# BD Project Document

## Team VII


| Name | Section | Bench No. |
| --- | --- | --- |
| Essam Wisam | 2 | 2 |
| Mohamed Saad | 2 | 15 |
| Hala Hamdy | 2 | 35 |
| Noran Hany | 2 | 34 |

Supervised by

Dr. Lydia Wahid

Eng. Omar Samir


May 2023

# Table of Contents

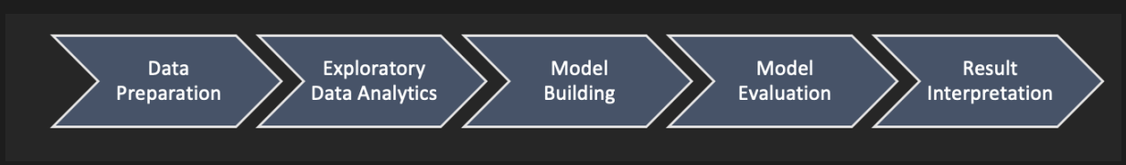
# Introduction

## Problem Definition

The airline industry is undeniably massive with an annual revenue exceeding $800B and 6M travelers per day. A team of four computer engineers from Cairo university have taken it upon themselves to find out a recipe for the perfect airline company by answering a question of paramount importance “What makes airline customers satisfied?”. The question is posed as both a data analysis problem and a machine learning problem that together answer it via exploratory analytics, association rule mining and predictive models that sets the scale for most important determinants of customer satisfaction.

## Project Pipeline

Our solution to the aforementioned problem considers the following pipeline




which corresponds to the folder structure shown below. Usually, we purely used notebook files for demonstration and Python files for the needed logic.

## Folder Structure

```
.
├── DataFiles
│   ├── airline-train.csv
│   └── airline-val.csv
├── DataPreparation
│   ├── DataPreparation.py
│   └── Visualization.ipynb
└── ModelPipelines
    ├── Apriori
    │   ├── Apriori.ipynb
    │   └── rules.txt
    ├── NaiveBayes
    │   ├── NaiveBayes.ipynb
    │   └── NaiveBayes.py
    ├── RandomForest
    │   ├── RandomForest.ipynb
    │   └── RandomForest.py
    └── SVC
        ├── SVC.ipynb
        └── SVC.py
```

## Data Preparation

Data preparation involves reading the data and putting in a suitable form using the PySpark library. Our data preparation module supported the following:

- Reading a specific split of the data (training or validation)
- Reading specific column types from the data (numerical, ordinal or categorical)
- Frequency Encoding
- Dropping missing values
- Imputing numerical outliers

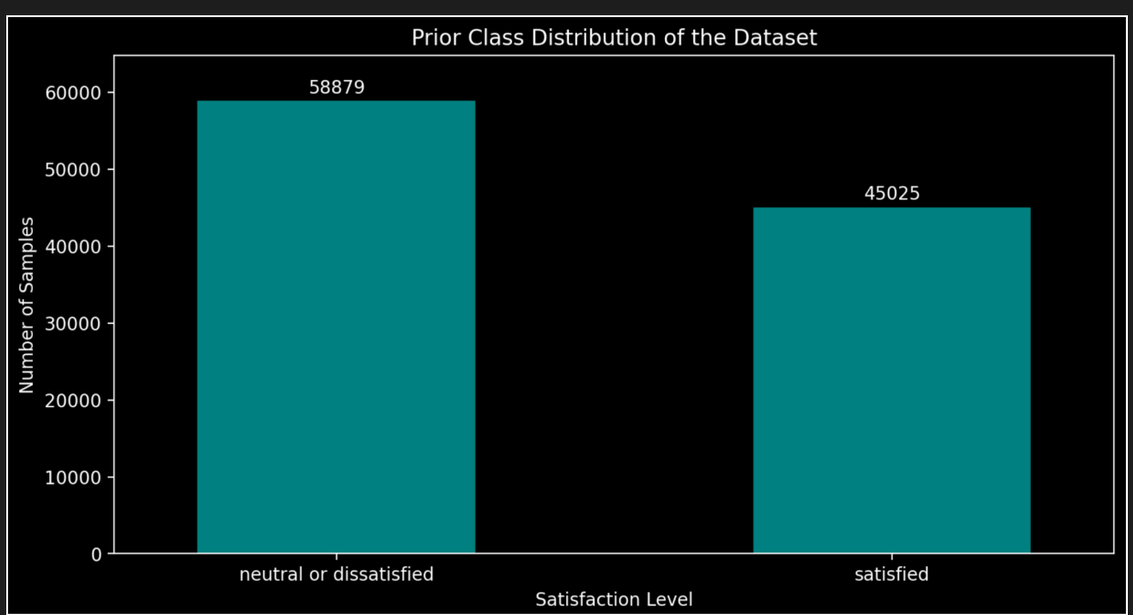
Alternatives for the function were implemented as well in case any model required further special preprocessing.

## Exploratory Data Analysis

To explore the effect of various travel variables on customer satisfaction and to guide model initiation we considered performing exploratory data analysis. In this, instead of querying the data for specific information we rather allowed the data to speak for itself all it had to say, then we gathered insights from that. In this, we employed three types of analysis: univariate analysis, correlations and associations and finally, multivariate analysis.

### Univariate Analysis

#### Prior Class Distribution



### Insights

✦ It seems that the majority of customers surveyed where potentially unsatisfied which makes work on this even more consequential.

✦ There is no severe imbalance. Machine learning models should be able to handle this as is

## Basics of each Variable

| Variable Name | Variable Description | Variable Type | Values |
|---|---|---|---|
| Gender | Gender of the passengers | Nominal | Female, Male |
| Customer Type | The customer type | Nominal | Loyal customer, Disloyal customer |
| Age | The actual age of the passengers | Numerical | - |
| Type of Travel | Purpose of the flight of the passengers | Nominal | Personal Travel, Business Travel |
| Class | Travel class in the plane of the passengers | Nominal | Business, Eco, Eco Plus |
| Flight Distance | The flight distance of this journey | Numerical | - |
| Inflight wifi service | Satisfaction level of the inflight wifi service | Ordinal | 1, 2, 3, 4, 5 |
| Departure/Arrival time convenient | Satisfaction level of Departure/Arrival time convenient | Ordinal | 1, 2, 3, 4, 5 |
| Ease of Online booking | Satisfaction level of online booking | Ordinal | 1, 2, 3, 4, 5 |
| Gate location | Satisfaction level of Gate location | Ordinal | 1, 2, 3, 4, 5 |
| Food and drink | Satisfaction level of Food and drink | Ordinal | 1, 2, 3, 4, 5 |
| Online boarding | Satisfaction level of online boarding | Ordinal | 1, 2, 3, 4, 5 |
| Seat comfort | Satisfaction level of Seat comfort | Ordinal | 1, 2, 3, 4, 5 |
| Inflight entertainment | Satisfaction level of inflight entertainment | Ordinal | 1, 2, 3, 4, 5 |
| On-board service | Satisfaction level of On-board service | Ordinal | 1, 2, 3, 4, 5 |
| Leg room service | Satisfaction level of Leg room service | Ordinal | 1, 2, 3, 4, 5 |
| Baggage handling | Satisfaction level of baggage handling | Ordinal | 1, 2, 3, 4, 5 |
| Check-in service | Satisfaction level of Check-in service | Ordinal | 1, 2, 3, 4, 5 |
| Inflight service | Satisfaction level of inflight service | Ordinal | 1, 2, 3, 4, 5 |
| Cleanliness | Satisfaction level of Cleanliness | Ordinal | 1, 2, 3, 4, 5 |
| Departure Delay in Minutes | Minutes delayed when departure | Numerical | - |
| Arrival Delay in Minutes | Minutes delayed when Arrival | Numerical | - |
| Satisfaction | Airline satisfaction level | Nominal | Satisfaction, Neutral, Dissatisfaction |

## Missing Values

| Age | Flight Distance | Departure Delay in Minutes | Arrival Delay in Minutes | Gender | Customer Type | Type of Travel | … |
|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 310 | 0 | 0 | 0 | 0 |

### Insights

✦ All columns in the dataset are free of missing values, except for the 'Arrival Delay' column.

✦ The presence of missing values in this column is likely due to data entry errors.

✦ Considering that over 50% of the 'Arrival Delay' values are 0, the missing values will be imputed with 0, which corresponds to the mode of the column.

## Central Tendency & Spread of each Variable

### I. Mean, Median, Mode, STD, Quartiles of Numerical Variables

| | Age | Flight Distance | Departure Delay in Minutes | Arrival Delay in Minutes |
|---|---|---|---|---|
| count | 103904.000000 | 103904.000000 | 103904.000000 | 103594.000000 |
| mean | 39.379706 | 1189.448375 | 14.815618 | 15.178678 |
| std | 15.114964 | 997.147281 | 38.230901 | 38.698682 |
| min | 7.000000 | 31.000000 | 0.000000 | 0.000000 |
| 25% | 27.000000 | 414.000000 | 0.000000 | 0.000000 |
| 50% | 40.000000 | 843.000000 | 0.000000 | 0.000000 |
| 75% | 51.000000 | 1743.000000 | 12.000000 | 13.000000 |
| max | 85.000000 | 4983.000000 | 1592.000000 | 1584.000000 |

#### Insights

✦ All ages are represented in the data with a balanced frequency

✦ The majority (75%) of delays are below 13 minutes, with very few cases exceeding this threshold.

✦ Such cases may indicate a disaster or a serious circumstance that requires investigation. As such, they can be considered as outliers, as they are rare and not representative of the normal cases.

### II. Mode, Median, IQR & Entropy of Ordinal Variables

| | 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 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 | 103904.0 |
| min | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 0.0 | 1.0 | 0.0 | 0.0 | 0.0 |
| 25% | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 2.0 | 3.0 | 3.0 | 3.0 | 2.0 |
| 50% | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 3.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 3.0 | 4.0 | 3.0 |
| 75% | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 5.0 | 4.0 | 4.0 | 4.0 | 5.0 | 4.0 | 5.0 | 4.0 |
| max | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 | 5.0 |
| mode | 3.0 | 4.0 | 3.0 | 3.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 | 4.0 |

#### Insights

✦ Seat comfort, inflight service, and baggage handling received high ratings, with 50% of the flights rating them as 4 or above.

### III. Mode, Entropy & Max Entropy of Nominal Variables

| | Gender | Customer Type | Type of Travel | Class |
|---|---|---|---|---|
| mode | Female | Loyal Customer | Business travel | Business |
| Entropy | 0.693036 | 0.47543 | 0.619399 | 0.901828 |
| Max Entropy | 0.693147 | 0.693147 | 0.693147 | 1.098612 |

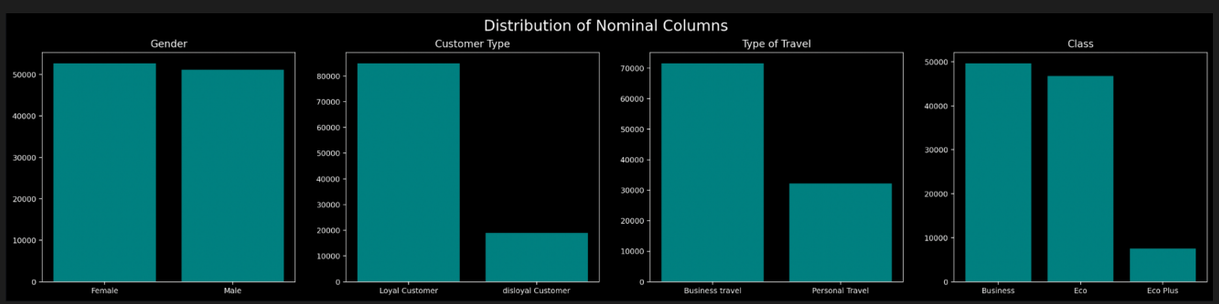
#### Insights

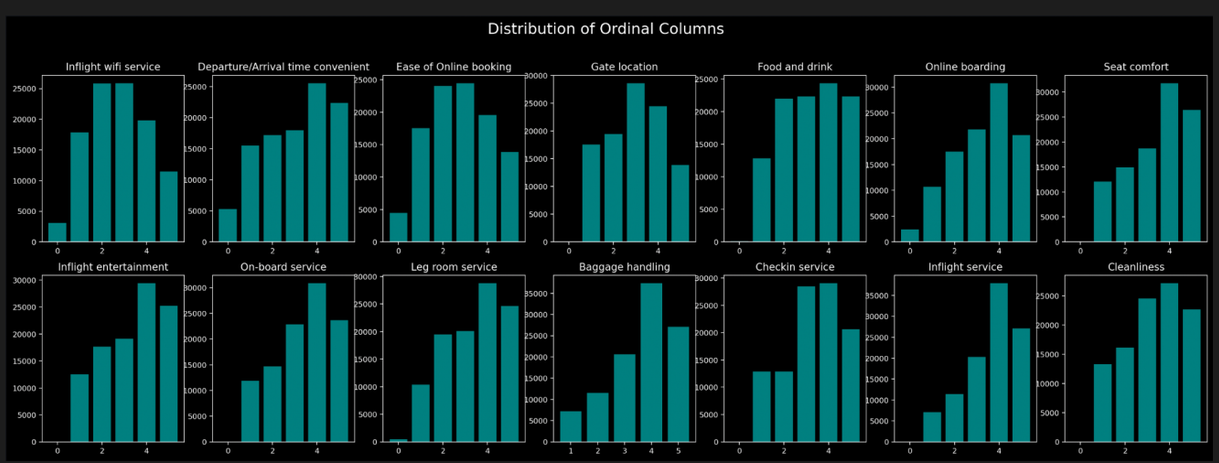
✦ All nominal variables has a large entropy value which indicate low predictability (similar as if the data is random which indicates variability).

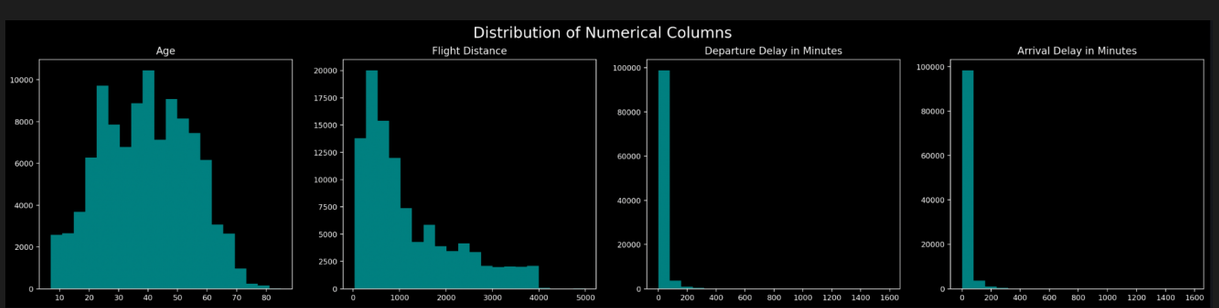
✦ This is likely due to the categorical variables being more evenly distributed across multiple categories, resulting in a higher entropy value.

✦ Applies to a smaller extent for loyal customers.

## Variable Distributions




## Insights

✦ Genders are equally represented through the dataset which is good

✦ The dataset seems to emphasize business travelers; meanwhile, personal travelers are expected to be more common in the real world

✦ The economy plus class is underrepresented in the dataset but that matches the real world as it's not always available

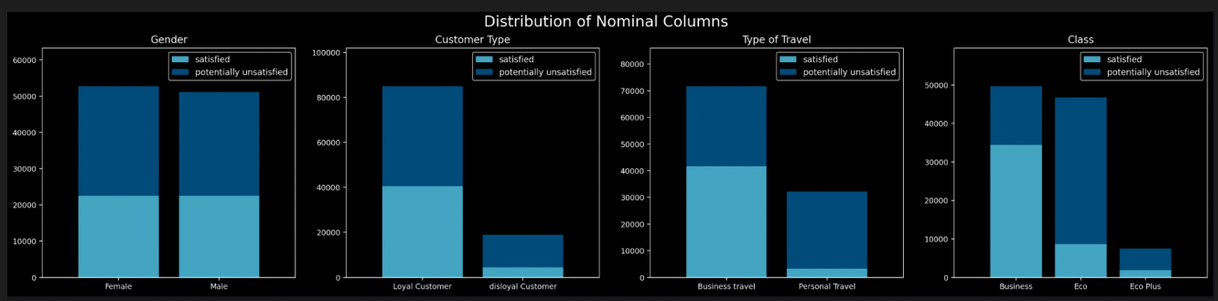
✦ It may be the case that only loyal customers where interested in taking the survey as they do outnumber disloyal ones

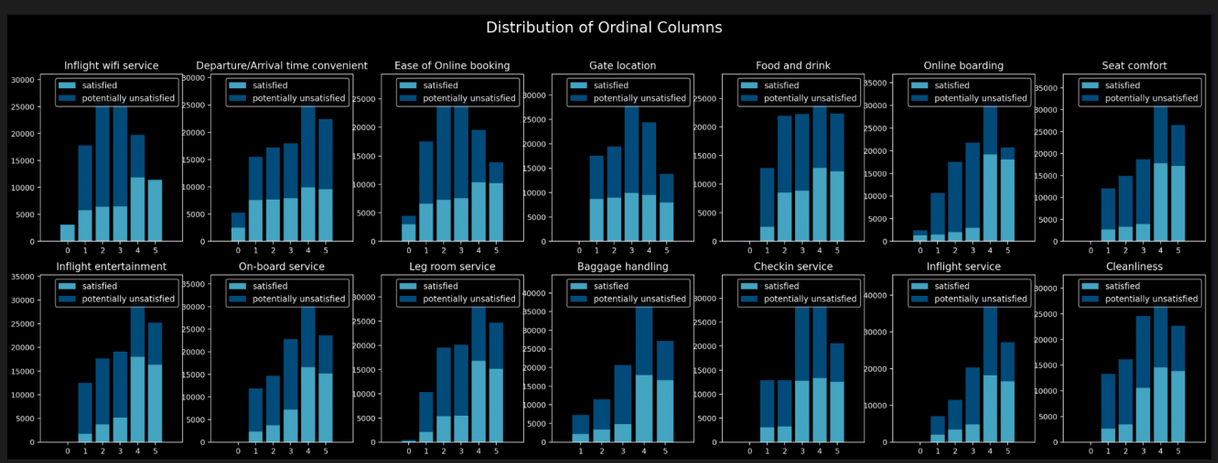
✦ Ordinal distributions are majorly classic, there is evident skew towards the higher ratings

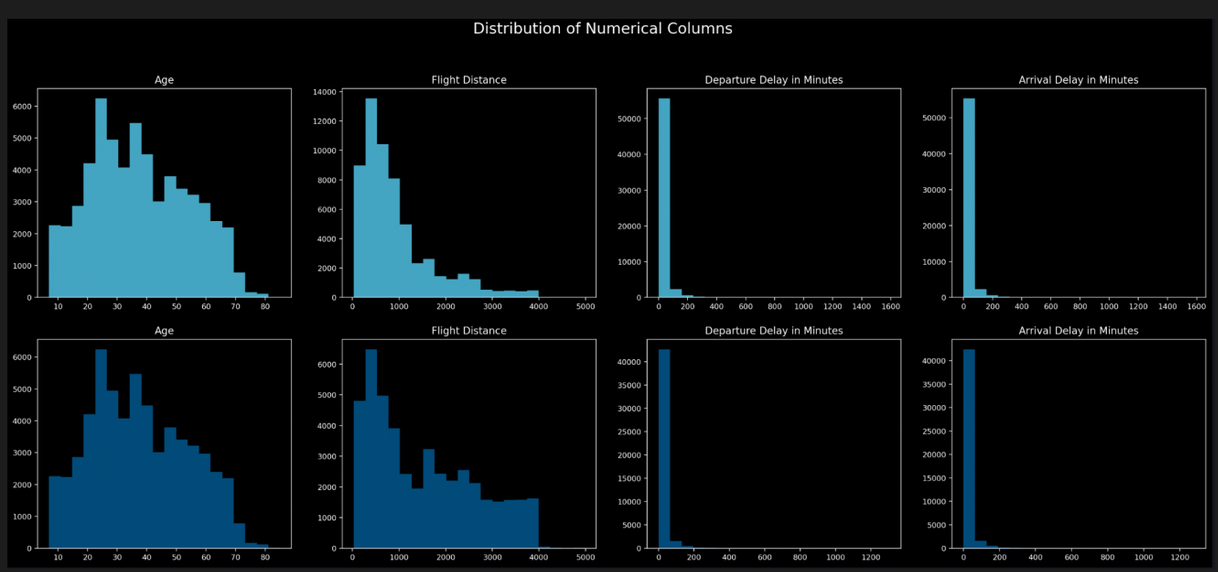
✦ Age follows a bell-shaped distribution

✦ Departure & Arrival delay follow very similar left-skewed distributions and seem to include outliers

## Variable Distributions per Class




## Insights

✦ A huge majority of personal travelers are potentially dissatisfied so they may need more focus

✦ As it logically follows, a huge majority of economy travelers are potentially dissatisfied

✦ It's obvious that many of those that have done online boarding where potentially dissatisfied

✦ Departure and arrival delay are noticeable for the dissatisfied class

## Correlations and Associations

### Dependence between Nominal Variables

The Chi-square test of independence tests if there is a relationship between two categorical variables. In particular, we have that

$H_0$: The two categorical variables are independent.

$H_1$: The two categorical variables are dependent.

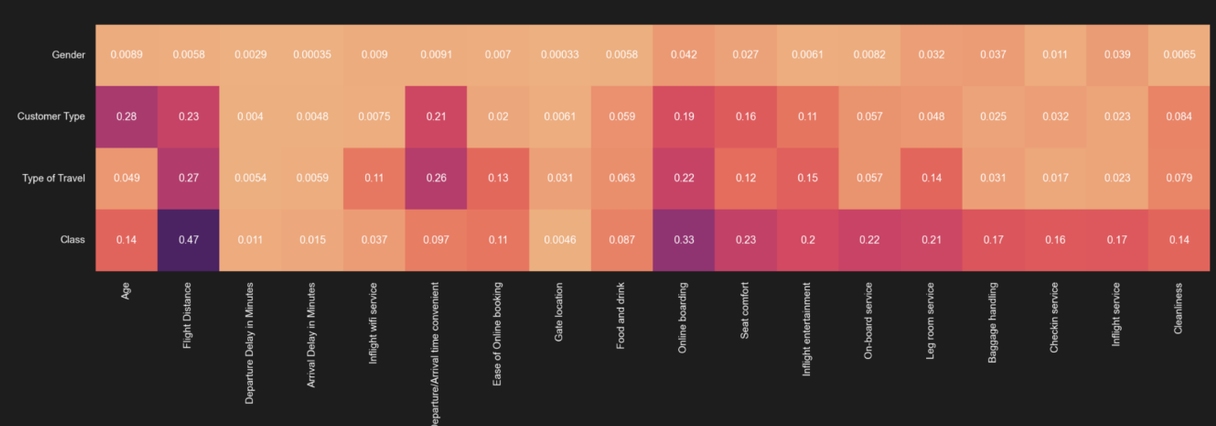
Here, we set $\alpha = 0.05$ and hence, if the p-value for the test done on two variables is less than $0.05$, we reject $H_0$ and conclude that the two variables are dependent.

The following shows the p-values for each possible nominal pair.

| | Gender | Customer Type | Type of Travel | Class |
|---|---|---|---|---|
| Gender | 0.0 | 0.0 | 0.026398 | 0.000119 |
| Customer Type | 0.0 | 0.0 | 0.0 | 0.0 |
| Type of Travel | 0.026398 | 0.0 | 0.0 | 0.0 |
| Class | 0.000119 | 0.0 | 0.0 | 0.0 |

This signifies that all nominal features are indeed dependent. We can set the significance level lower to ignore dependencies between gender and type of travel which means the dependence is less adverse.

### Correlations between Nominal & Numerical Variables




### Insights

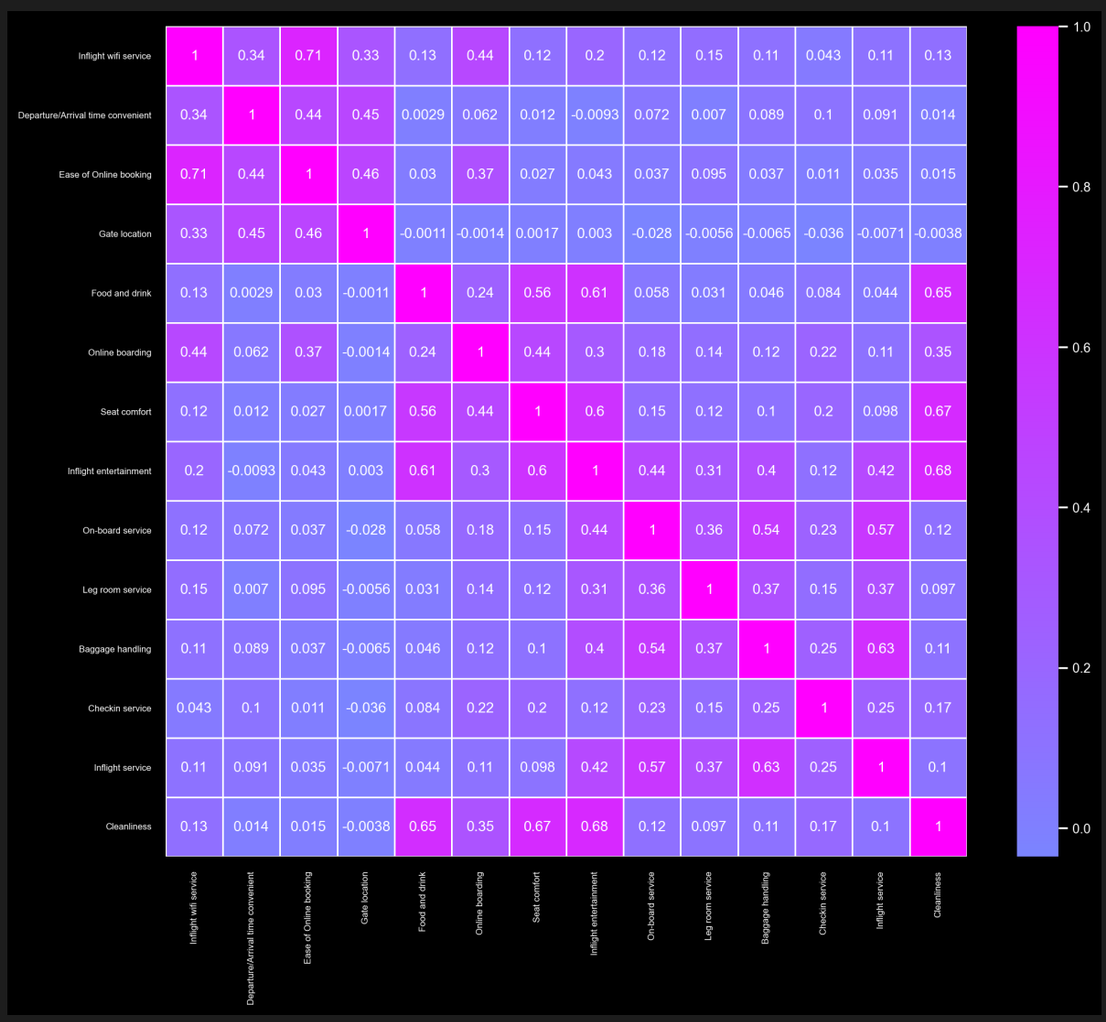
✦ There is a strong association between the class and flight distance; in other words, knowing the flight distance tells us something about the class

✦ In general, class seems to have extraordinary associations with most ratings. They are perhaps more likely to give better ratings

✦ Likewise, knowing the age tells us something about the customer type. Most business travelers wouldn't be young

✦ Gender seems to have no strong associations with anything

## Monotonic Association between Ordinal Variables



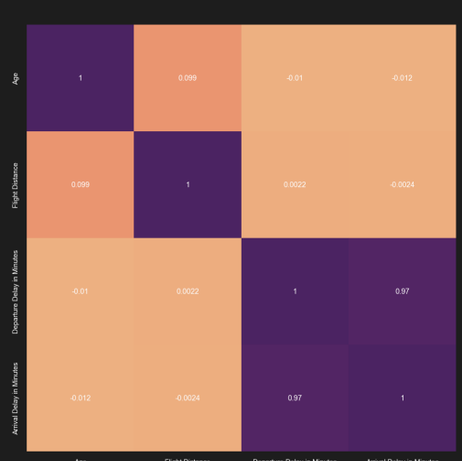

### Insights

✦ High correlation between Ease of online booking and quality of wifi service.

✦ It seems that flights that have high cleanliness also have good food, comfort seat and good entertainment

✦ There is also a high correlation between cleanliness and seat comfort, cleanliness and inflight entertainment

## Correlations between Numerical Variables




### Insights

✦ Numerical variables are mostly uncorrelated

✦ There is an extreme exception. It holds that the arrival and departure delay are almost perfectly correlated

✦ Notice that in this, we only consider linear correlations.

## Naive Bayes Assumption

We worked on a module that given a set of features and their values tested whether the Naive Bayes assumption holds by comparing the multinomial probability to that computed assuming conditional independence.

As expected, the Naive Bayes assumption does not hold. In particular, we have that

$$P(x_1, x_2, ...|C_1 = 0) = 0.16$$

as computed numerically using the definition of the probability. Meanwhile, applying the Naive Bayes assumption we have that

$$P(x_1, x_2, ...|C_1 = 0) = P(x_1|C_1 = 0)P(x_2|C_1 = 0)... = 0.08$$

which is different from the correct probability.

Likewise, for the class $C_1 = 1$ we have that

$$P(x_1, x_2, ...|C_1 = 1) = 0.07$$

but

$$P(x_1, x_2, ...|C_1 = 1) = P(x_1|C_1 = 1)P(x_2|C_1 = 1)... = 0.04$$

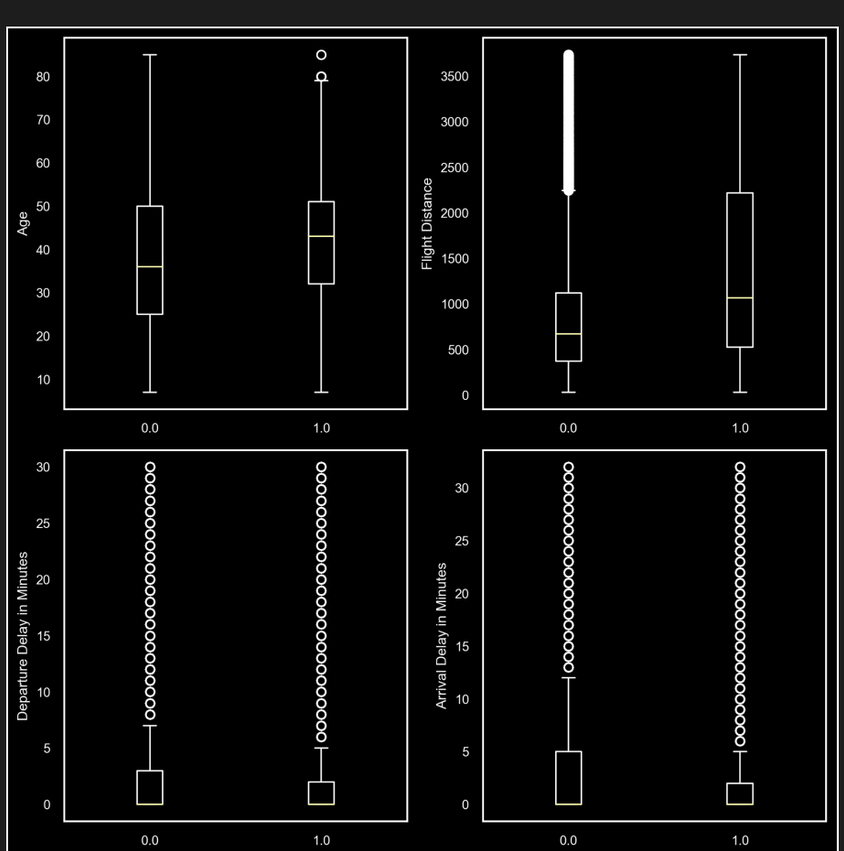
which is different from the correct probability.

**Regarding the optimality of Naive Bayes, from Zhang et al. (2004):**

*Therefore, no matter how strong the dependences among attributes are, naive Bayes can still be optimal if the dependences distribute evenly in classes, or if the dependences can cel each other out. We propose and prove a sufficient and necessary conditions for the optimality of naive Bayes.*

# Multivariate Analysis

## Separability & Distribution of Numerical Variables




### Insights

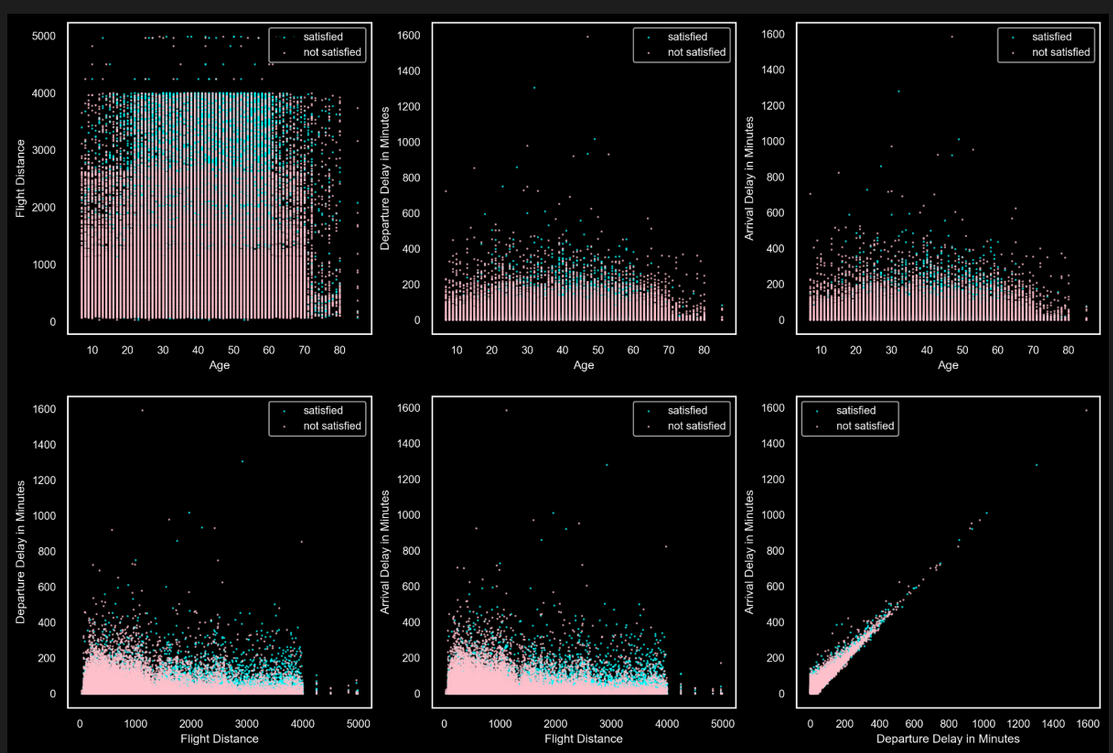
- ✦ The box plots for the departure and arrival delays appear to be compact or small.
- ✦ This is primarily due to the presence of outliers in both columns, which have significantly

higher values compared to the majority of values in the distribution of the two columns.

✦ Furthermore, it is evident that the outliers in the departure column correspond to outliers in the arrival column.

✦ This observation further supports the correlation between the two variables, as flights with a departure delay tend to experience a corresponding delay in arrival, along with additional minor delays during the flight.

## Separability & Distribution of Numerical Pairs



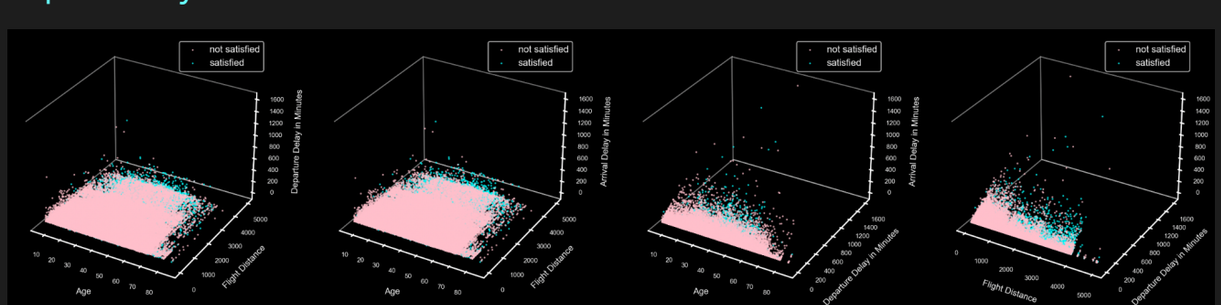

### Insights

✦ Overall, linear separability is not very good over numerical predictors

✦ Extreme correlation between departure and arrival time shows up again in the 6th plot

✦ The correlation between them is also evident in how the plot of each with age looks likeThe nonlinearity evident is not even obvious from the plots. The data complexity is not low

## Separability & Distribution of Numerical Trios

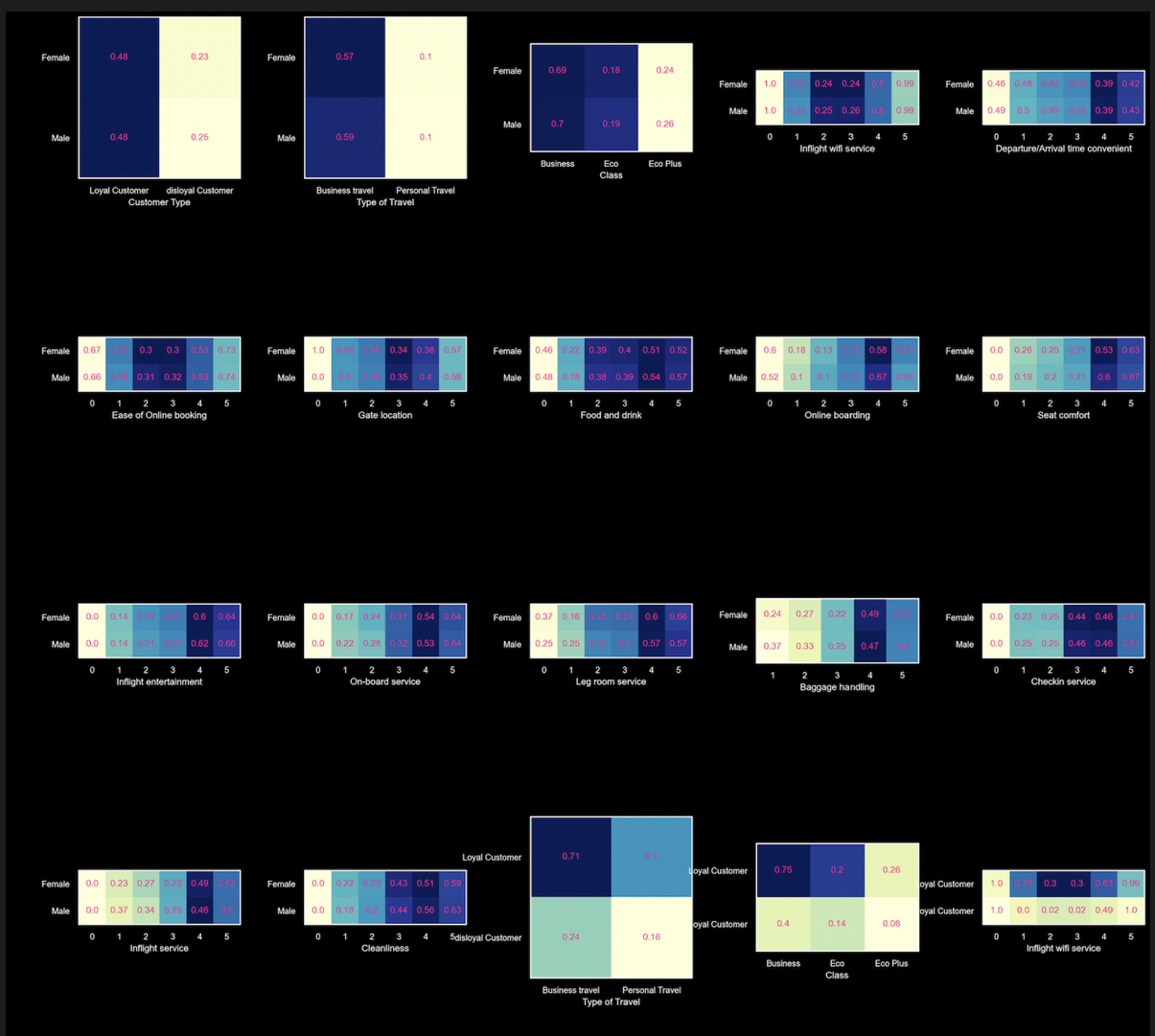



### Insights

✦ Separability issues persist in higher dimensions and collinearity is again obvious between

delay

## Separability & Distribution of Categorical (+Ordinal) Pairs




## Insights

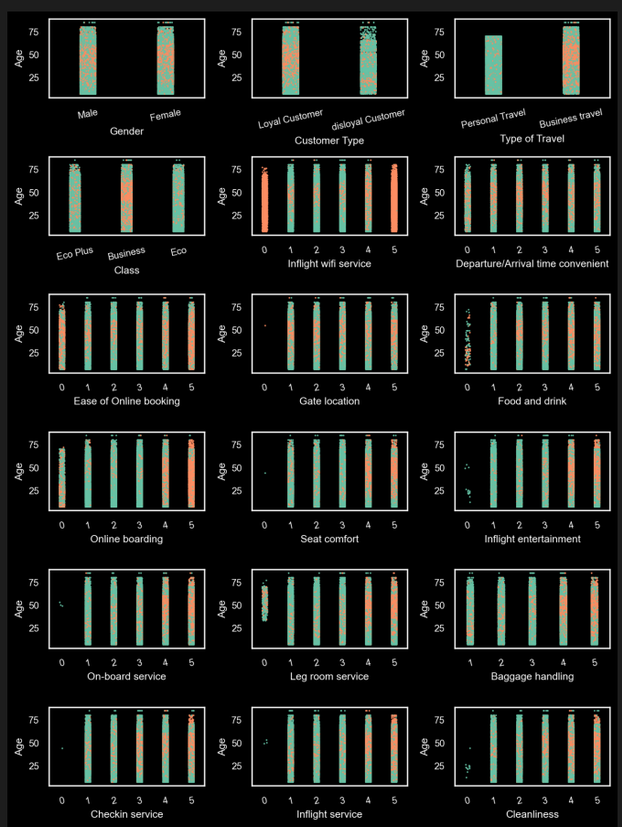
✦ The dark color evident over most of the plots inspire that there are many associations to learn

✦ The fact that many of those have very high or very low satisfaction rates also sheds light on that there is much to learn

✦ Manual insights may be also derived using the same fact.

*Note that this and the next plot are truncated for view in the document. For the full version, please check the corresponding Python notebook. It's also a good way to benchmark your RAM.*

## Separability and Distribution of Numerical and Categorical




### Insights

- ✦ In both cases where the wifi-service was rated as 0 or 5, customers were satisfied regardless of the numerical variable. This suggests that there may have been a data entry issue.
- ✦ For certain columns such as wifi-service, departure and arrival time convenience, and food and drink, as the flight distance increases, the satisfaction ratio also increases.
- ✦ This is because longer flights typically prioritize food services, offering a greater variety of hot meals and providing tablets for in-flight entertainment. However, airlines tend to allocate less attention to shorter flights, resulting in colder food options.
- ✦ Despite this, there is still a fraction of customers who are dissatisfied with longer flights, regardless of the quality of service. This dissatisfaction may be due to disturbances from high levels of noise during the flight or the discomfort of remaining seated for extended periods.

## Models Considered

The abundance of categorical features and their significance as shown above has inspired considering Naive Bayes and Random Forest as predictive models. We later follow up with an SVM model due its powerfulness and that ordinal features can be readily assumed as numerical as well. But before employing such models we considered topping off our exploratory data analytics with association rule learning.

### Apriori

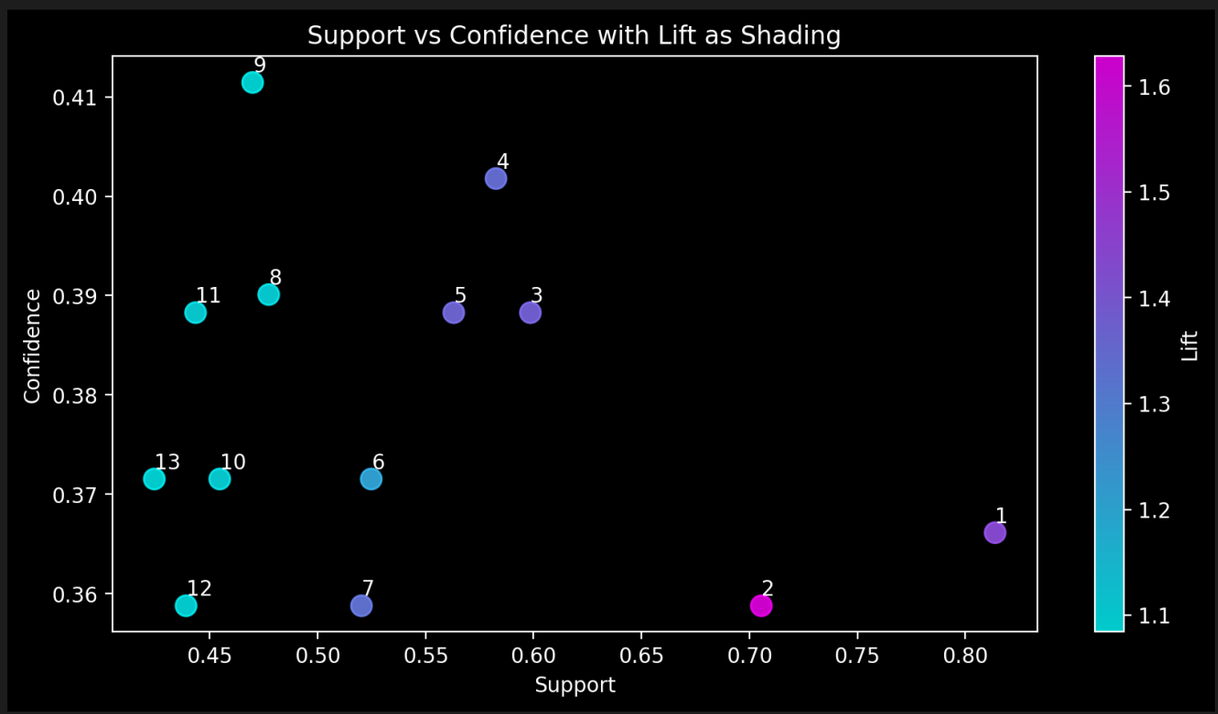
Apriori model has been used to help us see which features and factors that may result in satisfaction or dissatisfaction of the passenger. To do so, we have first considered some preprocessing which includes dropping the Nan values in Arrival delay in Minutes feature. Besides , from the EDA , we can see that the Arrival & Departure delays were very +vely skewed so we dropped these columns. Then the continuous features were first classified into categories.

```
# lets categories the age column into 3 groups
train['Age']=pd.cut(train['Age'],  bins=[0,  18,  65,  100],  labels=['child',
'adult', 'elderly'])
# lets categories the flight column into 3 groups each of equal size
train['Flight Distance']=pd.qcut(train['Flight Distance'], q=3, labels=['low',
'medium', 'high'])
```

Finally , we performed one hot encoding on the columns as this is the expected format for apriori.

We started building the feature set that will be used in generating the rules taking into consideration the min_support as the parameter that filters the feature_itemset. A good soft spot for this min_support landed ~ 0.35 which leads to a moderate  number of rules. Then, we started building the association rules and considering only those with lift >=1.

To interpret which features are the most influential. We have considered plotting a color plot of support against the confidence and the lift as the shading color.




| | antecedents | consequents | support | confidence | lift |
|---|---|---|---|---|---|
| 1 | Class = Eco | satisfaction = neutral or dissatisfied | 0.366146 | 0.813862 | 1.436226 |
| 2 | Type of Travel = Business travel, Customer Type = Loyal Customer | satisfaction = satisfied | 0.358793 | 0.705553 | 1.628201 |
| 3 | Type of Travel = Business travel, Age = adult | satisfaction = satisfied | 0.388281 | 0.598531 | 1.381228 |
| 4 | Type of Travel = Business travel | satisfaction = satisfied | 0.401775 | 0.582597 | 1.344457 |
| 5 | Type of Travel = Business travel | satisfaction = satisfied, Age = adult | 0.388281 | 0.563031 | 1.368449 |
| 6 | Customer Type = Loyal Customer, Age = adult | satisfaction = satisfied | 0.371545 | 0.524781 | 1.211035 |
| 7 | Type of Travel = Business travel | satisfaction = satisfied, Customer Type = Loyal Customer | 0.358793 | 0.520271 | 1.333684 |
| 8 | Customer Type = Loyal Customer | satisfaction = satisfied | 0.390100 | 0.477291 | 1.101443 |
| 9 | Age = adult | satisfaction = satisfied | 0.411437 | 0.469837 | 1.084241 |
| 10 | Customer Type = Loyal Customer | satisfaction = satisfied, Age = adult | 0.371545 | 0.454588 | 1.104878 |
| 11 | Age = adult | satisfaction = satisfied, Type of Travel = Business travel | 0.388281 | 0.443394 | 1.103589 |
| 12 | Customer Type = Loyal Customer | satisfaction = satisfied, Type of Travel = Business travel | 0.358793 | 0.438986 | 1.092617 |
| 13 | Age = adult | satisfaction = satisfied, Customer Type = Loyal Customer | 0.371545 | 0.424282 | 1.087622 |

## Insights

- ✦ In general it seems that "Type of Travel" the most potential attribute to distinguish between satisfaction and neutral or dissatisfied as it if frequent in most of the rules generated
- ✦ As seen, the productions are sorted from the most effective to the least in which all have lift>1.

## Naive Bayes

For Naive Bayes, we started with preprocessing the data by converting the string categories to integers and bucketizing the four numerical variables based on the 10 percentiles (10%, 20%, 30%,...) so that they can be treated similar to categorical variables

We then implemented the NaiveBayes training algorithm to compute the prior and posterior probabilities using MapReduce on Spark. The former is trivial and the latter is as shown below

✦ **NB on Big Data**

- Make $N_m$ splits of the data // *each split $D_i$ is is a set of $(A_i, C_i)$*
  - Let each mapper compute $N(a_m = v,\ C)$ over the data it has (three for loops)

  ⇒ Clearly, can't further perform $P(a_m = v|C) = \frac{N(a_m=v,\ C)}{N(C)}$ as other mappers may also have $(a_m = v,\ C)$
  - Formally, $D_i \rightarrow List([C, < a_m = v,\ N(a_m = v, C) >])$
  - Key is $C$ so for $K$ classes we have $K$ reduce tasks and each of them can group and sum the inner value $N(a_m = v, C)$ over $a_m = v$ to yield the real count

- Reducer gets a single pair $(C,\ List[a_m = v,\ N(a_m = v, C)])$ (all data belonging to class)
  - Produce a final sum of $N(a_m = v, C)$ for each $a_m = v$
  - Formally, $List(K_3 = C, V_3 = (a_m = v, \sum_{a_m=v} N(a_m = v, C)))$ which has the total number of times $a_m = v$ has occurred in class $C$ for each $m, v$

  ⇒ The sum will have at most $N_m$ terms where each mapper was able to find $a_m = v$
  - It can as well easily produce $N(C) = \sum_{all} N(a_i = v, C))$ // choose any $a_i$*
    - » Won't be accurate if there are missing values for $a_i$ (mappers don't emit)

  ⇒ At this point, $P(a_m = v|C) = \frac{N(a_m=v,\ C)}{N(C)}$ can be computed for each $m, v$

  - If we assume the size of the dataset $N(A)$ is known then as well $P(C) = \frac{N(C)}{N(A)}$

After training we have access to $P(a_m = v|C)$ for all features $a_m$ and values $v$ and classes $C$. Inference is then performed using the Naive assumption as follows given example $(a_1, a_2, ..., a_M) = (v_1, v_2, ..., v_M)$

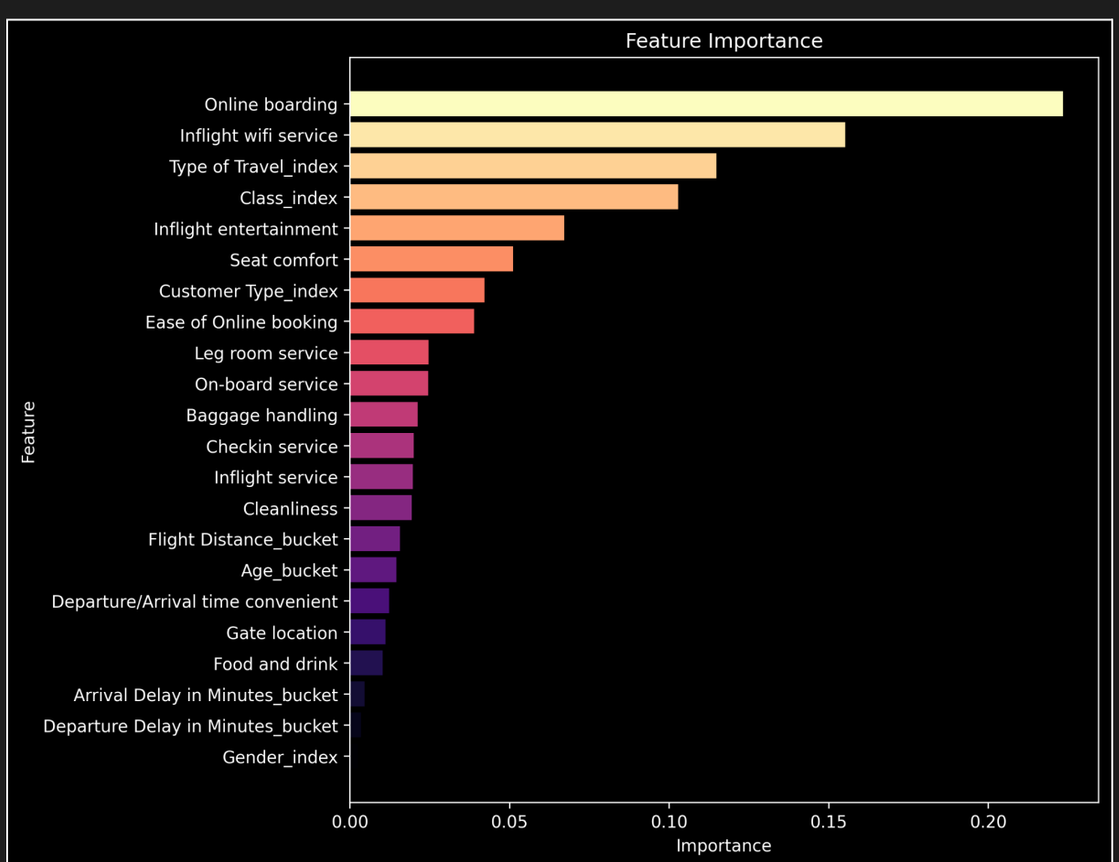
$$K = argmax_k \{P(a_1 = v_1|C_k)P(a_2 = v_2|C_k)...P(a_M = v_2|C_k)P(C_k)$$

The final validation accuracy for the Naive Bayes model was 89.1%.

## Random Forest

We also considered initiating a Random Forest model which did not further require any special processing (beyond NB). Luckily, PySpark's RandomForest inherently supports both categorical and numerical features after applying the model the perceived accuracy on the validation set was 96%

We further used Random Forest to assess the importance of different travel variables in relation to the target. Decision trees assign importance to a node (feature) through the corresponding decrease in impurity (gini importance) by splitting on it. Random Forests average such importance over all trees. The results are as follows




Clearly, type of travel and class play a crucial role in satisfaction; business customers are often more satisfied. Besides, internet-related variables such as online boarding, WiFi, entertainment seem to be even more important than food or comfort.

It also seems that arrival or departure delay are not key determinants of satisfaction given the presence of other factors.

### SVM

The SVM model was initiated using spark. As a preprocessing step, string indexer was done. Then, 2 SVM experiments were conducted for different extra preprocessing.

The first one considered standardizing the numerical values resulting in: Training Accuracy: 0.87467, Test Accuracy: 0.87131.

The second experiment was using the Buckertizers with a GridSearch which led to the following results: Best regParam: 0.01, Best maxIter: 100, Training Accuracy: 0.8779, Test Accuracy: 0.87336.

## Results & Evaluation

Below are shown the final results for the predictive models

| Model | Training Accuracy | Validation Accuracy |
|---|---|---|
| Multinomial Naive Bayes | 89.4% | 89.1% |
| Random Forest | 97.4% | 96% |
| Support Vector Machines (Linear) | 87.8% | 87.3% |

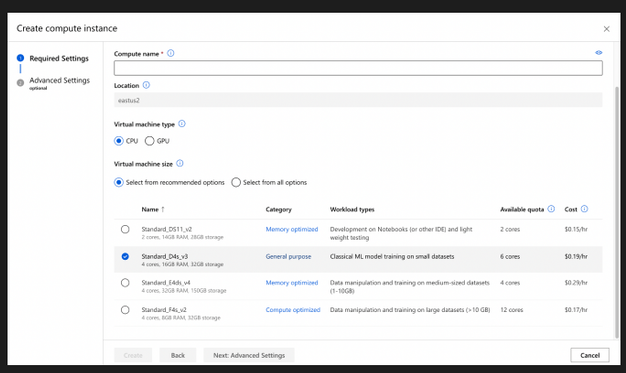
## Business Perspective

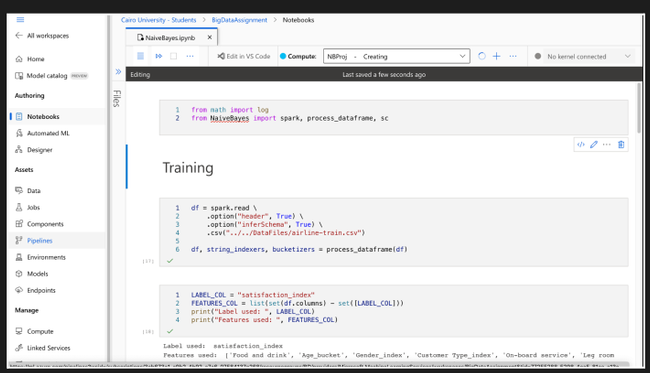
From a business perspective, we observe that:

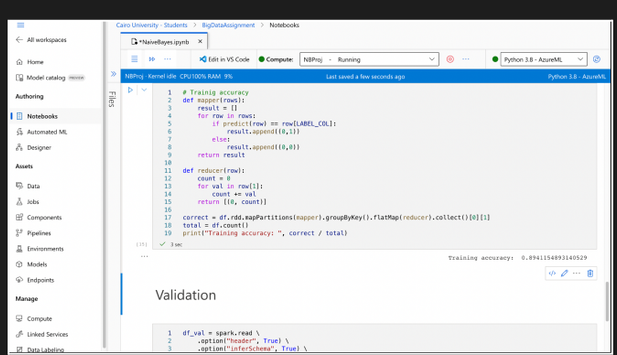
- There is a lack of satisfaction in airline travel experience (imbalance)
- Such lack is focused on economy travelers
- Wifi Service, Entertainment and Online Boarding are key determinants.
- Comfort and ease of booking also matter
- Distance and delays seem to have a less adverse effect

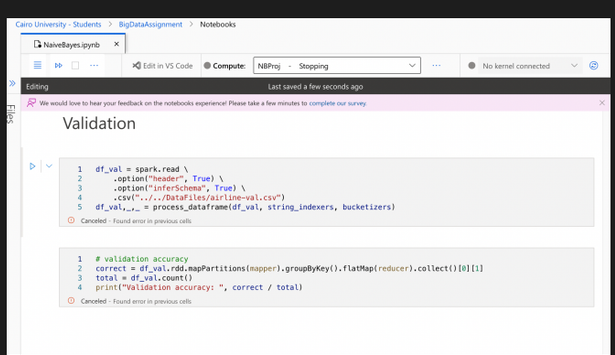
# Running on the Cloud

We utilized AutoML from Azure to run the whole project on a cloud environment.




# Enhancements & Future Work

- First-hand experience to test the real distribution of customer satisfaction
  - Consider traveling to the Louvre museum
  - Tokyo, Japan also has many nice skyscrapers
  - We may as well consider Sydney, Australia to try living upside down
- Extending Naive Bayes with the Gaussian assumption to deal with Gaussian data
- Considering nonlinear kernels for the SVM