Final Project: Finding Variables that Contribute to a Positive Shinkansen Experience

STAT 306: Group D4

Divyadarshan Punjabi, Tyler Wong, Safa Sajid

05/04/2024

Introduction

The Shinkansen, Japan's iconic bullet train system, is a network of high speed trains that span the country and are renowned for their speed, safety, punctuality, and comfort. Despite the quality of service, it is a never-ending battle to increase customer satisfaction, and understanding the passenger experience is crucial for maintaining quality and staying competitive.

In order to gain a deeper understanding of the factors influencing customer's perception of the train system, a comprehensive study was conducted on a random sample of passengers. The survey questioned travelers on their opinions of the train's service quality, entertainment, comfort, and more. The survey data was paired with a record of the on-time performance of the train each passenger had taken. The final data sets were 'Traveldata' and 'Surveydata', which were further separated into training and testing sets.

These 2 data sets when combined, encompassed 5 continuous variables and 20 categorical variables:

Age, Travel Class (Eco or Business), Travel Distance, Departure Delay, Arrival Delay, Seat Comfort, Seat Class (Green Car, Ordinary), Arrival Time Convenient, Catering, Platform Location, Onboard Wifi Service, Onboard Entertainment, Online Support and Overall Experience. With the response variable being 'Overall Experience' coded as a binary indicator of satisfication, we designed our project question to be:

What are the 3 strongest predictors of a passenger's overall trip satisfication? We propose to answer this question by exploring key features of our data through EDA, performing feature selection using Cramer's V, reducing dimension using Principle Component Analysis, and finally selecting a model of size 3 using LASSO, AIC values, and backwards selection.

The results of this research will allow us to see what factors need to be prioritized in order to provide the best quality of service to commuters. By identifying the key predictors of trip satisfaction, our study aims to generate actionable insights to fine-tune travel services, enhance customer loyalty, and fortify Shinkansen's prestigous reputation.

Exploratory Data Analysis/Visualizations

We will begin by checking to see if our response variable, 'Overall Experience' is balanced.

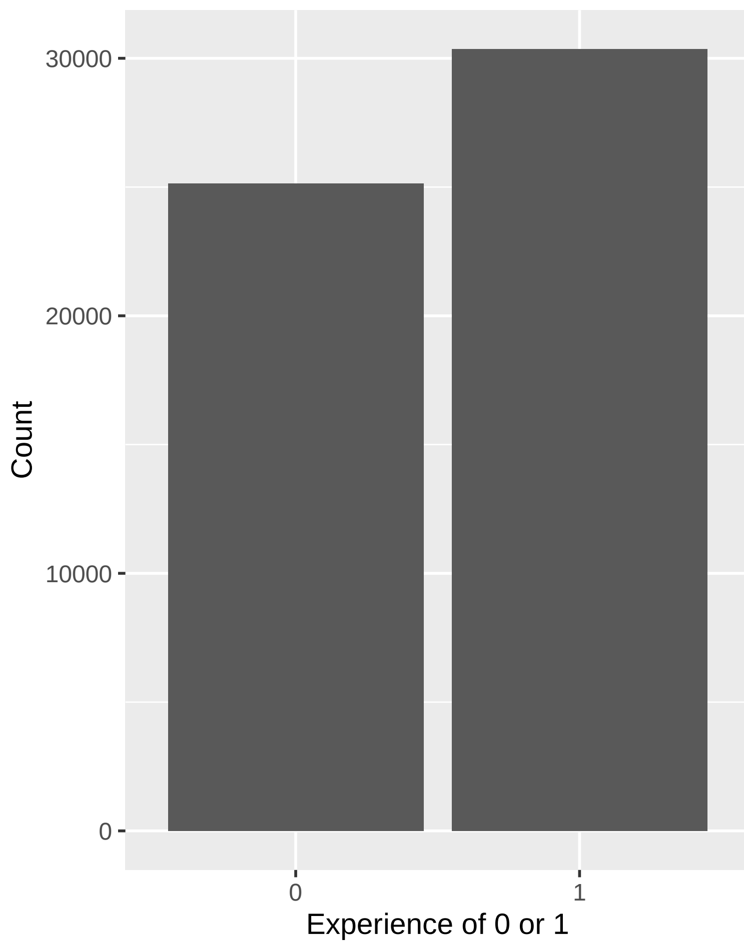


Figure 1: Distribution of Count of Overall Experience (Response Variable)

There is a relatively even distribution of overall experiences, with proportions of about 45% for an outcome of 0 (poor) and 55% for an outcome of 1 (good), so the dataset is balanced, and there is no need to undersample/oversample any data.

We move on to check the distributions of our numerical variables.

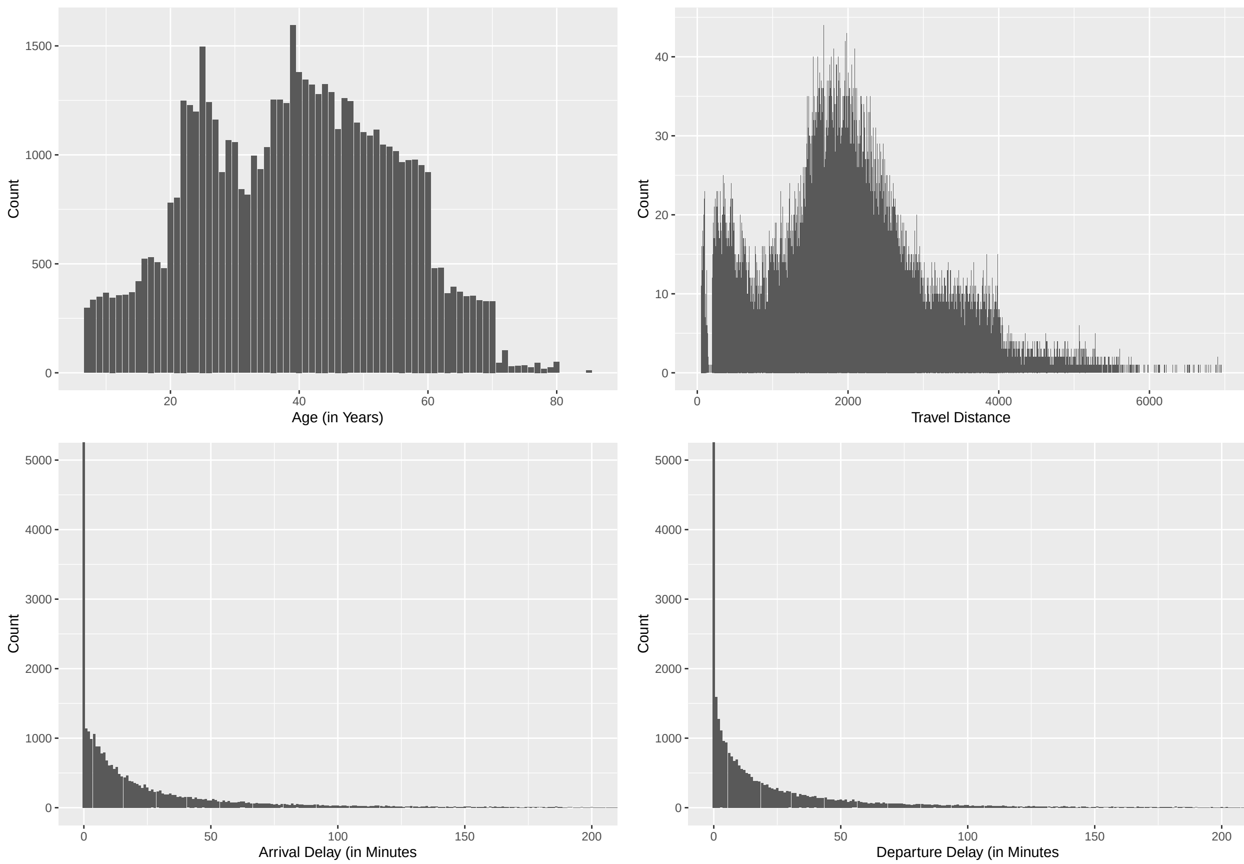


Figure 2: Distributions of Age, Travel Distance, Arrival Delay, Departure Delay

The age and travel distance values follow a roughly normal distribution, with the majority of ages being between 25-60 and the majority of distances being between 1000 and 3000.

The distributions of arrival delay and deparature delay appear right-skewed, however this is expected behaviour for a variable denoting delay. We also note that the distribution of arrival delay and departure delay appear very similar, so we can plot them against each other to check for a linear relationship, and to see if its likely one of them should be removed from the model.

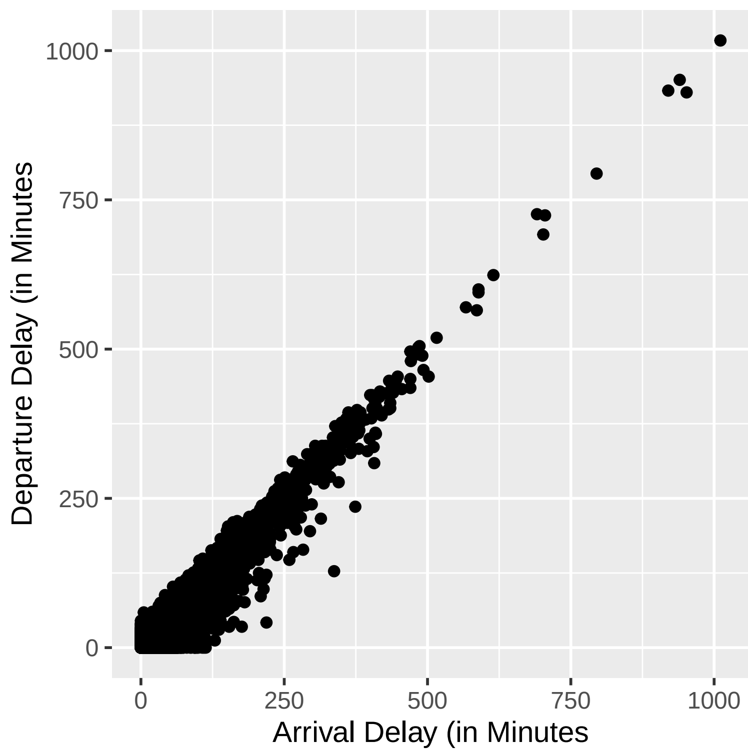


Figure 3: Scatterplot of Departure Delay vs ARrival Delay

These look quite positive correlated, so it is likely we will remove one of the variables during feature selection.

We must also check the distributions of the categorical variables denoting customer information, to check for balanced data.

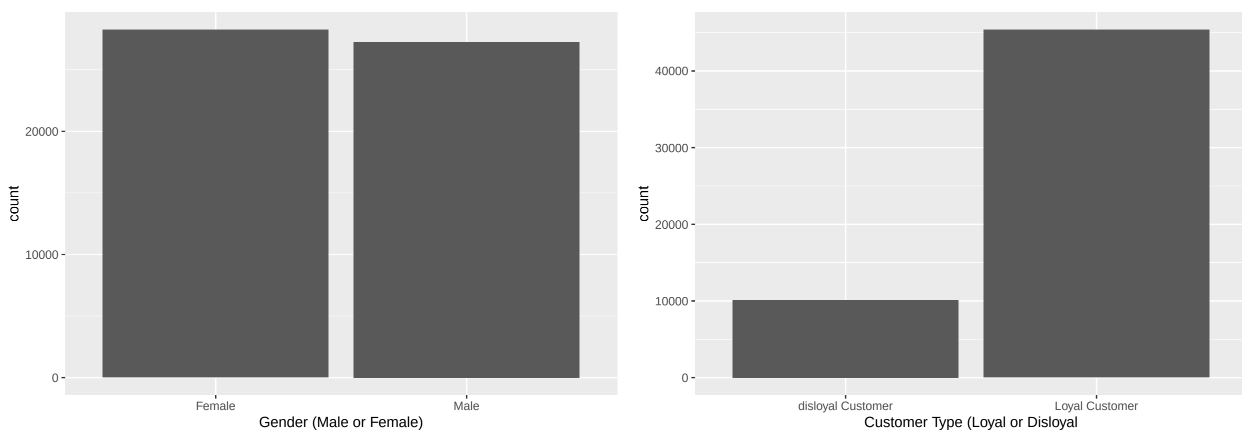


Figure 4: Distributions of Gender and Customer Type

There appears to be an imbalance in the type of customer. Having more loyal customers may skew data, in terms of the frequency of these customers giving their surveys, as well as their overall opinion of the system.

We can do a quick check to see if this imbalance is shown in the overall experience (response variable).

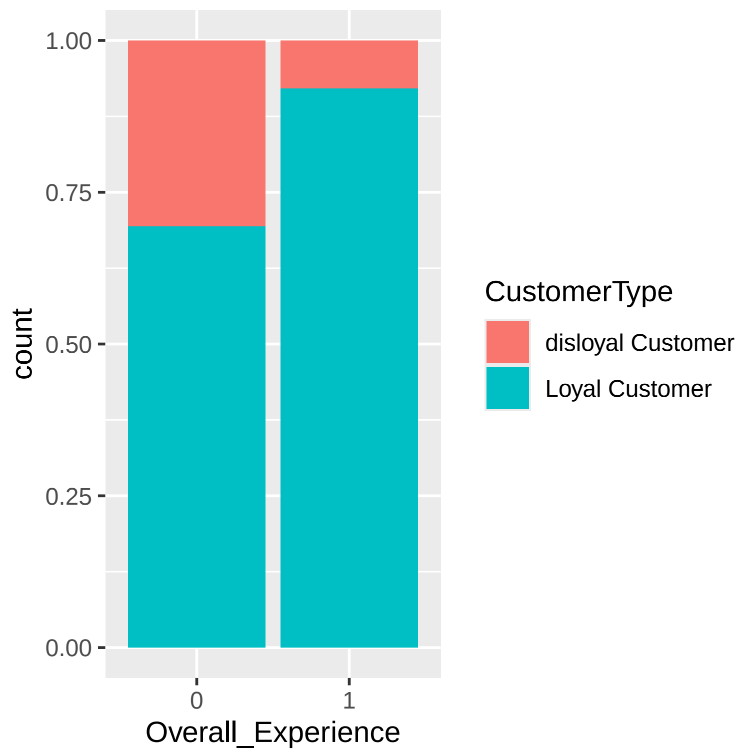


Figure 5: Proportions of customer type in overall experience

It does appear that more loyal customers have a better experience overall, but this isn't an enormous difference, so it appears the imbalance of customer type won't be skewing the results too badly.

Overall, it appeared the dataset is decently well balanced, and there wasn't a pressing need to undersample or oversample any variables. We moved on to perform our analysis.

Feature Selection

Our model involves 25 variables (including categorical and continous variables). Of which, the categorical variables involve 5-7 different levels of categories. Our overall data, after processing, involves over 90000 rows. This makes our model significantly complex.

Therefore, we first conducted feature selection to reduce model complexity.

But, what should our approach look like?

Understanding Cramer's V. Cramer's V is a statistic that will be used to measure the association between two categorical variables, offering a value from 0 to 1. It is calculated from the chi-squared statistic from a contingency table, which assesses the independence of two variables. The formula for Cramer's V is:

$$V = \sqrt{\frac{\chi^2/n}{\min(k-1, r-1)}}$$

where χ^2 is the chi-squared statistic, n is the total number of observations, k is the number of columns, and r is the number of rows in the table. A Cramer's V near 0 signifies a weak association, and one close to 1 indicates a strong associtomer satisfaction.

Cramer's V in the Context of Our Project. In our study, we applied Cramer's V to evaluate the relationship between various categorical predictors and our response variable, 'Overall_Experience'. This measure guided us in understanding the extent to which different factors affect the overall customer experience.

Additionally, we used Cramer's V for detecting multicollinearity among categorical predictors. Multicollinearity, where predictors are highly inter-correlated, can compromise the integrity of statistical inferences.

Significance of Cramer's V in feature selection process. High values of Cramer's V between pairs of variables will highlight redundancies and strategic associations, influencing our decision to remove features that exhibit multicollinearity, and choose variables that best explain 'Overall_Experience' of commuters.

```
##                            Var1                    Var2 Chi_Squared       P_Value
## X-squared  Overall_Experience            Seat_comfort          NA            NA
## X-squared1 Overall_Experience              Seat_Class   0.3558417  5.508247e-01
## X-squared2 Overall_Experience Arrival_time_convenient  13.5929257  3.452912e-02
## X-squared3 Overall_Experience                Catering 618.0773194 2.937796e-130
## X-squared4 Overall_Experience       Platform_location          NA            NA
## X-squared5 Overall_Experience     Onboardwifi_service 529.2869303 4.115076e-111
##               Cramers_V
## X-squared           NaN
## X-squared1 0.006153337
## X-squared2 0.038031088
## X-squared3 0.256450579
## X-squared4          NaN
## X-squared5 0.237316462
```

We see here that there are quite a chi-squared tests that failed. We therefore chose to remove the rows that failed. Generally, a Cramer's V association of 0.25 or above shows significant association. We have filtered the dataframe to check for this condition too.

```
##                                 Var1                    Var2 Chi_Squared
## X-squared3        Overall_Experience                Catering    618.0773
## X-squared6        Overall_Experience Onboard_entertainment   3808.0624
## X-squared8        Overall_Experience    Onlinebooking_Ease   1806.3510
## X-squared10       Overall_Experience                Leg_room    992.2061
## X-squared11       Overall_Experience      Baggage_handling    857.7691
## X-squared16       Overall_Experience            CustomerType    718.9507
## X-squared18       Overall_Experience            Travel_Class    927.8679
## X-squared54  Arrival_time_convenient                Catering  11754.2662
## X-squared72                 Catering Onboard_entertainment   8717.1552
## X-squared99      Onboardwifi_service Onboard_entertainment  10907.9394
## X-squared101     Onboardwifi_service    Onlinebooking_Ease  15203.0022
## X-squared123   Onboard_entertainment            Travel_Class    628.0131
## X-squared136      Onlinebooking_Ease                Leg_room   5293.4418
## X-squared137      Onlinebooking_Ease      Baggage_handling   7159.9584
## X-squared154                Leg_room      Baggage_handling   5528.5730
## X-squared189              TypeTravel            Travel_Class   2519.0932
##                    P_Value Cramers_V
## X-squared3   2.937796e-130 0.2564506
## X-squared6    0.000000e+00 0.6365526
## X-squared8    0.000000e+00 0.4384129
```

```
## X-squared10  4.335564e-211 0.3249251
## X-squared11  3.666017e-183 0.3021117
## X-squared16  7.617873e-157 0.2765871
## X-squared18  8.585566e-204 0.3142139
## X-squared54   0.000000e+00 0.4565669
## X-squared72   0.000000e+00 0.3931824
## X-squared99   0.000000e+00 0.4398231
## X-squared101  0.000000e+00 0.5192438
## X-squared123 2.110091e-132 0.2585036
## X-squared136  0.000000e+00 0.3063908
## X-squared137  0.000000e+00 0.3903485
## X-squared154  0.000000e+00 0.3430076
## X-squared189  0.000000e+00 0.5177313
```

We were now able to narrow down our search to the 7 predictors that show high association with `Overall_Experience` in the first 7 rows, when categorical variables are concerned.

Upon checking for multicollinearity between any of the 7 predictors - `Catering`, `Onboard_entertainment`, `Onlinebooking_Ease`, `Leg_room`, `Baggage_handling`, `CustomerType`, and `Travel_Class` - we find `Onboard_entertainment`, `Onlinebooking_Ease`, `Baggage_handling`, and `CustomerType` to be the most significant.

`Onboard_entertainment` trumps over `Catering` and `Travel_Class`, due to its stronger association of 0.6365526 with `Overall_Experience`. Similary, `Onlinebooking_Ease` trumps over `Leg_room`.

The chosen 4 predictors are not correlated to each other and demonstrate strong relationships with the response variable, `Overall_Experience`.

We moved on to check for multicollinearity amongst numerical variables using a correlation matrix:

```
##                                Age Travel_Distance DepartureDelay_in_Mins
## Age                    1.000000000      -0.2553502            0.003081416
## Travel_Distance       -0.255350203       1.0000000            0.107836205
## DepartureDelay_in_Mins 0.003081416       0.1078362            1.000000000
## ArrivalDelay_in_Mins   0.005747269       0.1051325            0.967571942
##                       ArrivalDelay_in_Mins
## Age                            0.005747269
## Travel_Distance                0.105132518
## DepartureDelay_in_Mins         0.967571942
## ArrivalDelay_in_Mins           1.000000000
```

Principle Component Analysis

In an attempt to further reduce the dimensions of our dataset, we performed principle component analysis on our numerical variables.

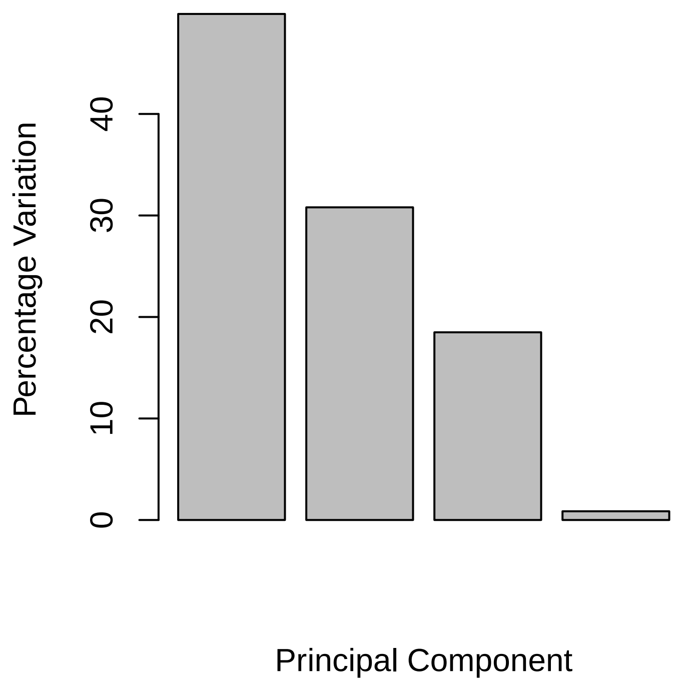


Figure 6: Princple Component Analysis Plot for Age, Travel Distance, Departure Delay, and Arrival Delay

We are more concerned about the `Overall_Experience` of onboarding passengers. Hence, we neglected `ArrivalDelay_in_Mins` since the survey collects this information after the passengers off-board.

Model Selection and Model Fitting

Now that we have successfully reduced data complexity, we moved forward to fit a logistic regression on our population data with our chosen variables.

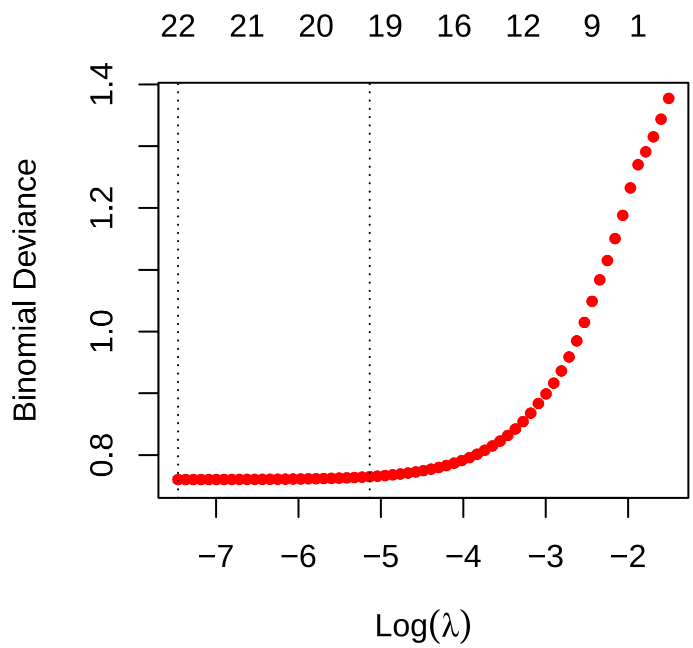
```
##
## Call:
## glm(formula = Overall_Experience ~ Onboard_entertainment + Onlinebooking_Ease +
##     Baggage_handling + CustomerType + Age + Travel_Distance +
##     DepartureDelay_in_Mins, family = binomial(), data = full_complete)
##
## Coefficients:
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          9.757e-01  6.337e-01   1.540 0.123630
## Onboard_entertainmentacceptable     -1.754e+00  5.361e-01  -3.272 0.001068
## Onboard_entertainmentexcellent       2.510e+00  5.368e-01   4.676 2.93e-06
## Onboard_entertainmentextremely poor  7.476e-01  5.382e-01   1.389 0.164801
## Onboard_entertainmentgood            3.849e-01  5.359e-01   0.718 0.472648
## Onboard_entertainmentneed improvement -1.841e+00  5.363e-01  -3.433 0.000596
```

```
## Onboard_entertainmentpoor            -1.607e+00  5.366e-01  -2.995 0.002743
## Onlinebooking_Easeacceptable         -1.458e+00  3.725e-01  -3.915 9.05e-05
## Onlinebooking_Easeexcellent          -6.004e-01  3.724e-01  -1.612 0.106868
## Onlinebooking_Easeextremely poor     -1.183e+01  4.563e+01  -0.259 0.795515
## Onlinebooking_Easegood               -5.206e-01  3.722e-01  -1.399 0.161811
## Onlinebooking_Easeneed improvement   -1.834e+00  3.726e-01  -4.923 8.53e-07
## Onlinebooking_Easepoor               -2.511e+00  3.732e-01  -6.727 1.73e-11
## Baggage_handlingacceptable           -2.771e-01  2.696e-01  -1.028 0.304112
## Baggage_handlingexcellent             1.203e+00  2.693e-01   4.468 7.91e-06
## Baggage_handlinggood                  5.610e-01  2.690e-01   2.085 0.037070
## Baggage_handlingneed improvement      1.402e-01  2.702e-01   0.519 0.604004
## Baggage_handlingpoor                  2.606e-01  2.713e-01   0.961 0.336735
## CustomerTypedisloyal Customer        -1.077e+00  3.906e-02 -27.566  < 2e-16
## CustomerTypeLoyal Customer            2.443e-01  3.174e-02   7.697 1.39e-14
## Age                                   4.780e-03  6.545e-04   7.303 2.81e-13
## Travel_Distance                      -2.546e-05  9.749e-06  -2.611 0.009021
## DepartureDelay_in_Mins               -4.527e-03  2.611e-04 -17.341  < 2e-16
##
## (Intercept)
## Onboard_entertainmentacceptable      **
## Onboard_entertainmentexcellent       ***
## Onboard_entertainmentextremely poor
## Onboard_entertainmentgood
## Onboard_entertainmentneed improvement ***
## Onboard_entertainmentpoor            **
## Onlinebooking_Easeacceptable         ***
## Onlinebooking_Easeexcellent
## Onlinebooking_Easeextremely poor
## Onlinebooking_Easegood
## Onlinebooking_Easeneed improvement   ***
## Onlinebooking_Easepoor               ***
## Baggage_handlingacceptable
## Baggage_handlingexcellent            ***
## Baggage_handlinggood                 *
## Baggage_handlingneed improvement
## Baggage_handlingpoor
## CustomerTypedisloyal Customer        ***
## CustomerTypeLoyal Customer           ***
## Age                                  ***
## Travel_Distance                      **
## DepartureDelay_in_Mins               ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 129475  on 93988  degrees of freedom
## Residual deviance:  71414  on 93966  degrees of freedom
## AIC: 71460
##
## Number of Fisher Scoring iterations: 10
```

We can see that since our categorical variables have several levels each, it seems difficult to understand which global predictors are essential. We chose to use other methods to better choose predictors.

Model Selection using Lasso, AIC and Backward Selection

We fit our full model to a Lasso regression, and gradually increase the strength of the regularization parameter. This way, we can easily visualize more significant parameters.




```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                                                 s1
## (Intercept)                             0.088100967
## Onboard_entertainment                   .
## Onboard_entertainmentacceptable        -1.674902229
## Onboard_entertainmentexcellent          2.181797525
## Onboard_entertainmentextremely poor     0.433887323
## Onboard_entertainmentgood               0.292962353
## Onboard_entertainmentneed improvement -1.747612091
## Onboard_entertainmentpoor              -1.485675173
## Onlinebooking_Easeacceptable            .
## Onlinebooking_Easeexcellent             0.766934408
## Onlinebooking_Easeextremely poor        .
## Onlinebooking_Easegood                  0.867598163
## Onlinebooking_Easeneed improvement     -0.331090251
## Onlinebooking_Easepoor                 -0.923401067
## Baggage_handlingacceptable             -0.623110349
## Baggage_handlingexcellent               0.648756855
## Baggage_handlinggood                    0.039625096
## Baggage_handlingneed improvement       -0.196384595
## Baggage_handlingpoor                   -0.037479102
## CustomerTypedisloyal Customer          -0.972206758
## CustomerTypeLoyal Customer              0.187027309
```

```
## Age                                  0.002356687
## Travel_Distance                      .
## DepartureDelay_in_Mins              -0.003143383
```

We observed the `Travel_Distance` variable being forced to zero. We then increased the strength parameter a bit further to observe its effects.

```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                                             s1
## (Intercept)                         -0.2817019
## Onboard_entertainment                .
## Onboard_entertainmentacceptable     -0.7193269
## Onboard_entertainmentexcellent       2.0032472
## Onboard_entertainmentextremely poor  .
## Onboard_entertainmentgood            0.7006682
## Onboard_entertainmentneed improvement -0.7581119
## Onboard_entertainmentpoor           -0.4145802
## Onlinebooking_Easeacceptable         .
## Onlinebooking_Easeexcellent          0.4512165
## Onlinebooking_Easeextremely poor     .
## Onlinebooking_Easegood               0.5478915
## Onlinebooking_Easeneed improvement  -0.1118917
## Onlinebooking_Easepoor              -0.4348406
## Baggage_handlingacceptable          -0.2683590
## Baggage_handlingexcellent            0.3613980
## Baggage_handlinggood                 .
## Baggage_handlingneed improvement     .
## Baggage_handlingpoor                 .
## CustomerTypedisloyal Customer       -0.6373816
## CustomerTypeLoyal Customer           .
## Age                                  .
## Travel_Distance                      .
## DepartureDelay_in_Mins               .
```

Here, we observed `Age` and `DepartureDelay_in_Mins` parameters also had null coefficients, along with `Travel_Distance`. We were able to remove these from our analysis too.

We also observed some critical information over here. For example, `Onboard_entertainmentexcellent` has the highest postitive regeression coefficient of 2.0032472. With a targeted strategic approach, Shinkansen trains maintenance team can work to ensure better onboard entertainment experience of its passengers to significantly improve their `Overall_Experience`.

Note: We decided to not perform stepwise selection (in any direction) due to the complex nature of our data, and the computational demands of running this algorithm. Instead of performing an exhaustive search and checking for all model combinations (even if redundant), we decided to optimize our search process.

Moving forward, we attempted to replicate the step wise selection process by simulating a function with a simple for loop that fits a logistic regression for some parameters and returns the models AIC values. The catch here is that we only perform this function for a parameter size of 3 predictors since that's what we are interested in. Then, we simply found the model with the lowest AIC value, and its predictor variable combination.

```
predictors <- c("Onboard_entertainment", "Onlinebooking_Ease", "Baggage_handling",
                "CustomerType")
combinations <- combn(predictors, 3)
```

```r
aic_values <- sapply(1:ncol(combinations), function(i) {
  formula_str <- paste("Overall_Experience ~", paste(combinations[, i], collapse = " + "))
  formula <- as.formula(formula_str)
  model <- glm(formula, data = full_complete, family = "binomial")
  AIC(model)
})
```

We then found the combination of variables with the lowest AIC value, which we found to be:

'Onboard_Entertainment, OnlineBooking_Ease, CustomerType'

Conclusion

From our analysis, the three variables that contribute the most to a person having a positive experience on a Shinkansen is the customer's loyalty, the quality of the onboard entertainment, and the ease of online booking. This insight can be used to help increase the amount of customers having a positive experience by creating stratetgies to push more people to become loyal customer and improving onboard entertainment as well as the online booking experience.

One thing that could have been done to improve the study would be to make use of the training and testing splits given on kaggle. Instead of only using the training data and fitting models on itself, we could have used a training/testing split to ensure the model is more suitable in general and lower the risk of overfitting. Additionally, given the large number of variables, we could have chosen more variables instead of the best three; there could be other variables that are also significant but are ignored due to choosing only three.