likely than if they were statistically independent.
    - The confidence value of 0.7499 indicates that 74.99% of the transactions that contain "Business travel" with a satisfaction level of "satisfied" also contain "Business" class and "Loyal Customer".
    - The support value of 0.3013 indicates that 30.13% of the transactions contain both "Business travel" with a satisfaction level of "satisfied", and "Business" class and "Loyal Customer".
- If a customer is classified as "Business" class and is a "Loyal Customer", then there is a moderate association with their type of travel being "Business travel" and their satisfaction level being "satisfied".
    - The lift value of 1.8417 indicates that the occurrence of the consequent given the antecedent is 1.8417 times more likely than if they were statistically independent.
    - The confidence value of 0.7400 indicates that 74.00% of the transactions that contain "Business" class and "Loyal Customer" also contain "Business travel" with a satisfaction level of "satisfied".
    - The support value of 0.3013 indicates that 30.13% of the transactions contain both "Business" class and "Loyal Customer", and "Business travel" with a satisfaction level of "satisfied".
- If a customer is classified as "Eco" class, then there is a moderate association with their type of travel being "Personal Travel".
    - The lift value of 1.8257 indicates that the occurrence of the consequent given the antecedent is 1.8257 times more likely than if they were statistically independent.
    - The confidence value of 0.5666 indicates that 56.66% of the transactions that contain "Eco" class also contain "Personal Travel".
    - The support value of 0.2549 indicates that 25.49% of the transactions contain both "Eco" class and "Personal Travel".
- If a customer's type of travel is "Personal Travel", then there is a strong association with the customer being classified as "Eco" class.

  - The lift value of 1.8257 indicates that the occurrence of the consequent given the antecedent is 1.8257 times more likely than if they were statistically independent.
  - The confidence value of 0.8214 indicates that 82.14% of the transactions that contain "Personal Travel" also contain "Eco" class.
  - The support value of 0.2549 indicates that 25.49% of the transactions contain both "Personal Travel" and "Eco" class.
- If a customer is classified as a "Loyal Customer" and their type of travel is "Personal Travel", then there is a strong association with the customer being classified as "Eco" class.
  - The lift value of 1.8247 indicates that the occurrence of the consequent given the antecedent is 1.8247 times more likely than if they were statistically independent.
  - The confidence value of 0.8209 indicates that 82.09% of the transactions that contain both "Loyal Customer" and "Personal Travel" also contain "Eco" class.
  - The support value of 0.2535 indicates that 25.35% of the transactions contain both "Loyal Customer" and "Personal Travel", and "Eco" class.
- If a customer is classified as "Eco" class, then there is a moderate association with the customer being classified as a "Loyal Customer" and their type of travel being "Personal Travel".
  - The lift value of 1.8247 indicates that the occurrence of the consequent given the antecedent is 1.8247 times more likely than if they were statistically independent.
  - The confidence value of 0.5635 indicates that 56.35% of the transactions that contain "Eco" class also contain both "Loyal Customer" and "Personal Travel".
  - The support value of 0.2535 indicates that 25.35% of the transactions contain "Eco" class, "Loyal Customer", and "Personal Travel".

(3) Preprocessing

1- Encode categorical variables.

2- Drop Arrival delay in minutes column
2- Drop unnecessary columns (columns that don’t affect satisfaction)

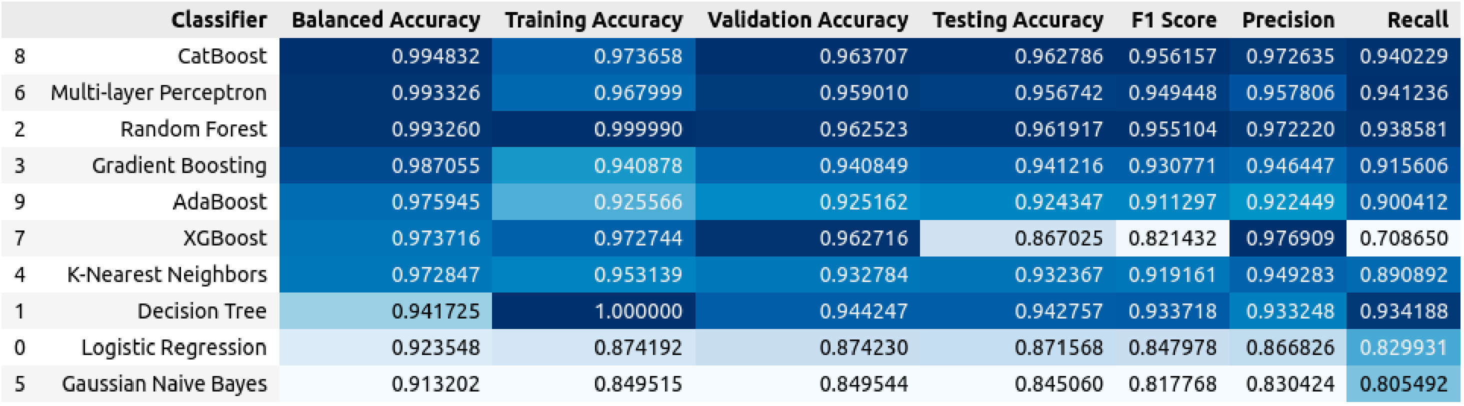
['Gender','Gate location','Departure/Arrival time convenient']

4- Apply grouping on features with continuous variables

5- Standardization: scaling features by subtracting the mean and then dividing by the standard deviation.

This results in features that have a mean of 0 and a standard deviation of 1.

(4) Model Building, Results and Evaluation:

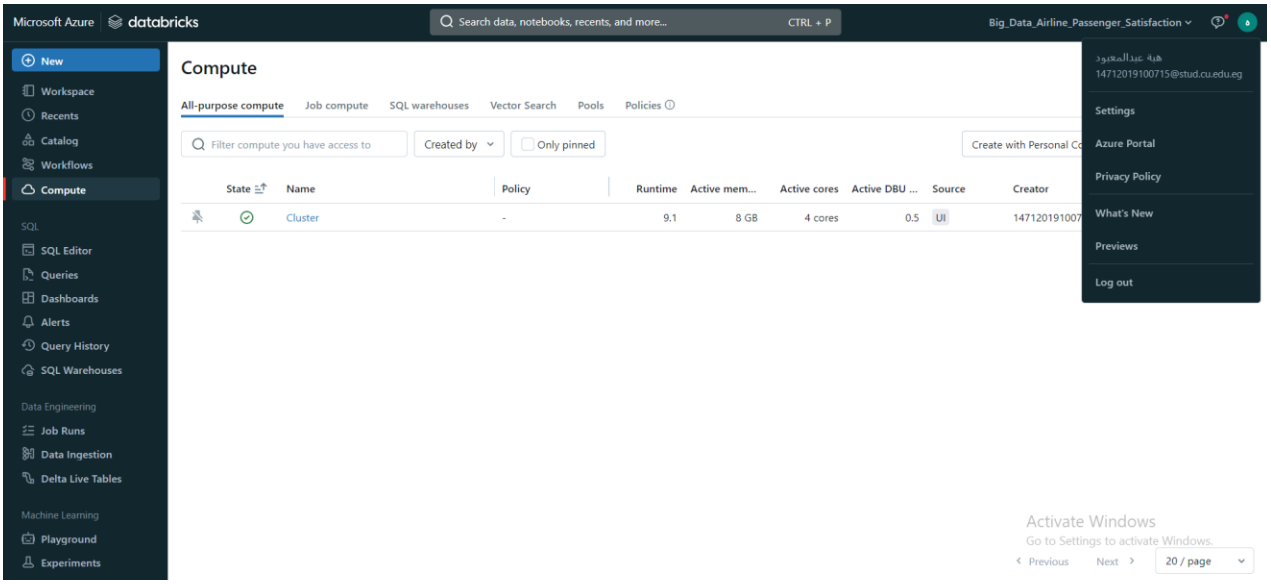


	Classifier	Balanced Accuracy	Training Accuracy	Validation Accuracy	Testing Accuracy	F1 Score	Precision	Recall
8	CatBoost	0.994832	0.973658	0.963707	0.962786	0.956157	0.972635	0.940229
6	Multi-layer Perceptron	0.993326	0.967999	0.959010	0.956742	0.949448	0.957806	0.941236
2	Random Forest	0.993260	0.999990	0.962523	0.961917	0.955104	0.972220	0.938581
3	Gradient Boosting	0.987055	0.940878	0.940849	0.941216	0.930771	0.946447	0.915606
9	AdaBoost	0.975945	0.925566	0.925162	0.924347	0.911297	0.922449	0.900412
7	XGBoost	0.973716	0.972744	0.962716	0.867025	0.821432	0.976909	0.708650
4	K-Nearest Neighbors	0.972847	0.953139	0.932784	0.932367	0.919161	0.949283	0.890892
1	Decision Tree	0.941725	1.000000	0.944247	0.942757	0.933718	0.933248	0.934188
0	Logistic Regression	0.923548	0.874192	0.874230	0.871568	0.847978	0.866826	0.829931
5	Gaussian Naive Bayes	0.913202	0.849515	0.849544	0.845060	0.817768	0.830424	0.805492

- **CatBoost** achieved the highest Testing accuracy, Training accuracy, F1 Score & Balanced Accuracy
- Multi Nominal Naive Bayes without applying grouping on features with continuous variables from sklearn:
  - Balanced Accuracy: 0.8741036230929793
  - Training Accuracy: 0.7680455035417308

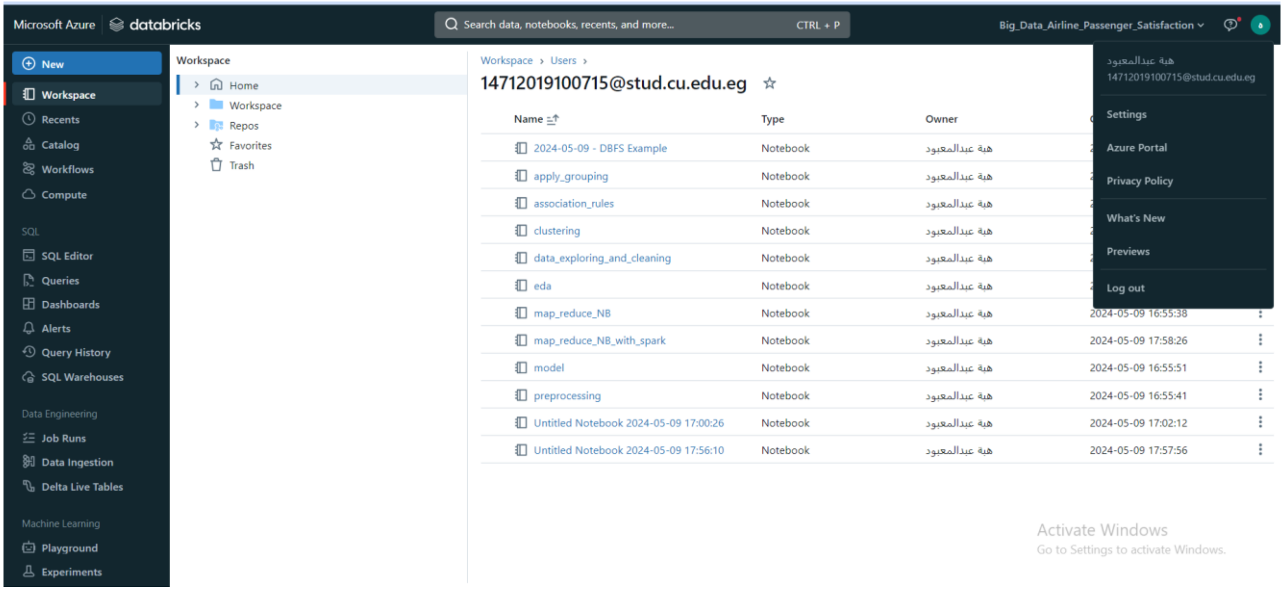
  - Testing Accuracy: 0.7649034093153716
  - Validation Accuracy: 0.7680455332217699
  - F1 Score: 0.7330791657322269
  - Precision: 0.7187335092348285
  - Recall: 0.7480091533180778
- Multi Nominal Naive Bayes with applying grouping on features with continuous variables from sklearn:
  - Balanced Accuracy: 0.8660401815904305
  - Training Accuracy: 0.7687865722205113
  - Testing Accuracy: 0.7656145063801209
  - Validation Accuracy: 0.7687288538824919
  - F1 Score: 0.7336236699142458
  - Precision: 0.7199506520972858
  - Recall: 0.7478260869565218
- Multi Nominal Naive Bayes without applying grouping on features with continuous variables from scratch using map reduce:
  - Balanced Accuracy: 0.7650422898817919
  - Training Accuracy: 0.8885220973206036
  - Testing Accuracy: 0.7626516019436653
  - f1_score: 0.7399809573271012
  - precision: 0.7018307199737296
  - recall: 0.7825171624713959
- Multi Nominal Naive Bayes with applying grouping on features with continuous variables from scratch using map reduce:
  - Balanced Accuracy: 0.7611859775085901
  - Training Accuracy: 0.9015629812134278
  - Testing Accuracy: 0.7623750641962628
  - f1_score: 0.7321547846996482
  - precision: 0.7128858827610128
  - recall: 0.7524942791762014

Azure Screenshots:



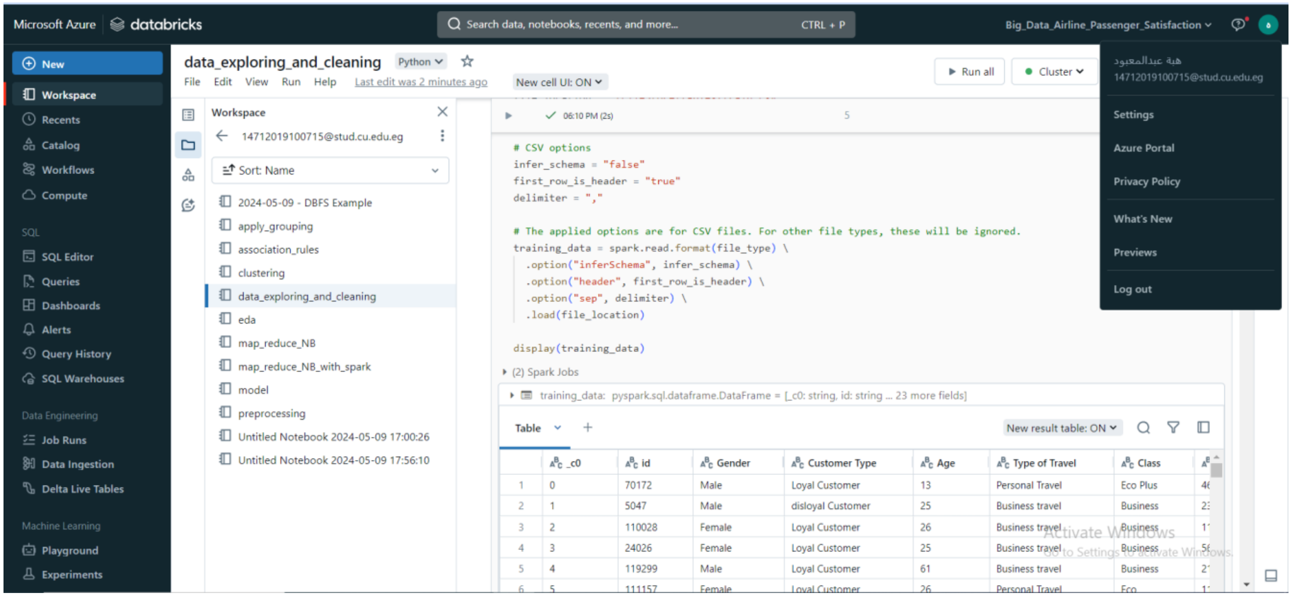
Compute

State	Name	Policy	Runtime	Active mem...	Active cores	Active DBU ...	Source	Creator
✓	Cluster	-	9.1	8 GB	4 cores	0.5	UI	147120191007



14712019100715@stud.cu.edu.eg ☆

Name	Type	Owner	C
2024-05-09 - DBFS Example	Notebook	هبة عبدالمعبود	2
apply_grouping	Notebook	هبة عبدالمعبود	2
association_rules	Notebook	هبة عبدالمعبود	2
clustering	Notebook	هبة عبدالمعبود	2
data_exploring_and_cleaning	Notebook	هبة عبدالمعبود	2
eda	Notebook	هبة عبدالمعبود	2
map_reduce_NB	Notebook	هبة عبدالمعبود	2024-05-09 16:55:38
map_reduce_NB_with_spark	Notebook	هبة عبدالمعبود	2024-05-09 17:58:26
model	Notebook	هبة عبدالمعبود	2024-05-09 16:55:51
preprocessing	Notebook	هبة عبدالمعبود	2024-05-09 16:55:41
Untitled Notebook 2024-05-09 17:00:26	Notebook	هبة عبدالمعبود	2024-05-09 17:02:12
Untitled Notebook 2024-05-09 17:56:10	Notebook	هبة عبدالمعبود	2024-05-09 17:57:56



```
# CSV options
infer_schema = "false"
first_row_is_header = "true"
delimiter = ","

# The applied options are for CSV files. For other file types, these will be ignored.
training_data = spark.read.format(file_type) \
  .option("inferSchema", infer_schema) \
  .option("header", first_row_is_header) \
  .option("sep", delimiter) \
  .load(file_location)

display(training_data)
```

training_data: pyspark.sql.dataframe.DataFrame = [_c0: string, id: string ... 23 more fields]

	_c0	id	Gender	Customer Type	Age	Type of Travel	Class
1	0	70172	Male	Loyal Customer	13	Personal Travel	Eco Plus
2	1	5047	Male	disloyal Customer	25	Business travel	Business
3	2	110028	Female	Loyal Customer	26	Business travel	Business
4	3	24026	Female	Loyal Customer	25	Business travel	Business
5	4	119299	Male	Loyal Customer	61	Business travel	Business
6	5	111157	Female	Loyal Customer	26	Personal Travel	Eco

Project Structure:

- data:
  - train.csv ==> original train data
  - test.csv ==> original test data
  - cleaned_train_data ==> train data after data exploratory and cleaning phase
  - cleaned_test_data ==> test data after data exploratory and cleaning phase
  - train_data_after_eda ==> train data after EDA phase
  - test_data_after_eda ==> test data after EDA phase
  - preprocessed_train_data ==> train data after preprocessing phase
  - preprocessed_test_data ==> test data after preprocessing phase
  - preprocessed_train_data_after_grouping ==> train data after preprocessing phase and applying grouping on features with continuous variables
  - preprocessed_test_data _after_grouping ==> test data after preprocessing phase and applying grouping on features with continuous variables
- code:
  - data_exploring_and_cleaning.ipynb ==> data exploratory and cleaning phase
  - eda.ipynb ==> Exploratory data analysis phase
  - clustering.ipynb ==> Applying clustering on the data and extracting insights from it
  - assocciation_rules.ipynb ==> Applying Apriori algorithm on the data and extracting insights from it
  - preprocessing.ipynb ==> Preprocessing phase
  - apply_grouping.ipynb ==> Apply grouping on the features with continuous values

    - model.ipynb ==> Training and evaluating different models
    - map_reduce_NB.ipynb ==> Multinominal Naive Bayes from scratch using map reduce
    - map_reduce_NB.ipynb ==> Multinominal Naive Bayes from scratch using map reduce with spark
    - map_reduce_NB_with_grouping.ipynb ==> Multinominal Naive Bayes from scratch using map reduce after applying grouping on the features with continuous values
- documents:
    - Project proposal
    - Project document
    - Report
    - Presentation

Unsuccessful trials:

Implementing Naive Bayes with map reduce without spark worked well but with spark we calculated classes prior probabilities and applied training but encountered an error while applying predictions so we couldn’t create the metrics

Enhancements and future work:

- Apply different clustering algorithms
- Apply map reduce with spark on different classification and clustering algorithms
- Apply grouping on features with continuous values with different number of groups