**Institute of Technology of Cambodia** 

**Department of Applied Mathematics and Statistic**

**Group:** I3-AMS-03

**Subject:** Advance Probability

**Topic**: Analyzing the factors that affect customer satisfaction in the airline industry using ordinal logistic regression Name: ID:

Tang Piseth e20201634

Kry Senghort e20200706

Sao Samarth e20200084

Hon Rathana e20201053

Pheng Sothea e20201734

Ek Vong Panharith e20200877

**Table of contents:**

1.

**Project Description and Objectives** 2.

**Dataset and Variables Description** 3.

**Data Preparation**

4.

**Exploratory Data Analysis(EDA)** 5.

**Feature Engineering**

6.

**Model Selection**

7.

**Decision Making**

**I.) Project Description and Objectives**

Airline Industry is the field of journey tourism or visiting to any places by using airline for travelling from one place to another one. On the other hand, We choose this topic beacause we want to use logistic linear regression to interpret and make prediction the result of satisfaction or disatifaction of passenger on airline industry. We will study on the variable that make effect to satisfaction or disatisfaction of customer.



**2). Dataset and Variables Description**

There are two data sets such as : test.csv and train.csv. There exist 129879 rows with 24 columns. There exists 24 variables as following:

Gender (object)

Customer Type (object)

Age (int64)

Type of Travel (object)

Class (object)

Flight Distance (int64)

Inflight wifi service (int64)

Departure/Arrival time convenient (int64) Ease of Online booking (int64)

Gate location (int64)

Food and drink (int64)

Online boarding (int64)

Seat comfort (int64)

Inflight entertainment (int64)

On-board service (int64)

Leg room service (int64)

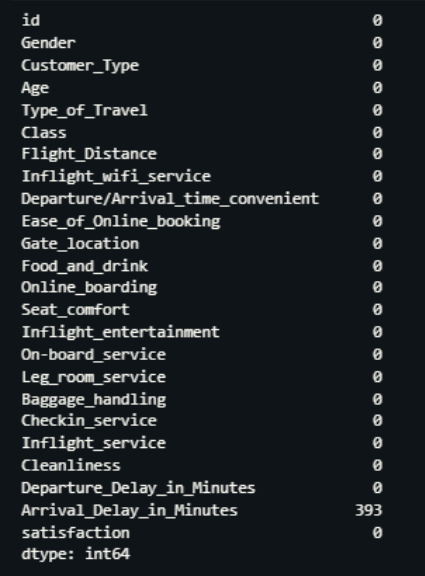
Baggage handling (int64)

Checkin service (int64)

Inflight service (int64)

Cleanliness (int64)

Departure Delay in Minutes (int64) Arrival Delay in Minutes (float) Satisfaction (object)(target variable)

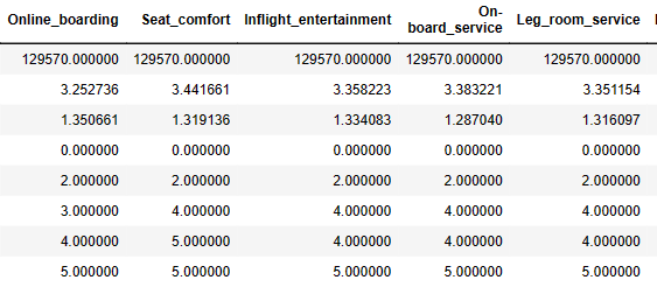
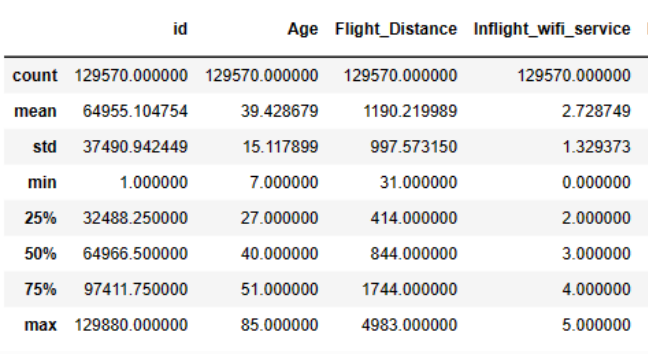
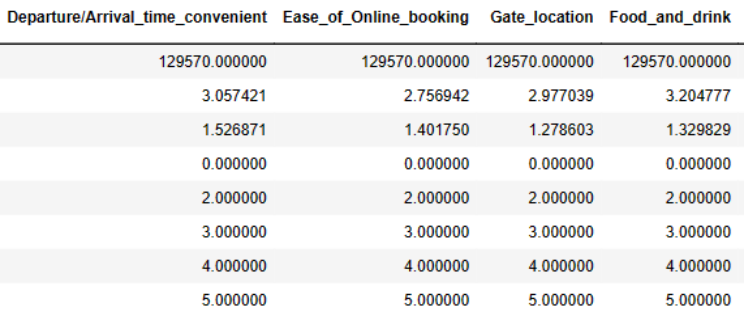
**3). Data Preparation** 

After checking to our dataset we observe that there exist some missing values and also we decide to handle it by substituting with the mean values for numerical variable and mode for the categorical variables. As the result, feature "Arrival\_Delay\_in\_Minutes" existing 393 missing values.

**Descriptive Statistics Before Handle Missing**

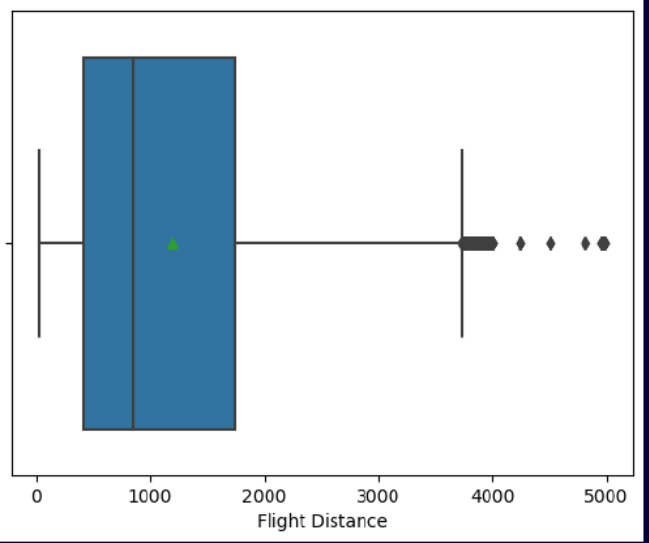
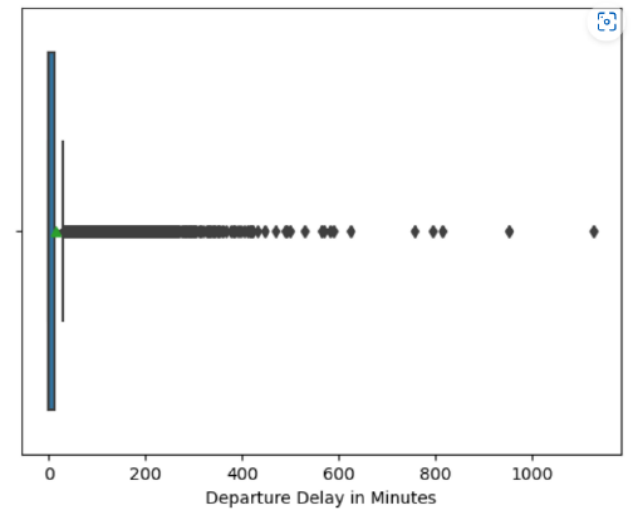
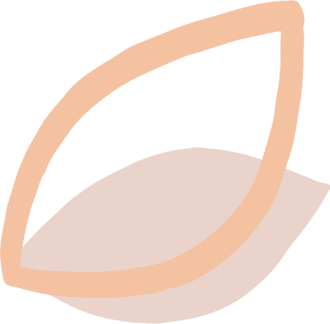
**4). Exploratory Data Analysis (EDA)**

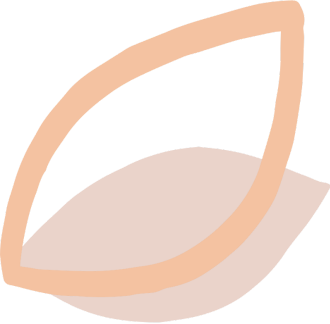
**Statistic Table**

****

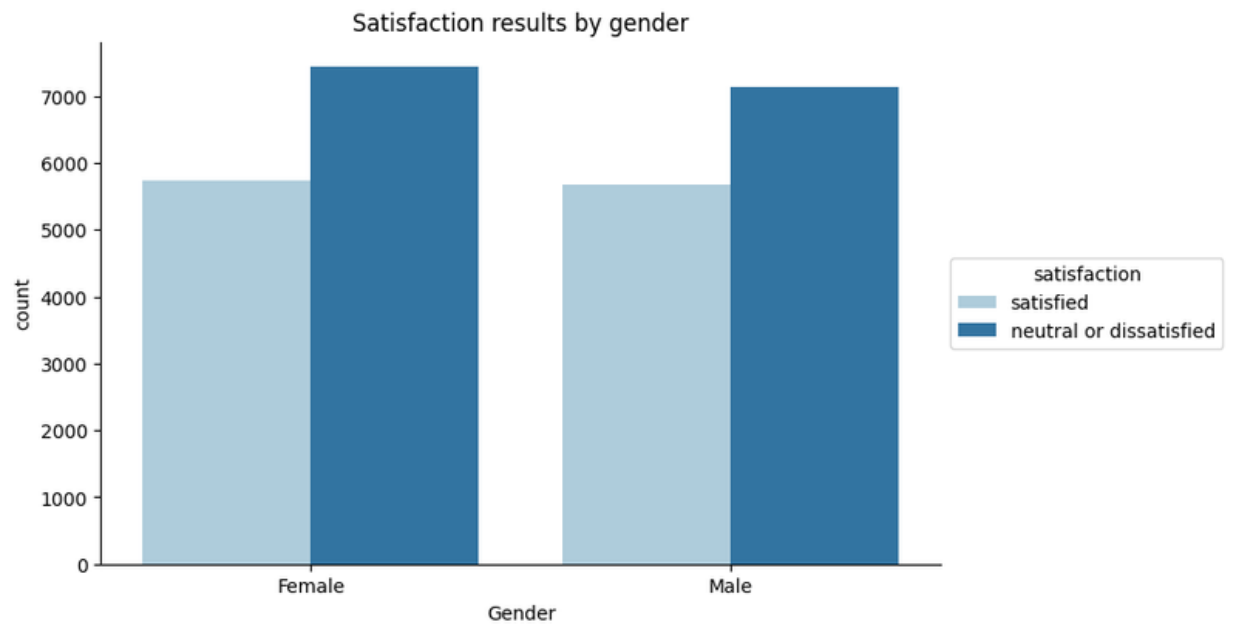
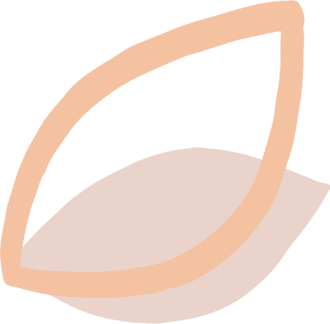
**Heat map**

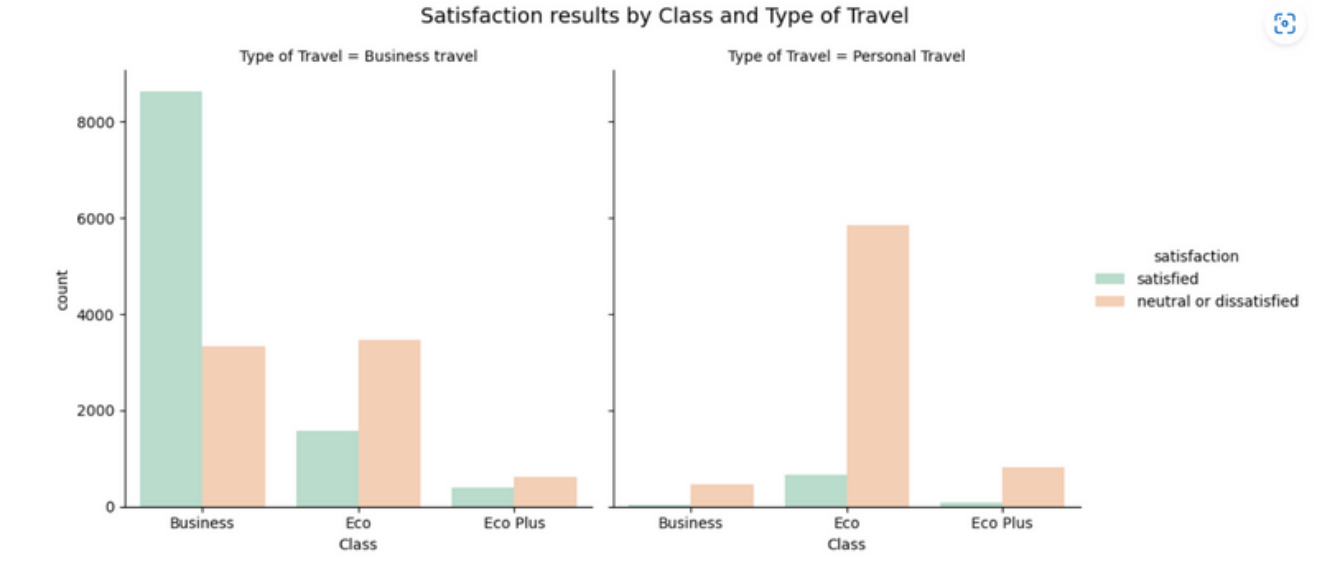
**Box plot**

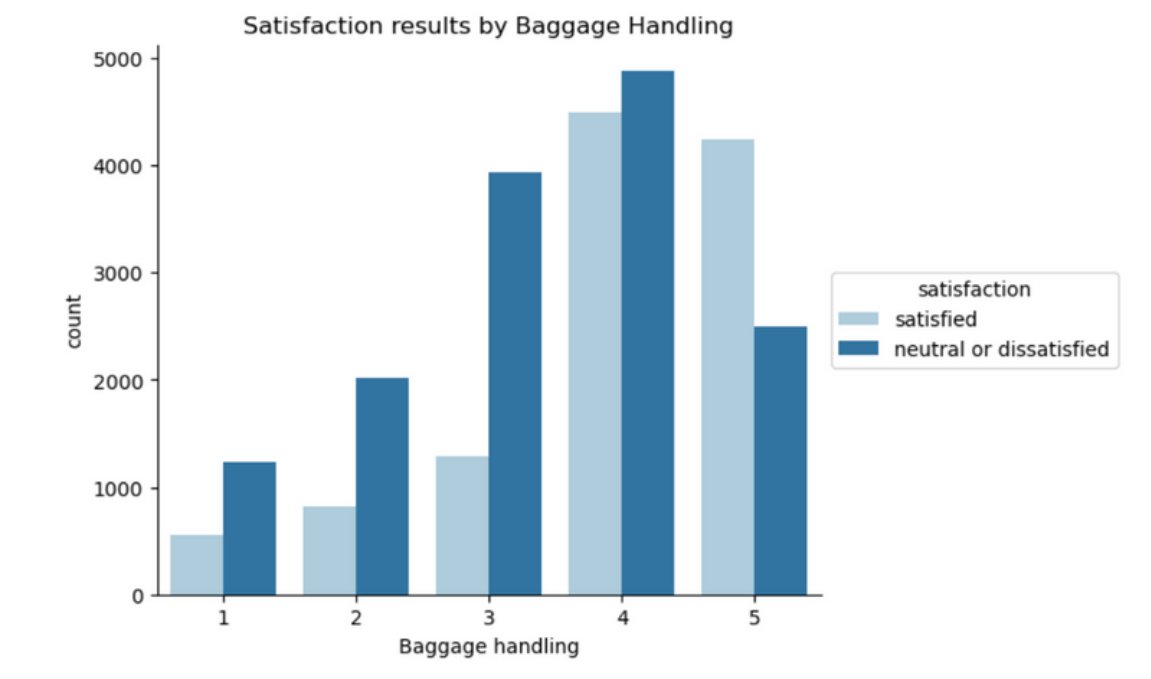
****

****

**Bar graph**

****

****

****

****

****

****

****

****

**Pair pot**

**Pie chat**

****

****

**5). Feature Engineering 5.1). Encoding Categorical Variables**

**In classification problems, the goal is to predict the category of a new observation based on a set of training data. However, machine learning algorithms cannot directly work with categorical variables. They can only work with numerical data. Therefore, it is necessary to encode categorical variables into numerical values before they can be used in a classification model.**

****

****

**5.2). Feature Selection and Feature Importance**

****

**Setting our categorical**

**& target variable** 

****

**Feature Selection by Chi-Sqaure**

**Test**

****

**Feature Importance by**

**Wrapper Method**

**(random forest)**

****

**Feature Importance by**

**Permutation Method**

****

****

****

**Feature Importance by Decision Tree**

****

**Feature Importance by**

**Recursive Feature Elimination**

****

**5.3).Dimentionality Reduction**

****

PerformModel Algorithm

**Linear regression analysis** 

satisfaction ~ Departure\_Delay\_in\_Minutes + Arrival\_Delay\_in\_Minutes

Kimmy 

satisfaction ~ Arrival\_Delay\_in\_Minutes









**More Details**

satisfaction ~ Departure\_Delay\_in\_Minutes



**Multicolinearity**

****Both 'Departure\_Delay\_in\_Minutes' and 'Arrival\_Delay\_in\_Minutes' have relatively high VIF factors of approximately 12.565103. VIF values above 5 or 10 are often considered indicative of significant multicollinearity





**ROC Curve of RFC**

****

**6). Model Selection**

Model1: Normal Logistic Regression

(all features)

Model2: Logistic Regression Penalized with Lasso (regularization-strength = 0.1)

Model3: Logistic Regression Penalized with Ridge (L2 penalty = 50%)

Model4: Logistic Regression Penalized with Elastic Net (L1 penalty = 50%,

L2 penalty = 50%)

Model5: Decision Tree

Medel6: Random Forest

**7). Decision Making**

**In terms of algorithmic complexity, it is best that we go we random forest classifier. However, in terms of logistic regression model, we select the mentioned features.**

**The important factor that make effect to customer satisfaction are :Seat\_comfort**

**Inflight\_entertainment**

**class**

**Type\_of\_Travel**

**Inflight\_wifi\_service**

**Online\_boarding**

Thank you