# STA303 Final Project

## How do specific factors influence airline passenger satisfaction in contemporary air travel?

Name: Yuyang Chen
Student number: 1007978191

## 1. Introduction

With globalization and technological advances, air travel has become the preferred mode of transportation for millions of people across the world, and the importance of passenger satisfaction has become a major concern. This study aims to utilize the results of an airline's mediation survey to develop a model for understanding and explaining the multiple factors that influence passenger satisfaction. Additionally, attention to this issue is supported by large amount of academic literature, which underscores the importance of investigating the factors that influence passenger satisfaction.

Firstly, Archana and Subha's study focused on discussing the connection between traditionally service quality and passenger satisfaction however, based on these results, this study will more focused on how modernized digital services will affect this outcome. Next, the study by Jiang et al. predicted passenger satisfaction by using advanced statistical models such as recursive feature elimination and random forests. This approach goes far beyond traditional statistical analysis in terms of technical complexity. In contrast, my study uses a generalized linear model try to analyze in more concise way, and makes the model easier to understand and interpret.

Finally, Park et al.'s study provides a macro perspective for understanding the impact of service quality through path analysis. In contrast to their study, my research will take a more micro perspective, which will more focusing on how individual service quality affect passenger experience.

# 2. Method

## 2.1 Choice of Method

Since our problem is to determine whether the passenger's final level of satisfaction is satisfied or not satisfied, which represents our response variable is a binary outcome either 0 or 1. Then, using a binary logistic regression model is an ideal choice because it can be more effective in predicting the probability of a passenger being satisfied or not satisfied.

We could use a generalized linear model (GLM) in the form:

$$log\frac{\pi}{1 - \pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots\cdots\cdot + \beta_q x_q = x\beta$$

## 2.2 Variable selection: BIC

Our choice of variables is based on the AIC (Akaike information criterion) and BIC (Bayesian Information Criterion), which base on:

$$AIC = -2 * ln(L) + K * 2.$$

$$BIC = -2 * ln(L) + K * ln(n).$$

$$\text{LASSO} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\left|\beta_j\right|.$$

With large samples, BIC can greatly reduce the risk of overfitting by its strict penalty on the number of parameters in the model, and ultimately drive the model to remain concise so that the variables ultimately chosen are truly relevant to the response. In addition, since BIC tends to penalize model complexity more severely than AIC, BIC selection will be used to picking a simpler more general model. Furthermore, Lasso's approach is also used in our variable selection (Table 7).

## 2.3 Model Diagnostics and validation

Validating binary logistic regression assumptions is crucial for inference reliability. These include:

- Binary Response: The response variable is dichotomous (two possible responses) or the sum of dichotomous responses.
- Independence: The observations must be independent of one another.
- Linearity: The log of the odds ratio, $log\frac{\pi}{1-\pi}$ must be a linear function of x

By looking at the Deviance Residuals Plots, we would like to see a random distribution of residuals, with no obvious large or small residuals, which means that our model captures the changes in the data great. The predictive power of the model can be shown by Model Calibration Plots, if the points on the plot are very close to the y=x line, it means that the predicted probability matches the actual probability, which shown that the prediction of model is accurate. Goodness of fit for binary responses can be evaluated using the ROC curve, and the key metric here is the Area Under the ROC Curve (AUC), where a larger AUC value signifies a better ability of the model to differentiate between the positive and negative classes accurately.

# 3. Result

## 3.1 Description of Data

The Airline Passenger Satisfaction Dataset includes comprehensive survey data collected from passengers of multiple airlines. There are 22 explanatory variables (X) represent different aspects of travel and service experience such as cleanliness, check-in service and type of travel. Dataset has a total of 129,880 observations, and none of them have NAs. The data variable summary is shown in Table 1.

Table 1: Summary of the Airline Passenger Satisfaction Data Set

| Variable | Variable Type | Min | 1st Qu | Median | Mean | 3rd Qu | Max |
|---|---|---|---|---|---|---|---|
| Age | continuous numerical | 7 | 26 | 40 | 39.390 | 51 | 85 |
| Flight Distance | continuous numerical | 67 | 404 | 859 | 1203.000 | 1773 | 4983 |
| Inflight Wifi Service | discrete numerical | 0 | 2 | 3 | 2.724 | 4 | 5 |
| Departure/Arrival Time Convenient | discrete numerical | 0 | 2 | 3 | 3.030 | 4 | 5 |
| Ease of Online Booking | discrete numerical | 0 | 2 | 3 | 2.749 | 4 | 5 |
| Gate Location | discrete numerical | 1 | 2 | 3 | 2.966 | 4 | 5 |
| Food and Drink | discrete numerical | 0 | 2 | 3 | 3.192 | 4 | 5 |
| Seat Comfort | discrete numerical | 1 | 2 | 4 | 3.425 | 5 | 5 |
| Inflight Entertainment | discrete numerical | 1 | 2 | 4 | 3.326 | 4 | 5 |
| Onboard Service | discrete numerical | 1 | 3 | 4 | 3.374 | 4 | 5 |
| Leg Room Service | discrete numerical | 0 | 2 | 4 | 3.314 | 4 | 5 |
| Baggage Handling | discrete numerical | 1 | 3 | 4 | 3.611 | 4 | 5 |
| Checkin Service | discrete numerical | 1 | 3 | 3 | 3.332 | 4 | 5 |
| Inflight Service | discrete numerical | 1 | 3 | 4 | 3.643 | 5 | 5 |
| Cleanliness | discrete numerical | 1 | 2 | 3 | 3.274 | 4 | 5 |
| Departure Delay in Minutes | continuous numerical | 0 | 0 | 0 | 14.390 | 12 | 435 |
| Arrival Delay in Minutes | continuous numerical | 0 | 0 | 0 | 15.070 | 13 | 470 |

| Variable | Categories | Description |
|---|---|---|
| Gender | Male, Female | Gender of the passenger |
| Customer Type | Loyal Customer, Disloyal Customer | Whether the customer is loyal or not |
| Type of Travel | Personal Travel, Business Travel | Purpose of the travel |
| Customer Class | Eco, Eco Plus, Business | Class of the flight ticket |

## 3.2 Process of Obtaining Final Model

The dataset is split into a 75% training set and a 25% test set. The train data is for fitting model and test data will be used for prediction. Then we do some exploratory data analysis on our data (Figure6). During the fitting model process, a full model was first built using all beta coefficients. Following this, variable selection was conducted using AIC, BIC, and LASSO methods (Table7). After employing these methods, we fitted three GLMs (Table 8), incorporating five common predictors and one unique predictor from each selection method. The unique predictor identified by the BIC method had the smallest VIF, indicating the least multicollinearity among the models considered. Therefore, we chose the BIC-based model as our final model. There are no influential point, and  every variable in the model was statistically significant, with specific coefficients and details demonstrated in Table 2.

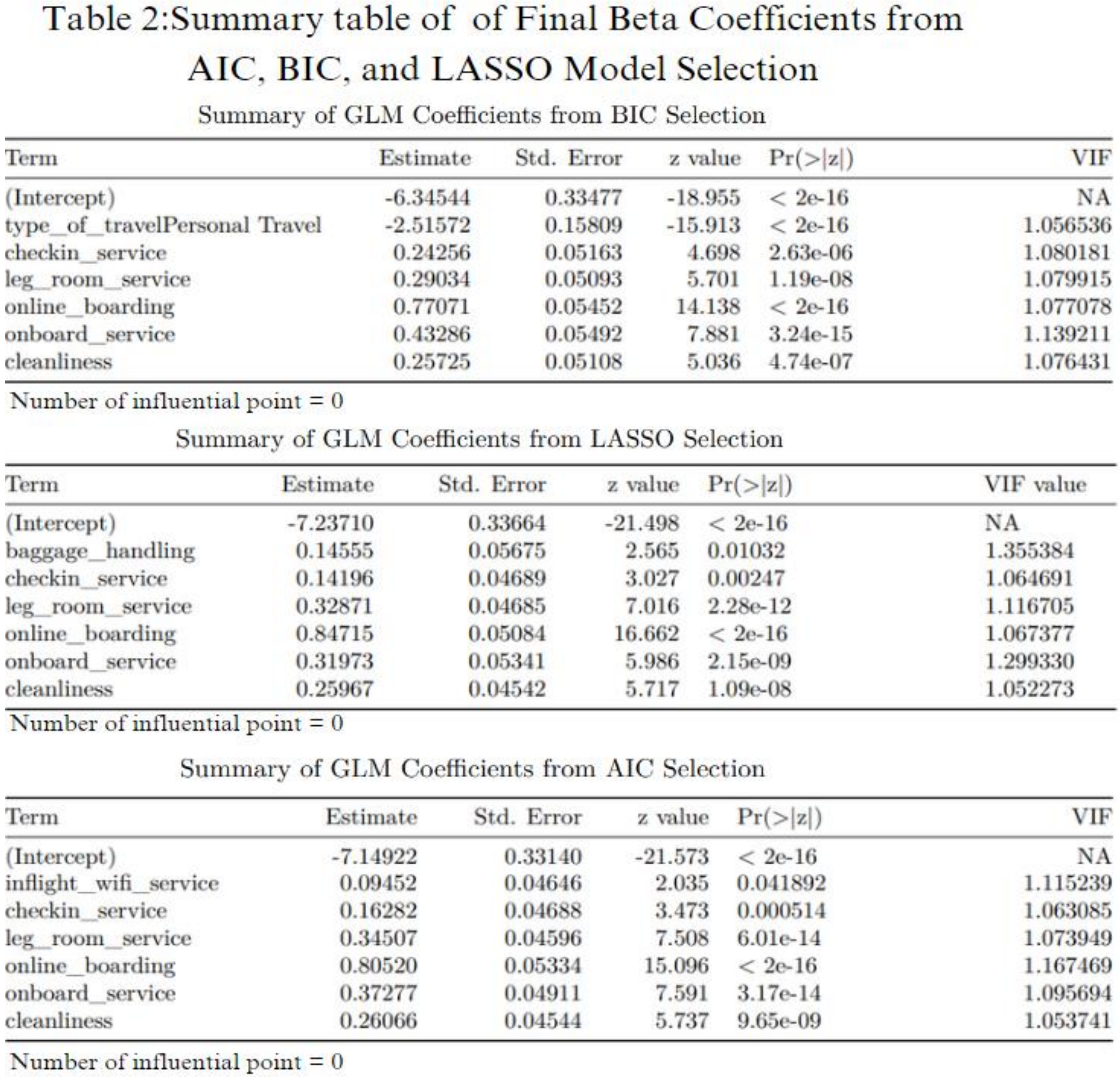


Table 2:Summary table of  of Final Beta Coefficients from AIC, BIC, and LASSO Model Selection

Summary of GLM Coefficients from BIC Selection

| Term | Estimate | Std. Error | z value | Pr(>\|z\|) | VIF |
|---|---|---|---|---|---|
| (Intercept) | -6.34544 | 0.33477 | -18.955 | < 2e-16 | NA |
| type_of_travelPersonal Travel | -2.51572 | 0.15809 | -15.913 | < 2e-16 | 1.056536 |
| checkin_service | 0.24256 | 0.05163 | 4.698 | 2.63e-06 | 1.080181 |
| leg_room_service | 0.29034 | 0.05093 | 5.701 | 1.19e-08 | 1.079915 |
| online_boarding | 0.77071 | 0.05452 | 14.138 | < 2e-16 | 1.077078 |
| onboard_service | 0.43286 | 0.05492 | 7.881 | 3.24e-15 | 1.139211 |
| cleanliness | 0.25725 | 0.05108 | 5.036 | 4.74e-07 | 1.076431 |

Number of influential point = 0

Summary of GLM Coefficients from LASSO Selection

| Term | Estimate | Std. Error | z value | Pr(>\|z\|) | VIF value |
|---|---|---|---|---|---|
| (Intercept) | -7.23710 | 0.33664 | -21.498 | < 2e-16 | NA |
| baggage_handling | 0.14555 | 0.05675 | 2.565 | 0.01032 | 1.355384 |
| checkin_service | 0.14196 | 0.04689 | 3.027 | 0.00247 | 1.064691 |
| leg_room_service | 0.32871 | 0.04685 | 7.016 | 2.28e-12 | 1.116705 |
| online_boarding | 0.84715 | 0.05084 | 16.662 | < 2e-16 | 1.067377 |
| onboard_service | 0.31973 | 0.05341 | 5.986 | 2.15e-09 | 1.299330 |
| cleanliness | 0.25967 | 0.04542 | 5.717 | 1.09e-08 | 1.052273 |

Number of influential point = 0

Summary of GLM Coefficients from AIC Selection

| Term | Estimate | Std. Error | z value | Pr(>\|z\|) | VIF |
|---|---|---|---|---|---|
| (Intercept) | -7.14922 | 0.33140 | -21.573 | < 2e-16 | NA |
| inflight_wifi_service | 0.09452 | 0.04646 | 2.035 | 0.041892 | 1.115239 |
| checkin_service | 0.16282 | 0.04688 | 3.473 | 0.000514 | 1.063085 |
| leg_room_service | 0.34507 | 0.04596 | 7.508 | 6.01e-14 | 1.073949 |
| online_boarding | 0.80520 | 0.05334 | 15.096 | < 2e-16 | 1.167469 |
| onboard_service | 0.37277 | 0.04911 | 7.591 | 3.17e-14 | 1.095694 |
| cleanliness | 0.26066 | 0.04544 | 5.737 | 9.65e-09 | 1.053741 |

Number of influential point = 0

The Final model is:

$$\begin{aligned}\operatorname{logit}(\widehat{\mathrm{P}}) = &-6.34544 - 2.51572 \cdot \text{Personal Travel} + 0.24256 \cdot \text{checkin service} \\ &+ 0.29034 \cdot \text{leg room service} + 0.77071 \cdot \text{online boarding} \\ &+ 0.43286 \cdot \text{onboard service} + 0.25725 \cdot \text{cleanliness}\end{aligned}$$

## 3.3 Goodness of Final Model

The Goodness of Final Model can be evaluated by several diagnostic plots to assess the performance of the models. In Figure 3, GLM has area under the curve in ROC plot is 0.92, which indicates that the model correctly predicts the outcome in 92% of the cases. This indicates that the model has strong discriminatory power. In terms of the calibration plot, there is a clear congruence between the model's predicted probabilities and the actual outcomes, demonstrating the model's accurate and reliable predictive power. For the deviance residual plots in Figure 4 service characteristics such as check-in service and Onboard services are all uniformly distributed, indicating that the model performs consistently with no systematic bias. The dfbeta plot presented in Figure 5 plots a detailed analysis for all predictor variables, with the majority of data points falling within the critical values( $\pm\frac{2}{\sqrt{n}}$).

Specially, for observations of type of travel, the box plots show a compact distribution, suggesting that the model also has stability and accuracy in predicting different groups.

Figure 3: Combined Calibration and ROC Curve Analysis

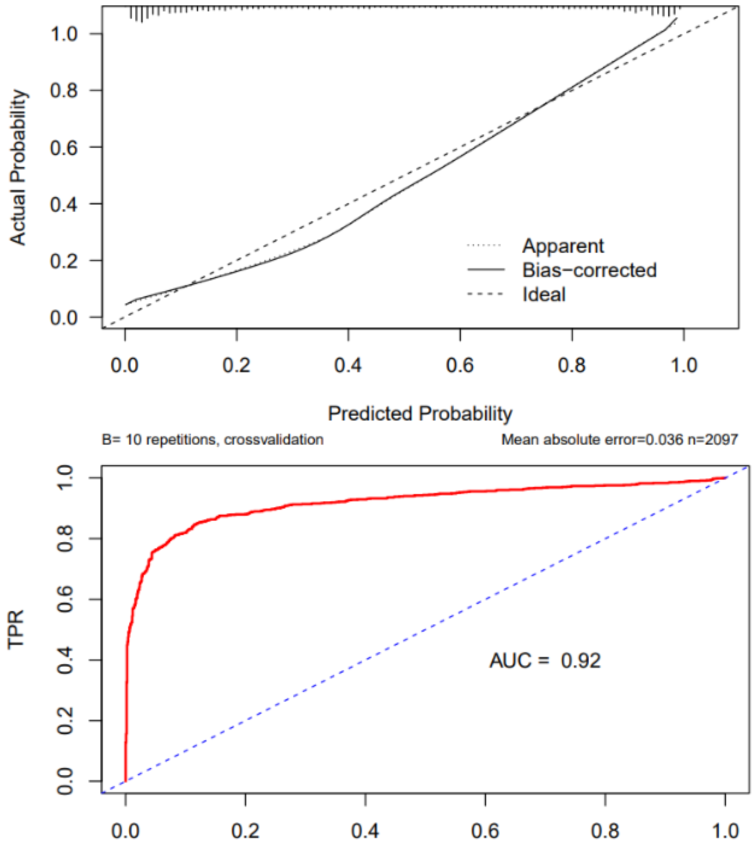


Figure 4: Deviance Residuals by All Predictors

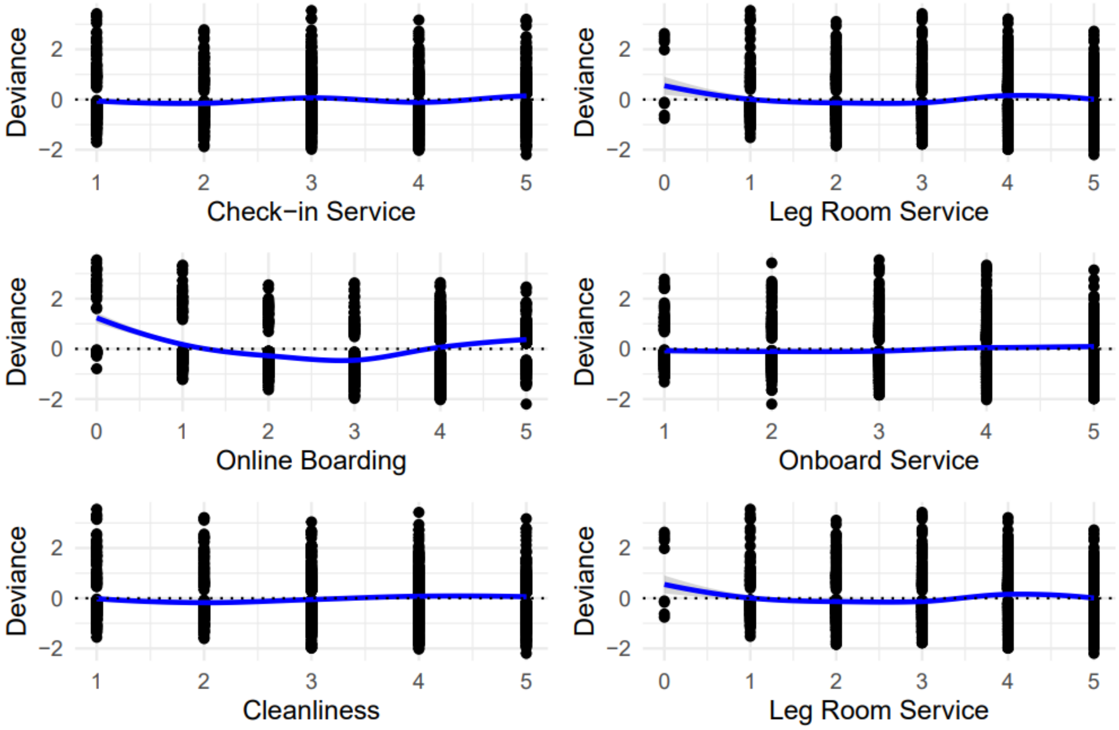
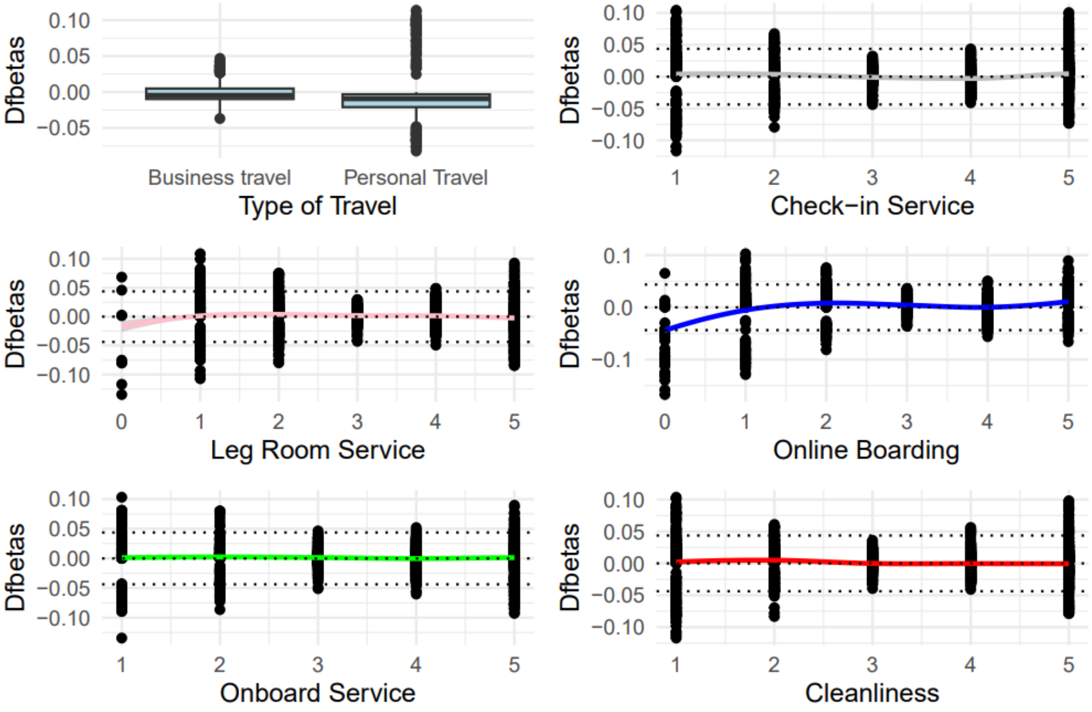


Figure 5: Dfbetas by All predictors



# 4. Discussion

## 4.1 Final Model Interpretation and Importance

The results of fitting the final GLM model are shown in Table 2. Regarding the categorical variables, the log ratio of predicted passenger satisfaction is 2.51572 lower for personal travel compared to business travel, implying that personal travel passengers are relatively less likely to be dissatisfied. For continuous variables such as cleanliness, the log ratio of passenger dissatisfaction increases by 0.25725 for each additional point, suggesting that higher levels of cleanliness will increase passenger satisfaction. Other continuous variables in the model can be interpreted in the same way. Finally, to check prediction ability, the test dataset is used, by comparing the results predicted by the model with the true value, the prediction accuracy is 83.09%, so our model has a strong ability to predict satisfaction level.

## 4.2 Limitation of Analysis

Due to the diversity of global air travel, especially among small or regional airlines, models may not be representative of all travelers. Additionally, GLM modeling may oversimplify passenger satisfaction factors and omit some potential variables due to its assumption of linear relationships and specific distributions. Finally, GLMs are usually not able to effectively deal with dynamic changes in time, so some other models can be used for long-term forecasting.

# Reference

- Archana, R., & Subha, M. V. (2012). A study on service quality and passenger satisfaction on Indian airlines. International Journal of Multidisciplinary Research, 2(2), 50-63.

- Jiang, X., Zhang, Y., Li, Y., & Zhang, B. (2022). Forecast and analysis of aircraft passenger satisfaction based on RF-RFE-LR model. Scientific Reports, 12(1), 11174.

- Park, J. W., Robertson, R., & Wu, C. L. (2004). The effect of airline service quality on passengers' behavioural intentions: a Korean case study. Journal of Air Transport Management, 10(6), 435-439.

# Appendix

## Figure 6: EDA about some variable

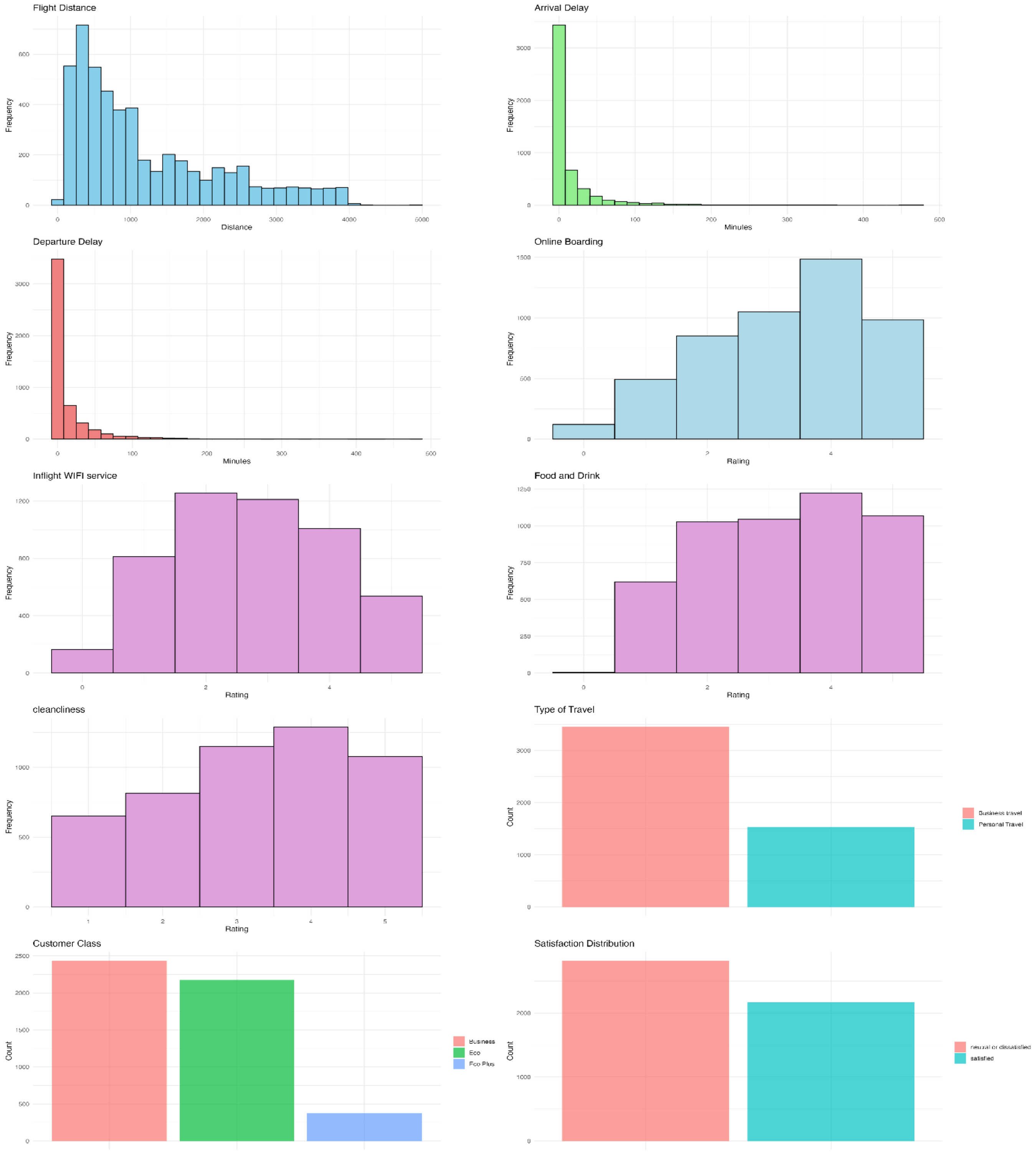
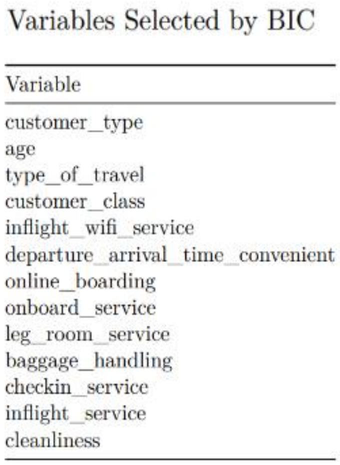
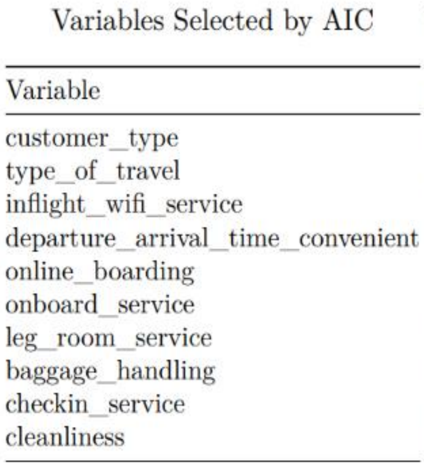


Table 7: Variable selection result from AIC,BIC,and Lasso



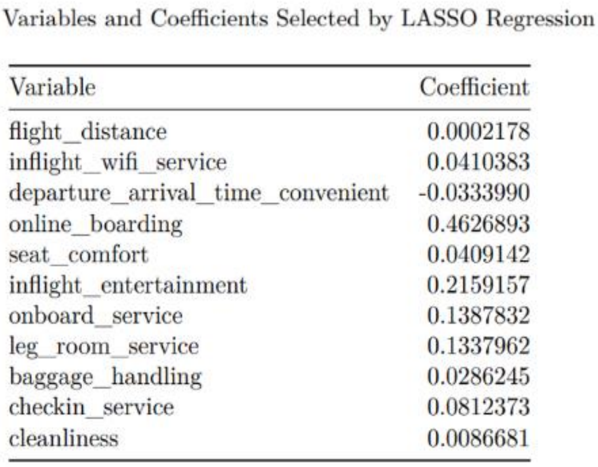
Variables Selected by BIC

| Variable |
|---|
| customer_type |
| age |
| type_of_travel |
| customer_class |
| inflight_wifi_service |
| departure_arrival_time_convenient |
| online_boarding |
| onboard_service |
| leg_room_service |
| baggage_handling |
| checkin_service |
| inflight_service |
| cleanliness |



Variables Selected by AIC

| Variable |
|---|
| customer_type |
| type_of_travel |
| inflight_wifi_service |
| departure_arrival_time_convenient |
| online_boarding |
| onboard_service |
| leg_room_service |
| baggage_handling |
| checkin_service |
| cleanliness |



Variables and Coefficients Selected by LASSO Regression

| Variable | Coefficient |
|---|---|
| flight_distance | 0.0002178 |
| inflight_wifi_service | 0.0410383 |
| departure_arrival_time_convenient | -0.0333990 |
| online_boarding | 0.4626893 |
| seat_comfort | 0.0409142 |
| inflight_entertainment | 0.2159157 |
| onboard_service | 0.1387832 |
| leg_room_service | 0.1337962 |
| baggage_handling | 0.0286245 |
| checkin_service | 0.0812373 |
| cleanliness | 0.0086681 |