Predicting Customer Satisfaction in the Airline Industry

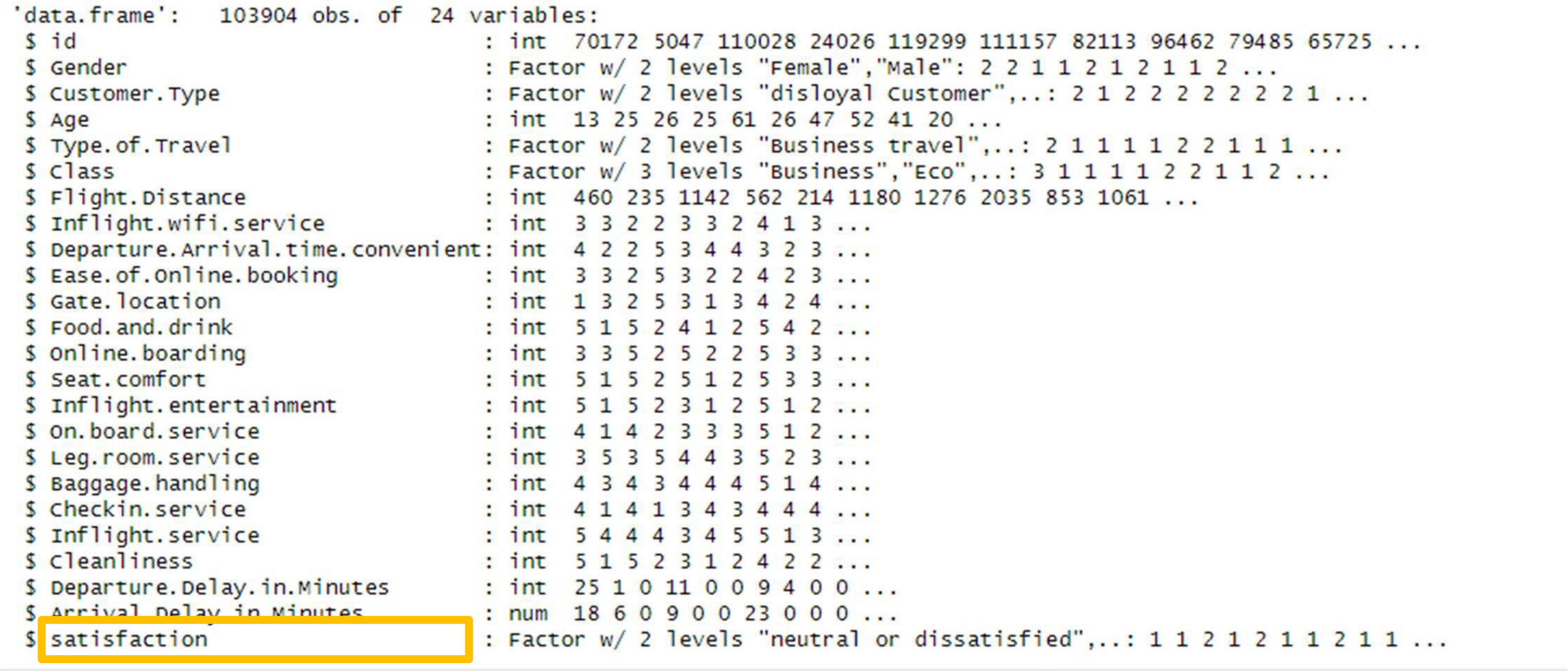
Team 6 – Joshua Farina (joshuapfarina), Henri Salomon (henri_salomon), Raajitha Middi (Raajitha_Middi), Ryan Chandler (zujin87), Alejandro Martinez (amzeta)

Problem Definition and Benefits

- We intend to use airline passenger satisfaction survey data to determine which factors drive customer satisfaction (or dissatisfaction) so that we can provide actionable recommendations on how money should be invested or re-allocated to keep passengers satisfied.
- Customer satisfaction is a key factor in attracting and retaining business. Identifying factors that have the strongest effect on customer satisfaction will provide key insights for airlines to enhance their services, improve the customer experience and optimize business operations.
- Some estimates indicate that increasing customer retention by as little as 5% increases profits by 25%-95%.[1]

Dataset Overview

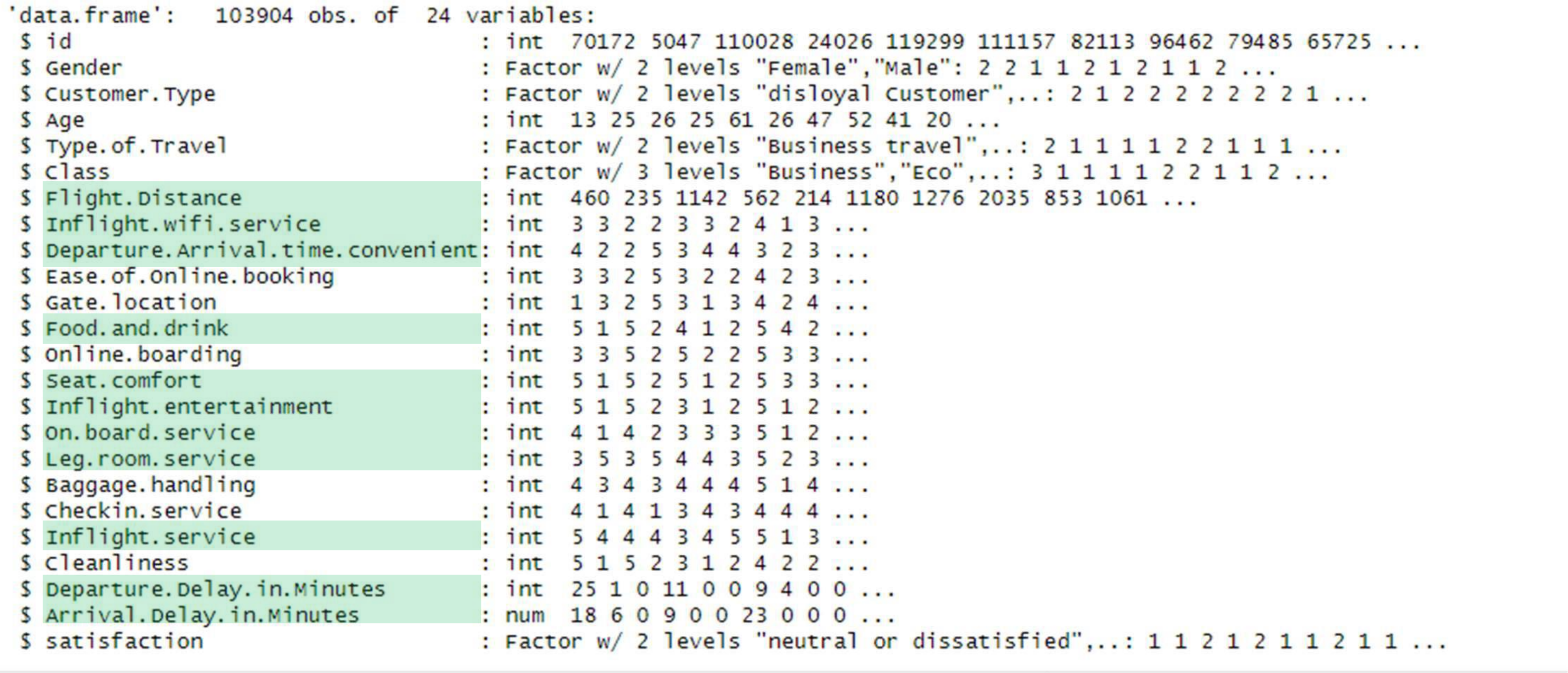
- The dataset is taken from surveys from a US airline. It includes user ratings for multiple aspects of air travel such as food and drink and seat comfort, as well as customer information such as age and gender.[2]



```
'data.frame':   103904 obs. of  24 variables:
 $ id                               : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
 $ 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 ...
 $ Type.of.Travel                   : Factor w/ 2 levels "Business travel",..: 2 1 1 1 1 2 2 1 1 1 ...
 $ 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 ...
 $ Departure.Arrival.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 ...
 $ Gate.location                    : int  1 3 2 5 3 1 3 4 2 4 ...
 $ 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 ...
 $ On.board.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         : num  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 ...
```

Dataset Overview

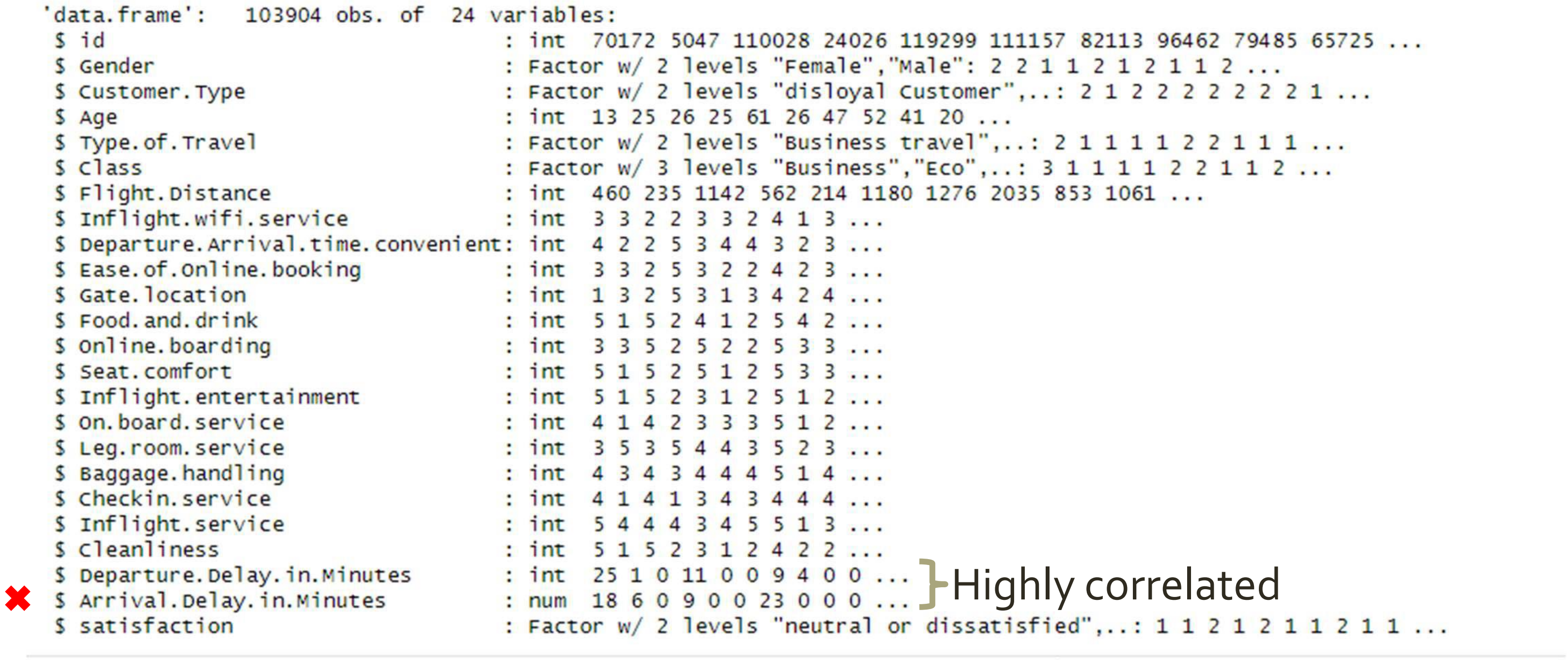
- The dataset is taken from surveys from a US airline. It includes user ratings for multiple aspects of air travel such as food and drink and seat comfort, as well as customer information such as age and gender.[2]



```
'data.frame':   103904 obs. of  24 variables:
 $ id                               : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
 $ 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 ...
 $ Type.of.Travel                   : Factor w/ 2 levels "Business travel",..: 2 1 1 1 1 2 2 1 1 1 ...
 $ 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 ...
 $ Departure.Arrival.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 ...
 $ Gate.location                    : int  1 3 2 5 3 1 3 4 2 4 ...
 $ 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 ...
 $ On.board.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         : num  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 ...
```

Dataset Overview

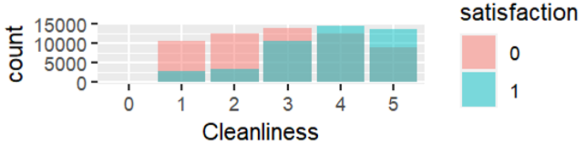
- The dataset is taken from surveys from a US airline. It includes user ratings for multiple aspects of air travel such as food and drink and seat comfort, as well as customer information such as age and gender.[2]



```
'data.frame':   103904 obs. of  24 variables:
 $ id                               : int  70172 5047 110028 24026 119299 111157 82113 96462 79485 65725 ...
 $ 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 ...
 $ Type.of.Travel                   : Factor w/ 2 levels "Business travel",..: 2 1 1 1 1 2 2 1 1 1 ...
 $ 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 ...
 $ Departure.Arrival.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 ...
 $ Gate.location                    : int  1 3 2 5 3 1 3 4 2 4 ...
 $ 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 ...
 $ On.board.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 ...  }Highly correlated
✖ $ Arrival.Delay.in.Minutes        : num  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 ...
```

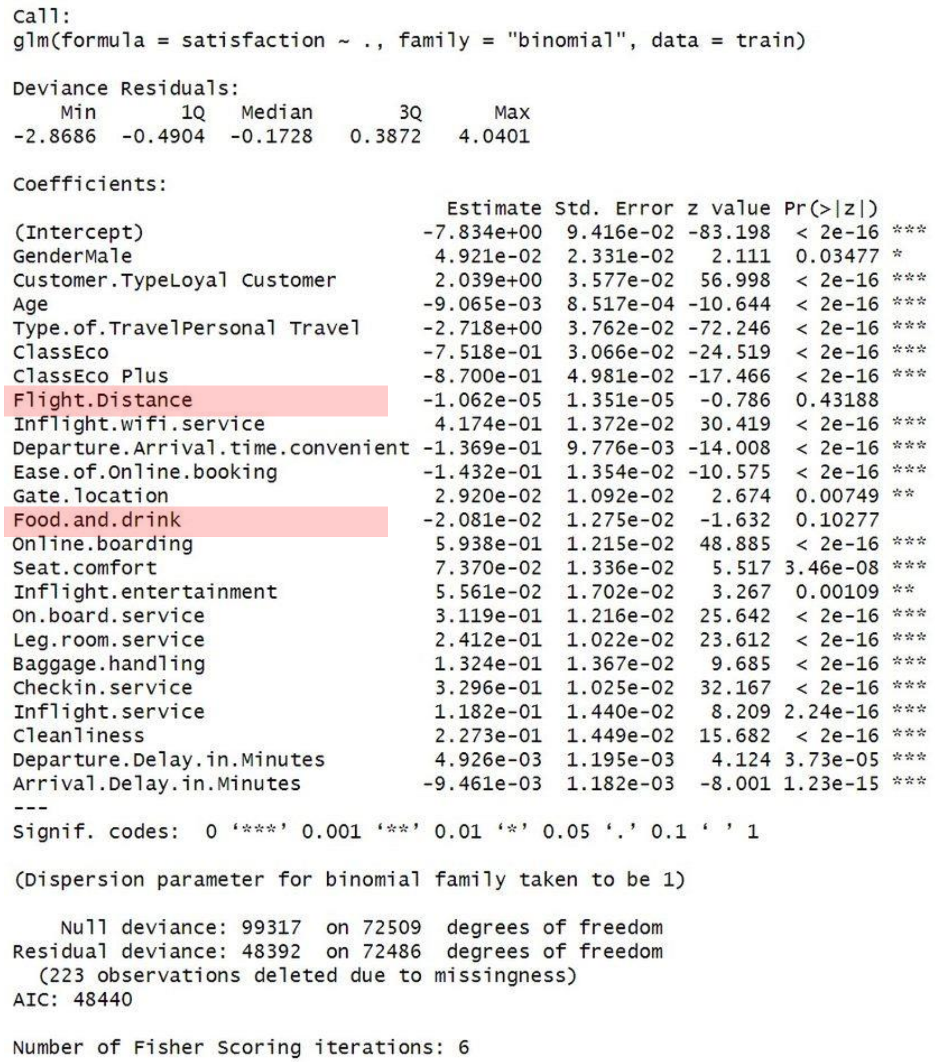
Progress and Challenges

- Data cleaning status
  - Missing values removed
  - Investigating values with a Z-score > 3
  - Manually one-hot encoding data to establish baselines
- Initial findings
  - Flight distance is not statistically significant
  - Business class is more likely to be satisfied
  - Some ratings seem irrelevant to customer satisfaction



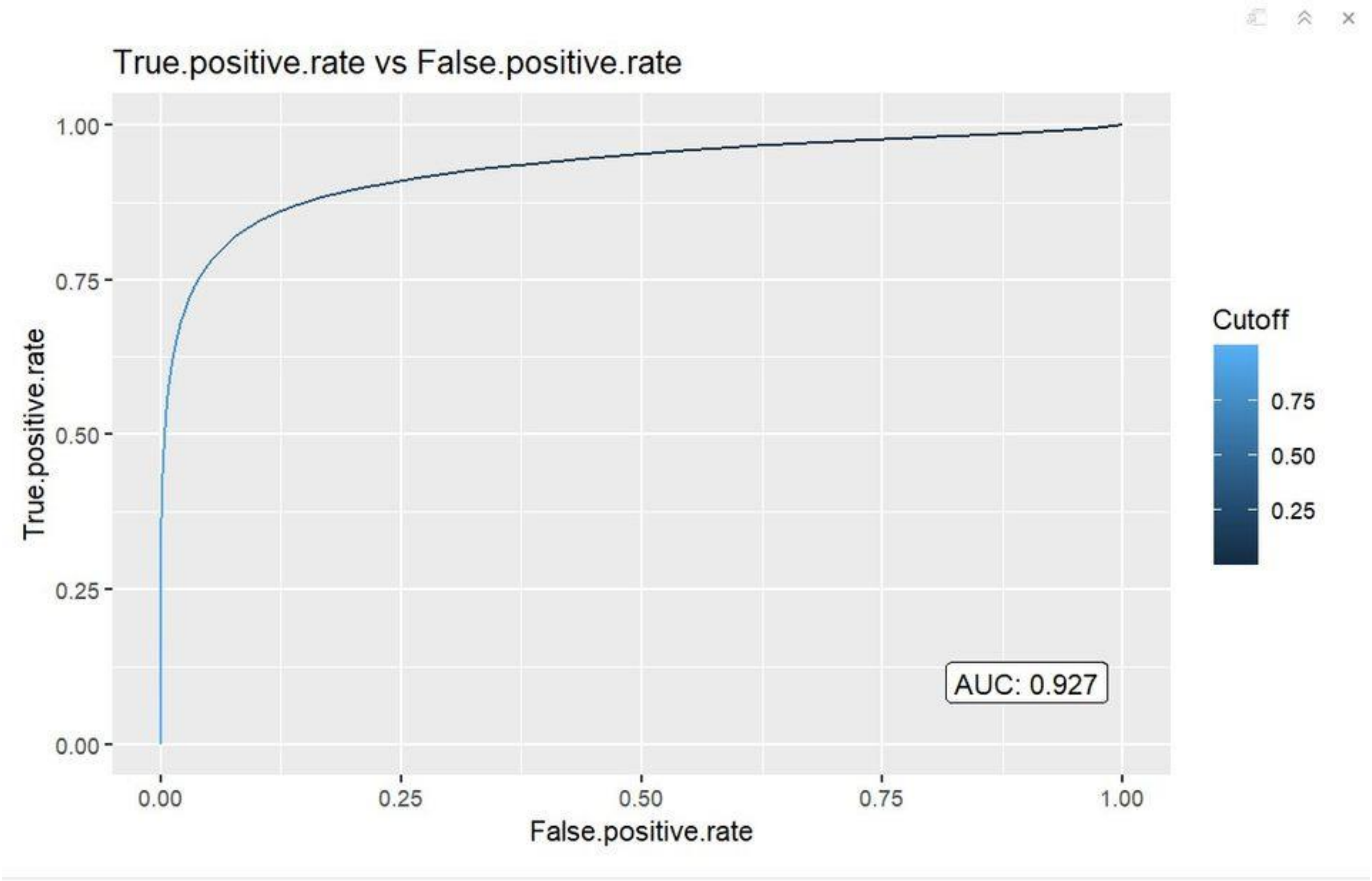
- Will not be able to merge flight path data to try to infer airport locations due to insufficient information:
  - Main dataset does not specify number of legs in a trip
  - Flight distances in main dataset are too similar in many instances which would make it impossible to narrow down with any certainty which airports were involved in the trip
- Understanding how to map a binary satisfaction variable to a Likert scale variable
- American Customer Satisfaction Index (ACSI) Dataset seems to contradict some of the findings from the primary dataset[3]

Modelling: First Pass Logistic Regression

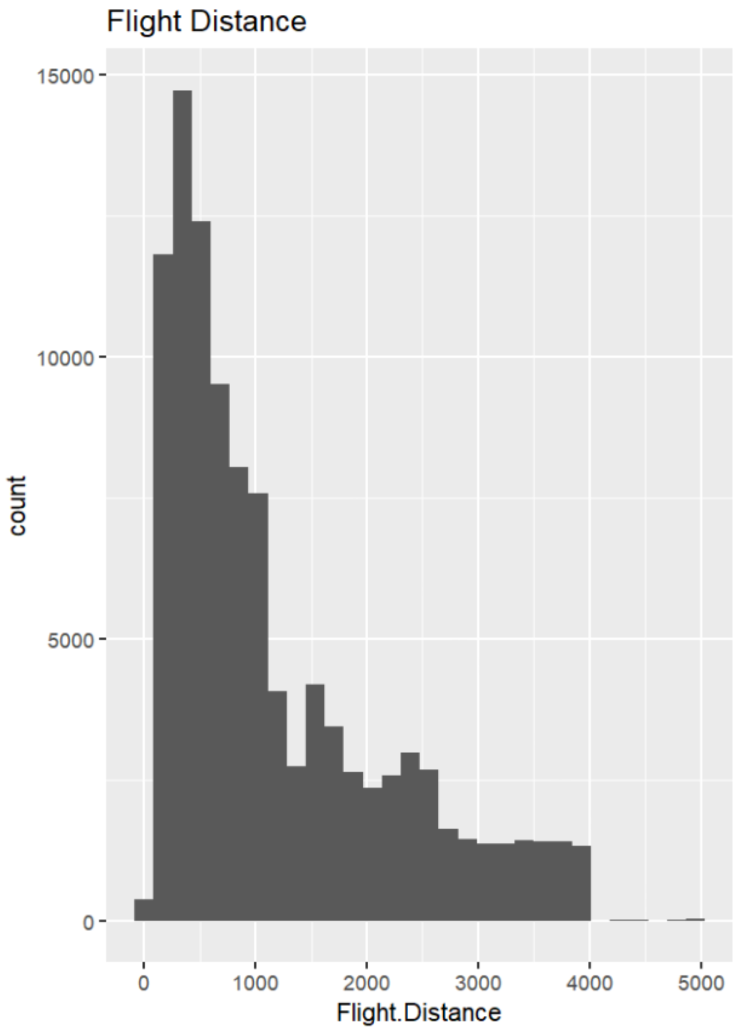


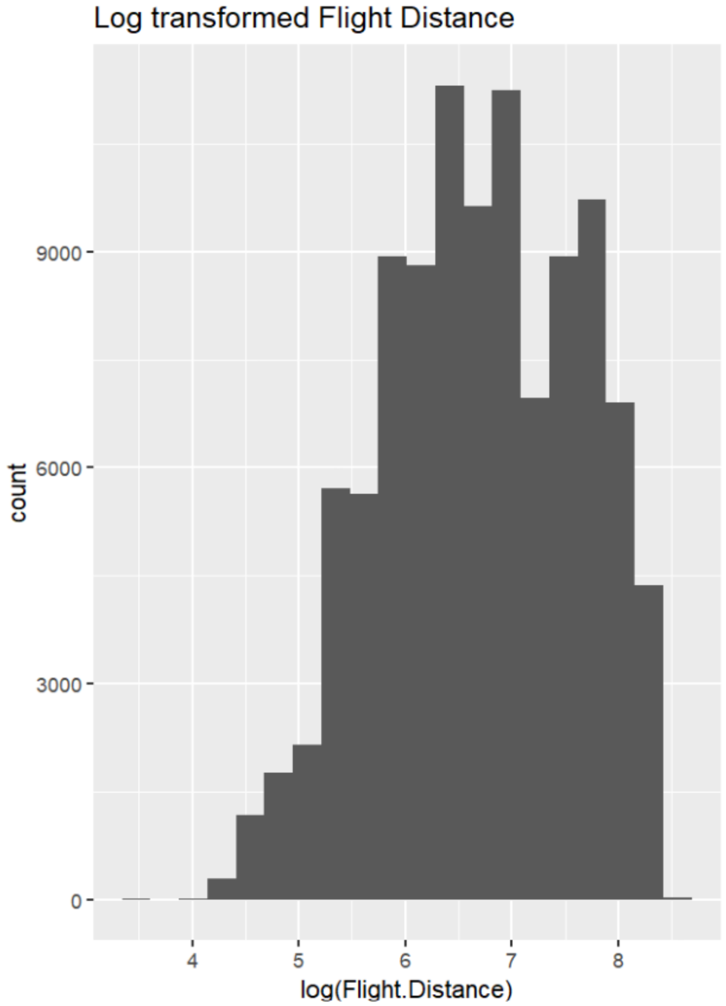
```
Call:
glm(formula = satisfaction ~ ., family = "binomial", data = train)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.8686  -0.4904  -0.1728   0.3872   4.0401  

Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)    
(Intercept)                       -7.834e+00  9.416e-02 -83.198  < 2e-16 ***
GenderMale                         4.921e-02  2.331e-02   2.111  0.03477 *  
Customer.TypeLoyal Customer        2.039e+00  3.577e-02  56.998  < 2e-16 ***
Age                               -9.065e-03  8.517e-04 -10.644  < 2e-16 ***
Type.of.TravelPersonal Travel     -2.718e+00  3.762e-02 -72.246  < 2e-16 ***
ClassEco                          -7.518e-01  3.066e-02 -24.519  < 2e-16 ***
ClassEco Plus                     -8.700e-01  4.981e-02 -17.466  < 2e-16 ***
Flight.Distance                   -1.062e-05  1.351e-05  -0.786  0.43188    
Inflight.wifi.service              4.174e-01  1.372e-02  30.419  < 2e-16 ***
Departure.Arrival.time.convenient -1.369e-01  9.776e-03 -14.008  < 2e-16 ***
Ease.of.Online.booking            -1.432e-01  1.354e-02 -10.575  < 2e-16 ***
Gate.location                      2.920e-02  1.092e-02   2.674  0.00749 ** 
Food.and.drink                    -2.081e-02  1.275e-02  -1.632  0.10277    
Online.boarding                    5.938e-01  1.215e-02  48.885  < 2e-16 ***
Seat.comfort                       7.370e-02  1.336e-02   5.517 3.46e-08 ***
Inflight.entertainment             5.561e-02  1.702e-02   3.267  0.00109 ** 
On.board.service                   3.119e-01  1.216e-02  25.642  < 2e-16 ***
Leg.room.service                   2.412e-01  1.022e-02  23.612  < 2e-16 ***
Baggage.handling                   1.324e-01  1.367e-02   9.685  < 2e-16 ***
Checkin.service                    3.296e-01  1.025e-02  32.167  < 2e-16 ***
Inflight.service                   1.182e-01  1.440e-02   8.209 2.24e-16 ***
Cleanliness                        2.273e-01  1.449e-02  15.682  < 2e-16 ***
Departure.Delay.in.Minutes         4.926e-03  1.195e-03   4.124 3.73e-05 ***
Arrival.Delay.in.Minutes          -9.461e-03  1.182e-03  -8.001 1.23e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99317  on 72509  degrees of freedom
Residual deviance: 48392  on 72486  degrees of freedom
  (223 observations deleted due to missingness)
AIC: 48440

Number of Fisher Scoring iterations: 6
```

Modelling: First Pass Logistic Regression



Transformations





Modelling

- Logistic Regression - Analysis of Distributions for possible transformations, Consider Regularization
  - Look at magnitude of coefficients
  - Compare results of transformed vs non-transformed data
  - Run diagnostics
  - Apply transformations as needed
- Random Forest
  - Discover relative feature importance.
- Support Vector Machines
  - Analysis of Distributions for possible transformations, Center and Scale Data, Optimal Choice of Kernel, Optimal Value of C/lambda.
- The models will be compared using ROC-AUC on a reserved validation set.

Other Datasets and Research Findings

- Pereira et al. (2023) provided an exhaustive and updated literature review about airlines' passenger satisfaction while analyzing the most influential factors on satisfaction before and during the COVID-19 pandemic.[3]
- Lucini et al. (2020) found practical implications to maximize customer satisfaction: focus on customer service for first class passengers, comfort for premium economy passengers, and checking luggage and waiting time for economy class travelers.[4]
- Morgeson et al. (2023) focused on the American Customer Satisfaction Index (ACSI) and helped us better explore and analyze it.[5]

Sources

1. https://www.forbes.com/sites/forbesbusinesscouncil/2022/12/12/customer-retention-versus-customer-acquisition/?sh=3cef46fc1c7d
2. https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction
3. https://www.sciencedirect.com/science/article/pii/S0969699723000844#bib20
4. https://www.sciencedirect.com/science/article/abs/pii/S0969699719302959
5. https://www.sciencedirect.com/science/article/pii/S2352340923002421
6. https://psycnet.apa.org/record/1989-10632-001
7. https://www.techscience.com/cmc/v75n1/51460/pdf