MWSUG 2019 RF-079

US Airline Passenger Satisfaction

Harish Reddy Patlolla, Oklahoma State University

ABSTRACT

In the past 20 years, the aviation industry has been growing rapidly. This growth of the industry provides opportunities as well as challenges. While the opportunities arise because of increasing demand, rival airlines pose threat.

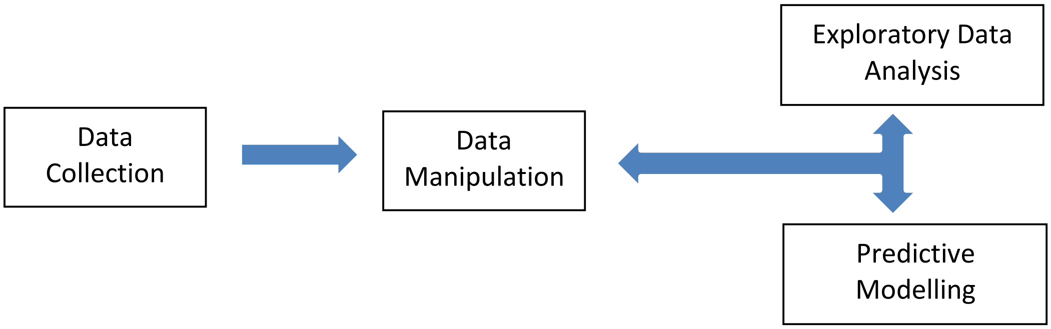
Apart from optimizing pricing, have you ever wondered what airlines do to overcome these threats? **Passenger Satisfaction**. Unhappy passengers mean fewer customers and less revenue. Therefore, it is important that passengers have a rich experience every time they travel. The satisfaction survey from 259,760 passengers, which is a combination of categorical and continuous variable has been used for this study. The models have been built using decision trees and logistic regression. This study seeks to not only explain the paramount factors which impact the passenger satisfaction in the US Airline industry but also change in those factors across different age groups. SAS 9.4 and SAS® Enterprise Miner™ have been used for data manipulation and predictive modelling respectively.

INTRODUCTION

In the past 20 years, the aviation industry has been growing rapidly. This growth of the industry provides opportunities as well as challenges to the airlines companies. The opportunities arise because of increasing demands. Whereas the challenges arise from the other airlines and also in making long-term relationship with the customers. To overcome these challenges, Airlines have to be remain on toes. Airlines try best to make passengers have a rich experience every time they travel. Factors such as departure and arrival time, inflight-entertainment, seat comfort could be very crucial in enhancing the customer experience. Also, the factors vary between different age groups. This study seeks to explain the paramount factors which impact the passenger satisfaction in the US Airline industry and change in those factors across different age groups.

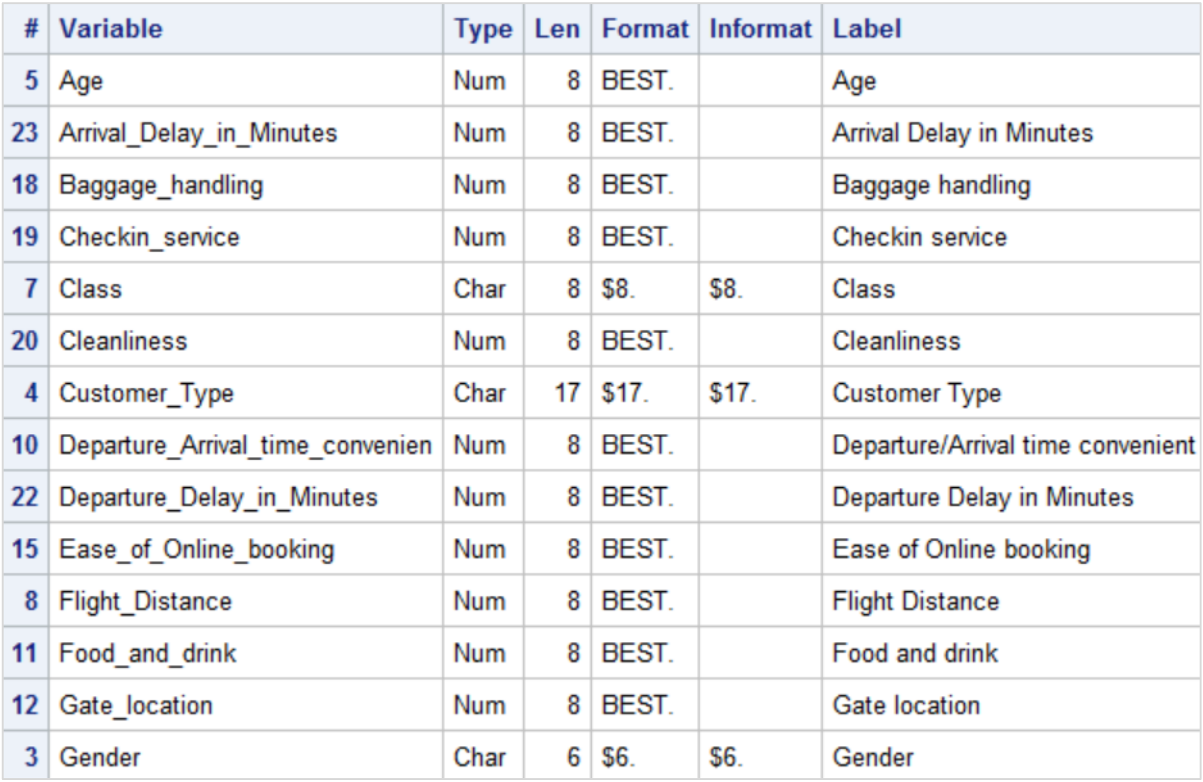
METHODOLOGY

The methodology used for this research has been divided into 3 stages shown in the diagram:



DATA COLLECTION

The dataset considered for this research paper contains 259,760 US Airline passenger satisfaction surveys taken from https://www.kaggle.com/johndddddd/customer-satisfaction. It contains 23 columns, out of which 5 are nominal, 4 are continuous and 14 are ordinal survey variables on scale of 1-5.



#	Variable	Type	Len	Format	Informat	Label
5	Age	Num	8	BEST.		Age
23	Arrival_Delay_in_Minutes	Num	8	BEST.		Arrival Delay in Minutes
18	Baggage_handling	Num	8	BEST.		Baggage handling
19	Checkin_service	Num	8	BEST.		Checkin service
7	Class	Char	8	$8.	$8.	Class
20	Cleanliness	Num	8	BEST.		Cleanliness
4	Customer_Type	Char	17	$17.	$17.	Customer Type
10	Departure_Arrival_time_convenien	Num	8	BEST.		Departure/Arrival time convenient
22	Departure_Delay_in_Minutes	Num	8	BEST.		Departure Delay in Minutes
15	Ease_of_Online_booking	Num	8	BEST.		Ease of Online booking
8	Flight_Distance	Num	8	BEST.		Flight Distance
11	Food_and_drink	Num	8	BEST.		Food and drink
12	Gate_location	Num	8	BEST.		Gate location
3	Gender	Char	6	$6.	$6.	Gender

DATA MANIPULATION

There were two data sets each containing 129880 observations. The excel datasets were imported into SAS using the SAS code below

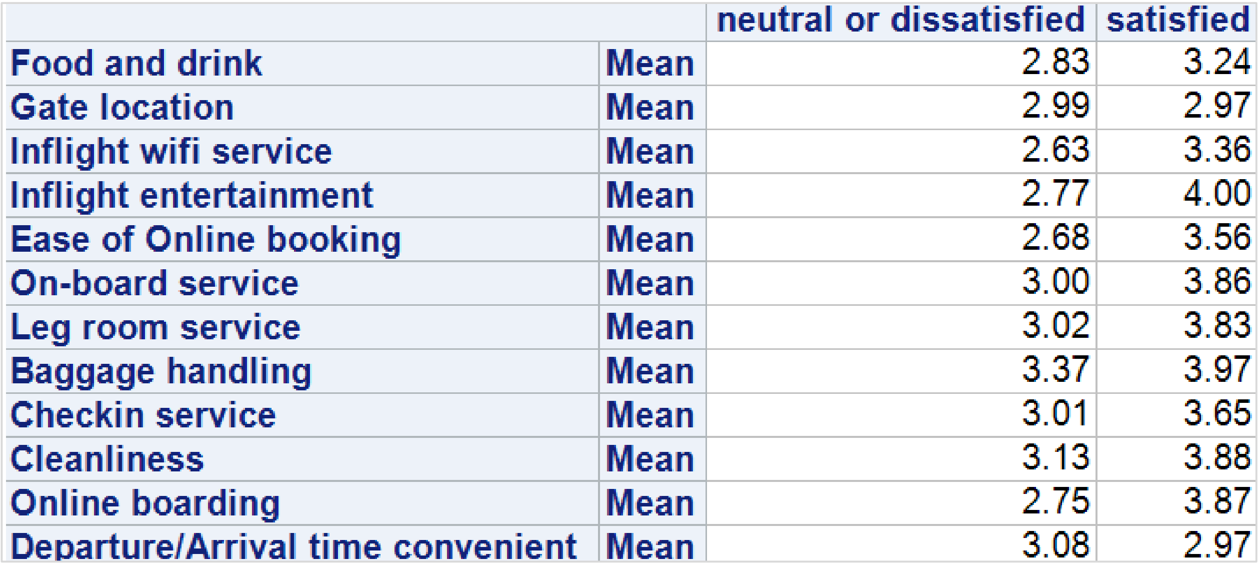
```
libname orion 'D:\Sunny_Personal\OSU MSBA\Spring 2019\Research
paper\customer-satisfaction';

PROC IMPORT OUT=orion.satisfaction_1
DATAFILE= "D:\Sunny_Personal\OSU MSBA\Spring 2019\Research
paper\customer-satisfaction/satisfaction.xlsx"
DBMS=xlsx REPLACE;
GETNAMES=YES;
RUN;

PROC IMPORT OUT=orion.satisfaction_2
DATAFILE= "D:\Sunny_Personal\OSU MSBA\Spring 2019\Research
paper\customer-satisfaction/satisfaction_2015.xlsx"
DBMS=xlsx REPLACE;
GETNAMES=YES;
RUN;
```

The below SAS code was then used to merge above two datasets to form a single dataset containing 259760 observations. Two columns Online_Support and Inflight_service were dropped because those two columns weren’t present in both the datasets.

```
DATA ORION.SATISFACTION_MERGED (DROP= ONLINE_SUPPORT
INFLIGHT_SERVICE);
SET ORION.SATISFACTION_1 ORION.SATISFACTION_2;
RUN;
```

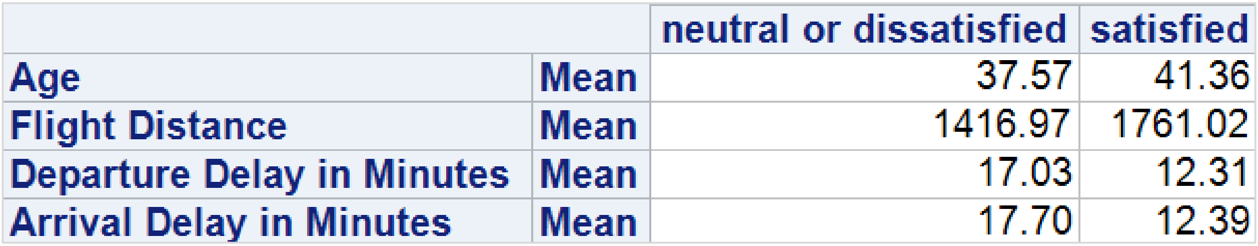
EXPLORATORY DATA ANALYSIS

As part of exploratory data analysis, means of the ordinal variables were reported between the satisfied and dissatisfied group of customers. The means of all the survey ordinal variables except for the Gate location, Departure/Arrival time convenient are higher for the satisfied group when compared to the dissatisfied group.

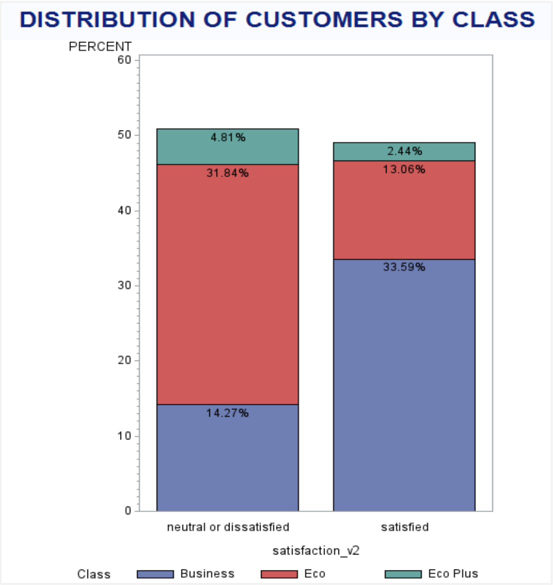


		neutral or dissatisfied	satisfied
Food and drink	Mean	2.83	3.24
Gate location	Mean	2.99	2.97
Inflight wifi service	Mean	2.63	3.36
Inflight entertainment	Mean	2.77	4.00
Ease of Online booking	Mean	2.68	3.56
On-board service	Mean	3.00	3.86
Leg room service	Mean	3.02	3.83
Baggage handling	Mean	3.37	3.97
Checkin service	Mean	3.01	3.65
Cleanliness	Mean	3.13	3.88
Online boarding	Mean	2.75	3.87
Departure/Arrival time convenient	Mean	3.08	2.97

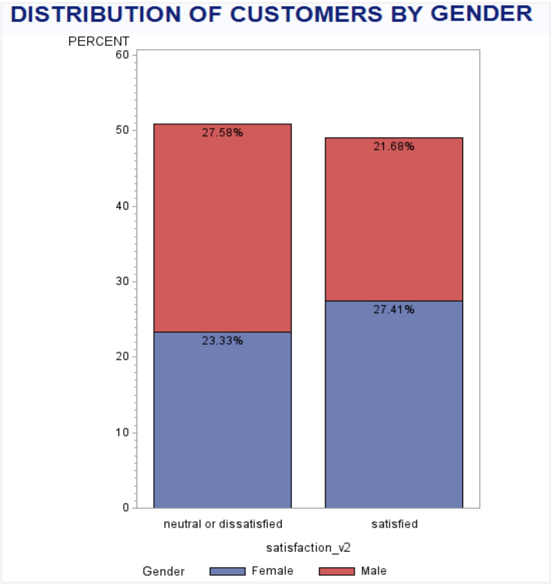
Means of the continuous variables were reported between the satisfied and dissatisfied group of customers. The average delay time for the satisfied customers is 12 minutes 31 seconds whereas 17 minutes and 3 seconds for the neutral or dissatisfied customers. The customers tend to be satisfied with the airlines if there is no arrival or departure delays.



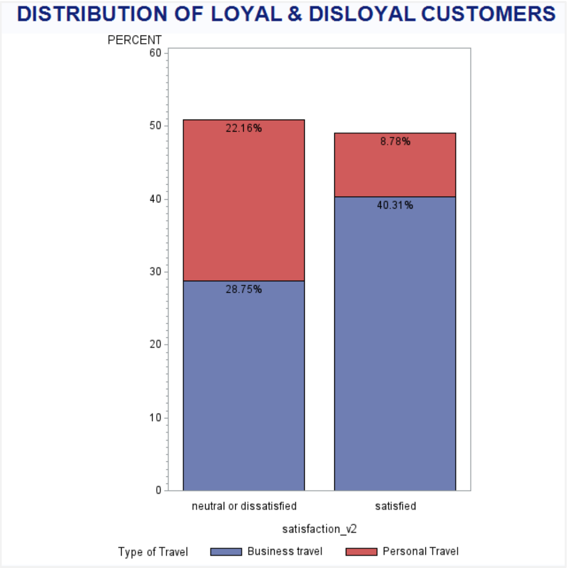
		neutral or dissatisfied	satisfied
Age	Mean	37.57	41.36
Flight Distance	Mean	1416.97	1761.02
Departure Delay in Minutes	Mean	17.03	12.31
Arrival Delay in Minutes	Mean	17.70	12.39



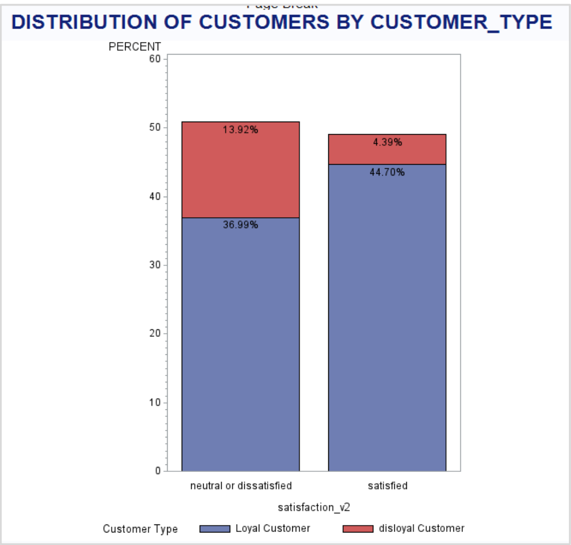
DISTRIBUTION OF CUSTOMERS BY CLASS



DISTRIBUTION OF CUSTOMERS BY GENDER

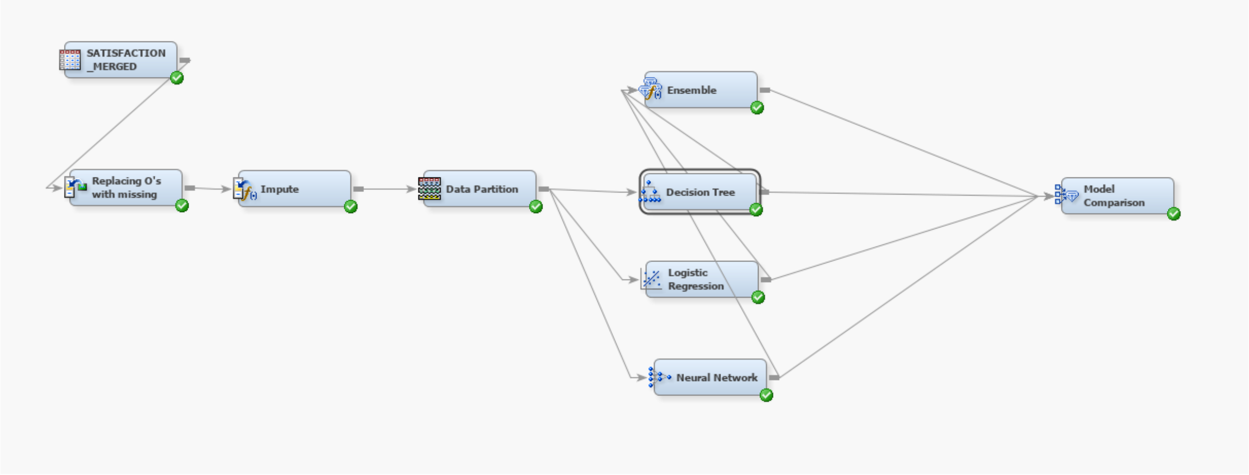


DISTRIBUTION OF LOYAL & DISLOYAL CUSTOMERS



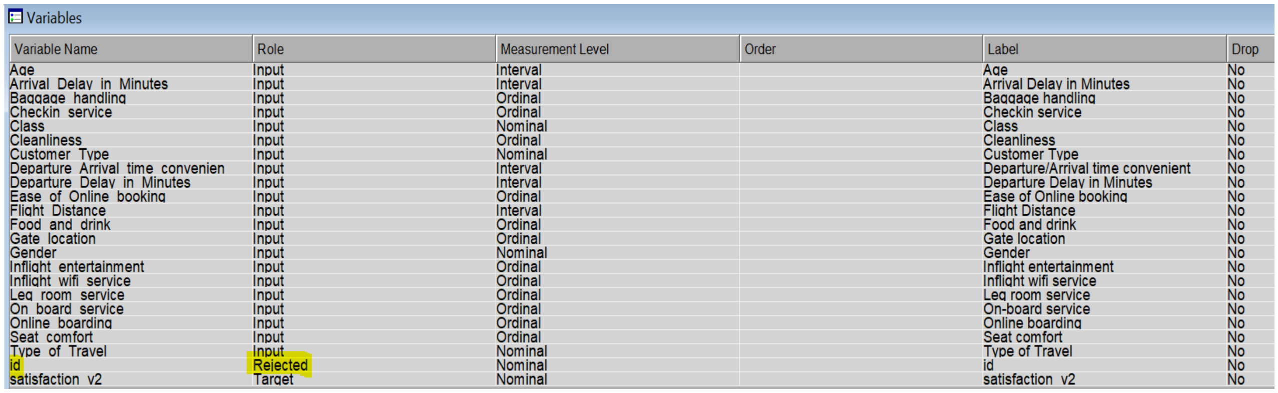
DISTRIBUTION OF CUSTOMERS BY CUSTOMER_TYPE

MODELLING



STEP 1: IMPORTING THE FILE

The data has been imported into SAS Miner using create data source wizard. The ID variable has been rejected. There are 12 ordinal, 4 nominal and 5 interval independent variables.

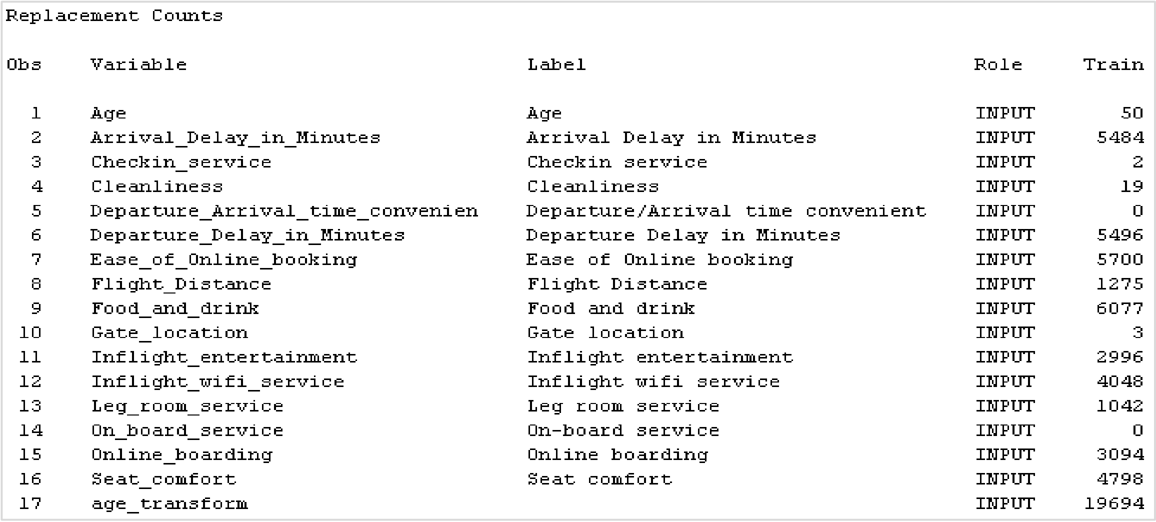


Variables

Variable Name	Role	Measurement Level	Order	Label	Drop
Age	Input	Interval		Age	No
Arrival Delay in Minutes	Input	Interval		Arrival Delay in Minutes	No
Baggage handling	Input	Ordinal		Baggage handling	No
Checkin service	Input	Ordinal		Checkin service	No
Class	Input	Nominal		Class	No
Cleanliness	Input	Ordinal		Cleanliness	No
Customer Type	Input	Nominal		Customer Type	No
Departure Arrival time convenien	Input	Interval		Departure/Arrival time convenient	No
Departure Delay in Minutes	Input	Interval		Departure Delay in Minutes	No
Ease of Online booking	Input	Ordinal		Ease of Online booking	No
Flight Distance	Input	Interval		Flight Distance	No
Food and drink	Input	Ordinal		Food and drink	No
Gate location	Input	Ordinal		Gate location	No
Gender	Input	Nominal		Gender	No
Inflight entertainment	Input	Ordinal		Inflight entertainment	No
Inflight wifi service	Input	Ordinal		Inflight wifi service	No
Leg room service	Input	Ordinal		Leg room service	No
On board service	Input	Ordinal		On-board service	No
Online boarding	Input	Ordinal		Online boarding	No
Seat comfort	Input	Ordinal		Seat comfort	No
Type of Travel	Input	Nominal		Type of Travel	No
id	Rejected	Nominal		id	No
satisfaction v2	Target	Nominal		satisfaction v2	No

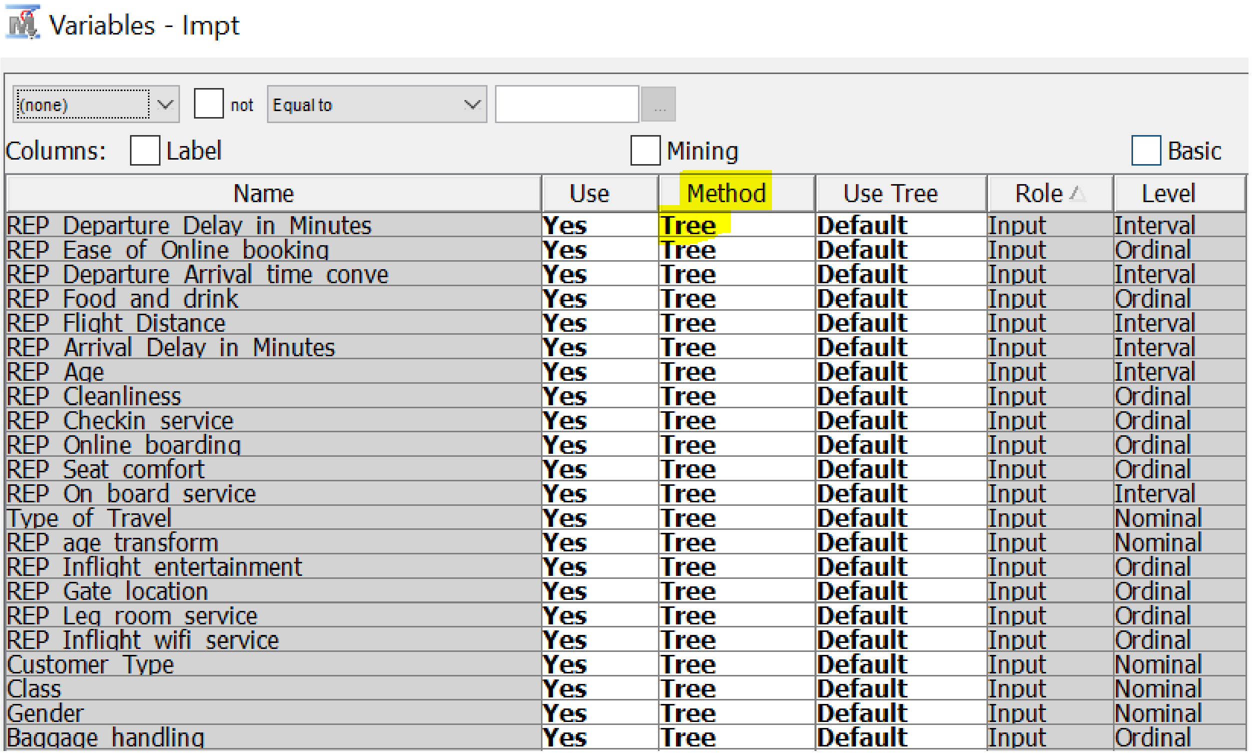
STEP 2: REPLACEMENT NODE

The ordinal survey variables had 0 as the invalid value. 0's has been replaced to a missing value. Below are the replacement counts



```
Replacement Counts

Obs    Variable                              Label                                Role      Train

  1    Age                                   Age                                  INPUT        50
  2    Arrival_Delay_in_Minutes              Arrival Delay in Minutes             INPUT      5484
  3    Checkin_service                       Checkin service                      INPUT         2
  4    Cleanliness                           Cleanliness                          INPUT        19
  5    Departure_Arrival_time_convenien      Departure/Arrival time convenient    INPUT         0
  6    Departure_Delay_in_Minutes            Departure Delay in Minutes           INPUT      5496
  7    Ease_of_Online_booking                Ease of Online booking               INPUT      5700
  8    Flight_Distance                       Flight Distance                      INPUT      1275
  9    Food_and_drink                        Food and drink                       INPUT      6077
 10    Gate_location                         Gate location                        INPUT         3
 11    Inflight_entertainment                Inflight entertainment               INPUT      2996
 12    Inflight_wifi_service                 Inflight wifi service                INPUT      4048
 13    Leg_room_service                      Leg room service                     INPUT      1042
 14    On_board_service                      On-board service                     INPUT         0
 15    Online_boarding                       Online boarding                      INPUT      3094
 16    Seat_comfort                          Seat comfort                         INPUT      4798
 17    age_transform                                                              INPUT     19694
```

STEP 3: DATA IMPUTATION



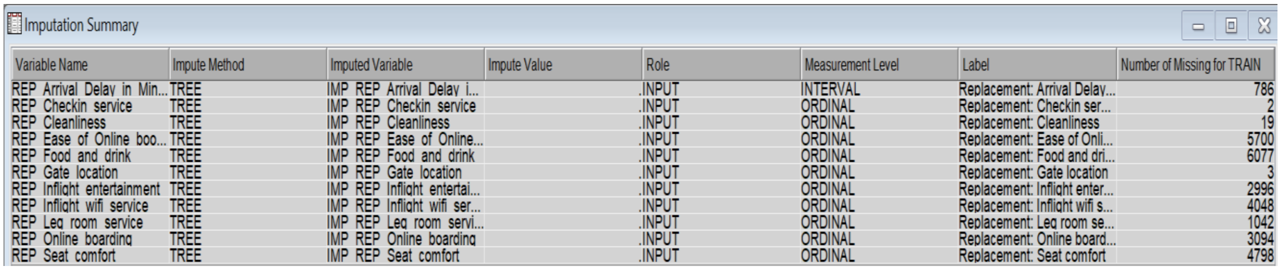
Variables - Impt

(none) □ not Equal to ...

Columns: □ Label □ Mining □ Basic

Name	Use	Method	Use Tree	Role	Level
REP Departure Delay in Minutes	Yes	Tree	Default	Input	Interval
REP Ease of Online booking	Yes	Tree	Default	Input	Ordinal
REP Departure Arrival time conve	Yes	Tree	Default	Input	Interval
REP Food and drink	Yes	Tree	Default	Input	Ordinal
REP Flight Distance	Yes	Tree	Default	Input	Interval
REP Arrival Delay in Minutes	Yes	Tree	Default	Input	Interval
REP Age	Yes	Tree	Default	Input	Interval
REP Cleanliness	Yes	Tree	Default	Input	Ordinal
REP Checkin service	Yes	Tree	Default	Input	Ordinal
REP Online boarding	Yes	Tree	Default	Input	Ordinal
REP Seat comfort	Yes	Tree	Default	Input	Ordinal
REP On board service	Yes	Tree	Default	Input	Interval
Type of Travel	Yes	Tree	Default	Input	Nominal
REP age transform	Yes	Tree	Default	Input	Nominal
REP Inflight entertainment	Yes	Tree	Default	Input	Ordinal
REP Gate location	Yes	Tree	Default	Input	Ordinal
REP Leg room service	Yes	Tree	Default	Input	Ordinal
REP Inflight wifi service	Yes	Tree	Default	Input	Ordinal
Customer Type	Yes	Tree	Default	Input	Nominal
Class	Yes	Tree	Default	Input	Nominal
Gender	Yes	Tree	Default	Input	Nominal
Baggage handling	Yes	Tree	Default	Input	Ordinal

The missing values have been imputed using Tree method. Below is the summary of imputation.

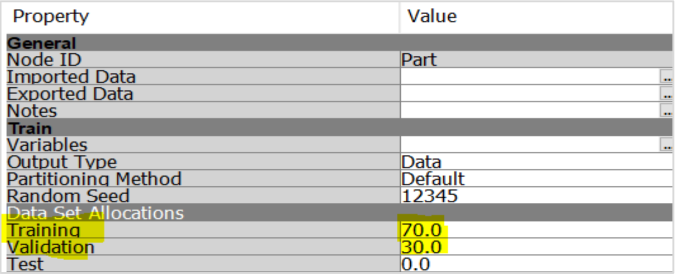


Imputation Summary

Variable Name	Impute Method	Imputed Variable	Impute Value	Role	Measurement Level	Label	Number of Missing for TRAIN
REP Arrival Delay in Min...	TREE	IMP REP Arrival Delay i...	.	INPUT	INTERVAL	Replacement: Arrival Delay...	786
REP Checkin service	TREE	IMP REP Checkin service	.	INPUT	ORDINAL	Replacement: Checkin ser...	2
REP Cleanliness	TREE	IMP REP Cleanliness	.	INPUT	ORDINAL	Replacement: Cleanliness	19
REP Ease of Online boo...	TREE	IMP REP Ease of Online...	.	INPUT	ORDINAL	Replacement: Ease of Onli...	5700
REP Food and drink	TREE	IMP REP Food and drink	.	INPUT	ORDINAL	Replacement: Food and dri...	6077
REP Gate location	TREE	IMP REP Gate location	.	INPUT	ORDINAL	Replacement: Gate location	3
REP Inflight entertainment	TREE	IMP REP Inflight entertai...	.	INPUT	ORDINAL	Replacement: Inflight enter...	2996
REP Inflight wifi service	TREE	IMP REP Inflight wifi ser...	.	INPUT	ORDINAL	Replacement: Inflight wifi s...	4048
REP Leg room service	TREE	IMP REP Leg room servi...	.	INPUT	ORDINAL	Replacement: Leg room se...	1042
REP Online boarding	TREE	IMP REP Online boarding	.	INPUT	ORDINAL	Replacement: Online board...	3094
REP Seat comfort	TREE	IMP REP Seat comfort	.	INPUT	ORDINAL	Replacement: Seat comfort	4798

STEP 4: DATA PARTITION

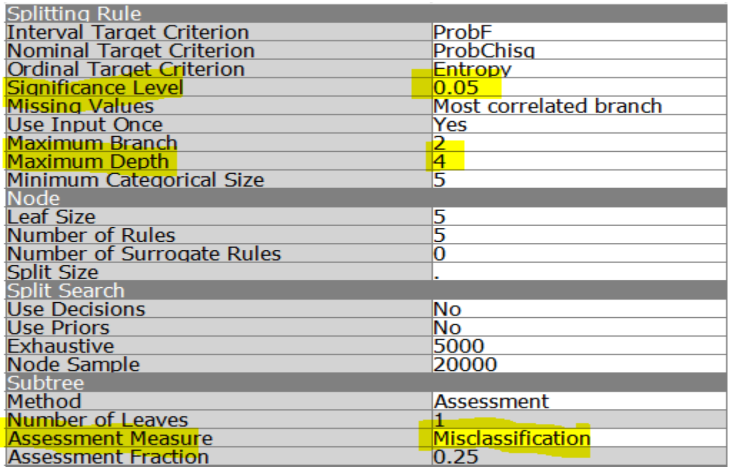
The Data has been partitioned into 70-30 split for training and validation. The model has been built using the training data and the same has been evaluated using the validation data.



Property	Value
General	
Node ID	Part
Imported Data	...
Exported Data	...
Notes	...
Train	
Variables	...
Output Type	Data
Partitioning Method	Default
Random Seed	12345
Data Set Allocations	
Training	70.0
Validation	30.0
Test	0.0

STEP 5: DECISION TREE

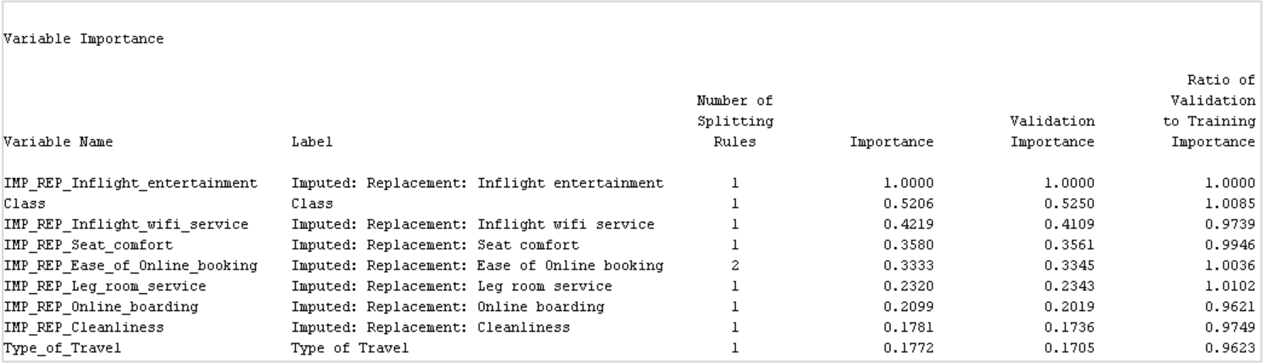
Below are the options selected for decision tree. The misclassification rate was the assessment measure to prune the tree.



Splitting Rule	
Interval Target Criterion	ProbF
Nominal Target Criterion	ProbChisq
Ordinal Target Criterion	Entropy
Significance Level	0.05
Missing Values	Most correlated branch
Use Input Once	Yes
Maximum Branch	2
Maximum Depth	4
Minimum Categorical Size	5
Node	
Leaf Size	5
Number of Rules	5
Number of Surrogate Rules	0
Split Size	.
Split Search	
Use Decisions	No
Use Priors	No
Exhaustive	5000
Node Sample	20000
Subtree	
Method	Assessment
Number of Leaves	1
Assessment Measure	Misclassification
Assessment Fraction	0.25

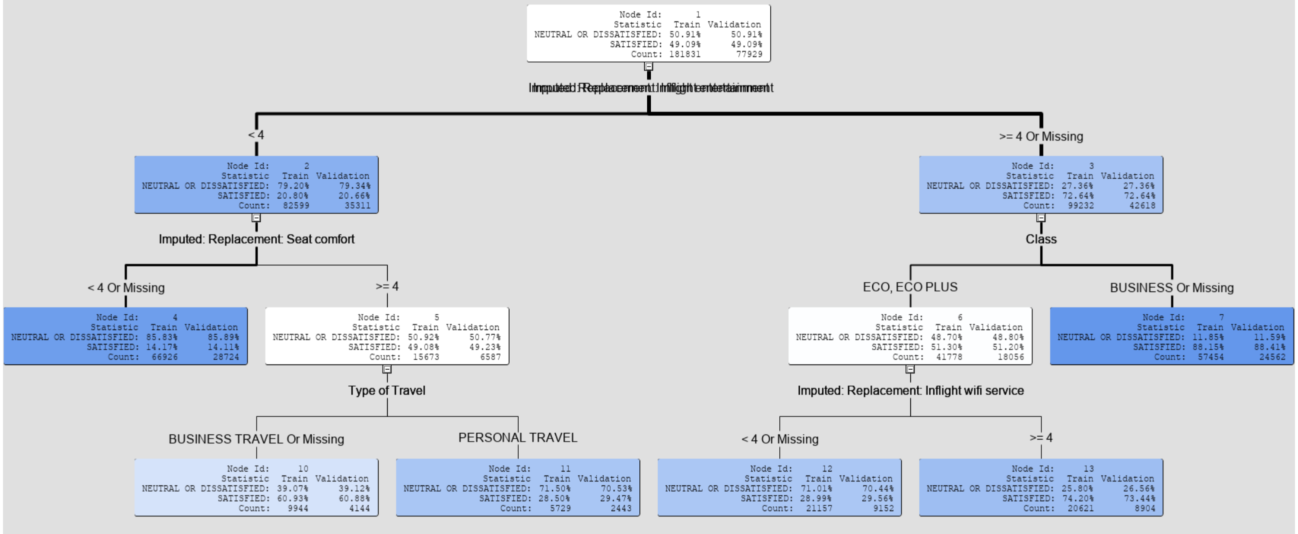
RESULTS

Inflight Entertainment was most important variable in predicting the customer satisfaction index. Inflight Wi-Fi service, Seat comfort, Ease of online booking, Leg room were the other important variables in the prediction.



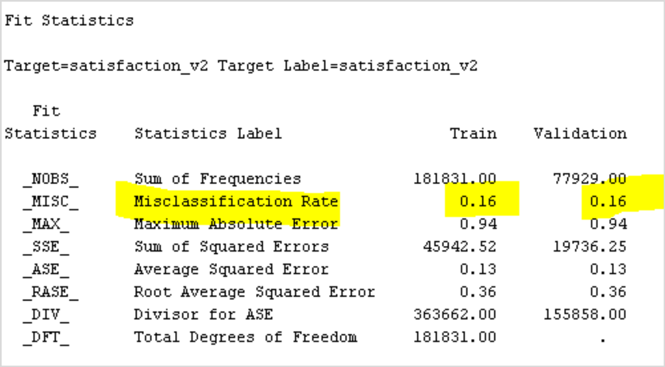
Variable Importance

Variable Name	Label	Number of Splitting Rules	Importance	Validation Importance	Ratio of Validation to Training Importance
IMP_REP_Inflight_entertainment	Imputed: Replacement: Inflight entertainment	1	1.0000	1.0000	1.0000
Class	Class	1	0.5206	0.5250	1.0085
IMP_REP_Inflight_wifi_service	Imputed: Replacement: Inflight wifi service	1	0.4219	0.4109	0.9739
IMP_REP_Seat_comfort	Imputed: Replacement: Seat comfort	1	0.3580	0.3561	0.9946
IMP_REP_Ease_of_Online_booking	Imputed: Replacement: Ease of Online booking	2	0.3333	0.3345	1.0036
IMP_REP_Leg_room_service	Imputed: Replacement: Leg room service	1	0.2320	0.2343	1.0102
IMP_REP_Online_boarding	Imputed: Replacement: Online boarding	1	0.2099	0.2019	0.9621
IMP_REP_Cleanliness	Imputed: Replacement: Cleanliness	1	0.1781	0.1736	0.9749
Type_of_Travel	Type of Travel	1	0.1772	0.1705	0.9623

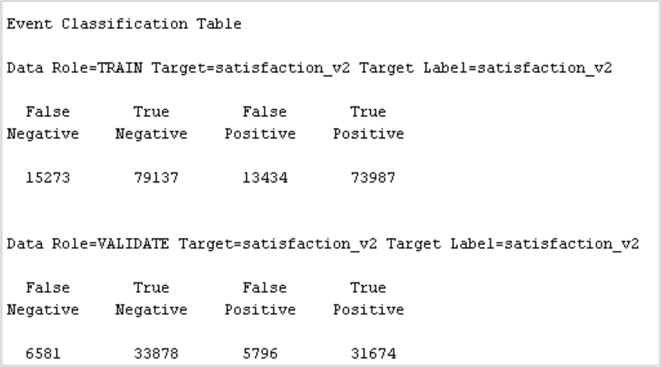

Node Id: 1
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 50.91% 50.91%
SATISFIED: 49.09% 49.09%
Count: 181831 77929
Imputed: Replacement: Inflight entertainment
< 4
>= 4 Or Missing
Node Id: 2
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 79.20% 79.34%
SATISFIED: 20.80% 20.66%
Count: 82599 35311
Imputed: Replacement: Seat comfort
Node Id: 3
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 27.36% 27.36%
SATISFIED: 72.64% 72.64%
Count: 99232 42618
Class
< 4 Or Missing
>= 4
ECO, ECO PLUS
BUSINESS Or Missing
Node Id: 4
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 85.83% 85.89%
SATISFIED: 14.17% 14.11%
Count: 66926 28724
Node Id: 5
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 50.92% 50.77%
SATISFIED: 49.08% 49.23%
Count: 15673 6587
Type of Travel
Node Id: 6
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 48.70% 48.80%
SATISFIED: 51.30% 51.20%
Count: 41778 18056
Imputed: Replacement: Inflight wifi service
Node Id: 7
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 11.85% 11.59%
SATISFIED: 88.15% 88.41%
Count: 57454 24562
BUSINESS TRAVEL Or Missing
PERSONAL TRAVEL
< 4 Or Missing
>= 4
Node Id: 10
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 39.07% 39.12%
SATISFIED: 60.93% 60.88%
Count: 9944 4144
Node Id: 11
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 71.50% 70.53%
SATISFIED: 28.50% 29.47%
Count: 5729 2443
Node Id: 12
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 71.01% 70.44%
SATISFIED: 28.99% 29.56%
Count: 21157 9152
Node Id: 13
Statistic Train Validation
NEUTRAL OR DISSATISFIED: 25.80% 26.56%
SATISFIED: 74.20% 73.44%
Count: 20621 8904

Below are the few rules from the decision tree

- When Inflight Entertainment is greater than or equal to 4, there is 72.64% chance that the customer is satisfied with the airline service. The chances of getting satisfied with the airlines goes to 88.15% when the customer opts for Business class
- Similarly, when Inflight Entertainment is less than 4, there is 79.2% chance that the customer will not be satisfied with the airlines. The chances of dissatisfaction go even further to 85.83% when the seat comfort is less than 4.
- The customers belonging to Eco and Eco plus category looks for inflight Wi-Fi service. When the inflight Wi-Fi service is greater than or equal to 4, there is 74.2% chance that the customer will be satisfied with the airlines.



```
Fit Statistics

Target=satisfaction_v2 Target Label=satisfaction_v2

  Fit
Statistics    Statistics Label                    Train    Validation

  _NOBS_      Sum of Frequencies              181831.00      77929.00
  _MISC_      Misclassification Rate               0.16          0.16
  _MAX_       Maximum Absolute Error               0.94          0.94
  _SSE_       Sum of Squared Errors            45942.52      19736.25
  _ASE_       Average Squared Error                0.13          0.13
  _RASE_      Root Average Squared Error           0.36          0.36
  _DIV_       Divisor for ASE                 363662.00     155858.00
  _DFT_       Total Degrees of Freedom        181831.00             .
```



```
Event Classification Table

Data Role=TRAIN Target=satisfaction_v2 Target Label=satisfaction_v2

  False       True       False       True
Negative    Negative    Positive    Positive

  15273       79137       13434       73987


Data Role=VALIDATE Target=satisfaction_v2 Target Label=satisfaction_v2

  False       True       False       True
Negative    Negative    Positive    Positive

  6581        33878       5796        31674
```

The misclassification rate of the training and validation data using decision tree is 16% i.e., the model's accuracy is 84%.

CONCLUSION

This paper gave us information about what are the paramount factors driving the passenger satisfaction. The most important factor being the In-Flight Entertainment followed by the seat comfort. The model has predicted about 84% of the cases correctly which implies that it is a good model.

REFERENCES

https://blogs.perficient.com/2018/05/14/customer-satisfaction-in-the-airline-industry/

ACKNOWLEDGMENTS

I sincerely thank Dr. Goutam Chakraborty for his valuable guidance and motivation for accomplishing this paper. I also thank Dr. Miriam McGaugh for her constant support and suggestions throughout this study.

CONTACT INFORMATION

Harish Reddy Patlolla

Oklahoma State University

405-385-1049

hpatlol@okstate.edu

APPENDIX

```
TITLE 'Comparison of means of Continuous Variables';

PROC TABULATE DATA=ORION.SATISFACTION_MERGED_AGE_TRF;
CLASS SATISFACTION_V2;
VAR AGE FLIGHT_DISTANCE DEPARTURE_DELAY_IN_MINUTES
ARRIVAL_DELAY_IN_MINUTES;
TABLE (AGE FLIGHT_DISTANCE DEPARTURE_DELAY_IN_MINUTES
ARRIVAL_DELAY_IN_MINUTES) *(MEAN),
SATISFACTION_V2;
RUN;

TITLE 'Comparison of means of Survey Ordinal Variables';

PROC TABULATE DATA=ORION.SATISFACTION_MERGED_AGE_TRF;
CLASS SATISFACTION_V2;
VAR FOOD_AND_DRINK GATE_LOCATION INFLIGHT_WIFI_SERVICE
INFLIGHT_ENTERTAINMENT EASE_OF_ONLINE_BOOKING
ON_BOARD_SERVICE LEG_ROOM_SERVICE BAGGAGE_HANDLING
CHECKIN_SERVICE CLEANLINESS ONLINE_BOARDING
DEPARTURE_ARRIVAL_TIME_CONVENIEN;
TABLE (FOOD_AND_DRINK GATE_LOCATION INFLIGHT_WIFI_SERVICE
INFLIGHT_ENTERTAINMENT EASE_OF_ONLINE_BOOKING
ON_BOARD_SERVICE LEG_ROOM_SERVICE BAGGAGE_HANDLING
CHECKIN_SERVICE CLEANLINESS ONLINE_BOARDING
DEPARTURE_ARRIVAL_TIME_CONVENIEN)  *(MEAN),
SATISFACTION_V2;
RUN;


TITLE 'DISTRIBUTION OF CUSTOMERS BY CLASS';
PROC GCHART DATA=ORION.SATISFACTION_MERGED;
VBAR SATISFACTION_V2/SUBGROUP=CLASS
TYPE=PERCENT
INSIDE=PERCENT;
RUN;

TITLE 'DISTRIBUTION OF CUSTOMERS BY GENDER';
PROC GCHART DATA=ORION.SATISFACTION_MERGED;
VBAR SATISFACTION_V2/SUBGROUP=GENDER
```

```
TYPE=PERCENT
INSIDE=PERCENT;
RUN;

TITLE 'DISTRIBUTION OF LOYAL & DISLOYAL CUSTOMERS';
PROC GCHART DATA=ORION.SATISFACTION_MERGED;
VBAR SATISFACTION_V2/SUBGROUP=TYPE_OF_TRAVEL
TYPE=PERCENT
INSIDE=PERCENT;
RUN;

TITLE 'DISTRIBUTION OF CUSTOMERS BY CUSTOMER_TYPE';
PROC GCHART DATA=ORION.SATISFACTION_MERGED;
VBAR SATISFACTION_V2/SUBGROUP=CUSTOMER_TYPE
TYPE=PERCENT
INSIDE=PERCENT;
RUN;
```