# Big Data project

## Team: 6


| Name | ID | SEC | BN |
|---|---|---|---|
| Asmaa Adel Abdelhamed kawashty | 9202285 | 1 | 13 |
| Samaa Hazem Mohamed Abdel-latif | 9202660 | 1 | 31 |
| Norhan Reda Abdelwahed Ahmed | 9203639 | 2 | 31 |
| Hoda Gamal Hamouda Ismail | 9203673 | 2 | 33 |

**Supervisor: Eng. Omar Samir**

# 1. Motivation

- Airline companies strive to provide the best possible travel experience for their passengers to maintain customer satisfaction and loyalty

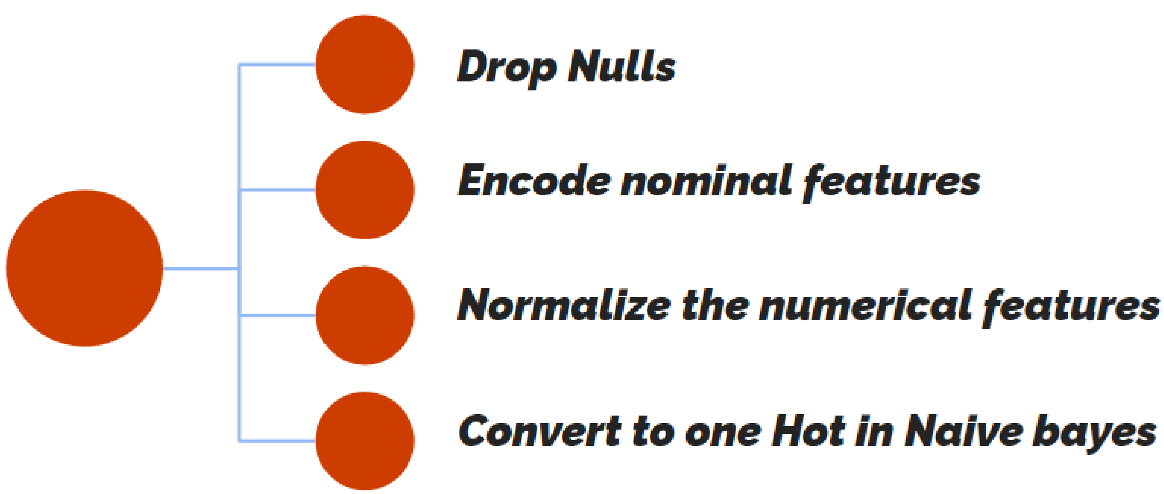
- Understanding the factors that contribute to passenger satisfaction is crucial for airlines to improve their services, enhance customer experience, and stay competitive in the industry

- By accurately predicting passenger satisfaction, airlines can identify areas for improvement and tailor their services to meet customer expectations more effectively

- Additionally, satisfied passengers are more likely to become repeat customers and recommend the airline to others, leading to increased revenue and growth

- The problem can be framed as a binary classification task, where the goal is to predict whether a passenger is satisfied or dissatisfied based on the input features.

# 2. Data Preprocessing



# 3. EDA

## Explore the dataset:

**Dataset is divided to 2 parts:**

Train dataset: 103904 rows
Test dataset: 25976 rows
Each with 25 column

**the dataset consists of mainy two parts:**

**1-user info which** [Age , Gender, Customer Type , Class ,Type of Travel]

**2-user input data which includes some info about the flight which** [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]

**also the dataset has this data types:**

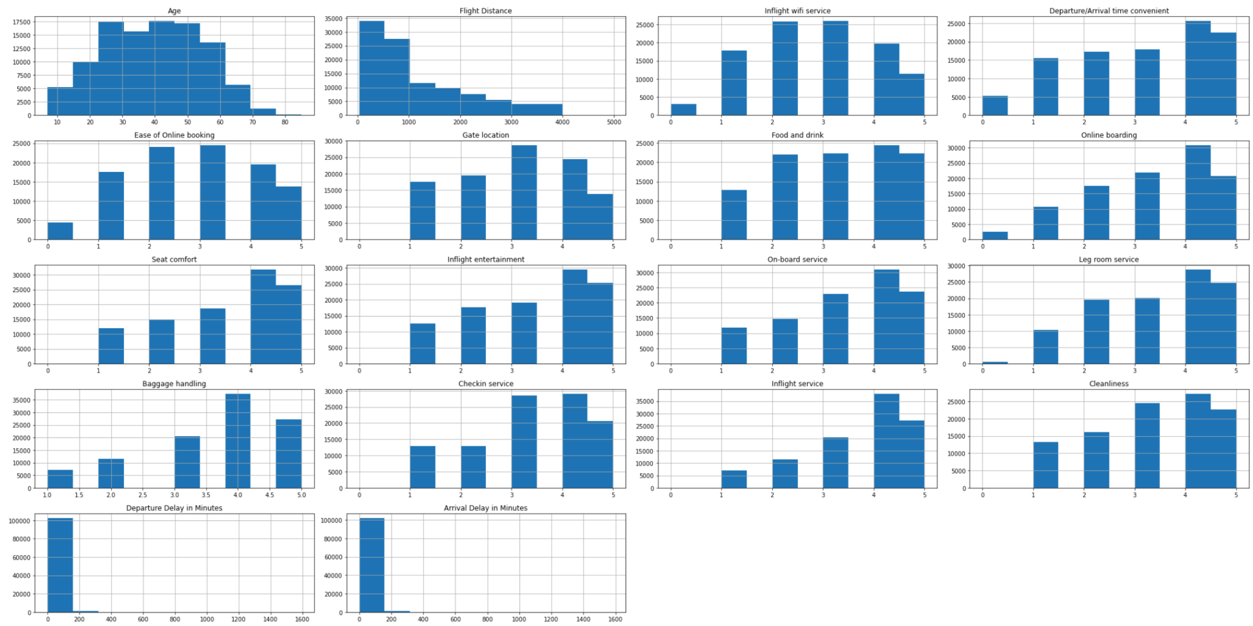
**1-numerical**[Age, Flight Distance, Departure Delay in Minutes, Arrival Delay in Minutes]

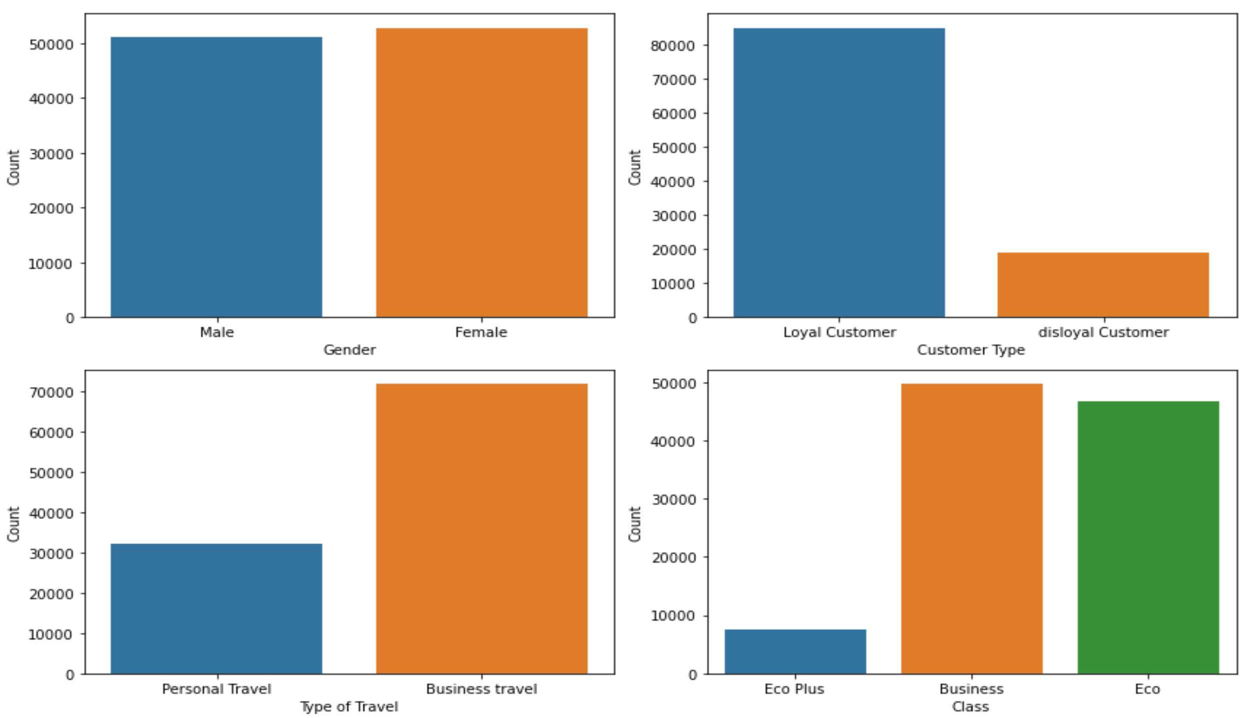
**2-ordinal** [ Inflight wifi service, Departure/Arrival time convenient, Ease of Online booking, Gate location, Food and drink, Online boarding, Seat comfort, Inflight entertainment, On-board service, Leg room service, Baggage handling, Checkin service, Inflight service, Cleanliness]

**3-nominal**[Gender, Customer Type, Type of Travel, Class]

**Target column:**
Satisfaction (satisfied/ neutral or dissatisfied)

# Visualize the data - Histograms



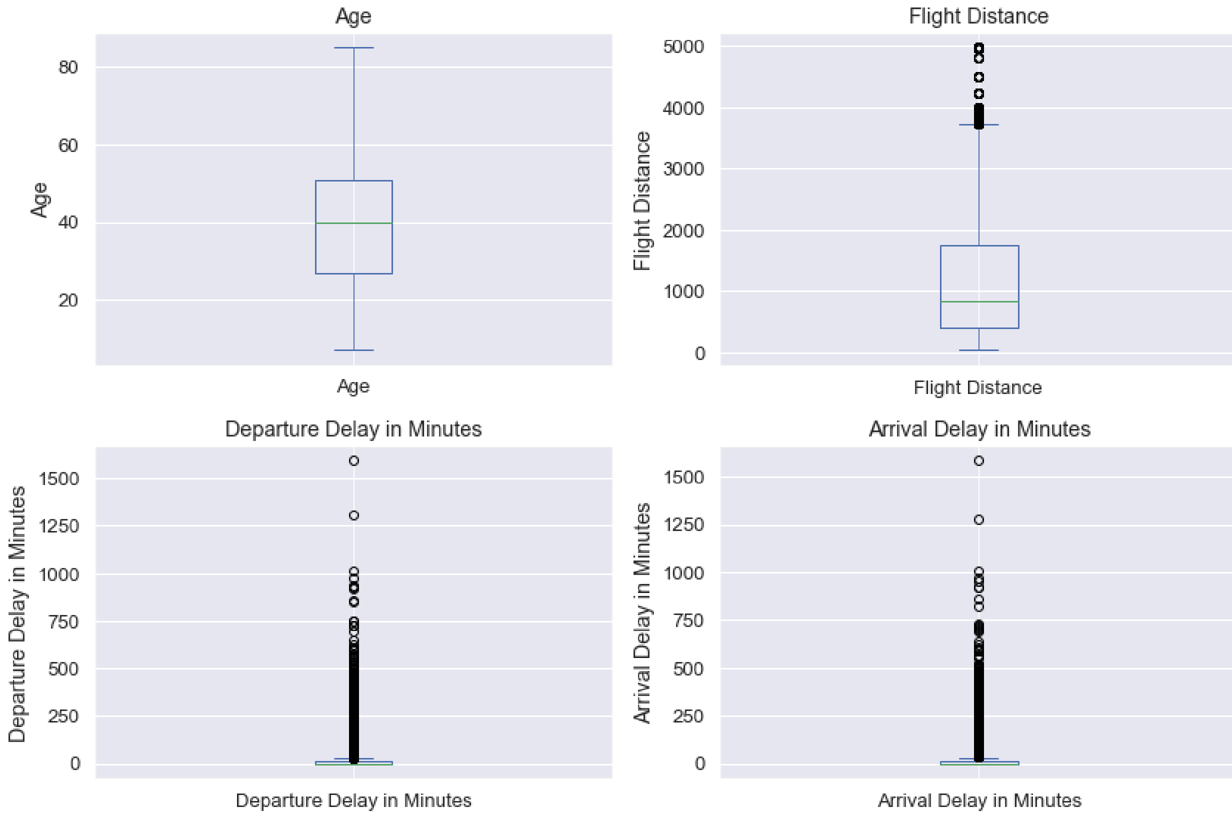



# view missing data (null values)

```
Missing values count:
 Gender                               0
Customer Type                        0
Age                                  0
Type of Travel                       0
Class                                0
Flight Distance                      0
Inflight wifi service                0
Departure/Arrival time convenient    0
Ease of Online booking               0
Gate location                        0
Food and drink                       0
Online boarding                      0
Seat comfort                         0
Inflight entertainment               0
On-board service                     0
Leg room service                     0
Baggage handling                     0
Checkin service                      0
Inflight service                     0
Cleanliness                          0
Departure Delay in Minutes           0
Arrival Delay in Minutes           310
satisfaction                         0
dtype: int64
```

**Arrival Delay in Minutes seems to has alot of values of nulls we can handle them by removing them**

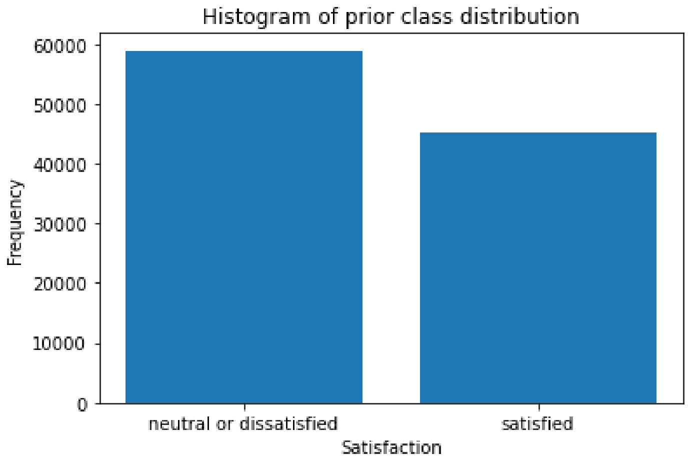
# Identify outliers - Box plots




**Arrival Delay in Minutes and Departure Delay in Minutes seems to has alot of outliers but they make sense because they are the delay of the flight**

**flight disytance also has some outliers but they make sense also**

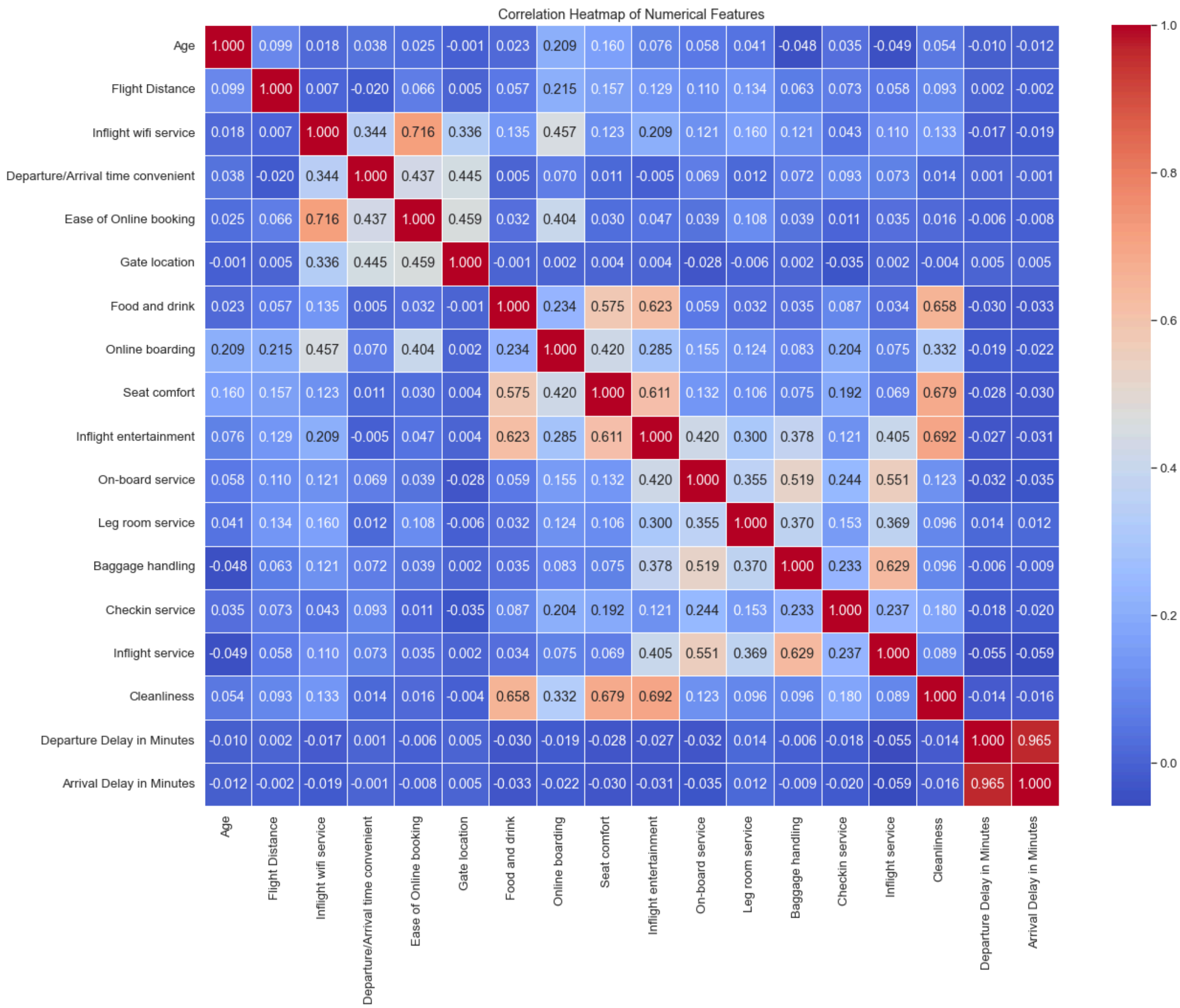
# prior class distributions




**it seems that the customers that are neutral or dissatisfied is more than the satisfied customers**

**the classes seems to be balanced as the difference between them is very small**

# correlations




**it seems that Departure Delay in Minutes and Arrival Delay in Minutes are stongly positive correlated**

**Inflight wifi service and Ease of Online booking are positively correlated**

**Cleanliness is correlated by same degree with Food and drink , Seat comfort , Inflight entertainment** which may have this meaning if the flight is clean the Food and drink , Seat comfort , Inflight entertainment may be good

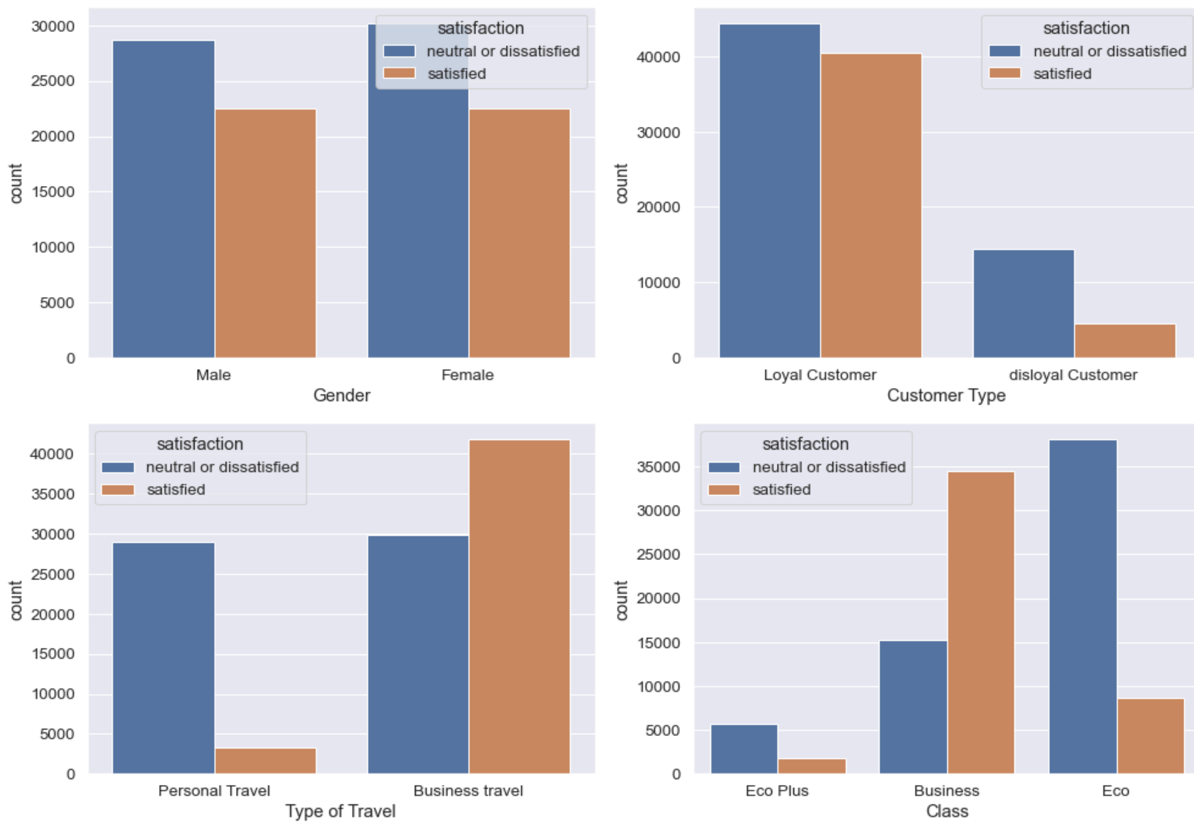
this may have a good conclusion cleanliness is a very good factor

**Food and drink , Seat comfort , Inflight entertainment are correlated with each other**

**from the correlation matrices there exists correlated elements but not in very high degree the require to remove any of them**

# analze the data set features with the satisfaction

# user info




**we can conclude that**

**1-there is a balance between gender and satisfaction so gender don't affect satisfaction**

**2-the majority of Cusomer Type is Loyal Customer**

**3-the majority of Disloyal Customers are neural of disatisfied**

**4-the majority of Type of Travel is a Busines Travel**

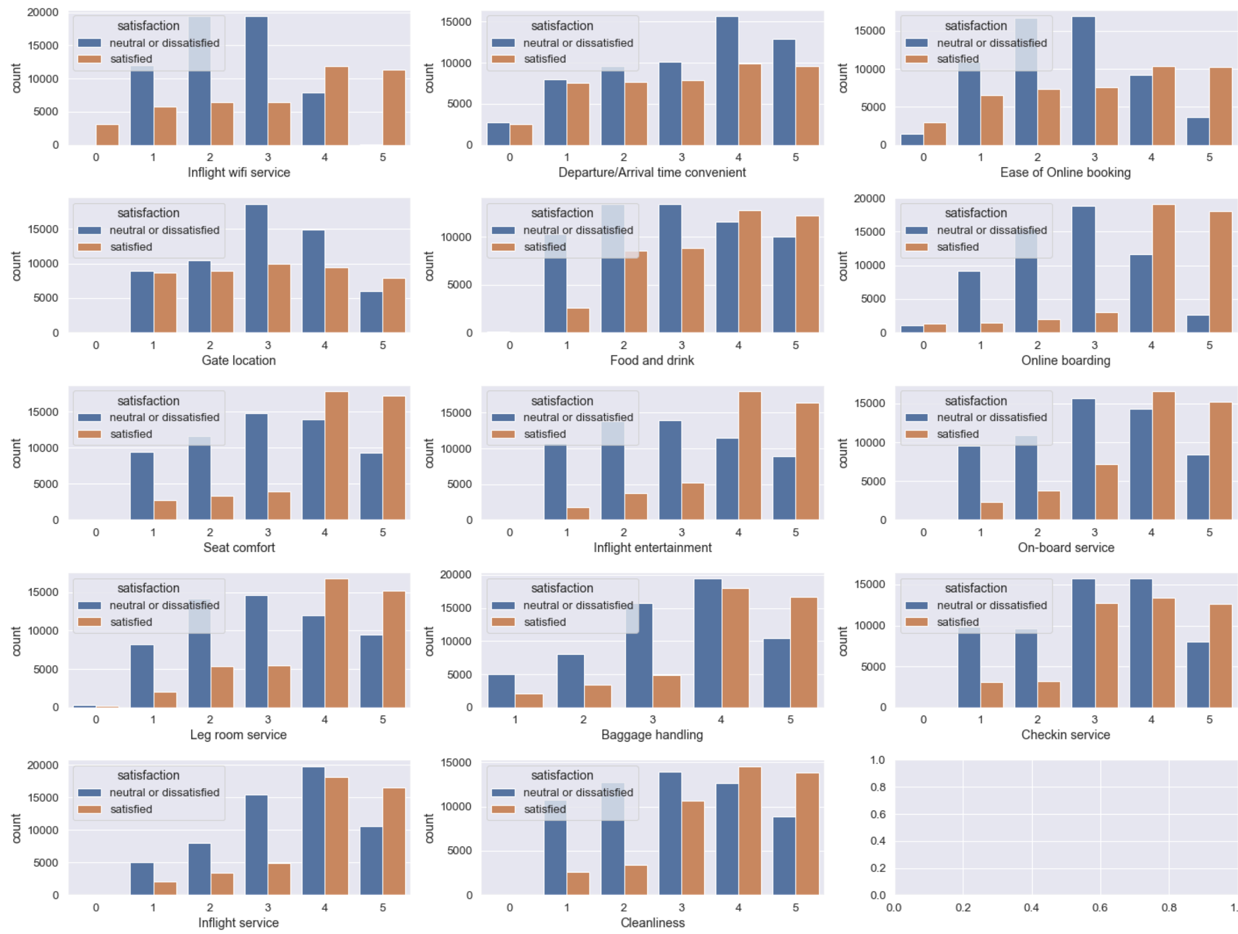
**5-the majority of Type Personal Travel are neural of disatisfied**

**6-the majority of Class is Business or Eco**

**7-the majority of Class Business are satisfied**

**6-the majority of Class Eco are neural of disatisfied**

# user input data



1- the majority of Inflight wifi service are 2,3,4 and the majority of them are neural or dissatisfied

2-the majority of customers who gave Departure/Arrival time convenient

3-the majority of customers who gave Ease of Online booking values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

4-the majority of customers who gave Gate location values 3,4 are neural or dissatisfied

5-the majority of customers who gave food and drink values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

6-the majority of customers who gave Online boarding values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

7-the majority of customers who gave Seat comfort values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

8-the majority of customers who gave Inflight entertainment values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

9-the majority of customers who gave On-board service values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

10-the majority of customers who gave Leg room service values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

11-the majority of customers who gave Baggage handling values 1,2,3,4 are neural or dissatisfied and the majority who gave 5 are satisfied

12-the majority of customers who gave Checkin service values 1,2,3,4 are neural or dissatisfied and the majority who gave 5 are satisfied

13-the majority of customers who gave Inflight service values 1,2,3,4 are neural or dissatisfied and the majority who gave 5 are satisfied

14-the majority of customers who gave Cleanliness values 1,2,3 are neural or dissatisfied and the majority who gave 4,5 are satisfied

**at the end we can conclude that the majority of satisfied cusomers those who gave the most of services values 4,5 and the majority of disatisfied who gave 1,2,3**

## 4. Models used

**Random Forest**
pyspark mapreduce

**SVM**
sklearn

**KNN**
sklearn

**Naive Bayes**
-our mapreduce
-pyspark mapreduce
-sklearn

**Aprori**
mlxtend

**K Means cluster**
sklearn

## Classification Models Accuracy

| Model | Train set accuracy | Test set accuracy |
|---|---|---|
| Random Forest (pyspark mapreduce) | 99.98% | 96.34% |
| SVM | 94.60% | 94.58% |
| KNN (sklearn) | 94.58% | 93.14% |
| Naive Bayes (our mapreduce) | 89.35% | 89% |
| Naive Bayes (pyspark mapreduce) | 89.35% | 89% |
| Naive Bayes (sklearn) | 89.35% | 89% |

- ***Random Forest***

  Here we used random forest from with `numTrees=80,maxDepth=30`
  We have tried many values of them but those gives us the max accuracy
- ***SVM***

  Here we used the default SVM model of sk-learn
- ***KNN***

  Here we used KNN with k=7 from sklearn
- ***Naive Bayes***

  - **NaiveBayes (Multinomial) implementation using MapReduce:**

    It consists of 3 stages, each stage consists of mapper and reducer.

    **Stage 1 for calculating the prior probabilities:**

    **Map:**
    **Input:** row of data
    **Output**: key value pair of (class label, 1)

    **Reduce:**
    **Input:** key value pair of (class label, list of 1s)
    **Output**: key value pair of (class label, sum of the list as the count of this class )
    prior probability = log(count/total_size)

    **Stage 2 for calculating the likelihood probabilities:**

    **Map:**
    **Input:** row of data
    **Output:** key value pair of (class label, (feature, feature value, count)

    **Reduce:**
    **Input:** key value pair of (class label, list of tuples (feature, feature value, count) )
    **Output:** key value pair of (class label, feature, feature value,

probability)
Probability = log(count / total count of the class)

**Stage 3 for calculating the accuracy:**

**Map:**
**Input:** row of data
**Output:** key value pair of (0,0) if predicted class is wrong
or (0,1) if predicted class is correct

**Reduce:**
**Input:** key value pair of (0, list of 0s and 1s)
**Output:** key value pair of (0, sum of this list as number of correct predictions)
Accuracy = number of correct predictions / total data size

- **NaiveBayes using pyspark:**

With modelType="multinomial" and smoothing= 0

- **NaiveBayes using sklearn:**

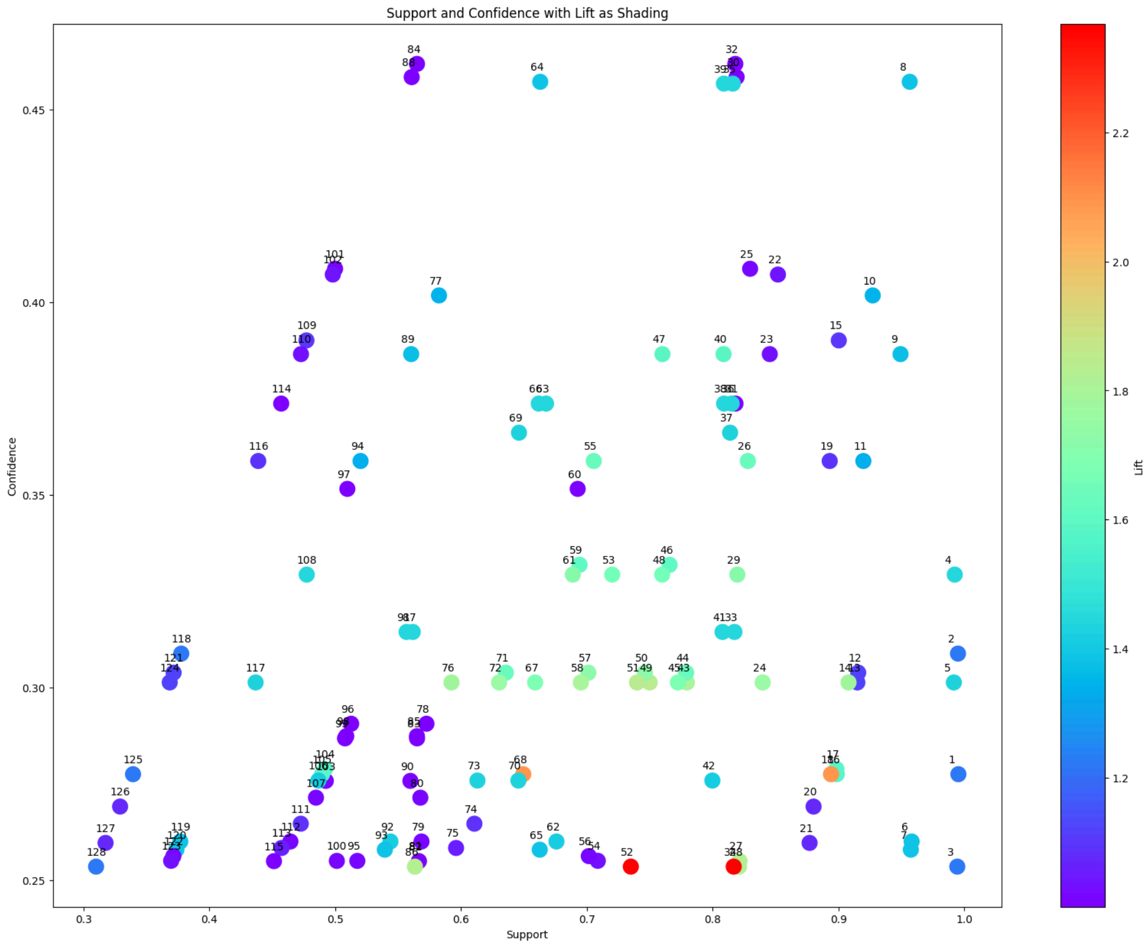
MultinomialNB model
With alpha = 1.0e-10

The parameters are chosen in NaiveBayes using pyspark and sklearn to be similar as MapReduce implementation.

- ## *Aprori*

 Here we get the most frequent items by setting a minimum support of 25%  and then get the rules base on lift of threshold =1
And for the result rules we sort them based on lift and confidence and that what we got for the most top rules

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 73 | (Type of Travel_Personal Travel, satisfaction_... | (Customer Type_Loyal Customer) | 0.278815 | 0.817322 | 0.277487 | 0.995236 | 1.217680 | 0.049605 | 38.349193 | 0.247879 |
| 10 | (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 |
| 66 | (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 |
| 106 | (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 |
| 117 | (Customer Type_Loyal Customer, Class_Business,... | (Type of Travel_Business travel) | 0.303848 | 0.689627 | 0.301307 | 0.991638 | 1.437934 | 0.091765 | 37.116618 | 0.437487 |
| 100 | (Arrival Delay in Minutes_0.0, Class_Business) | (Type of Travel_Business travel) | 0.271395 | 0.689627 | 0.260019 | 0.958084 | 1.389278 | 0.072858 | 7.404576 | 0.384573 |
| 95 | (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 |
| 25 | (Class_Business) | (Type of Travel_Business travel) | 0.477989 | 0.689627 | 0.457230 | 0.956569 | 1.387082 | 0.127595 | 7.146350 | 0.534591 |
| 54 | (Customer Type_Loyal Customer, Class_Business) | (Type of Travel_Business travel) | 0.407193 | 0.689627 | 0.386539 | 0.949278 | 1.376509 | 0.105728 | 6.119093 | 0.461406 |
| 31 | (satisfaction_satisfied) | (Type of Travel_Business travel) | 0.433333 | 0.689627 | 0.401775 | 0.927174 | 1.344457 | 0.102937 | 4.261832 | 0.452126 |

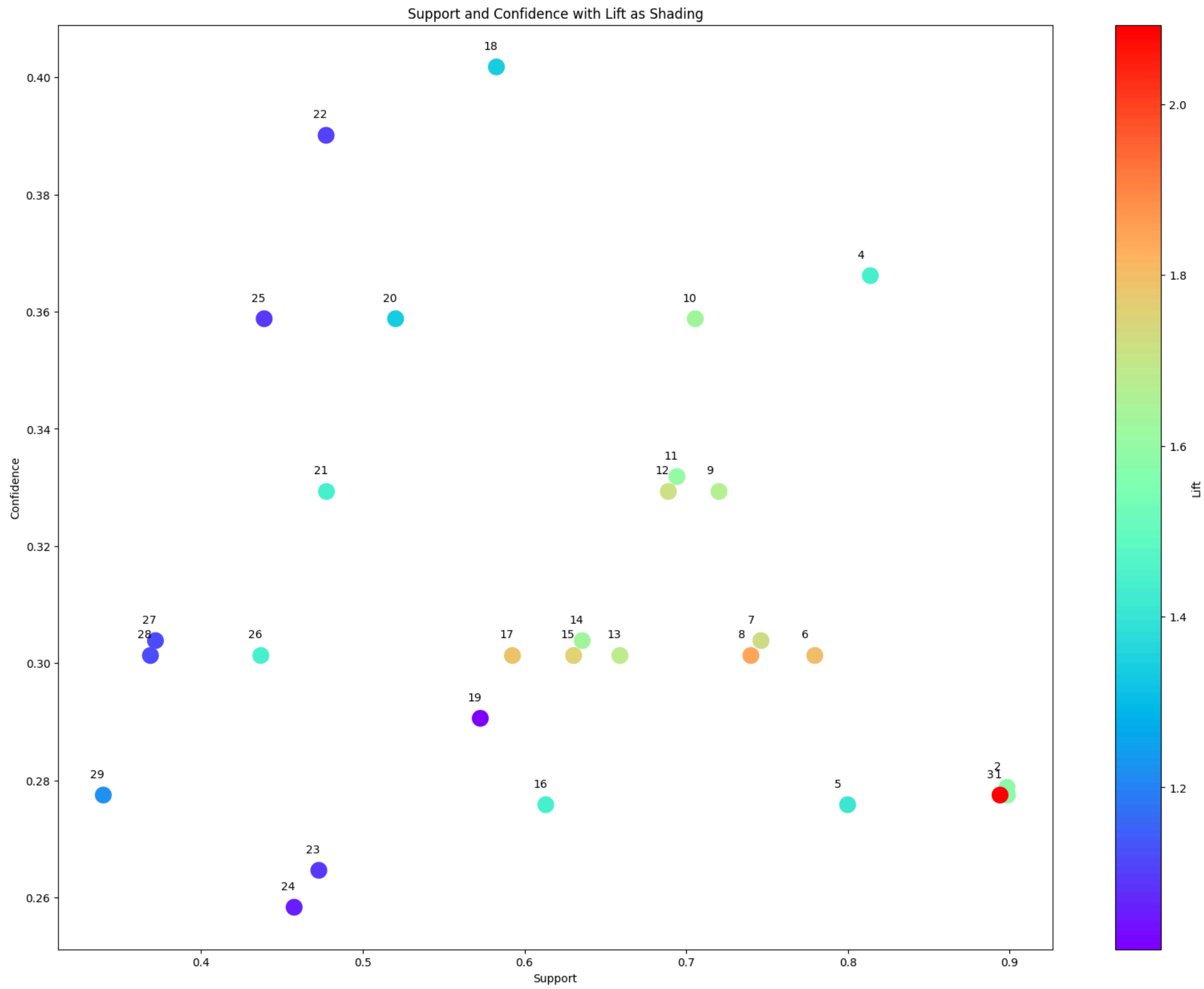



## Conclusion

- We can see that (Type of Travel, Customer Type, Class) exist in many rules together, which means that there is a significant relation between them.
- Customers whose flight for a personal travel are most likely loyal customers.
- Customers whose flight in a business class are most likely have a business travel.

And another thing that we do is to get the rules that has the satisfaction label of the consequent to know which rule lead to satisfaction and dissatisfaction

| | antecedents | consequents | antecedent support | consequent support | support | confidence | lift | leverage | conviction | zhangs_metric |
|---|---|---|---|---|---|---|---|---|---|---|
| 72 | (Type of Travel_Personal Travel, Customer Type... | (satisfaction_neutral or dissatisfied) | 0.308795 | 0.566667 | 0.277487 | 0.898613 | 1.585786 | 0.102503 | 4.274048 | 0.534426 |
| 34 | (Type of Travel_Personal Travel) | (satisfaction_neutral or dissatisfied) | 0.310373 | 0.566667 | 0.278815 | 0.898322 | 1.585273 | 0.102937 | 4.261832 | 0.535353 |
| 75 | (Type of Travel_Personal Travel) | (Customer Type_Loyal Customer, satisfaction_ne... | 0.310373 | 0.427221 | 0.277487 | 0.894043 | 2.092694 | 0.144889 | 5.405777 | 0.757144 |
| 42 | (Class_Eco) | (satisfaction_neutral or dissatisfied) | 0.449886 | 0.566667 | 0.366146 | 0.813862 | 1.436226 | 0.111210 | 2.328024 | 0.552124 |
| 84 | (Class_Eco, Customer Type_Loyal Customer) | (satisfaction_neutral or dissatisfied) | 0.344886 | 0.566667 | 0.275841 | 0.799805 | 1.411418 | 0.080406 | 2.164549 | 0.444950 |
| 114 | (Customer Type_Loyal Customer, Class_Business,... | (satisfaction_satisfied) | 0.386539 | 0.433333 | 0.301307 | 0.779499 | 1.798845 | 0.133807 | 2.569903 | 0.723906 |
| 78 | (Customer Type_Loyal Customer, Class_Business) | (satisfaction_satisfied) | 0.407193 | 0.433333 | 0.303848 | 0.746201 | 1.722004 | 0.127398 | 2.232737 | 0.707281 |
| 121 | (Customer Type_Loyal Customer, Class_Business) | (Type of Travel_Business travel, satisfaction_... | 0.407193 | 0.401775 | 0.301307 | 0.739961 | 1.841731 | 0.137707 | 2.300519 | 0.770963 |
| 104 | (Class_Business, Type of Travel_Business travel) | (satisfaction_satisfied) | 0.457230 | 0.433333 | 0.329304 | 0.720216 | 1.662038 | 0.131171 | 2.025371 | 0.733882 |
| 60 | (Customer Type_Loyal Customer, Type of Travel_... | (satisfaction_satisfied) | 0.508527 | 0.433333 | 0.358793 | 0.705553 | 1.628201 | 0.138431 | 1.924513 | 0.785039 |

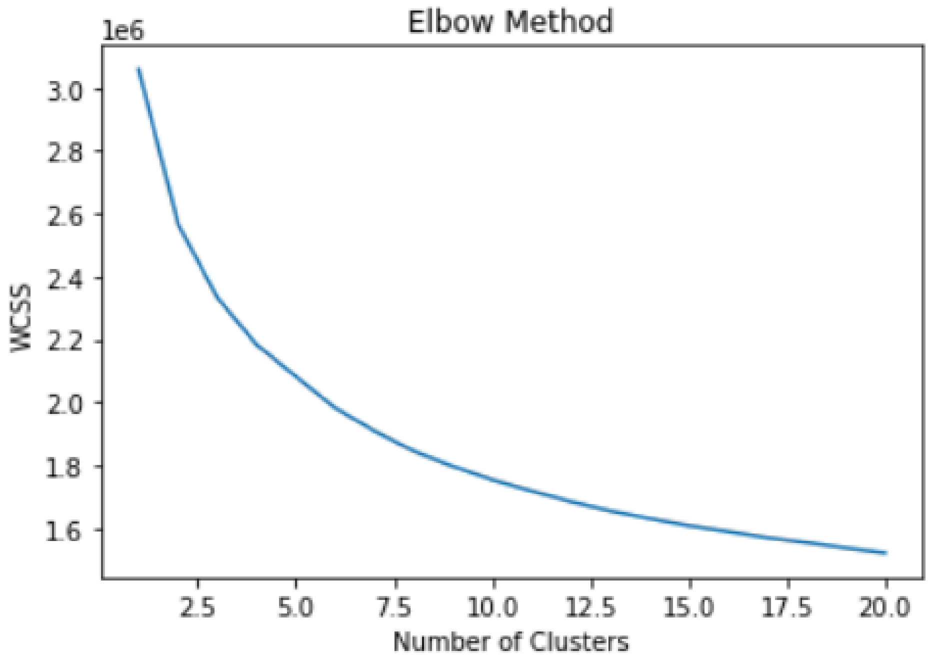



### Conclusion

- We still can see (Type of Travel, Customer Type, Class) exist in many rules together,
  which means that there is a significant relation between them and also a significant effect on customer satisfaction.
- Customers with a personal travel are most likely dissatisfied, even if they are loyal customer
- Customers with an Eco class are most likely dissatisfied, even if they are loyal customer
- Loyal customers with a business travel or a business class flight are most likely satisfied

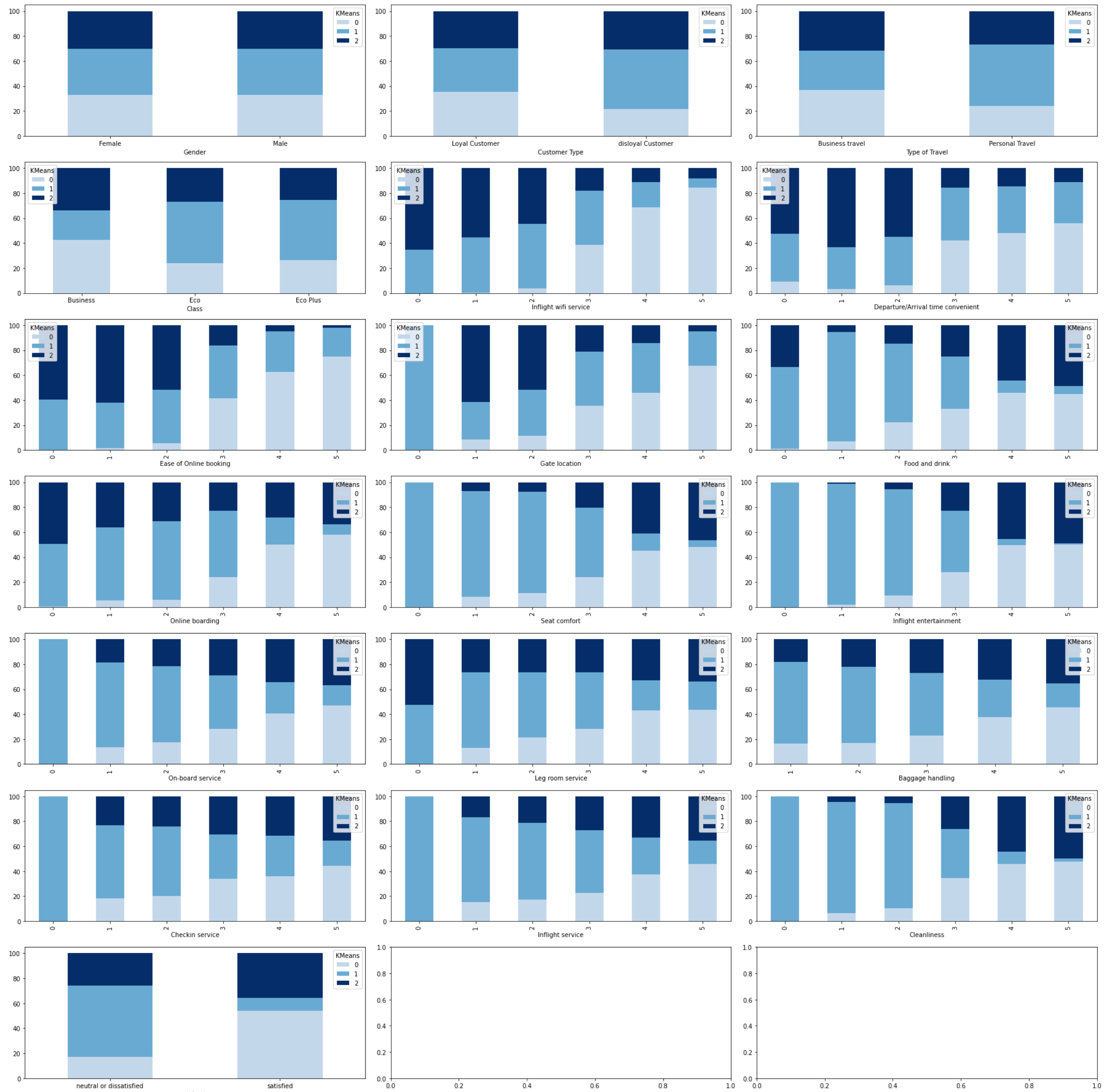
## • *K-means cluster*

Here we used the Elbow method to get the optimal number of clusters




From the graph the optimal number of clusters is k =3

# Grouping all features and label with clusters




what we can conclude is that for most of services the majority of customers who give rating 0,1,2,3 falls in cluster 1 and who give 4,5 falls in cluster 0 and cluster 2 and we can see that the majority of cluster 0 is satisfied ✅ and the majority of cluster 1 is neural or dissatisfied ❌ the majority of cluster 2 is satisfied ✅

## Conclusion

### Cluster 0

- Satisfied customers percentage is a little bit more than unsatisfied (Quite satisfied)
- Some features have a high rate (4,5) in it:
  - Cleanliness
  - Inflight entertainment
  - Seat comfort
  - Food and drink
  - On-board service
  - Online boarding
- Some features have a low rate (0:3) in it:
  - Inflight wifi service
  - Departure/Arrival time convenient
  - Ease of Online booking
  - Gate Location
  - Online boarding
  - Leg room service

### Cluster 1

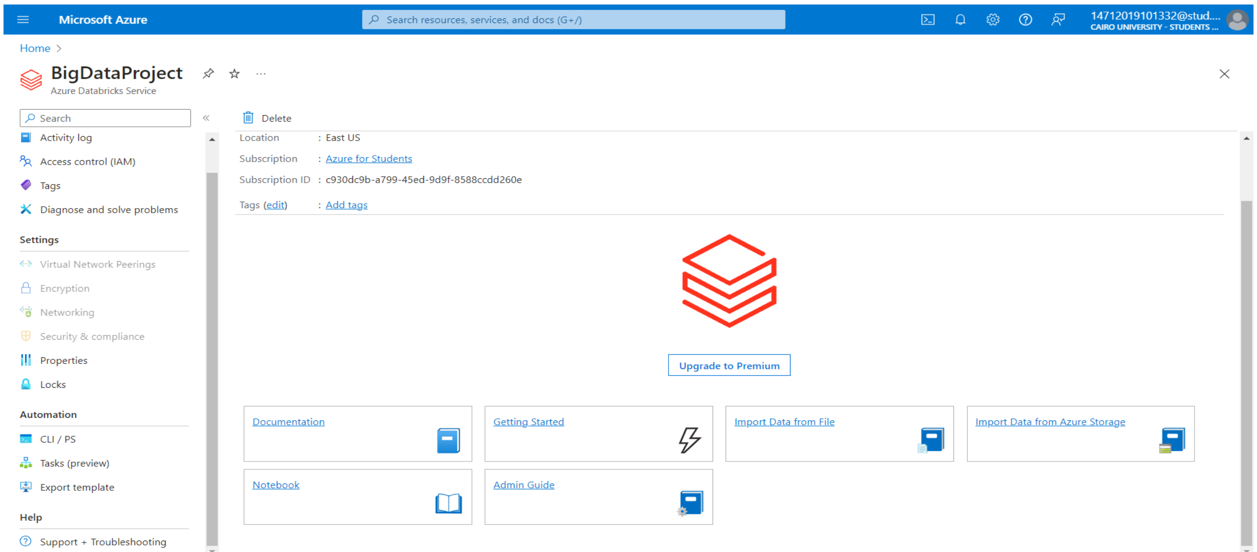
- The majority in it is unsatisfied customers
- The majority in it is disloyal customers
- The majority is in Eco & Eco plus class, and for a personal travel
- Most of the services have low rate in this cluster (0:3)

### Cluster 2

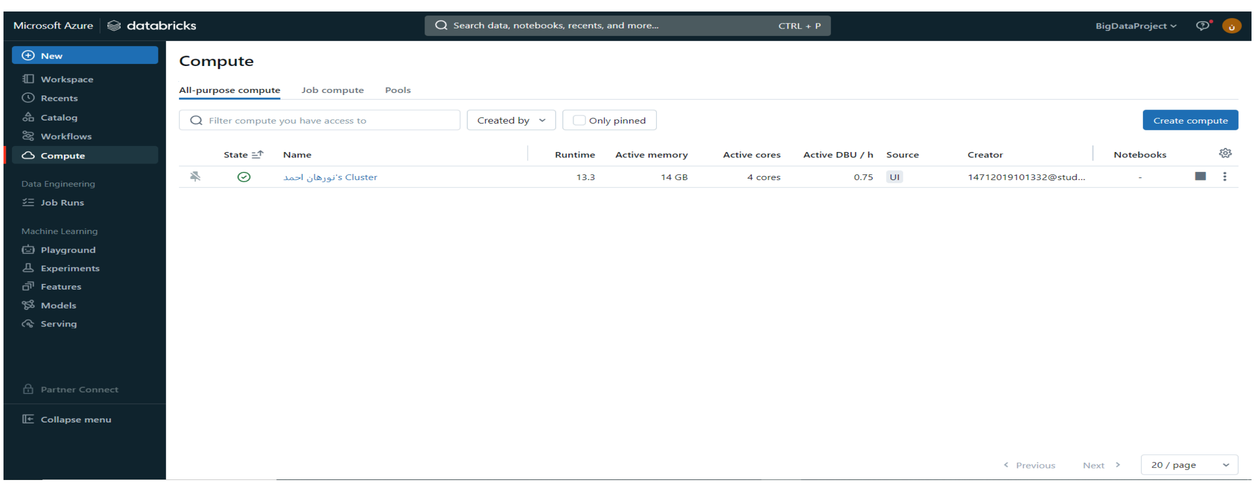
- The majority in it is satisfied customers
- The majority in it is loyal customers
- The majority is in Business class, and for a business travel
- Some features have a high rate (4,5) in it with very high percentage:
  - Inflight wifi service
  - Departure/Arrival time convenient
  - Ease of Online booking
  - Gate Location
- The rest of the services have also a high rate (4,5) with a good percentage

# 5. Cloud part

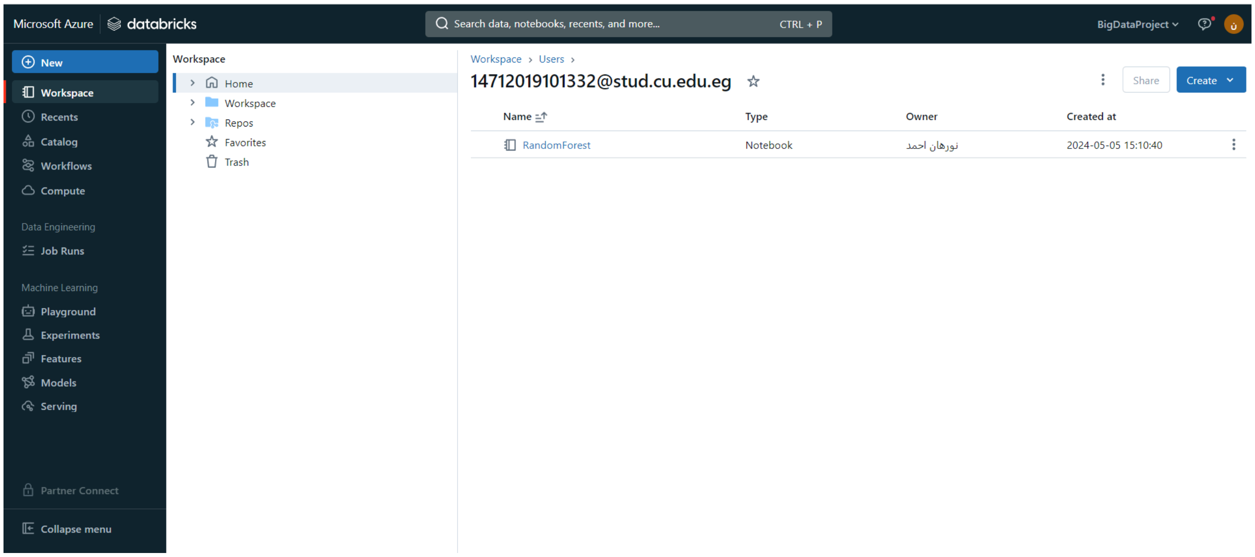
We create a databricks workspace



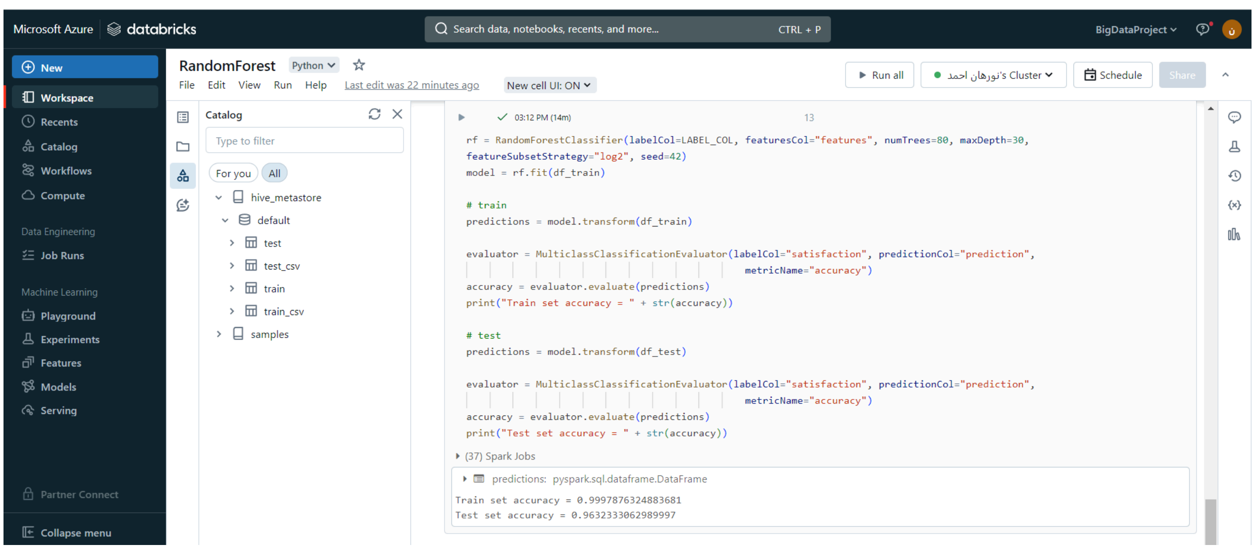

Then we create a Compute unit



Then we upload the model notebooks and dataset




run this model



## 6. Business Part

- There is a significant percentage of dissatisfaction between customers, as more than 50% of customers are neutral or dissatisfied.

- Gender of the customer has no effect on the satisfaction.

- Loyal customer percentage is more than 80%, so the loyalty of customers doesn't make them satisfied.

- More than 68% of the customers have the flight for business travels.

- The majority of flight class is the Business class, and after it Eco class, Eco plus class percentage is very small.

- Lack of customer satisfaction on Departure/Arrival time convenient feature, So they should also give this more attention to the Departure/Arrival time.

- Most dissatisfied customers are the customers whose flight class is Eco or Eco plus, and their flight for personal purposes.

  On the other hand, most satisfied customers are the customers whose flight class is Business, and their flight for business purposes.

  So Type of Travel, Flight Class play a significant role in customer satisfaction.

  So They should give more attention to Eco class services.

- Arrival delay and departure delay are very related to each other positively.

- Cleanliness, Food and drink , Seat comfort , Inflight entertainment are related to each other.

- Inflight wifi service and Ease of Online booking are related to each other positively.