Airline Passenger Satisfaction

Data Exploration

```
> str(df)
'data.frame':   103904 obs. of  21 variables:
 $ Gender                          : Factor w/ 2 levels "Female","Male": 2 2 1 1 2 1 2 1 1 2 ...
 $ Customer.Type                   : Factor w/ 2 levels "disloyal Customer",..: 2 1 2 2 2 2 2 2 2 1 ...
 $ Age                             : int  13 25 26 25 61 26 47 52 41 20 ...
 $ Class                           : Factor w/ 3 levels "Business","Eco",..: 3 1 1 1 1 2 2 1 1 2 ...
 $ Flight.Distance                 : int  460 235 1142 562 214 1180 1276 2035 853 1061 ...
 $ Inflight.wifi.service           : int  3 3 2 2 3 3 2 4 1 3 ...
 $ DepartureArrival.time.convenient: int  4 2 2 5 3 4 4 3 2 3 ...
 $ Ease.of.Online.booking          : int  3 3 2 5 3 2 2 4 2 3 ...
 $ Food.and.drink                  : int  5 1 5 2 4 1 2 5 4 2 ...
 $ Online.boarding                 : int  3 3 5 2 5 2 2 5 3 3 ...
 $ Seat.comfort                    : int  5 1 5 2 5 1 2 5 3 3 ...
 $ Inflight.entertainment          : int  5 1 5 2 3 1 2 5 1 2 ...
 $ Onboard.service                 : int  4 1 4 2 3 3 3 5 1 2 ...
 $ Leg.room.service                : int  3 5 3 5 4 4 3 5 2 3 ...
 $ Baggage.handling                : int  4 3 4 3 4 4 4 5 1 4 ...
 $ Checkin.service                 : int  4 1 4 1 3 4 3 4 4 4 ...
 $ Inflight.service                : int  5 4 4 4 3 4 5 5 1 3 ...
 $ Cleanliness                     : int  5 1 5 2 3 1 2 4 2 2 ...
 $ Departure.Delay.in.Minutes      : int  25 1 0 11 0 0 9 4 0 0 ...
 $ Arrival.Delay.in.Minutes        : int  18 6 0 9 0 0 23 0 0 0 ...
 $ satisfaction                    : Factor w/ 2 levels "neutral or dissatisfied",..: 1 1 2 1 2 1 1 2 1 1 ...
```

- 103904 observations (rows) and 21 variables
- 4 Factor/Categorical variables (Gender, Customer Type, Class, and Satisfaction)
- Several integer variables (survey ratings out of 1-5, Age, Miles travelled)

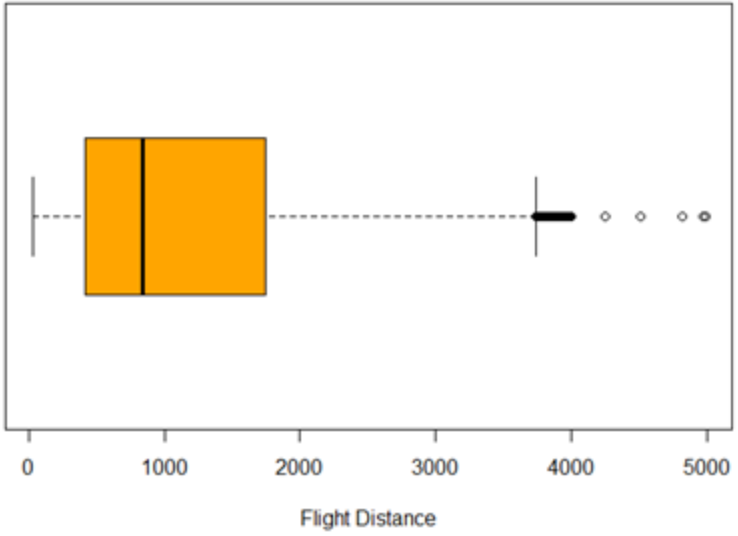
Data source: Airline Passenger Satisfaction (kaggle.com)

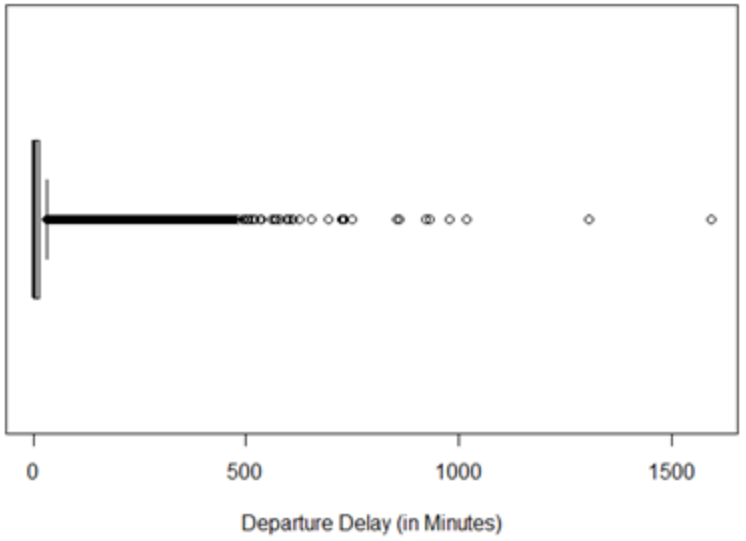
	Variable	Definition	Mean	Std. Dev.	Min	Max
IVs	Gender	Gender of the passengers (Female=1, Male=2)	1.49259	0.499947	1	2
	Customer Type	The customer type (Loyal customer=2, disloyal customer=1)	1.8168619	0.3867812	1	2
	Age	The actual age of the passengers	39.3876	15.1176	7	85
	Class	Travel class in the plane of the passengers (Eco=2, Eco Plus=3 or Business=1)	1.59	0.62	1	3
	Flight Distance	The flight distance of this journey (Miles)	1189.451	997.561	31	4983
	Inflight Wi-Fi service	Satisfaction level of the inflight Wi-Fi service (0: Not Applicable;1-5)	2.728544	1.329235	0	5
	Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenience (0: Not Applicable;1-5)	3.057349	1.526787	0	5
	Ease of online booking	Satisfaction level of online booking (0: Not Applicable;1-5)	2.756786	1.401662	0	5
	Food and drink	Satisfaction level of Food and drink (0: Not Applicable;1-5)	3.204685	1.329905	0	5
	Online boarding	Satisfaction level of online boarding (0: Not Applicable;1-5)	3.25272	1.350651	0	5
	Seat comfort	Satisfaction level of Seat comfort (0: Not Applicable;1-5)	3.441589	1.319168	0	5
	Inflight entertainment	Satisfaction level of inflight entertainment (0: Not Applicable;1-5)	3.358067	1.334149	0	5
	On-board service	Satisfaction level of On-board service (0: Not Applicable;1-5)	3.383204	1.287032	0	5
	Leg room service	Satisfaction level of Leg room service (0: Not Applicable;1-5)	3.351078	1.316132	0	5
	Baggage handling	Satisfaction level of baggage handling (1-5)	3.631886	1.180082	1	5
	Check-in service	Satisfaction level of Check-in service (0: Not Applicable;1-5)	3.306239	1.266146	0	5
	Inflight service	Satisfaction level of inflight service (0: Not Applicable;1-5)	3.642373	1.176614	0	5
	Cleanliness	Satisfaction level of Cleanliness (0: Not Applicable;1-5)	3.286222	1.313624	0	5
	Departure Delay	Minutes delayed when departed (In Minutes)	14.82339	38.23287	0	1592
	Arrival Delay	Minutes delayed when Arrival (In Minutes)	15.18113	38.6565	0	1584
DV	Satisfaction	Airline satisfaction level (Satisfaction= 2, neutral or dissatisfaction = 1)	1.4344992	0.495693	1	2

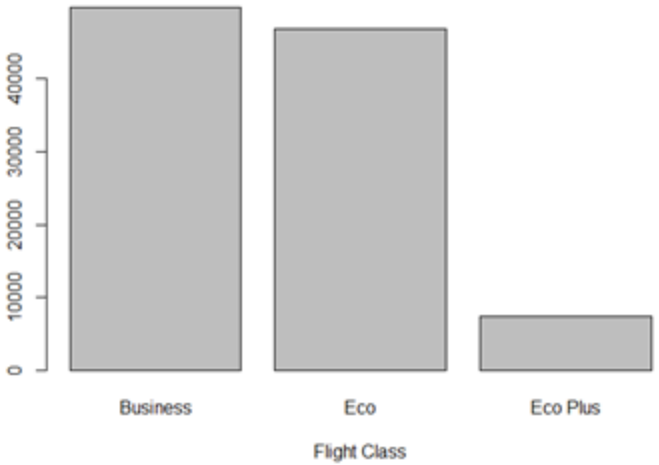
- Independent Variables (IVs)
  20 independent variables
- Dependent Variable (DV)
  Satisfaction level

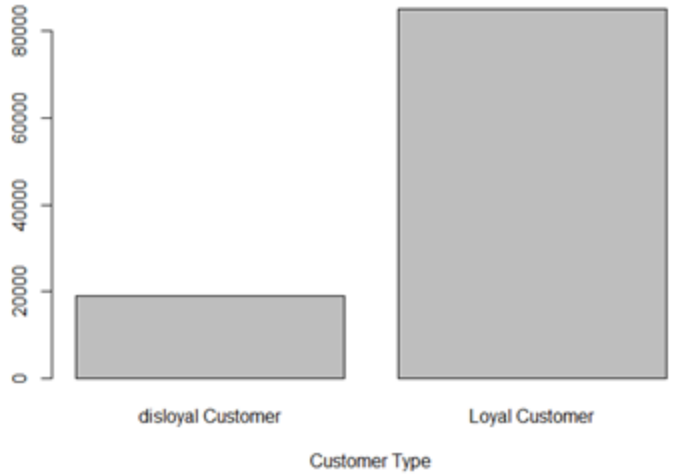
Data source: Airline Passenger Satisfaction (kaggle.com)

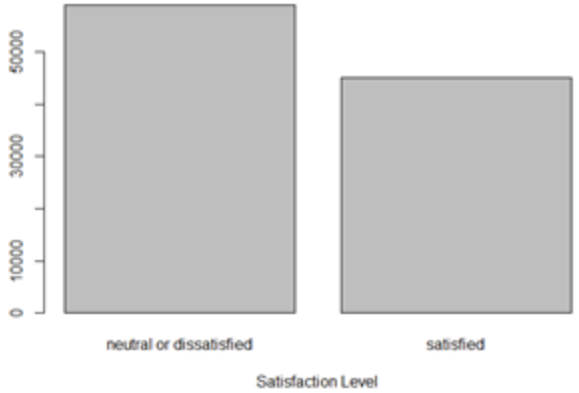
Data Visualization






0
10000
20000
30000
40000
Business
Eco
Eco Plus
Flight Class


0
20000
40000
60000
80000
disloyal Customer
Loyal Customer
Customer Type


0
10000
30000
50000
neutral or dissatisfied
satisfied
Satisfaction Level

Data Preprocessing

- There were 310 NA values in Arrival Delay in Minutes column only which were replaced with 0.
- Various different pre-processing techniques such as standardization and one-hot encoding were used.
- Outliers were not removed as they represented natural variations in airline industry.

Classification- Decision Tree: Test/Train Data

- Train Data: 83123 observations

```
> set.seed(5)
> train = sample(1:nrow(df), nrow(df)*(4/5))
> train
```

- Test Data: 20781 observations

```
> df.train = df[train,]
> df.test = df[-train,]
```

Classification - Decision Tree

```
n= 83123

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 83123 36139 neutral or dissatisfied (0.565234652 0.434765348)
   2) Online.boarding< 3.5 41866  6305 neutral or dissatisfied (0.849400468 0.150599532)
     4) Inflight.wifi.service>=0.5 40417  4864 neutral or dissatisfied (0.879654601 0.120345399)
       8) Inflight.wifi.service< 3.5 36624  2404 neutral or dissatisfied (0.934359983 0.065640017) *
       9) Inflight.wifi.service>=3.5 3793  1333 satisfied (0.351436857 0.648563143) *
     5) Inflight.wifi.service< 0.5 1449     8 satisfied (0.005521049 0.994478951) *
   3) Online.boarding>=3.5 41257 11423 satisfied (0.276874227 0.723125773)
     6) Class=Eco,Eco Plus 14468  6241 neutral or dissatisfied (0.568634227 0.431365773)
      12) Inflight.wifi.service< 4.5 11555  3355 neutral or dissatisfied (0.709649502 0.290350498) *
      13) Inflight.wifi.service>=4.5 2913    27 satisfied (0.009268795 0.990731205) *
     7) Class=Business 26789  3196 satisfied (0.119302699 0.880697301) *
```

Classification - Decision Tree: Confusion Matrix & Accuracy

- Training Data Accuracy:

flight.pred	flight.actual neutral or dissatisfied	satisfied
neutral or dissatisfied	42420	5759
satisfied	4564	30380

```
> pt <- prop.table(confusion.matrix)
> pt[1,1] + pt[2,2]
[1] 0.8758105
```

- Testing Data Accuracy:

flight_test.pred	flight_test.actual neutral or dissatisfied	satisfied	Sum
neutral or dissatisfied	10756	1403	12159
satisfied	1139	7483	8622
Sum	11895	8886	20781

```
> pt <- prop.table(confusion.matrix)
> pt[1,1] + pt[2,2]
[1] 0.8776767
```

Summary:

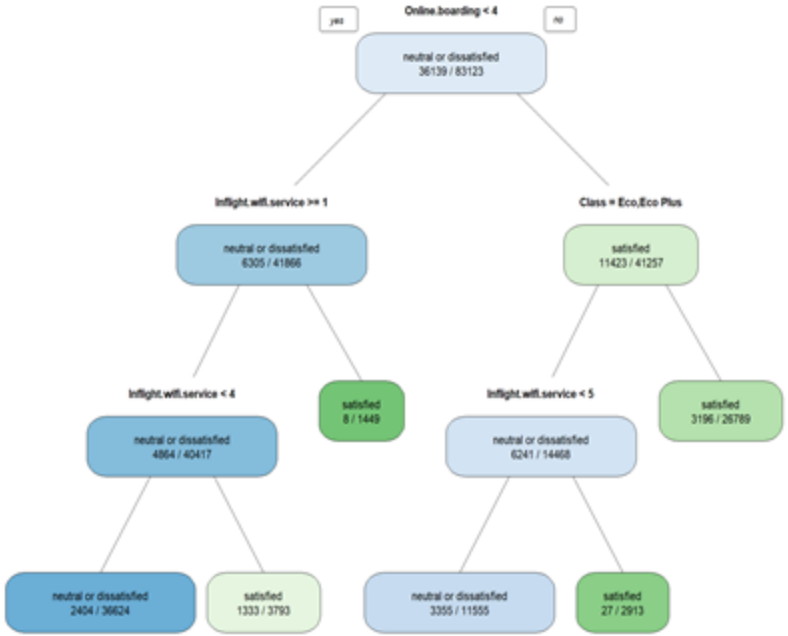
True Positive: Prediction and Actual + 1449+2913+26789= 31151

False Positive: Prediction + and Actual - 8+27+3916= 3951

True Negative: Prediction - and Actual - 2404+3355= 5769

False Negative: Prediction - and Actual + 36624+11555=48179

Classification - Decision Tree: Figure



- If Online.boarding is less than 4, the next split is based on the variable Inflight.wifi.service.
- If Inflight.wifi.service is greater than or equal to 1, then the customer is predicted to be neutral or dissatisfied. Otherwise, the customer is predicted to be satisfied.
- If Inflight.wifi.service is less than 4, the customer is predicted to be neutral or dissatisfied.
- If Online.boarding is greater than or equal to 4, the next split is based on the variable Class. If Class is Eco or Eco Plus, then the customer is predicted to be neutral or dissatisfied.
- If Class is Business, then the customer is predicted to be satisfied.
- If Class is Eco or Eco Plus, the next split is based on the variable Inflight.wifi.service. If Inflight.wifi.service is less than 5, then the customer is predicted to be neutral or dissatisfied. Otherwise, the customer is predicted to be satisfied.

Logistic Regression - Summary

```
Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -6.3574091  0.0754900 -84.215  < 2e-16 ***
GenderMale                      0.0667818  0.0202645   3.296 0.000982 ***
Customer.TypeLoyal Customer     0.9531699  0.0302759  31.483  < 2e-16 ***
Age                            -0.0025450  0.0007101  -3.584 0.000338 ***
ClassEco                       -1.9164264  0.0242069 -79.169  < 2e-16 ***
ClassEco Plus                  -1.6859107  0.0420006 -40.140  < 2e-16 ***
Flight.Distance                 0.0001042  0.0000117   8.906  < 2e-16 ***
Inflight.wifi.service           0.5250466  0.0120888  43.432  < 2e-16 ***
DepartureArrival.time.convenient -0.2465566  0.0079058 -31.187  < 2e-16 ***
Ease.of.Online.booking          -0.1627893  0.0116616 -13.959  < 2e-16 ***
Food.and.drink                  -0.0587177  0.0109972  -5.339 9.33e-08 ***
Online.boarding                 0.6046663  0.0103191  58.597  < 2e-16 ***
Seat.comfort                    0.0396856  0.0115057   3.449 0.000562 ***
Inflight.entertainment          0.2651587  0.0143179  18.519  < 2e-16 ***
Onboard.service                 0.2279178  0.0105319  21.641  < 2e-16 ***
Leg.room.service                0.2677585  0.0087668  30.542  < 2e-16 ***
Baggage.handling                0.0451249  0.0115520   3.906 9.37e-05 ***
Checkin.service                 0.2201116  0.0087848  25.056  < 2e-16 ***
Inflight.service                -0.0033884  0.0122351  -0.277 0.781822
Cleanliness                     0.1435976  0.0124169  11.565  < 2e-16 ***
Departure.Delay.in.Minutes      0.0031439  0.0010167   3.092 0.001986 **
Arrival.Delay.in.Minutes        -0.0068628  0.0009997  -6.865 6.65e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Highest estimate is Customer Loyalty at 0.95317
  - Odds Ratio: exp(0.95317)= 2.59
  - Being a loyal customer is associated with around 159% increase in odds of satisfaction

	Neutral or Dissatisfied	Satisfied
Neutral or Dissatisfied	10563 (TN)	1664 (FN)
Satisfied	1293 (FP)	7205 (TP)

Logistic Regression - Confusion Matrix

```
> logitPredict <- predict(logit.reg, df.test, type = "response")
> logitPredictClass <- ifelse(logitPredict > 0.5, 1, 0)
> actual <- df.test$satisfaction
> predict <- logitPredictClass
> cm <- table(predict, actual)
> tp <- cm[2,2]
> tn <- cm[1,1]
> fp <- cm[2,1]
> fn <- cm[1,2]
> (tp + tn)/(tp + tn + fp + fn)
[1] 0.8573221
> tp/(fn+tp)
[1] 0.8123802
> tn/(fp+tn)
[1] 0.8909413
> fp/(fp+tn)
[1] 0.1090587
> fn/(fn+tp)
[1] 0.1876198
```

Accuracy= 0.8573

TPR/Sensitivity= 0.8124

TNR/Specificity= 0.8909

FPR= 0.1091

FNR= 0.1876

Conclusion and Implications in Business Operations

- Likert scale based ratings on surveys allowed for consistent results
- Highest Effect Size Variables based on Logistic Regression were:
  - Customer Loyalty, Flight Class, Distance, Inflight WiFi, Departure and Arrival time convenience, Ease of Online booking, Online boarding, Inflight entertainment, onboard service, leg room, cleanliness and, delay
- Medium Effect Size Variables:
  - Age, Seat Comfort, Check- In Service,
- Lower Effect Size Variables:
  - Food and Drink, Inflight Service

Conclusion and Implications in Business Operations - Cont.

- Loyal Customers more prone to expecting better services, hence higher effect size on satisfaction
- Customer Satisfaction begins well before flight boarding, based on Likert scale ratings and significance on online booking experience and departure wait times.
    - Airline must also invest in proper digitized booking experiences to increase customer satisfaction
- Surprisingly, In Flight service does not contribute to customer experience as much as the other tested variables as one may expect.
- This data analysis was conducted on data from one airline but these findings may not extrapolated to other airline businesses to measure customer satisfaction.

- Developing a loyal customer base is key to higher level of customer satisfaction.
- Long term, these variables and significance can help track improvements made in operations by comparing scores before and after operational changes.
- Recommendations for further research include:
    - Balancing out customers surveyed from loyal and disloyal customers and cabin class
    - Include more datasets from other airlines for more impartial analysis
    - Use more sophisticated Machine Learning Algorithms to possibly achieve higher accuracy scores

Thank you