# MARKET BASKET INSIGHITS

# Introduction

**In this notebook, I will complete an effective binomial logistic regression.**

**I am interested in knowing if a better in-flight entertainment experience leads to higher customer satisfaction. I would like to construct and evaluate a model that predicts whether a future customer would be satisfied with airline's services given previous customer feedback about their flight experience.**

**The data for this notebook is for a sample size of 129,880 customers. It includes data points such as class, flight distance, and in-flight entertainment, among others. My goal will be to utilize a binomial logistic regression model to help the airline model and better understand this data.**

**I will need to conduct basic EDA, data cleaning, and other manipulations to prepare the data for modeling.**

## 1: Imports

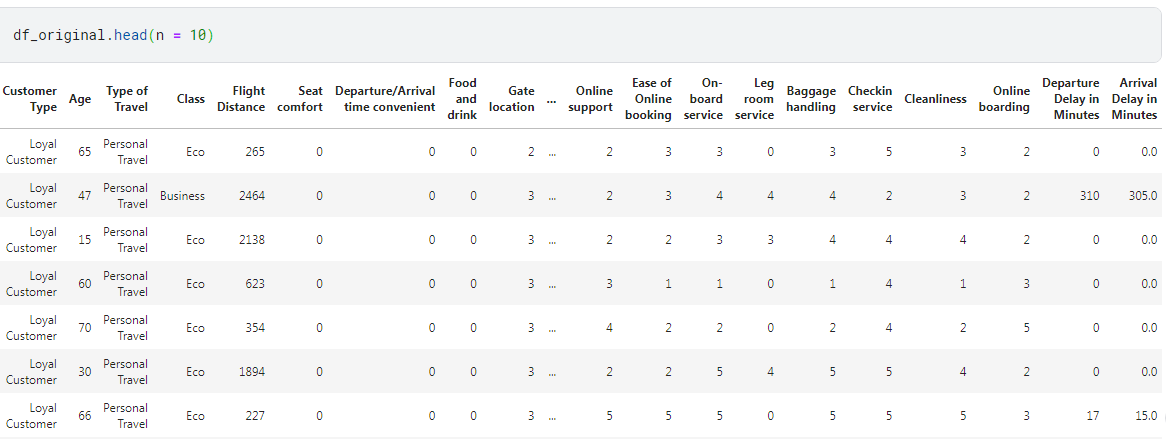
### Import packages[¶](https://kkb-production.jupyter-proxy.kaggle.net/static/dist/jupyterlab/jupyterlab-index-7d6feb0f1d697b8f72d0.html?session=eyJhbGciOiJub25lIiwidHlwIjoiSldUIn0.#Import-packages)



### Load the dataset

df\_original = pd.read\_csv("/kaggle/input/customer-satisfaction-in-airline/Invistico\_Airline.csv")

### Output the first 10 rows



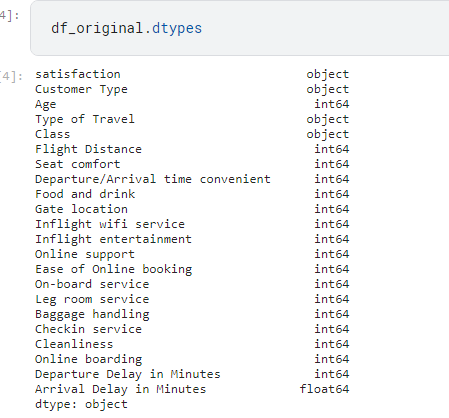
## 2: Data exploration, data cleaning, and model preparation[¶](https://kkb-production.jupyter-proxy.kaggle.net/static/dist/jupyterlab/jupyterlab-index-7d6feb0f1d697b8f72d0.html?session=eyJhbGciOiJub25lIiwidHlwIjoiSldUIn0.#2:-Data-exploration,-data-cleaning,-and-model-preparation)

### Prepare the data[¶](https://kkb-production.jupyter-proxy.kaggle.net/static/dist/jupyterlab/jupyterlab-index-7d6feb0f1d697b8f72d0.html?session=eyJhbGciOiJub25lIiwidHlwIjoiSldUIn0.#Prepare-the-data)

After loading the dataset, I prepare the data to be suitable for a logistic regression model. This includes:

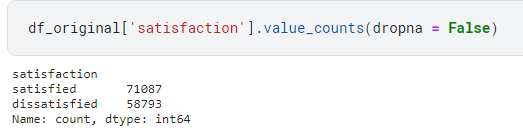
* Exploring the data
* Checking for missing values
* Encoding the data
* Renaming a column
* Creating the training and testing data

### Explore the data



### Check the number of satisfied customers in the dataset

To predict customer satisfaction, I need to know how many customers in the dataset are satisfied before modeling.

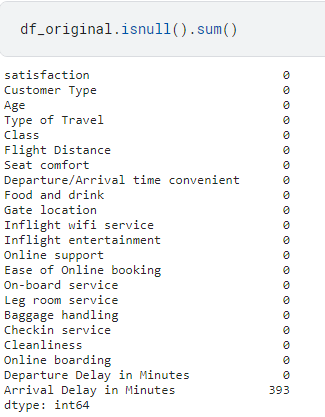


**Key insights:**

* There were 71,087 satisfied customers and 58,793 dissatisfied customers.
* 54.7 percent (71,087/129,880) of customers were satisfied. While this is a simple calculation, this value can be compared to a logistic regression model's accuracy.

### Check for missing values

An assumption of logistic regression models is that there are no missing values.



**Key insight:**

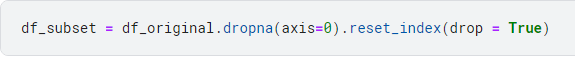
For this activity, I am specifically interested in knowing if a better in-flight entertainment experience leads to higher customer satisfaction. The Arrival Delay in Minutes column won't be included in the binomial logistic regression model; however, someone might become interested in this column in the future.

For now, the missing values should be removed for two reasons:

* There are only 393 missing values out of the total of 129,880, so these are a small percentage of the total.
* This column might impact the relationship between entertainment and satisfaction.

### Drop the rows with missing values[¶](https://kkb-production.jupyter-proxy.kaggle.net/static/dist/jupyterlab/jupyterlab-index-7d6feb0f1d697b8f72d0.html?session=eyJhbGciOiJub25lIiwidHlwIjoiSldUIn0.#Drop-the-rows-with-missing-values)

I drop the rows with missing values and save the resulting pandas DataFrame in a variable named df\_subset



### Prepare the data

I want to create a plot (sns.regplot) of my model to visualize results later in the notebook. So the independent variable Inflight entertainment cannot be "of type int" and the dependent variable satisfaction cannot be "of type object."

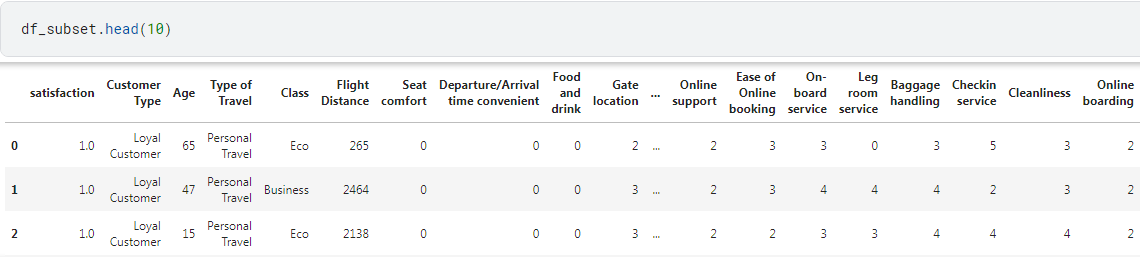


### Convert the categorical column satisfaction into numeric

I convert the categorical column satisfaction into numeric through one-hot encoding.

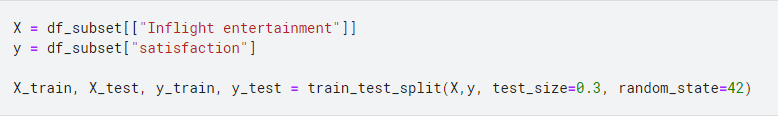


### Output the first 10 rows of df\_subset



### Create the training and testing data

I put 70% of the data into a training set and the remaining 30% into a testing set.



**Considerations**

**My findings:**

* Logistic regression accurately predicted satisfaction 80.2 percent of the time.
* The confusion matrix is useful, as it displays a similar amount of true positives and true negatives.

**Recommendations:**

* Customers who rated in-flight entertainment highly were more likely to be satisfied. Improving in-flight entertainment should lead to better customer satisfaction.
* The model is 80.2 percent accurate. This is an improvement over the dataset's customer satisfaction rate of 54.7 percent.
* The success of the model suggests that the airline should invest more in model developement to examine if adding more independent variables leads to better results. Building this model could not only be useful in predicting whether or not a customer would be satisfied but also lead to a better understanding of what independent variables lead to happier customers.