S.K Somaiya College of Arts, Science and Commerce
Somaiya Vidyavihar University
Vidyavihar, Mumbai 400 077

MSc Statistics Part-2
SEM - III

Submitted by
Varsha Jadhav (31031820020)
Niki Mehta (31031820027)
Jyothi Puligilla (31031820038)
Vedang Sawant (31031820041)
Moheed Tai (31031820044)
Vishal Yadav (31031820046)

MENTORED BY
Prof. Jaishankar Singh

INDEX:-

I. INTRODUCTION:-

Airlines industry is currently the biggest industry of transportation in the world. It focuses in service strategy to get costumers, because a great quality of service is a way to get customer loyalty. The objectives of this study are to examine and to identify the ways to increase airline customer satisfaction with service quality dimension, consisting of tangibility, reliability, responsiveness, assurance, and empathy. The main purpose of this research is to achieve customer satisfaction in order to get loyalty from airline customers by improving the service quality dimensions, especially in Responsiveness, Reliability and Empathy.

Transportation services have become the basic needs of the community both for daily activities and travel needs. For a long-distance journey, most people prefer air transportation for the efficiency and effectiveness of time. Air transport can reach places that cannot be reached by other modes of transport such as land and sea, in addition to being able to move faster and have a straight, practically barrier-free path.

Satisfaction is not only considered as a customer's goal to be derived as a result of degrading services, but also as a company's goal, as a way of getting higher customer retention rates and ways of generating profits . If the service / product provided in accordance with customer expectations, he will feel satisfied that increase the level of consumer loyalty. Conversely, if service delivery is lower than customer expectation, service quality will be considered bad and decrease of customer loyalty. One way to increase customer loyalty is to improve the quality of service. Customer loyalty is considered as commitment or costumer principle to always choose the service / product is continuous and consistent in the future .

In theory, if a customer is satisfied with the service or product provided, he will be loyal to use the product and even tells others the benefits of the product or service. Satisfied customers will continue to buy the product again. Although , customer satisfaction is not the main goal, customer satisfaction is the key to the success of a company to maintain the quality of product / service and maintain the image of the company's brand so that customers will repurchase. That is why, the company must provide superior service quality to win the business competition among Airlines.

This study aims to examine and analyse the impact of service quality dimensions consisting of tangibility, reliability, responsiveness, assurance, and empathy in improving the quality of service of the airlines. It also aims to analyse which quality dimensions dominantly affect the service quality in the airlines. The main purpose of this study is to know and analyse to what extend the suitability between the level of importance of the elements of service according to the customer / passengers and what has been done by the airline is considered satisfactory and to find out and analyse the elements of dimensions that need to be improved in order to improve customer satisfaction impact on customer loyalty.

II. OBJECTIVES:-

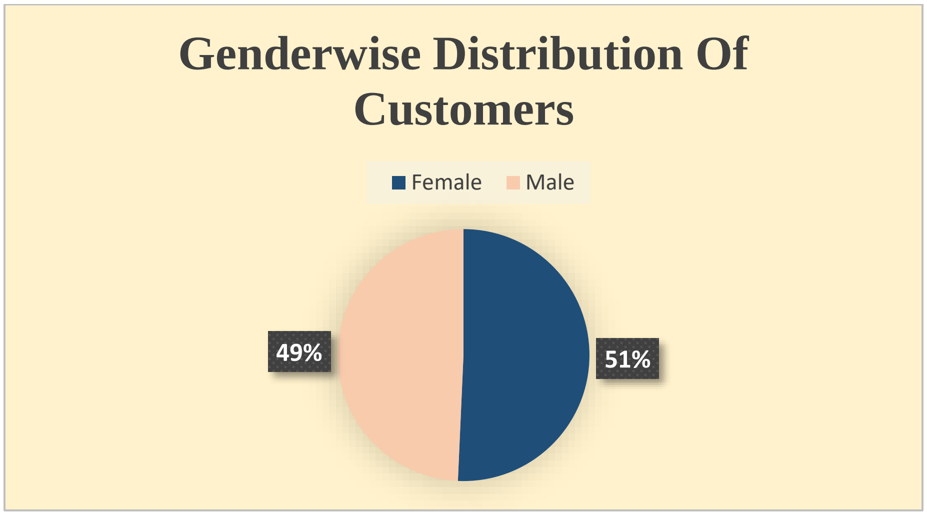
- To study the distribution of customers with respect to their gender
- To study the number of loyal and disloyal customers to the airline
- To study the number of satisfied and dissatisfied customers based on their loyalty
- To study the number of satisfied and dissatisfied customers with the service
- To study the distribution of customers based on type of travel
- To check whether there is a relation between Leg room service and Seat comfort with respect to satisfaction
- To test the independence between the Customer type and the class of travel
- To fit a predictive model to check whether a customer is satisfied with the service
- Prediction of passenger satisfaction using ANN

III. METHODOLOGY:-

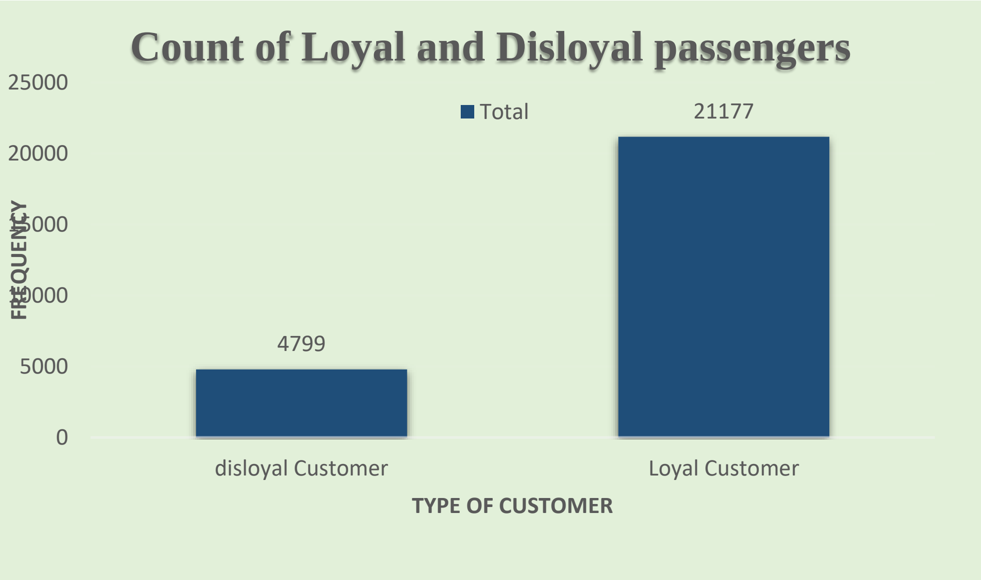
The First Step was to get the data . The Airlines data was extracted from Kaggle. It consisted of 25975 data points having 24 variables. There were 83 missing values . A Sample of 10% i.e, 2598 data points was taken using the technique of simple random sampling. Then the missing values were replaced with median value. After that Outliers were treated.

Finally, we got dataset of 2057 points.

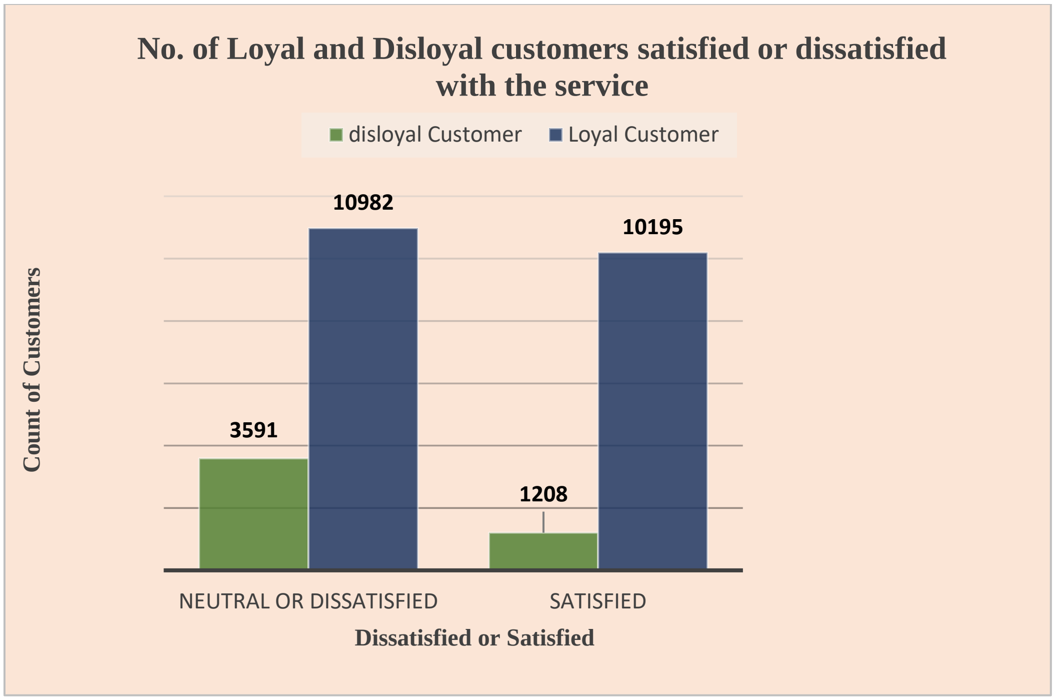
IV. EXPLORATORY DATA ANALYSIS:-



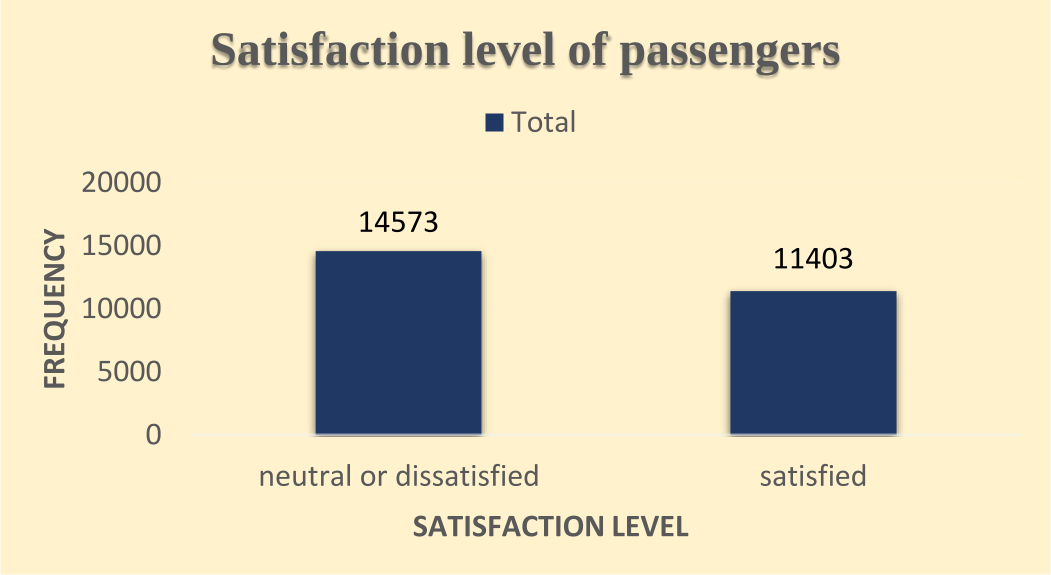
From above figure , we see that there are more number of female passengers i.e. 51% than male passengers i.e. 49%.



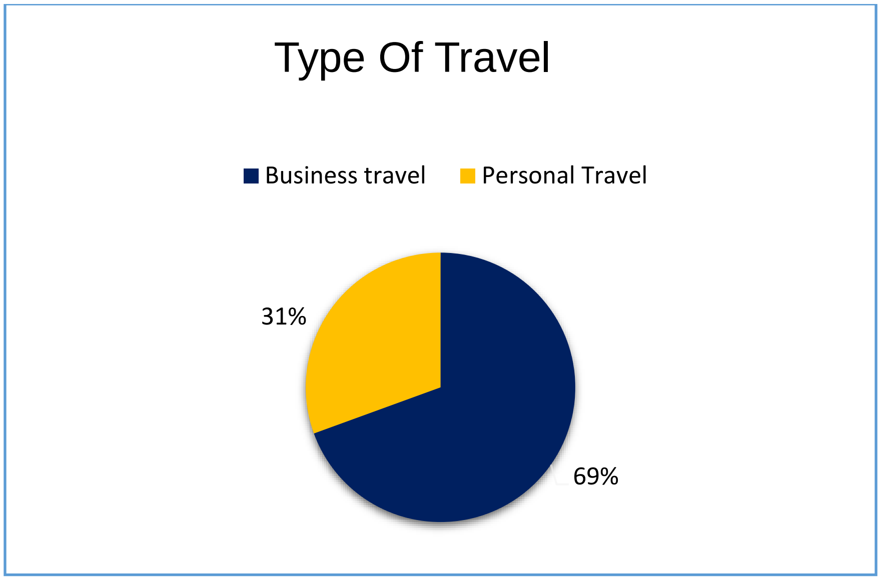
From above figure, we see that there are more number of loyal customers than disloyal customers.



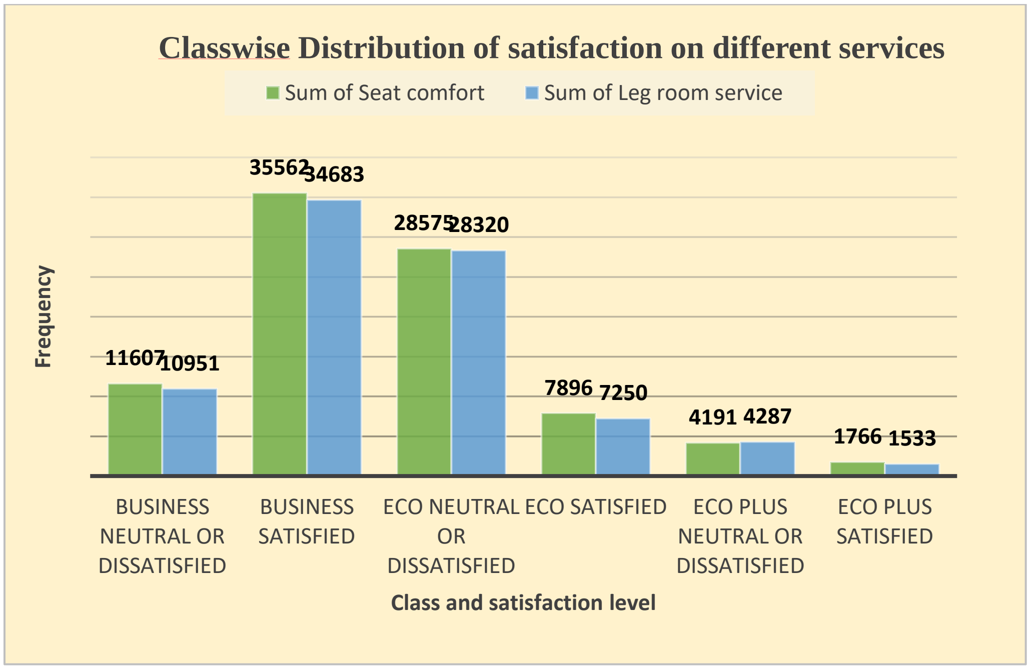
From above figure ,we see that in both the cases of satisfied and dissatisfied the no. of loyal customers are more than the no of disloyal customers.



From above figure ,we see that there are more dissatisfied customers than satisfied customers with the service.

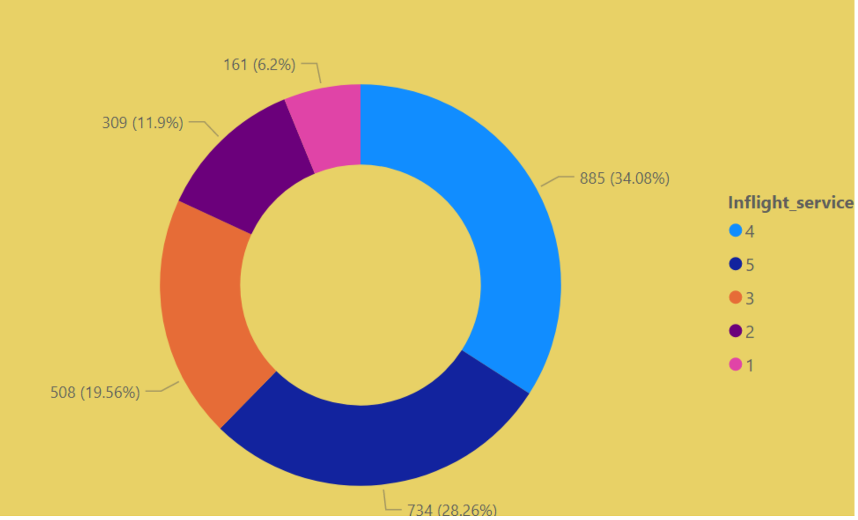


From above figure, we see that there are more number of customers for Business travel i.e. 69% than the number of customers for Personal travel i.e. 31%.



From above figure, we see that in all the classes the number of customers satisfied or dissatisfied with seat comfort is similar to the number of customers satisfied or dissatisfied with the leg room service. So we can say that there is a relation between these two services.

RATING GIVEN BY THE PASSENGER

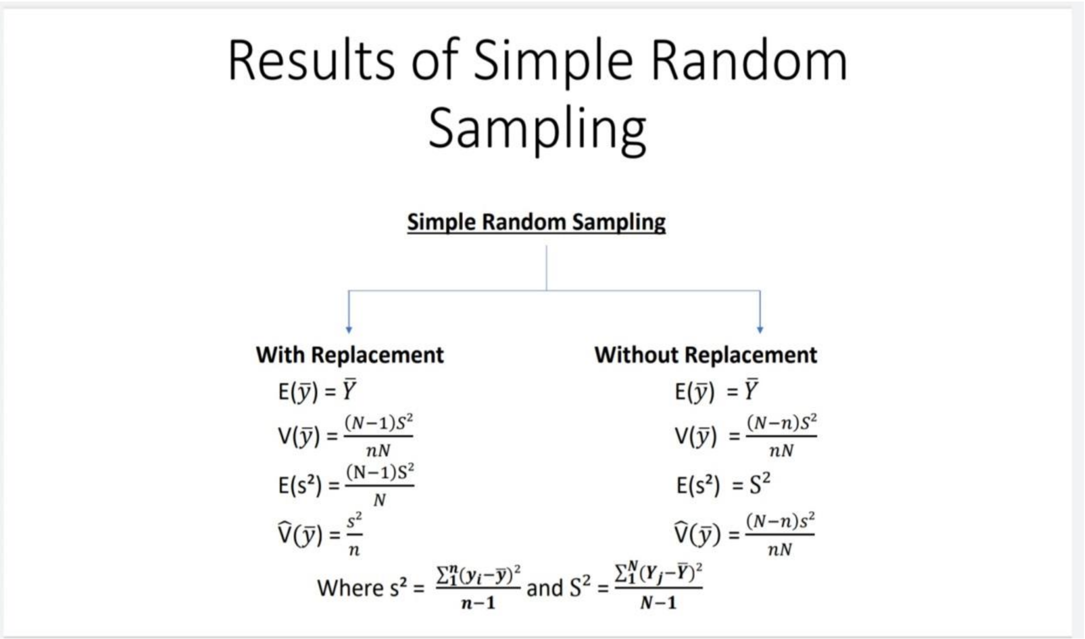


The above figure tells us that out of all the customers, 34.08% of them gave 4 point rating for inflight service.

IV. ANALYSIS AND MODELS

a) SIMPLE RANDOM SAMPLING:-

- If each population unit has the same probability (chance) of being selected in the sample then the procedure of selecting sample is called as Simple Random Sampling. Sample thus selected is called as Simple Random Sample.
- In practice, units in the sample are selected one by one.
- If a unit of the population is not allowed to appear more than once in the sample, the sampling is termed as Simple Random Sampling Without Replacement (SRSWOR).



Results of Simple Random Sampling

Simple Random Sampling

With Replacement

E($\bar{y}$) = $\bar{Y}$

V($\bar{y}$) = $\frac{(N-1)S^2}{nN}$

E(s^2) = $\frac{(N-1)S^2}{N}$

$\hat{V}(\bar{y})$ = $\frac{s^2}{n}$

Without Replacement

E($\bar{y}$) = $\bar{Y}$

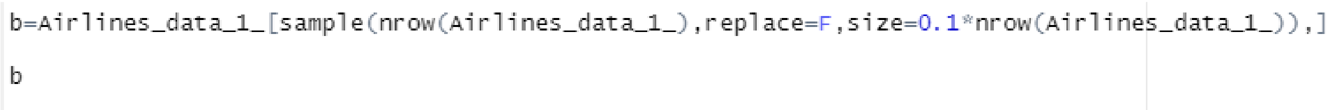
V($\bar{y}$) = $\frac{(N-n)S^2}{nN}$

E(s^2) = S^2

$\hat{V}(\bar{y})$ = $\frac{(N-n)s^2}{nN}$

Where s^2 = $\frac{\sum_1^n (y_i-\bar{y})^2}{n-1}$ and S^2 = $\frac{\sum_1^N (Y_j-\bar{Y})^2}{N-1}$

By the following r code we generated a sample of 10%



```
b=Airlines_data_1_[sample(nrow(Airlines_data_1_),replace=F,size=0.1*nrow(Airlines_data_1_)),]
b
```

A sample of size 2598 was generated from 25975 data points randomly.

b) Chi-square Distribution:-

The chi-squared distribution (also chi-square or χ^2-distribution) with *k* degrees of freedom is the distribution of a sum of the squares of *k* independent standard normal random variables. The chi-square distribution is a special case of the gamma distribution and is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing or in construction of confidence intervals. The chi-squared distribution is used in the common chi-squared tests for goodness of fit of an observed distribution to a theoretical one, the independence of two criteria of classification of qualitative data, and in confidence interval estimation for a population standard deviation of a normal distribution from a sample standard deviation. Many other statistical tests also use this distribution, such as Friedman's analysis of variance by ranks. It is also a component of the definition of the t-distribution and the F-distribution used in t-tests, analysis of variance, and regression analysis.

Test for Independence of Attributes:-

In probability theory, two events are independent, statistically independent, or stochastically independent if the occurrence of one does not affect the probability of occurrence of the other. Similarly, two random variables are independent if the realization of one does not affect the probability distribution of the other. The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables. The chi square test for independence of two variables uses a cross classification table to examine the nature of the relationship between these variables. These tables are sometimes referred to as contingency tables. These tables show the manner in which two variables are either related or are not related to each other. The test for independence examines whether the observed pattern between the variables in the table is strong enough to show that the two variables are dependent on each other or not. The chi square test of independence is very general, and can be used with variables measured on any type of scale, nominal, ordinal, interval or ratio. The only limitation on the use of this test is that the sample sizes must be sufficiently large to ensure that the expected number of cases in each category is five or more. This rule can be modified somewhat, but as with all approximations, larger sample sizes are preferable to smaller sample sizes. There are no other limitations on the use of the test, and the chi square statistic can be used to test any contingency or cross classification table for independence of the two variables.

HYPOTHESIS:-

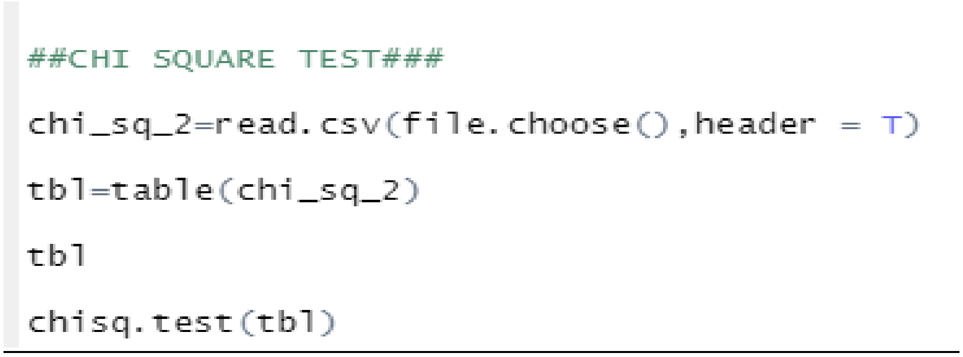
The chi square test for independence is conducted by assuming that there is no relationship between the two variables being examined. The alternative hypothesis is that there is some relationship between the variables.

H_0 : There is no association between Customer type and Class

v/s

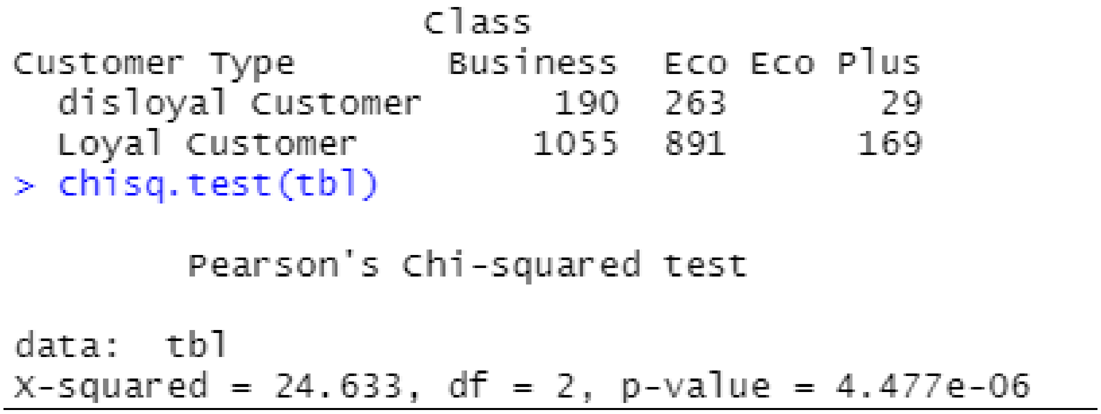
H_1 : There is an association between Customer type and Class

CALCULATIONS:-



```
##CHI SQUARE TEST###

chi_sq_2=read.csv(file.choose(),header = T)

tbl=table(chi_sq_2)

tbl

chisq.test(tbl)
```

OUTPUT:-



```
                   Class
Customer Type       Business  Eco Eco Plus
  disloyal Customer      190  263       29
  Loyal Customer        1055  891      169
> chisq.test(tbl)

        Pearson's Chi-squared test

data:  tbl
X-squared = 24.633, df = 2, p-value = 4.477e-06
```

CONCLUSION:-

From our R-code, we get the p-value to be 4.477×10^{-6}; which is less than 0.05. Therefore, we reject our null hypothesis.

Thus, there is an association between the Customer type and Class.

c) LOGISTIC REGRESSION:-

Logistic regression is a statistical model that in its basic form uses a logistic function to model a binary dependent variable, although many more complex extensions exist. In regression analysis, **logistic regression** (or **logit regression**) is estimating the parameters of a logistic model (a form of binary regression). Mathematically, a binary logistic model has a dependent variable with two possible values, such as pass/fail which is represented by an indicator variable, where the two values are labelled "0" and "1". In the logistic model, the log-odds (the logarithm of the odds) for the value labelled "1" is a linear combination of one or more independent variables ("predictors"); the independent variables can each be a binary variable (two classes, coded by an indicator variable) or a continuous variable (any real value).

Outputs with more than two values are modelled by multinomial logistic regression and, if the multiple categories are ordered, by ordinal logistic regression

The unit of measurement for the log-odds scale is called a logit, from ***log***_istic un_***it***, hence the alternative names

Logistic regression is an extension of simple linear regression. Where the dependent variable is dichotomous or binary in nature, we cannot use simple linear regression. Logistic regression is a statistical technique used to predict the relationship between predictors (our independent variables) and a predicted variable (the dependent variable) where the dependent variable is binary.

Model:

$$ln\left(\frac{\text{Prob}(Y_i = 1)}{1 - \text{Prob}(Y_i = 1)}\right) = \beta_0 + \beta_1 X_{ij} + \cdots + \beta_k X_{ik}.$$

In our model,

Dependent variable:

Satisfaction or Dissatisfaction with the service.

Values: ‘1’ , ‘0’

Independent Variables:

1. **Inflight wifi service:** Wifi service of the flight.

   Values: 0 (dissatisfied) – 5 (satisfied)

2. **Leg room service:** The space your legs have between your seat and the seat in front of you.

   Values: 0 (dissatisfied) – 5 (satisfied)

3. **Online boarding:** Online check-in in which passengers confirm their presence on a flight via the internet.

   Values: 0 (dissatisfied) – 5 (satisfied)

4. **Departure/Arrival time convenient:** Convience of departure and arrival time.

   Values: 0 (dissatisfied) – 5 (satisfied)

5. **Gate location:** Location of gate at the airport.

   Values: 1 (dissatisfied) – 5 (satisfied)

6. **Inflight entertainment:** Entertainment available to aircraft during a flight.

   Values: 1 (dissatisfied) – 5 (satisfied)

7. **On-board service:** Service of in-flight food or beverages purchased on board, or ordered in advance as an optional extra during or after booking process.

   Values: 2 (dissatisfied) – 5 (satisfied)

8. **Checkin service:** Ensuring that the passenger will be on board with the airline.

   Values: 2 (dissatisfied) – 5 (satisfied)

d) MULTICOLLINEARITY THEORY:-

Multicollinearity refers to a situation in which two or more explanatory variable multiple regression model are highly linearly related. An easy way to detect multicollinearity is to calculate correlation coefficients for all pairs of predictor variables. If the correlation coefficient, r, is exactly +1 or -1, this is called perfect multicollinearity. If r is close to or exactly -1 or +1, one of the variables should be removed from the model if at all possible.
The classical linear regresssion model assumes that regressors are not correlated i.e. there is no milticollinearity among the regressors included in the regression model. The term multicollinearity originally meant the existence of exact linear relationship among some or all explanatory variables in the model i.e. the following condition is satisfied $\beta_1X_{1+}\ \beta_2X_2 +\ldots+ \beta_kX_k$ =0 where lambda's are the constants. However, the term multicollinearity is used in a broader sense to include explanatory variables that are inter correlated but not perfectly so. In such a situation $\beta_1X_{1+}\ \beta_2X_2 +\ldots+ \beta_kX_k+v= 0$. If multicollinearity exists, the regression coefficients possess large standard errors which means coefficients cannot be estimated with great precision or accuracy.

DETECTION OF MULTICOLLINEARITY:-

Indicators that multicollinearity may be present in a model include the following:

1) Some authors have suggested a formal detection-tolerance or the variance inflation factor (VIF) for multicollinearity:

   VIF=1/(1-R^2)

   where R^2 is the coefficient of determination of a regression of explanator on all the explanators. A tolerance of less than 0.20 or 0.10 and/or a VIF of 5 or 10 and above indicates a multicollinearity problem.

2) R^2 and t ratios: If R^2 is high (>0.8) then F test will generally reject the null hypothesis ($H_0:\beta_1= \beta_2=\ldots=\beta_k=0$) but the individual t tests ($H_0:\beta_i=0$ v/s $H_1:\beta_i\neq0$) show that none or very few of the regression coefficients are significant then multicollinearity may exist.

3) REMEDIAL MEASURES

- Collect more data to reduce multicollinearity
- Remove one of the regressors having high VIF from the model and again check for multicollinearity

- Combine two or more regressors having high correlation

SPEARMAN’S CORRELATION:-

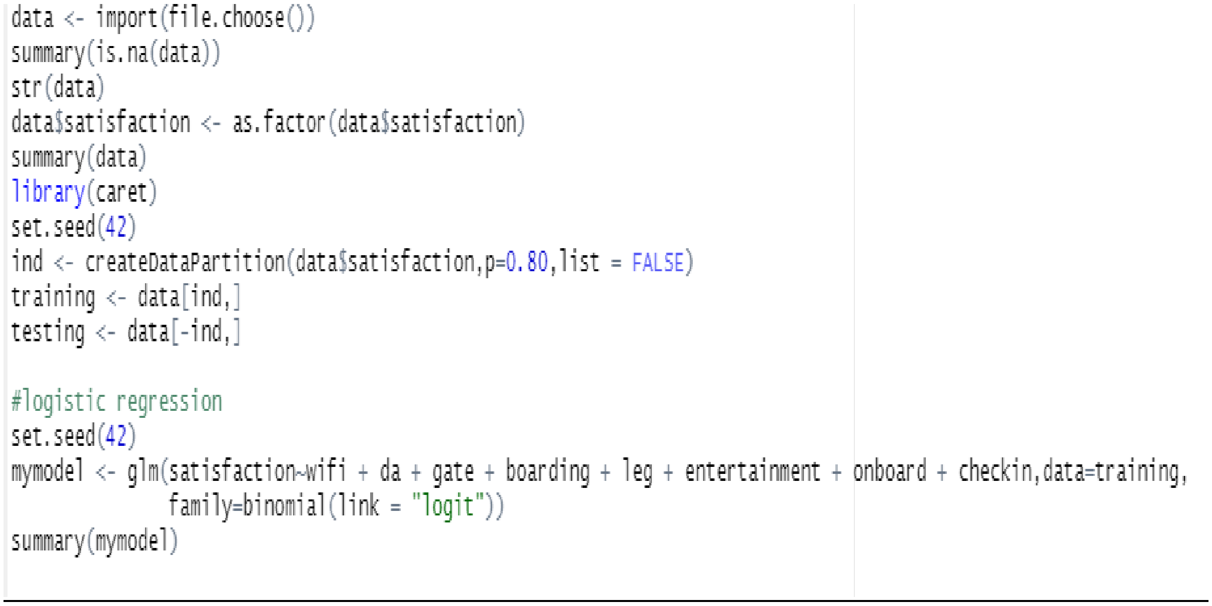
- **Spearman's rank correlation coefficient** or **Spearman's *ρ*** is a non-parametric measure of rank correlation.
- **Spearman's Rank correlation coefficient** is a technique which can be **used** to summarise the strength and direction (negative or positive) of a relationship between two variables. The result will always be between 1 and minus 1.
- The Spearman's Rank Correlation Coefficient is used to discover the strength of a link between two sets of data.
- Spearman's coefficient is appropriate for both continuos and discrete ordinal variable

<u>Testing multicollinearity: Spearman’s Correlation</u>

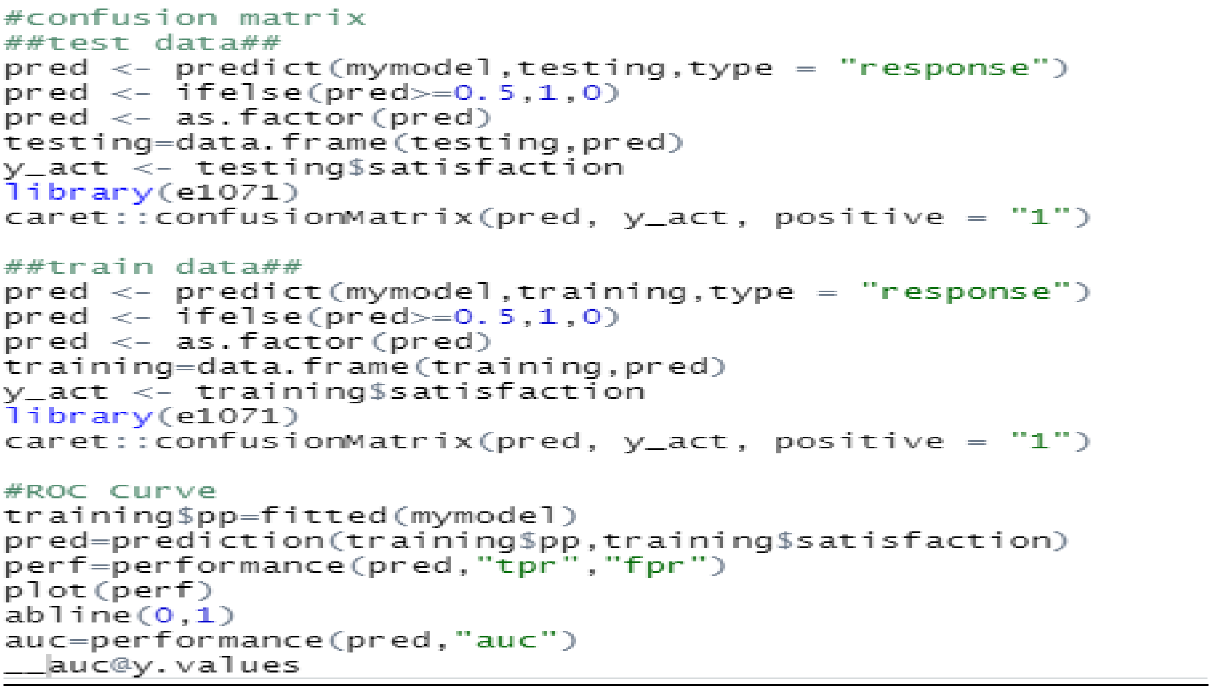
	Inflight wifi service	Leg room service	Online boarding	Departure/ Arrival time convenient	Gate location	Inflight entertain ment	On-board service	Checkin service
Inflight wifi service	1							
Leg room service	0.1117	1						
Online boarding	0.4363	0.1380	1					
Departure/Arrival time	0.3794	-0.0067	0.0421	1				
Gate location	0.4007	0.0273	0.0075	0.4286	1			
Inflight entertainment	0.1534	0.2806	0.3156	-0.0374	0.0378	1		
On-board service	0.1209	0.3511	0.1299	0.0605	0.0199	0.4064	1	
Checkin service	0.1091	0.0939	0.1371	0.0319	-0.0638	0.0438	0.1425	1

We see that there is very weak collinearity among the independent variables. Hence, we can go ahead with our logistic model.

CODES:-

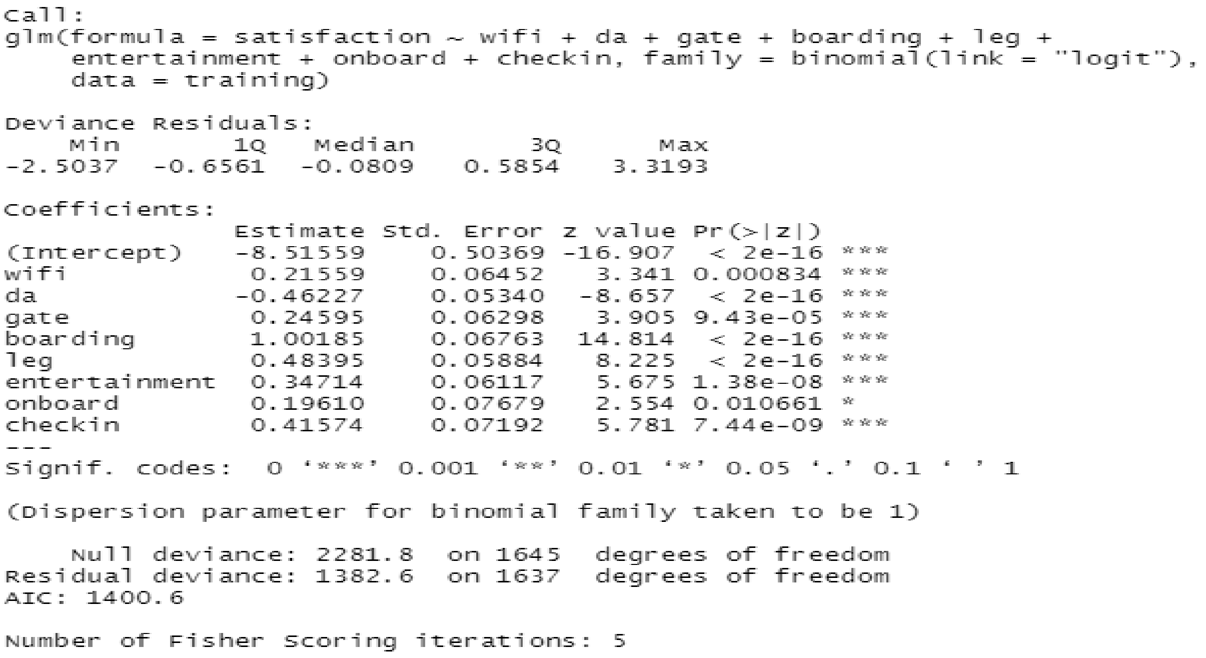


```
data <- import(file.choose())
summary(is.na(data))
str(data)
data$satisfaction <- as.factor(data$satisfaction)
summary(data)
library(caret)
set.seed(42)
ind <- createDataPartition(data$satisfaction,p=0.80,list = FALSE)
training <- data[ind,]
testing <- data[-ind,]

#logistic regression
set.seed(42)
mymodel <- glm(satisfaction~wifi + da + gate + boarding + leg + entertainment + onboard + checkin,data=training,
               family=binomial(link = "logit"))
summary(mymodel)
```



```
#confusion matrix
##test data##
pred <- predict(mymodel,testing,type = "response")
pred <- ifelse(pred>=0.5,1,0)
pred <- as.factor(pred)
testing=data.frame(testing,pred)
y_act <- testing$satisfaction
library(e1071)
caret::confusionMatrix(pred, y_act, positive = "1")

##train data##
pred <- predict(mymodel,training,type = "response")
pred <- ifelse(pred>=0.5,1,0)
pred <- as.factor(pred)
training=data.frame(training,pred)
y_act <- training$satisfaction
library(e1071)
caret::confusionMatrix(pred, y_act, positive = "1")

#ROC Curve
training$pp=fitted(mymodel)
pred=prediction(training$pp,training$satisfaction)
perf=performance(pred,"tpr","fpr")
plot(perf)
abline(0,1)
auc=performance(pred,"auc")
auc@y.values
```

SUMMARY OF MODEL:-



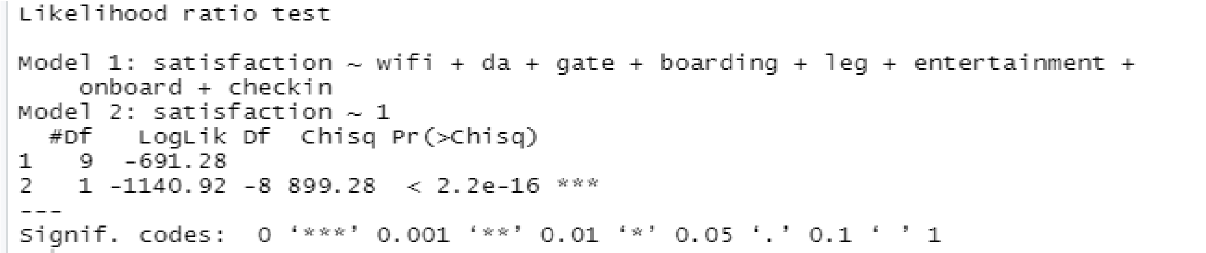
```
Call:
glm(formula = satisfaction ~ wifi + da + gate + boarding + leg +
    entertainment + onboard + checkin, family = binomial(link = "logit"),
    data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5037  -0.6561  -0.0809   0.5854   3.3193

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -8.51559    0.50369 -16.907  < 2e-16 ***
wifi           0.21559    0.06452   3.341 0.000834 ***
da            -0.46227    0.05340  -8.657  < 2e-16 ***
gate           0.24595    0.06298   3.905 9.43e-05 ***
boarding       1.00185    0.06763  14.814  < 2e-16 ***
leg            0.48395    0.05884   8.225  < 2e-16 ***
entertainment  0.34714    0.06117   5.675 1.38e-08 ***
onboard        0.19610    0.07679   2.554 0.010661 *
checkin        0.41574    0.07192   5.781 7.44e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2281.8  on 1645  degrees of freedom
Residual deviance: 1382.6  on 1637  degrees of freedom
AIC: 1400.6

Number of Fisher Scoring iterations: 5
```

GLOBAL TESTING:-

HYPOTHESIS:-

H_o: None of the independent variables has significant impact on the dependent variable.

H_1: At least one independent variable is significant.

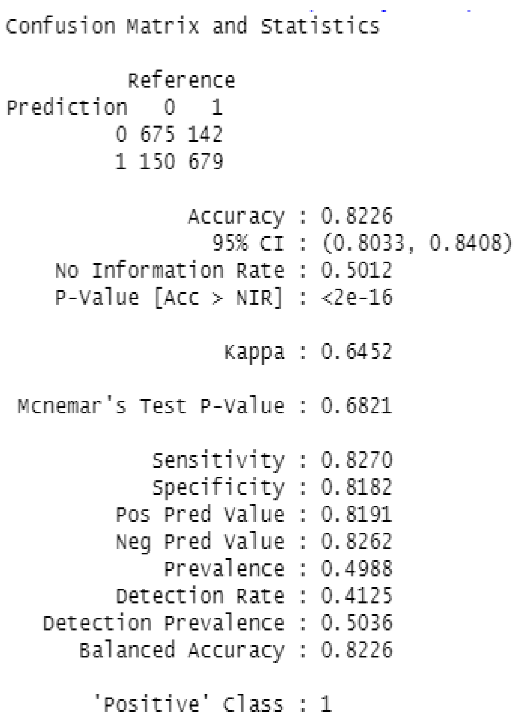


```
Likelihood ratio test

Model 1: satisfaction ~ wifi + da + gate + boarding + leg + entertainment +
    onboard + checkin
Model 2: satisfaction ~ 1
  #Df   LogLik Df  Chisq Pr(>Chisq)
1   9  -691.28
2   1 -1140.92 -8 899.28  < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

CONCLUSION:-We see that the p-value is 2.2e^-16. Hence, we reject our null hypothesis and conclude that our model is significant.

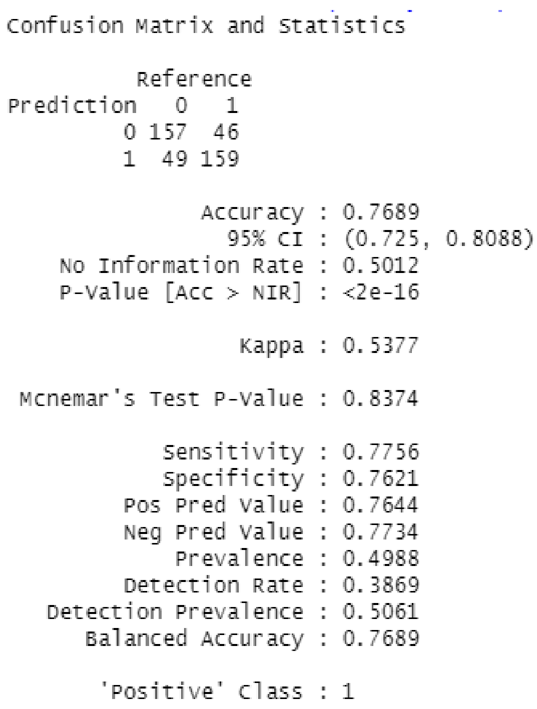
The data was divided into 80% training data and 20% testing data

With Train data:-



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 675 142
         1 150 679

               Accuracy : 0.8226
                 95% CI : (0.8033, 0.8408)
    No Information Rate : 0.5012
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.6452

 Mcnemar's Test P-Value : 0.6821

            Sensitivity : 0.8270
            Specificity : 0.8182
         Pos Pred Value : 0.8191
         Neg Pred Value : 0.8262
             Prevalence : 0.4988
         Detection Rate : 0.4125
   Detection Prevalence : 0.5036
      Balanced Accuracy : 0.8226

       'Positive' Class : 1
```

With Test data:-



```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 157  46
         1  49 159

               Accuracy : 0.7689
                 95% CI : (0.725, 0.8088)
    No Information Rate : 0.5012
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.5377

 Mcnemar's Test P-Value : 0.8374

            Sensitivity : 0.7756
            Specificity : 0.7621
         Pos Pred Value : 0.7644
         Neg Pred Value : 0.7734
             Prevalence : 0.4988
         Detection Rate : 0.3869
   Detection Prevalence : 0.5061
      Balanced Accuracy : 0.7689

       'Positive' Class : 1
```

ROC CURVE:-

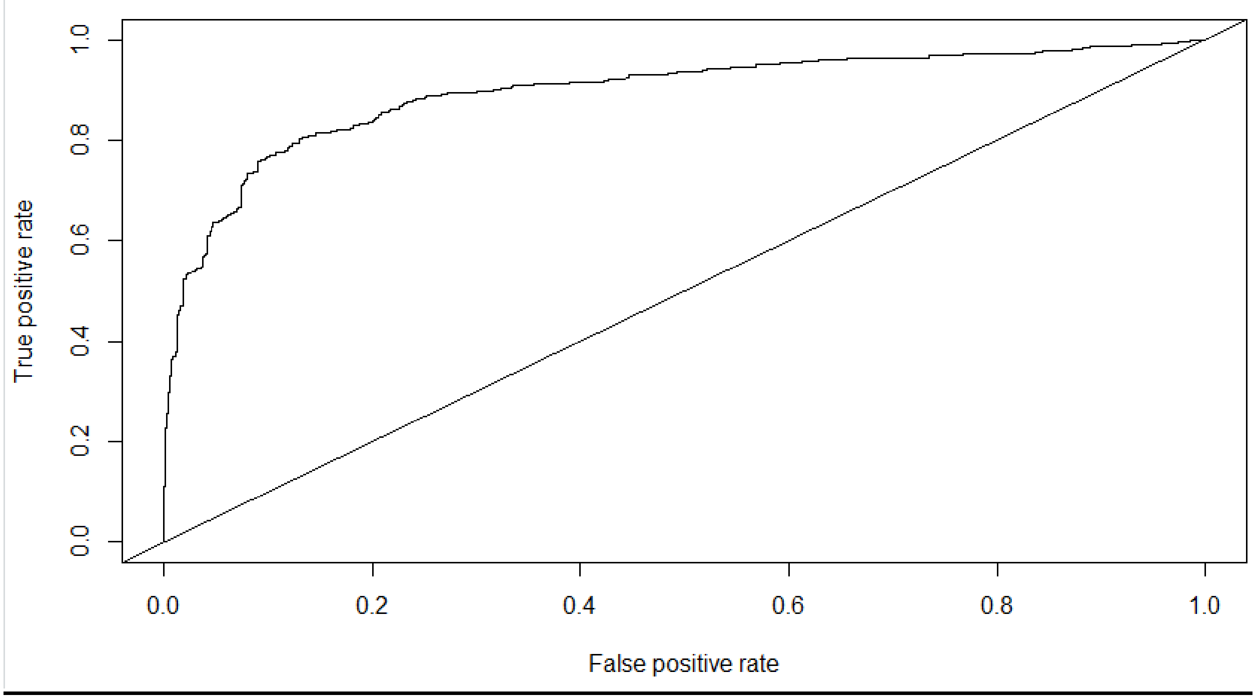
THEORY:-

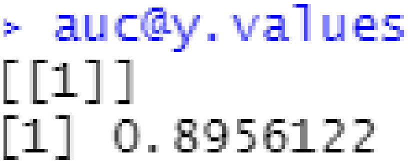
ROC:-
A ROC curve is constructed by plotting the true positive rate (TPR) against the false positive rate (FPR). The true positive rate is the proportion of observations that were correctly predicted to be positive out of all positive observations (TP/(TP + FN)). Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/(TN + FP)).

The ROC curve shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR). Classifiers that give curves closer to the top-left corner indicate a better performance.

AUC:-

To compare different classifiers, it can be useful to summarize the performance of each classifier into a single measure. One common approach is to calculate the area under the ROC curve, which is abbreviated to AUC. A classifier with high AUC can occasionally score worse in a specific region than another classifier with lower AUC. But in practice, the AUC performs well as a general measure of predictive accuracy.





```
> auc@y.values
[[1]]
[1] 0.8956122
```

CONCLUSION:-

Using global testing, we see that our model is significant. We get an accuracy of 82.26% on our train data and 76.89% on the test data.

From the AUC value we see that the area under the curve is 0.8956122. This means that 89.5% variability in the data is explained by our model.

e) Artificial Neural Network

Artificial neural networks (ANNs), usually simply called neural networks (NNs), are computing systems inspired by the biological neural networks that constitute animal brains.

An ANN is based on a collection of connected units or nodes called artificial neurons, which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit a signal to other neurons. An artificial neuron receives a signal then processes it and can signal neurons connected to it. The "signal" at a connection is a real number, and the output of each neuron is computed by some non-linear function of the sum of its inputs. The connections are called *edges*. Neurons and edges typically have a *weight* that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Neurons may have a threshold such that a signal is sent only if the aggregate signal crosses that threshold. Typically, neurons are aggregated into layers. Different layers may perform different transformations on their inputs. Signals travel from the first layer (the input layer), to the last layer (the output layer), possibly after traversing the layers multiple times.

VARIABLES ENCODED

1.Gender (female-0 , male-1)

2.Customer Type (disloyal customer-0 , loyal customer-1)

3.Type of travel (Personal travel-0 Business travel-1)

4.Class (Economic-0 , business-1 , economic plus-2)

5.Satisfaction (Neutral or dissatisfied-0 satisfied-1)

The above variables were in object form they are converted to numeric form by encoding.

PARAMETERS USED FOR NEURAL NETWORKING

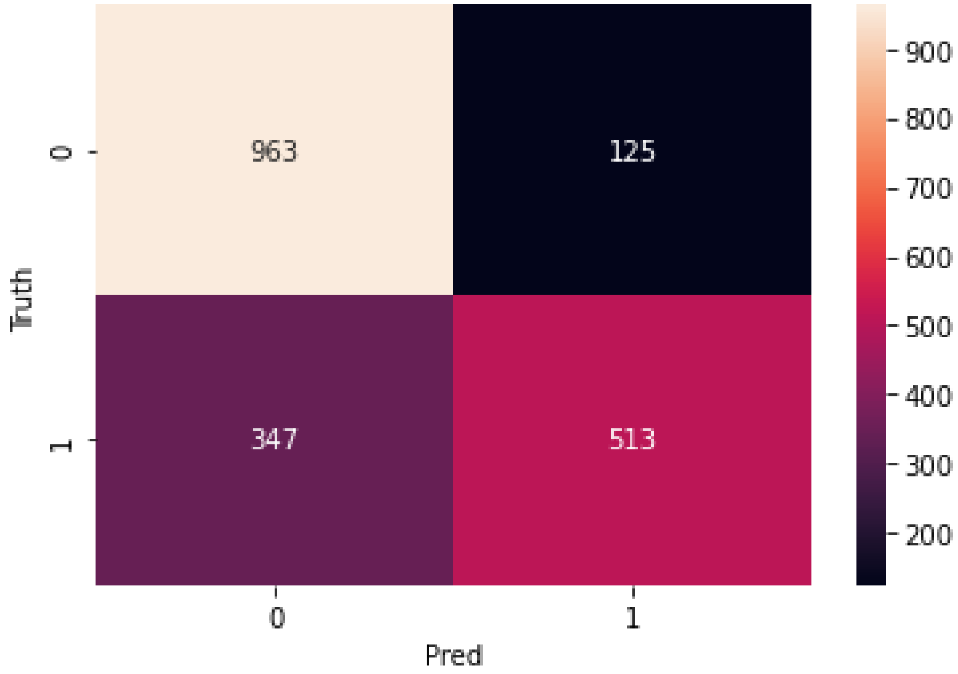
Neural Network used	Dense
Activation Functions Used	Relu and sigmoid
Optimizer used	Adam
Loss function used	Binary crossentropy
Metric	Accuracy
Number of epochs	100

CLASSIFICATION EVALUATION TABLE

```
              precision    recall  f1-score   support

           0       0.74      0.89      0.80      1088
           1       0.80      0.60      0.68       860

    accuracy                           0.76      1948
   macro avg       0.77      0.74      0.74      1948
weighted avg       0.77      0.76      0.75      1948
```

The above table tells that 0 was predicted with 74% accuracy and 1 was predicted with 80% accuracy.

CONFUSION MATRIX



The above matrix states that 963 and 513 were correctly predicted as 0 and 1 respectively

And 347 and 125 were wrongly predicted as 0 and 1 respectively.

VI. FINDINGS AND CONCLUSION:-

- Using chi square test of independence, we found there is a significant relationship between the type of customer and Class of travel.

- Using logistic regression, we see that Satisfaction level of customers depends significantly on the Inflight services, gate location at the airport, Check in Services.
- Using global testing, we see that our model is significant.
- We get an accuracy of 82.2% on our train data and 76.8% on the test data.
- From the ROC curve we see that the area under the curve is 0.8956122. This means that 89.5% variability in the data is explained by our model.

- From ANN we get that it's precision of prediction for (dissatisfaction or neutral) is 74% and for (satisfaction) it is 80%

VII. LIMITATIONS AND FUTURE SCOPE:-

LIMITATIONS:-

- Due to time constraint we were restricted to use the analysis we were aware about and were not able to explore more types of analysis suitable for the data.
- Due to the current restrictions we were only able to interact virtually leading to some drawbacks which could have overcome with one to one interaction with each other.

FUTURE SCOPE:-

- Bigger sample could be undertaken as we have used only 10% of the sample for analysis.
- With the help of other softwares more analysis could be performed.
- Will be able to learn the techniques that we performed in depth
- Making of presentations, reports can be done more precisely

VIII. REFERENCES:-

- https://www.kaggle.com/sjleshrac/airlines-customer-satisfaction

- https://www.researchgate.net/publication/322913951_AN_ANALYSIS_OF_AIRLINES_CUSTOMER_SATISFACTION_BY_IMPROVING_CUSTOMER_SERVICE_PERFORMANCE
- http://ojbe.steconomiceuoradea.ro/wp-content/uploads/2017/03/OJBE_vol-21_01_Ganiyu.pdf
- https://decisionstats.com/2012/03/22/random-sampling-a-dataset-in-r/
- https://www.displayr.com/what-is-a-roc-curve-how-to-interpret-it/#:~:text=The%20ROC%20curve%20shows%20the,diagonal%20(FPR%20%3D%20TPR).
- https://www.analyticsvidhya.com
- https://en.wikipedia.org/wiki/Logistic_regression
- https://www.statisticshowto.com/multicollinearity/
- **https://www.towardsdatascience.com**
- **https://www.quora.com**
- **https://www.statisticssolutions.com**
- SOFTWARES USED :- R,EXCEL,POWER BI AND PYTHON