AIRLINE PASSENGER SATISFACTION

Group Number 9

Poorva Joshi

Rupam Kalita

Jaya Mundre

College of Professional Studies, Northeastern University

Final Project Report

ALY6015 - Intermediate Analytics

Prof. Zhi He

Feb 14th, 2024

INTRODUCTION

This study utilizes advanced statistical tools and data visualization techniques in R programming to conduct a preliminary analysis of an Airline Passenger Satisfaction (APS) dataset. By exploring various variables encompassing in-flight experiences and demographic data, the aim is to extract actionable insights to enhance customer experiences, improve services, and ultimately boost passenger satisfaction in the competitive airline industry. This report also employs a comprehensive approach, utilizing linear regression models (lm()), generalized linear models (glm()), and analysis of variance (anova()) to delve into airline customer satisfaction and flight distance prediction. By leveraging these statistical techniques, the study aims to uncover the underlying factors influencing both flight distance and customer happiness within the airline industry. Through linear regression, correlations between predictor variables and flight distance are examined, shedding light on the determinants of travel distances. Generalized linear models are utilized to predict binary customer satisfaction outcomes, offering insights into passenger contentment or dissatisfaction. Additionally, analysis of variance plays a crucial role in assessing the significance of model components and interactions, providing a nuanced understanding of the variables impacting flight distance and customer satisfaction levels. Ultimately, the research endeavors to glean valuable insights that can inform strategic decision-making and enhance the overall consumer experience in the aviation sector.

DATASET: https://mavenanalytics.io/data-playground?search=airline

METHODS

1. *ANOVA:* Analysis of Variance, or ANOVA, is a statistical technique used to compare averages among several groups and determine whether or not they show statistically significant differences. ANOVA is a key tool in our analysis to determine the factors that influence customer happiness and discontent. ANOVA is used to analyze categorical variables and determine how they affect the continuous end variable of customer satisfaction. Examples of these variables include class, kind of trip, and in-flight amenities. ANOVA facilitates the discovery of major determinants of customer perceptions by comparing variances between various levels of these categorical components. This information informs strategic decisions aimed at improving overall customer satisfaction levels in the airline sector.

2. *Linear Model "lm()":* Fitting linear regression models is the main application for the "linear model" function, or lm(). By fitting a linear equation to observed data, linear regression seeks to model the relationship between one or more independent variables (predictors) and a continuous dependent variable (response). Within the code, the function lm() is utilized to construct linear regression models that forecast the flight distance by taking into account multiple variables, including customer preferences and satisfaction levels.

3. *Generalized Linear Model "glm()":* The "generalized linear model" (glm()) function is a more

adaptable tool that can manage several response variable kinds, such as binary, count, and categorical variables. Unlike lm(), glm() may simulate binary outcomes and other non-linear interactions by allowing the selection of a particular probability distribution and link function. This code uses glm() to train a logistic regression model based on predictor variables like age, travel distance, and seat comfort to predict binary consumer satisfaction (satisfied or not satisfied). When there are only two possible outcomes for the response variable in a binary classification problem, logistic regression is frequently utilized.

ANALYSIS

1. Read the dataset and conduct perform Exploratory Data Analysis. Several functions were employed to load and examine the dataset in order to gain an understanding of its variables, structure, and summary statistics. For relevant variables, descriptive statistics are computed, such as mean, median, standard deviation, minimum, and maximum.

```
> colnames(APS)
 [1] "gender"                                  "age"
 [3] "customer_type"                           "type_of_travel"
 [5] "class"                                   "flight_distance"
 [7] "departure_delay"                         "arrival_delay"
 [9] "departure_and_arrival_time_convenience" "ease_of_online_booking"
[11] "check_in_service"                        "online_boarding"
[13] "gate_location"                           "on_board_service"
[15] "seat_comfort"                            "leg_room_service"
[17] "cleanliness"                             "food_and_drink"
[19] "in_flight_service"                       "in_flight_wifi_service"
[21] "in_flight_entertainment"                 "baggage_handling"
[23] "satisfaction"                            "age_group"
[25] "flight_distance_km"
```

```
> summary(APS)
   gender           age           customer_type     type_of_travel          class       flight_distance
Female:65899   Min.   : 7.00   First-time: 23780   Business:89693   Business    :62160   Min.   :  31
Male  :63981   1st Qu.:27.00   Returning :106100   Personal:40187   Economy     :58309   1st Qu.: 414
               Median :40.00                                        Economy Plus: 9411   Median : 844
               Mean   :39.43                                                             Mean   :1190
               3rd Qu.:51.00                                                             3rd Qu.:1744
               Max.   :85.00                                                             Max.   :4983
 departure_delay   arrival_delay     departure_and_arrival_time_convenience ease_of_online_booking
 Min.   :   0.00   Min.   :   0.00   Min.   :0.000                          Min.   :0.000
 1st Qu.:   0.00   1st Qu.:   0.00   1st Qu.:2.000                          1st Qu.:2.000
 Median :   0.00   Median :   0.00   Median :3.000                          Median :3.000
 Mean   :  14.71   Mean   :  15.05   Mean   :3.058                          Mean   :2.757
 3rd Qu.:  12.00   3rd Qu.:  13.00   3rd Qu.:4.000                          3rd Qu.:4.000
 Max.   :1592.00   Max.   :1584.00   Max.   :5.000                          Max.   :5.000
```

The below table in the excel sheet describes the Descriptive Statistics of the 24 variables present in the Airline Passenger Satisfaction dataset.

vars	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	se
1	129880	64940.5	37493.27	64940.5	64940.5	48140.02	1	129880	129879	0	-1.20003	104.0357
2	129880	1.492616	0.499947	1	1.49077	0	1	2	1	0.029538	-1.99914	0.001387
3	129880	39.42796	15.11936	40	39.44993	17.7912	7	85	78	-0.00361	-0.71919	0.041953
4	129880	1.816908	0.386743	2	1.896135	0	1	2	1	-1.63884	0.685804	0.001073
5	129880	1.309416	0.462255	1	1.26177	0	1	2	1	0.824576	-1.32008	0.001283
6	129880	1.593864	0.621378	2	1.526755	1.4826	1	3	2	0.54723	-0.61866	0.001724
7	129880	1190.316	997.4525	844	1044.009	767.9868	31	4983	4952	1.108117	0.265396	2.767713
8	129880	14.71371	38.07113	0	5.797554	0	0	1592	1592	6.821823	100.639	0.105639
9	129487	15.09113	38.46565	0	6.095394	0	0	1584	1584	6.66997	95.11188	0.106896
10	129880	3.057599	1.526741	3	3.136299	1.4826	0	5	5	-0.33246	-1.04093	0.004236
11	129880	2.756876	1.40174	3	2.75078	1.4826	0	5	5	-0.01878	-0.91357	0.00389
12	129880	3.306267	1.266185	3	3.382844	1.4826	0	5	5	-0.36656	-0.82995	0.003513
13	129880	3.252633	1.350719	3	3.345434	1.4826	0	5	5	-0.4569	-0.69871	0.003748
14	129880	2.976925	1.27852	3	2.971166	1.4826	0	5	5	-0.05826	-1.03159	0.003548
15	129880	3.383023	1.287099	4	3.478827	1.4826	0	5	5	-0.42131	-0.88889	0.003571
16	129880	3.441361	1.319289	4	3.551711	1.4826	0	5	5	-0.48581	-0.92288	0.003661
17	129880	3.350878	1.316252	4	3.444352	1.4826	0	5	5	-0.34841	-0.98305	0.003652
18	129880	3.286326	1.313682	3	3.358042	1.4826	0	5	5	-0.30092	-1.01484	0.003645
19	129880	3.204774	1.329933	3	3.257237	1.4826	0	5	5	-0.15506	-1.14541	0.00369
20	129880	3.642193	1.176669	4	3.763128	1.4826	0	5	5	-0.69156	-0.35833	0.003265
21	129880	2.728696	1.32934	3	2.698558	1.4826	0	5	5	0.040464	-0.84864	0.003689
22	129880	3.358077	1.334049	4	3.447769	1.4826	0	5	5	-0.36638	-1.06115	0.003702
23	129880	3.632114	1.180025	4	3.752031	1.4826	1	5	4	-0.67738	-0.38385	0.003274
24	129880	1.434463	0.495688	1	1.418078	0	1	2	1	0.264428	-1.93009	0.001375

2. Clean the data by using clean_names() function. And identify how many missing values are present. In the 'arrival_delay' column, values that are missing are substituted with 0.

```
> #Cleaning the data
> p_load(janitor)
> APS <- clean_names(APS)
> #Missing Values
> missing_values <- sum(is.na(APS))
> missing_values
[1] 393
> #We want to replace missing values in a specific column arrival_delay.
> APS$arrival_delay[is.na(APS$arrival_delay)] <- 0
```

3. Factors are suitably created from categorical variables. To offer insights into the distribution of categorical variables, frequency distributions for the following variables were presented: "Class", "Type of Travel", "Customer Type", and "Gender".

Frequency distribution of the above variables have been conducted-

```
[1] "Frequency distribution for Gender:"
> print(gender_summary)
  Gender Count Percentage.Var1 Percentage.Freq
1 Female 65899          Female        50.73837
2   Male 63981            Male        49.26163
```

```
[1] "Frequency distribution for Customer Type:"
> print(customer_type_summary)
  Customer.Type  Count Percentage.Var1 Percentage.Freq
1    First-time  23780      First-time        18.30921
2     Returning 106100       Returning        81.69079
```

```
[1] "Frequency distribution for Type of Travel:"
> print(type_of_travel_summary)
  Type.of.Travel Count Percentage.Var1 Percentage.Freq
1       Business 89693        Business        69.05836
2       Personal 40187        Personal        30.94164
```

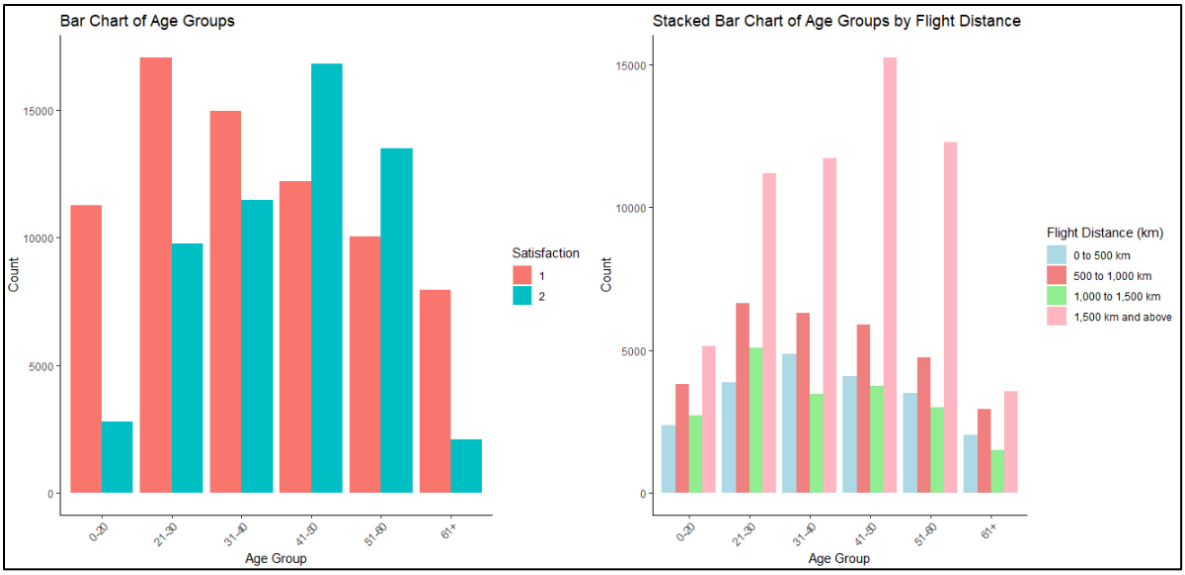
```
[1] "Frequency distribution for Class:"
> print(class_summary)
         Class Count Percentage.Var1 Percentage.Freq
1     Business 62160        Business       47.859563
2      Economy 58309         Economy       44.894518
3 Economy Plus  9411    Economy Plus        7.245919
```

4. Again, the Descriptive Statistics of the above variables (Gender, Customer Type, Travel and Class) have been calculated and stored in an Excel sheet.

	age	flight_distance	departure_delay	arrival_delay
mean	39.43	1190.32	14.71	15.05
median	40	844	0	0
sd	15.12	997.45	38.07	38.42
min	7	31	0	0
max	85	4983	1592	1584

5. Visualizations:-

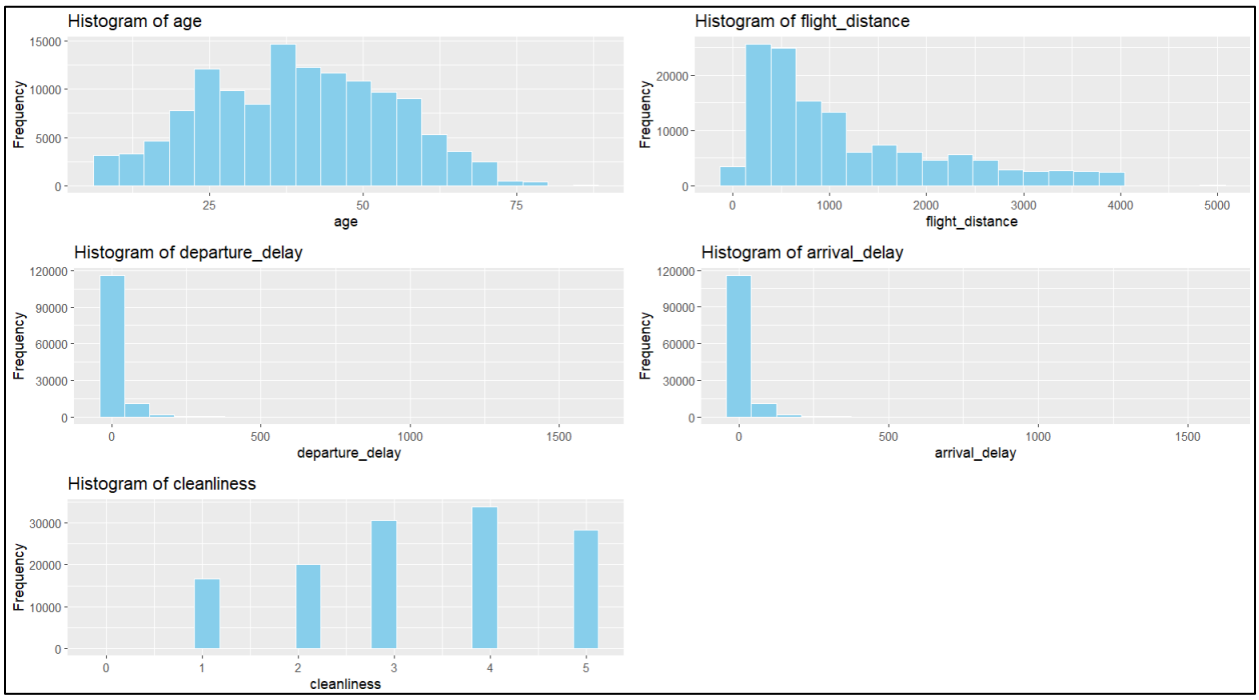
a) The below visualizations show the distribution of travel lengths across different age groups and the association between age groups and passenger satisfaction using exploratory data visualization with the ggplot2 package.



The chart provides a clear overview of how satisfaction varies across different age groups.

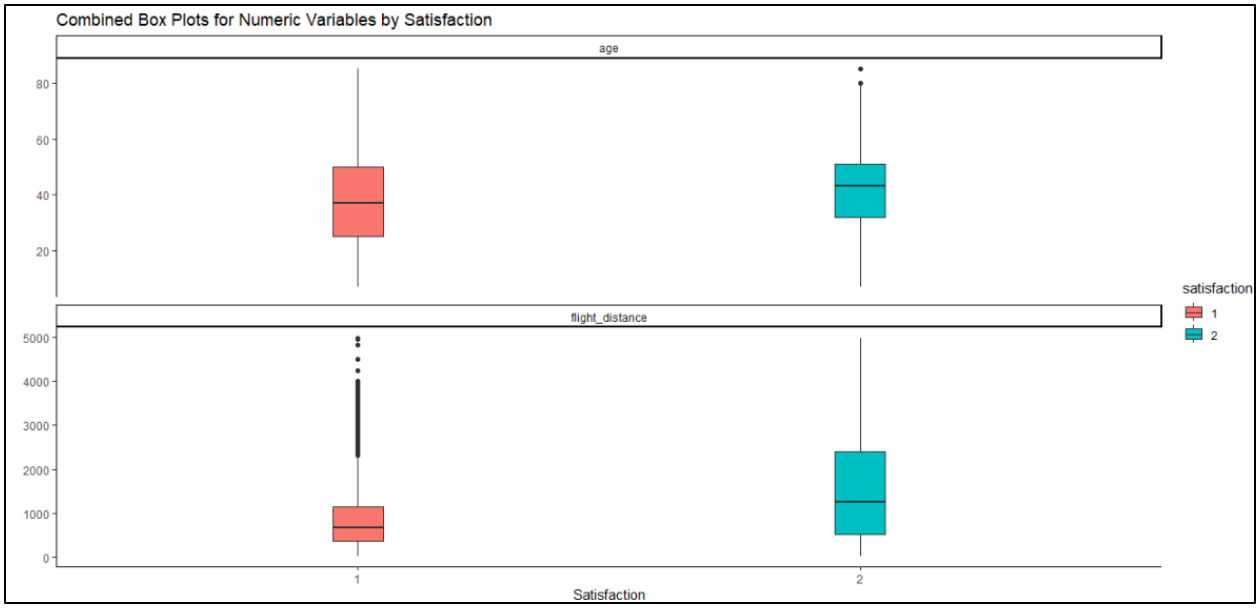
The colors in the chart represent specific distance categories, providing insights into how flight distance varies within each age group.

b) Using the Airline Passenger Satisfaction (APS) dataset, the R code creates histograms for important numerical variables, allowing for a visual examination of their corresponding distributions. The variables that have been chosen are "age," "flight_distance," "arrival_delay," "departure_delay," and "cleanliness." Each histogram offers information about the frequency distribution of the corresponding numerical attribute through the usage of the ggplot2 tool.



The resulting two-column grid layout makes for a more structured presentation and enables a thorough analysis of the distributional properties of these numerical variables in the APS dataset. Understanding the underlying patterns and any outliers in the dataset is crucial, and this visual examination adds important context to the first analysis.

c) The below Box Plot shows the distribution of the variables "age" and "flight_distance" among satisfaction levels by starting the analysis with individual box plots for each. A clear comparison of the primary tendencies and spread is provided by the color-coding of each box plot according to satisfaction. The ability to examine several numerical variables at once with this consolidated plot facilitates the detection of possible trends or variances in passenger satisfaction.



The resulting combined Box Plot provides a thorough overview of the dataset's numerical properties and offers insightful information on the effects of age, flight distance, and departure delay on overall satisfaction.

6. Research Question:- Which percentage of airline passengers are satisfied? Does it vary by customer type? What about type of travel?

We examined potential differences in happiness levels based on client type and kind of travel in this area of the analysis, as well as the overall satisfaction percentage among airline passengers. The'satisfaction' variable was first cleaned by eliminating leading and following whitespaces. A binary representation of satisfaction, with "1" denoting contentment and "2" denoting displeasure, was revealed by the unique values. A calculation of the overall satisfaction rate revealed that roughly 56.55% of passengers were satisfied.

```
> # Calculate the overall satisfaction percentage
> overall_satisfaction_percentage <- APS %>%
+   summarize(percentage_satisfied = mean(satisfaction == "1") * 100)
> overall_satisfaction_percentage
  percentage_satisfied
1             56.55374
```

Next, we segmented the consumer base and calculated satisfaction percentages for each customer type and travel category. First-time passengers exhibited 76.0% satisfaction, whereas repeat customers showed 52.2% contentment. These results demonstrated significant disparities in consumer satisfaction between the two groups. In a similar vein, there was a notable difference in satisfaction depending on the kind of travel, with 89.9% of passengers reporting satisfaction on

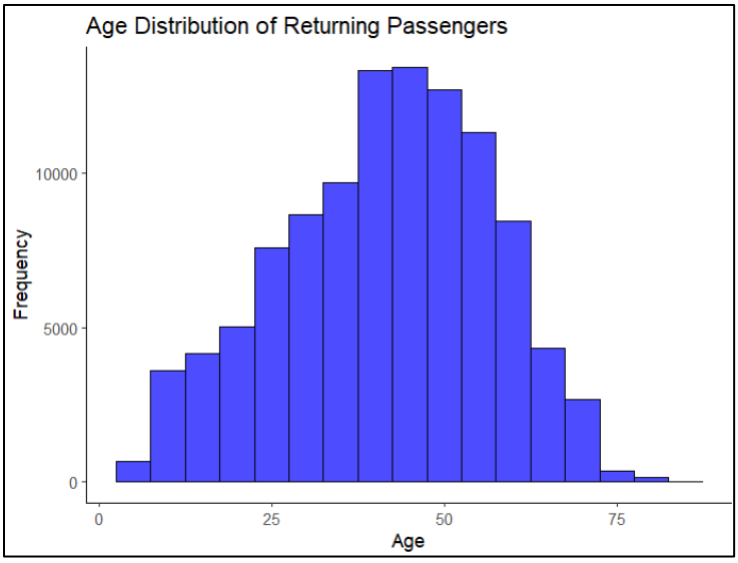
personal trips and 41.6% reporting pleasure on business travels. These findings enable airlines to more effectively customize their services to meet the varying expectations of their passengers by offering insightful information on the subtle elements impacting passenger happiness within different segments.

```
> # Calculate satisfaction percentages by customer type
> satisfaction_by_customer_type <- APS %>%
+   group_by(customer_type) %>%
+   summarize(percentage_satisfied = mean(satisfaction == "1") * 100)
> cat("\nSatisfaction percentages by customer type:\n")

Satisfaction percentages by customer type:
> print(satisfaction_by_customer_type)
# A tibble: 2 × 2
  customer_type percentage_satisfied
  <fct>                        <dbl>
1 First-time                    76.0
2 Returning                     52.2
```

```
> # Calculate satisfaction percentages by type of travel
> satisfaction_by_type_of_travel <- APS %>%
+   group_by(type_of_travel) %>%
+   summarize(percentage_satisfied = mean(satisfaction == "1") * 100)
> cat("\nSatisfaction percentages by type of travel:\n")

Satisfaction percentages by type of travel:
> print(satisfaction_by_type_of_travel)
# A tibble: 2 × 2
  type_of_travel percentage_satisfied
  <fct>                         <dbl>
1 Business                       41.6
2 Personal                       89.9
```

7. Research Question:- What is the customer profile for a repeating airline passenger?

Isolating returning passengers from the APS dataset is the first step in the analysis of this qestion. Three different visuals are then used to explore important components of their consumer profile. The first histogram presents the age distribution of passengers who are returning, offering valuable insights into the age groups that are most prevalent in this particular segment.



It can be observed that passengers between the age group of 40-45 years travel frequently.

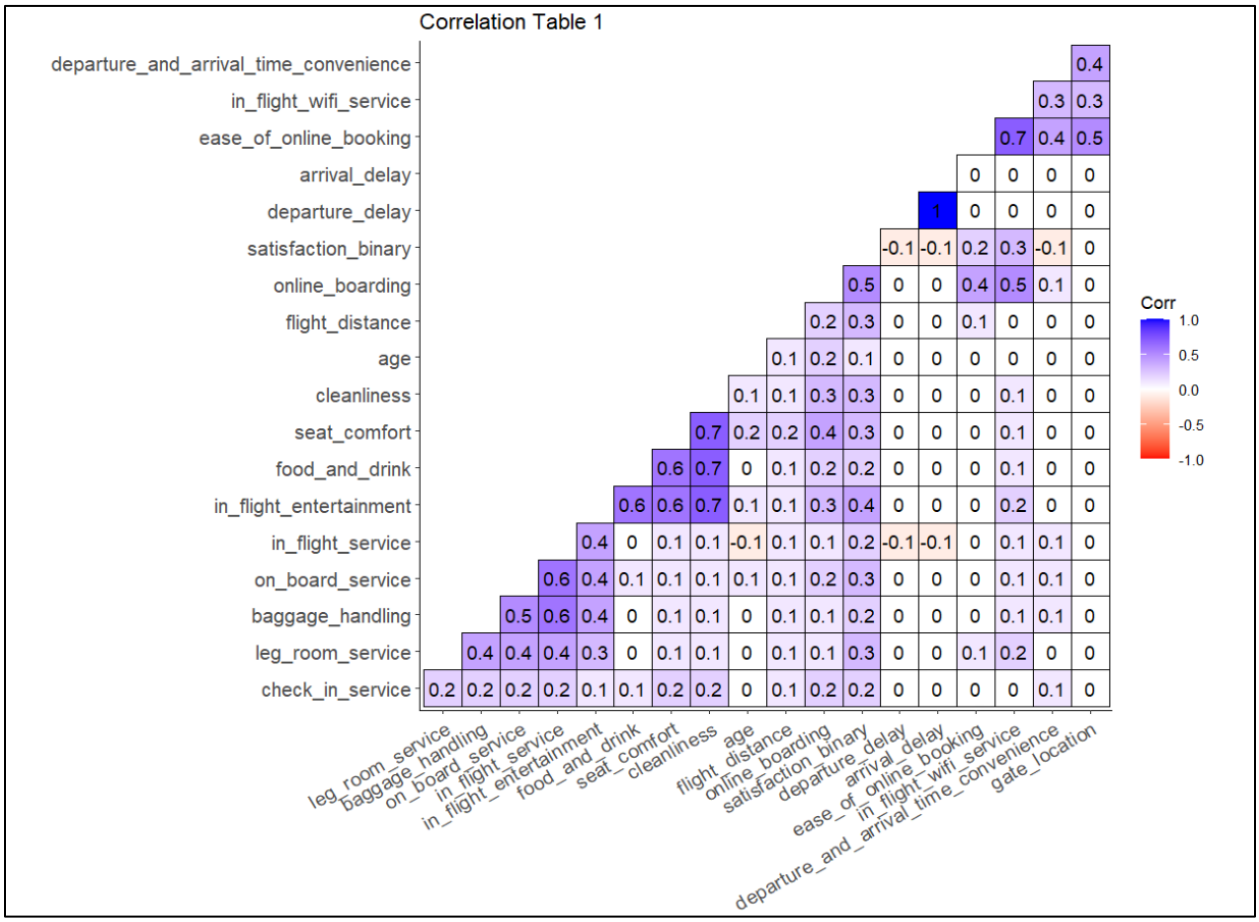
The number of male and female passengers who are returning is shown in the second graph, which explores the gender distribution. The graph is easier to understand when text labels are included. There is not much difference between the number of passengers returning between the two genders.

Finally, the third graph investigates the distribution of satisfaction levels among returning passengers, revealing the frequency of each satisfaction category.

Together, these visualizations help to provide a thorough picture of the satisfaction and demographic-related traits of returning travelers, providing insightful information for tactics aimed at improving both marketing and customer service.

8. Research Question:- Does flight distance affect customer preferences or flight patterns?

The correlation matrix between each of the numerical columns in the APS dataset is computed as shown below. The sapply() and is.numeric() functions are used to extract the numerical columns, while the cor() function is used to build the correlation matrix. The correlation coefficients between two sets of numerical variables are shown in the correlation matrix.



The correlation matrix in the above image shows that there are some strong positive correlations between some of the variables. For example, there is a correlation of 0.7 between food and drink and in-flight entertainment. This means that these two variables are positively correlated, so if one variable increases, the other variable is also likely to increase.

There are also some strong negative correlations between some of the variables. For example, there is a correlation of -1 between seat comfort and food and drink. This means that these two variables are negatively correlated, so if one variable increases, the other variable is likely to decrease.

The code uses two separate functions, lm() and glm(), to investigate the link between predictor variables and responder variable(s). Whereas glm() is used to predict binary outcomes (satisfaction), lm() is used to predict continuous outcomes (flight distance). The algorithm fits these regression models in an effort to get insight into the various aspects influencing flight

distance and customer happiness, which can help the airline sector make decisions.

Using the lm() function, this code snippet uses multiple linear regression to predict satisfaction based on a number of characteristics. The dependent variable in the APS dataset is "satisfaction_binary," which most likely indicates a satisfaction level. The independent variables include things like flight distance, travel type, check-in service, online boarding, seat comfort, and more. The model is trained on this dataset.

```
# Multi-variable Linear Regression to predict satisfaction
lm_model <- lm(satisfaction_binary ~ flight_distance + type_of_travel + check_in_service + online_boarding +
                 on_board_service + seat_comfort + leg_room_service + ease_of_online_booking + food_and_drink +
                 online_boarding + in_flight_wifi_service + class + leg_room_service +
                 seat_comfort + food_and_drink, data = APS)

summary(lm_model)
coef(lm_model)
broom::glance(lm_model)

# Make predictions with the linear regression model
predicted_satisfaction <- predict(lm_model)

# Display the summary and predicted values
print(summary(lm_model))
head(predicted_satisfaction)

plot(lm_model)
```

summary(lm_model) :-

```
Residual standard error: 0.3485 on 129867 degrees of freedom
Multiple R-squared:  0.5057,    Adjusted R-squared:  0.5056
F-statistic: 1.107e+04 on 12 and 129867 DF,  p-value: < 0.00000000000000022
```

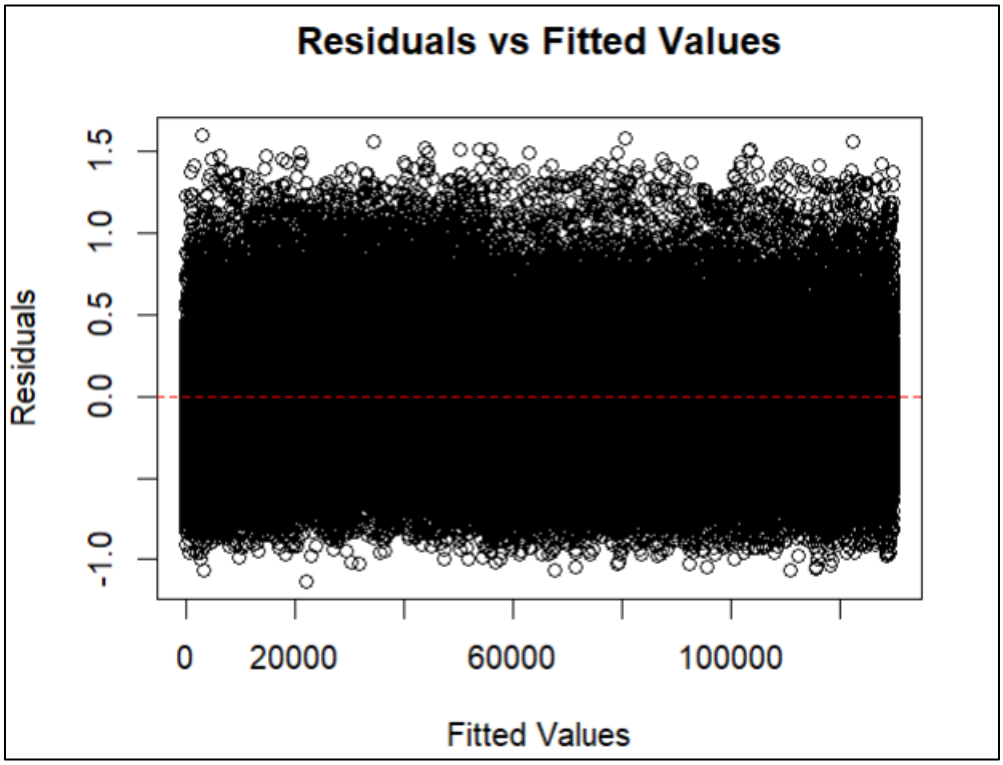
The R-squared value of 0.5057 suggests that the independent variables (flight distance, travel type, check-in service, online boarding, etc.) included in the linear regression model can account for roughly 50.57% of the variability in satisfaction among the data. This indicates that around half of the variation in passenger satisfaction that has been observed may be explained by these factors taken together.

```
> # Make predictions with the linear regression model
> predicted_satisfaction <- predict(lm_model)
> # Display the summary and predicted values
> head(predicted_satisfaction)
        1         2         3         4         5         6
0.5884937 0.9138750 0.8424181 0.8869565 0.8647355 0.9287107
```

The above code snippet applies the linear regression model (lm_model) to predict passenger satisfaction based on various factors such as flight distance, type of travel, check-in service, and others. The output suggests that the model predicts satisfaction levels for the first six observations in the dataset. These predicted values range between 0 and 1, where higher values indicate higher predicted satisfaction.

```r
#Identify the outliers
residuals <- residuals(lm_model)
residuals
plot(residuals, main = "Residuals vs Fitted Values", xlab = "Fitted Values", ylab = "Residuals")
abline(h = 0, col = "red", lty = 2)
```

The above code helps to visually assess how well the linear regression model fits the data and whether the model's assumptions are met.




```r
#Using GLM - predict satisfaction
library(caret)
set.seed(3456)
trainindex <- createDataPartition(APS$satisfaction_binary, times = 1, p = 0.7, list = FALSE)
train <- APS[trainindex,]
test <- APS[-trainindex,]

model <- glm(satisfaction_binary ~ flight_distance + type_of_travel + check_in_service + online_boarding +
               on_board_service + seat_comfort + leg_room_service + ease_of_online_booking + food_and_drink +
               online_boarding + in_flight_wifi_service + class + leg_room_service +
               seat_comfort + food_and_drink, data = train, family = binomial)
summary(model)
```

Based on the above glm(), the output shows that the AIC is equal to 66747. This indicates the trade-off between model complexity (number of parameters) and goodness of fit for the logistic regression model fitted to the data.

```
> # Confusion matrix for train set
> train_prob <- predict(model, newdata = train, type = "response")
> train_predictions <- ifelse(train_prob > 0.5, 1, 0)
> conf_matrix_train <- table(Actual = train$satisfaction_binary, Predicted = train_predictions)
> print(conf_matrix_train)
      Predicted
Actual     0     1
     0 45491  6000
     1  7173 32252
> # Confusion matrix for test set
> predictions <- predict(model, newdata = test, type = "response")
> predicted_classes <- ifelse(predictions > 0.5, 1, 0)
> conf_matrix_test <- table(Actual = test$satisfaction_binary, Predicted = predicted_classes)
> print(conf_matrix_test)
      Predicted
Actual     0     1
     0 19430  2531
     1  3105 13898
```

The training set confusion matrix: Indicates the model's performance on data it was trained on, showing counts of true negatives (45491), false positives (6000), false negatives (7173), and true positives (32252).

Test set confusion matrix: Shows the model's performance on unseen data, with counts of true negatives (19430), false positives (2531), false negatives (3105), and true positives (13898).4

```
> cat("\nAccuracy on the training set:", accuracy_train)

Accuracy on the training set: 0.855108
> cat("\nSensitivity on the training set:", sensitivity_train)

Sensitivity on the training set: 0.8180596
> cat("\nSpecificity on the training set:", specificity_train)

Specificity on the training set: 0.8834748
> cat("\nRecall on the training set:", recall_train)

Recall on the training set: 0.8180596
```

Accuracy: The percentage of accurate predictions made compared to all forecasts made is known as accuracy. This value, which is roughly 85.5%, shows how accurate the model is overall.

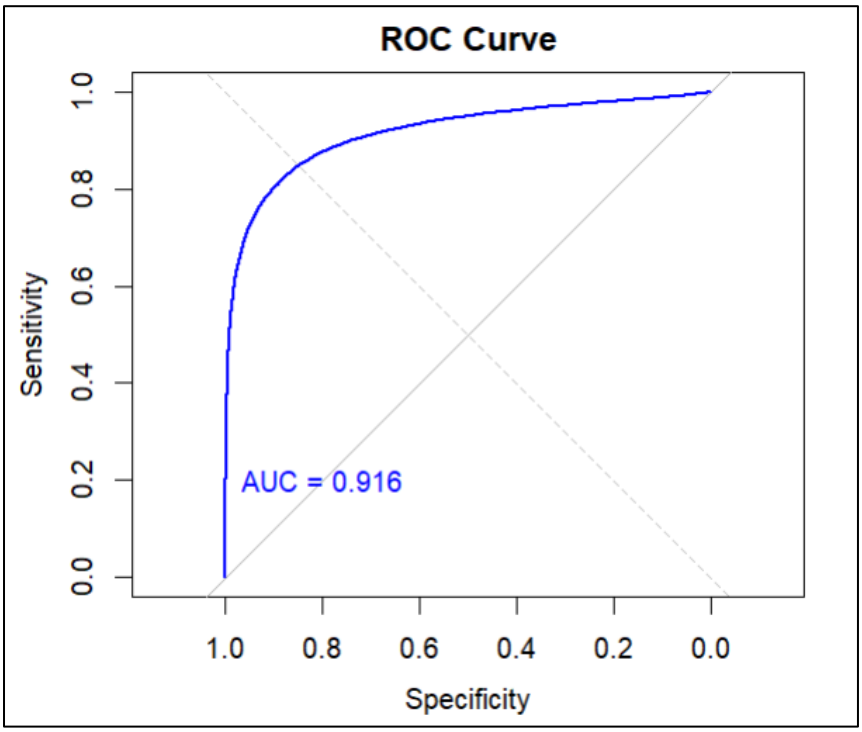
Sensitivity (Recall): The percentage of actual positives out of all true positives that the model properly detected. It's roughly 81.8% in this instance, showing that the model can account for the majority of true positives.

Specificity: The percentage of real negatives out of all genuine negatives that the model properly detected. It is roughly 88.3%, demonstrating the model's capacity to recognize negatives accurately.

Precision: The percentage of actual positives that the model accurately detected from all occurrences that were predicted to be positive. It is roughly 84.3%, reflecting how well the model makes favorable predictions.

```r
# Create ROC curve object
roc_curve <- roc(test$satisfaction_binary, predictions)
plot(roc_curve, main = "ROC Curve", col = "blue", lwd = 2)

#Calculate and interpret the AUC
auc_value <- round(auc(roc_curve), 3)
text(0.8, 0.2, paste("AUC =", auc_value), col = "blue")
abline(a = 0, b = 1, lty = 2, col = "gray")
cat("AUC for LR on the test set:", auc_value)
```

In order to assess a logistic regression model's performance on the test set, this code creates and plots a Receiver Operating Characteristic (ROC) curve. It also computes and displays the Area Under the Curve (AUC) value, which indicates the discriminatory power of the model.



Based on the above ROC curve, the logistic regression model has a strong discriminatory ability to differentiate between the two classes (satisfied and unsatisfied passengers), as indicated by an AUC (Area Under the Curve) value of 0.916. Better performance is indicated by an AUC score that is closer to 1.0. Thus, in this particular case, an AUC of 0.916 indicates that the logistic regression model is performing well.

Result: The logistic regression model (glm()) outperformed the linear regression model (lm()) in this scenario for predicting passenger satisfaction, as evidenced by the significantly higher value (0.916) of AUC, which is a more appropriate metric for classification models like logistic regression. The R-squared value of the linear regression model is only 0.5057.

9. Research Question:- Which factors contribute to customer satisfaction the most? What about dissatisfaction?

We were able to determine the F-values and accompanying p-values by doing ANOVA tests for each of the dataset's variables. The ratio of variance between groups to variance within groups is represented by the F-value. Greater variances between group means are indicated by a higher F-value, which implies a bigger effect of the factor on customer satisfaction. On the other hand, a smaller p-value suggests a higher level of statistical significance and suggests that the observed differences are not likely to be the result of chances.

```
# Perform ANOVA for each factor
factors <- c("type_of_travel", "class", "in_flight_service", "food_and_drink",
             "ease_of_online_booking", "on_board_service", "seat_comfort",
             "leg_room_service", "cleanliness", "departure_and_arrival_time_convenience",
             "check_in_service", "online_boarding", "gate_location",
             "in_flight_wifi_service", "in_flight_entertainment", "baggage_handling")

# Create an empty data frame to store ANOVA results
anova_results <- data.frame(Factor = character(), F_value = numeric(), p_value = numeric(), stringsAsFactors = FALSE)

# Perform ANOVA for each factor and store results
for (factor in factors) {
  anova_result <- aov(satisfaction_binary ~ get(factor), data = APS)
  f_value <- summary(anova_result)[[1]]$`F value`[1]
  p_value <- summary(anova_result)[[1]]$"Pr(>F)"[1]
  anova_results <- rbind(anova_results, data.frame(Factor = factor, F_value = f_value, p_value = p_value))
}

# Sort results by p-value
anova_results <- anova_results %>% arrange(p_value)
```

Output:-

```
> # Factors contributing to customer satisfaction the most
> satisfaction_factors <- anova_results$Factor[1:5]  # Top 5 factors with lowest p-values
> print("Factors contributing to customer satisfaction the most:")
[1] "Factors contributing to customer satisfaction the most:"
> print(satisfaction_factors)
[1] "type_of_travel"        "class"                 "in_flight_service"     "food_and_drink"
[5] "ease_of_online_booking"
> # Factors contributing to customer dissatisfaction the most
> dissatisfaction_factors <- anova_results$Factor[tail(seq_along(anova_results$Factor), 5)]# Bottom 5 factors with highest p-values
> print("Factors contributing to customer dissatisfaction the most:")
[1] "Factors contributing to customer dissatisfaction the most:"
> print(dissatisfaction_factors)
[1] "in_flight_wifi_service"               "in_flight_entertainment"
[3] "baggage_handling"                     "departure_and_arrival_time_convenience"
[5] "gate_location"
```

Factors Affecting Customer Satisfaction: Based on the results of our ANOVA, a number of factors are very important in influencing customer satisfaction. The following are the top five factors with the lowest p-values:

1. Type of Travel
2. Class
3. In-flight Service
4. Food and Drink Quality
5. Ease of Online Booking

These results imply that important factors influencing consumer satisfaction levels include the kind of travel, service class, in-flight service, food and drink selections, and simplicity of online

booking. Airlines should concentrate on enhancing these elements in order to improve consumer loyalty and overall experience.

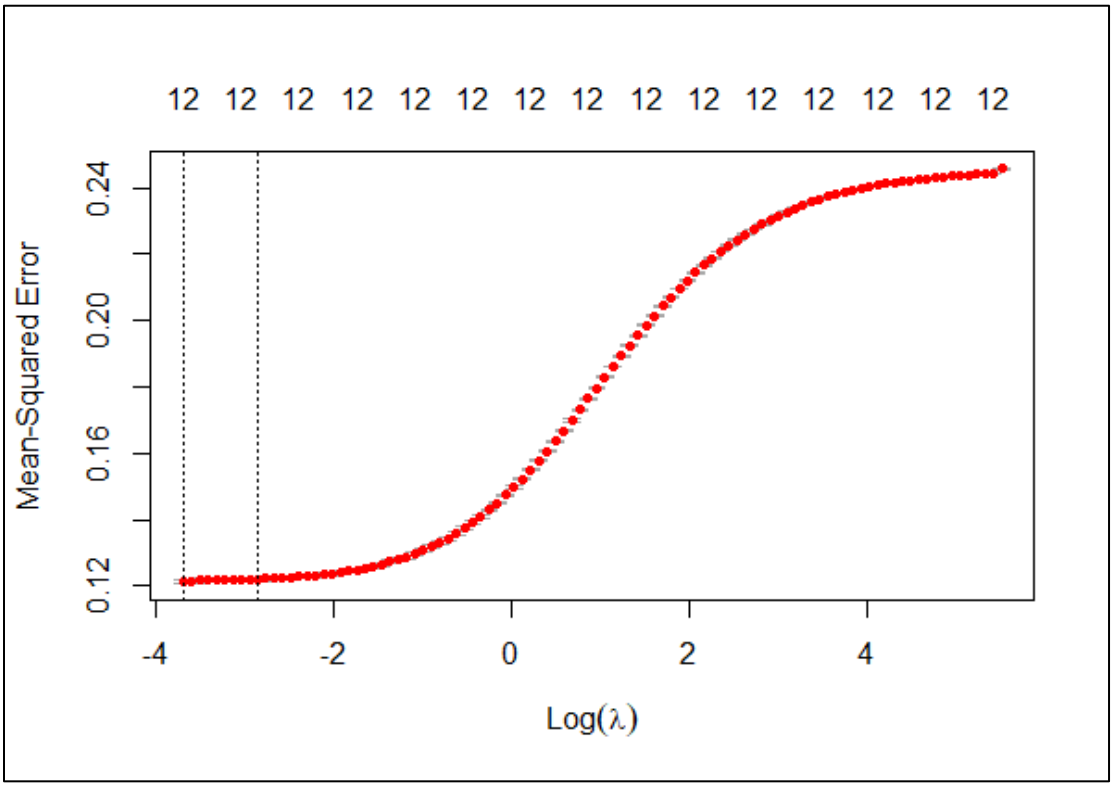
Factors Affecting Client Dissatisfaction: On the other hand, several elements were discovered to have a negligible effect on consumer pleasure, suggesting possible areas of unhappiness. The following are the lowest five factors with the highest p-values:

1. In-flight Wi-Fi Service
2. In-flight Entertainment
3. Baggage Handling
4. Departure and Arrival Time Convenience
5. Gate Location

Although these characteristics might not have as much of an impact on customer satisfaction as other aspects, airlines should nonetheless pay attention to them because they may lead to unfavorable experiences and passenger unhappiness. By addressing these issues, service quality can be increased overall and customer discontent can be reduced.

3. Regularization

Ridge Regression (L2 regularization):



This plot will show the cross-validated mean squared error (MSE) against the log(lambda) values. You'll notice that it shows the values of lambda at the minimum and one standard error away.

```
> # Compare and discuss the values
> lambda_min_r
[1] 0.02490988
> lambda_1se_r
[1] 0.05754506
> log(cv_fit$lambda.min)
[1] -3.692491
> log(cv_fit$lambda.1se)
[1] -2.855187
```

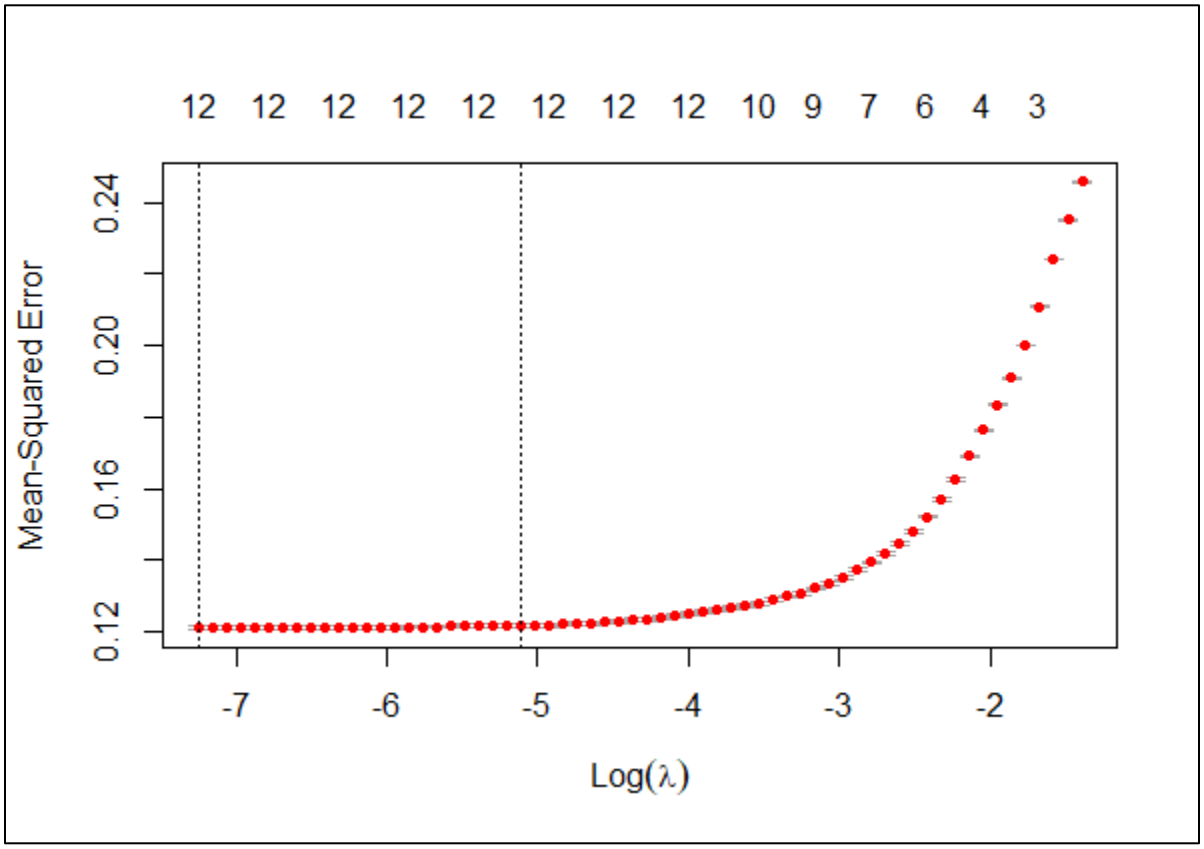
The minimum retains all 12 predictor models and the maximum value is within 1 standard error of the minimum. This model has 12 non-zero coefficients.

Is our model Overfit? (Ridge)

```
> #compare RMSE value
> train_RMSE_r_1se
[1] 0.3493719
> test_RMSE_r_1se
[1] 0.3498079
> train_RMSE_r_min
[1] 0.3486311
> test_RMSE_r_min
[1] 0.3491347
```

The RMSE values on both the training and test sets are close to each other for both lambda.min and lambda.1se. The difference is not substantial, therefore, the Ridge regression model is **not** strongly overfitting.

LASSO Regression (L1 regularization):



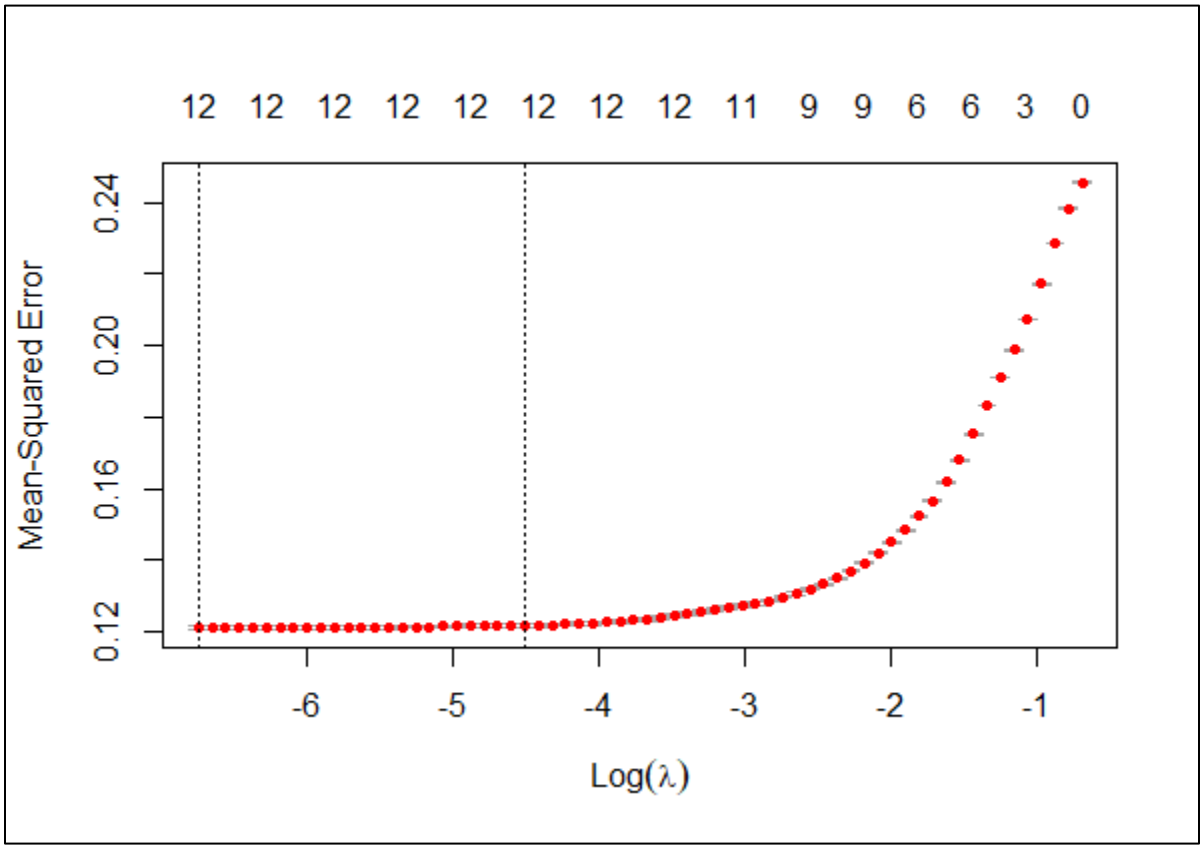
The plot indicates a model with 12 non-zero coefficients out of 15 which stands at -7.251031 lambda.min), and 12 non-zero coefficients out of 15 at -5.111255 (lambda.1se).

Is our model Overfit? (Lasso)

```
> #compare RMSE values
> train_RMSE_l
[1] 0.3490013
> test_RMSE_l
[1] 0.3494512
```

The RMSE values on both the training and test sets are very close to each other, indicating that the LASSO regression model is **not exhibiting** a significant overfitting issue. This alignment between training and test performance suggests that the model is generalizing well to unseen data.

ElasticNet (L1 + L2):



The plot indicates a model with 12 non-zero coefficients out of 15 which stands at -6.743952 lambda.min), and 12 non-zero coefficients out of 15 at -4.511142 (lambda.1se).

Is our model Overfit? (ElasticNet)

```
> train_RMSE_enet
[1] 0.3490248
> test_RMSE_enet
[1] 0.3494712
```

The RMSE values on both the training and test sets for ElasticNet are close to each other. This suggests that the ElasticNet model is also **not** overfitting, as the performance on the test set is comparable to the performance on the training set. The regularization introduced by ElasticNet helps in preventing overfitting and promotes generalization to new, unseen data.

Compare the RMSE Train and Test values of Stepwise, Ridge, Lasso and ElasticNet:

	ColumnName	RMSE_Train	RMSE_Test
1	Stepwise	0.000000000000000005238035	0.000000000000000005198808
2	Ridge	0.3486309486258186152163 3	0.3491345649994742617039 4
3	Lasso	0.3484062595434249676174 0	0.3489656072723466562202 1
4	ElasticNet	0.3484054383517002362324 9	0.3489653140320324586909 4

Which model performs well?

From the provided RMSE values, it seems that all the models have similar performance, and the differences in RMSE are quite small. However, the Stepwise Regression model has an extremely low RMSE of approximately 0.000000000000000005198808 which is essentially zero. This might indicate a perfect fit in the test data, but it raises concerns about overfitting, especially if the data set is not extremely large.

CONCLUSION

In conclusion, the exploratory data analysis (EDA) process has provided us with a comprehensive understanding of the APS dataset. Through data summaries and visualizations, we have uncovered patterns related to passenger satisfaction and demographics. These insights serve as valuable guidance for airline firms to tailor their services according to the specific needs and preferences of their passengers, ultimately aiming to enhance overall customer satisfaction. Additionally, the findings from this analysis can serve as a foundational basis for further exploration and decision-making processes within the industry.

This report presents a thorough examination of airline customer satisfaction and flight distance prediction, employing various analytical techniques such as ANOVA, regularization methods, logistic regression, and linear regression. Specifically, we addressed four key questions pertaining to airline customer satisfaction and flight distance prediction. By utilizing lm() and glm() models, we delved into the influence of flight distance on customer preferences and flight patterns, uncovering essential insights into the factors impacting passenger satisfaction levels. Notably, we found that glm() outperformed lm() in our analysis.

Through ANOVA analysis, we identified significant contributors to customer satisfaction, including factors such as travel type, class, amenities, food quality, and online booking convenience. Conversely, we also examined factors contributing to dissatisfaction, such as in-flight Wi-Fi service, entertainment options, baggage handling, and departure/arrival time convenience. These findings offer actionable insights for airlines to optimize overall customer experience, refine service quality, and make well-informed decisions within the aviation sector.

Furthermore, regularization techniques such as ElasticNet, Ridge, and LASSO were employed to enhance model generalization. Despite similar performance, we observed that the Stepwise Regression model exhibited low RMSE, suggesting a near-perfect fit to the training data. Overall, this comprehensive approach provides valuable guidance for airlines to elevate customer satisfaction levels and facilitate strategic decision-making within the dynamic aviation industry.

INDIVIDUAL CONTRIBUTIONS

Poorva Joshi - Took a leading role in the statistical analysis and interpretation of findings, focusing on EDA, data cleaning, transformation, and visualization using R. Additionally, actively contributed to implementing and evaluating statistical models, particularly ANOVA, to analyze airline customer satisfaction and flight distance prediction. The involvement extended to the implementation and analysis of regularization techniques, such as Ridge and LASSO regression models, aimed at mitigating overfitting and improving model generalization.

Rupam Kalita - Primary contribution to the project involved interpreting descriptive statistics, correlation analysis, and data visualizations to understand variable relationships and identify patterns within the dataset. Also participated in implementing, comparing and evaluating statistical models, such as linear regression and logistic regression. This included fitting the models, assessing their performance, and interpreting the results to gain valuable insights into airline customer satisfaction and flight distance prediction. The engagement also encompassed the implementation and analysis of regularization techniques, notably the ElasticNet regression model.

Jaya Mundre - Led the statistical analysis and interpretation, focusing on exploratory data analysis (EDA), including data cleaning, transformation, and visualization using R. Also contributed to implementing and analyzing stepwise regression models as part of regularization techniques to prevent overfitting and enhance predictive model generalization. The goal was to provide actionable insights for improving customer experience and informing strategic decision-making in the aviation industry through data visualization and statistical analysis.

REFERENCES

1. https://www.statology.org/glm-vs-lm-in-r/
2. https://northeastern.instructure.com/courses/164833/modules
3. Bluman, A. G. (2009). Elementary statistics: A step by step approach (7th ed.).
4. Regularization: https://rpubs.com/mpfoley73/521922
5. An, M., & Noh, Y. (2009). Airline customer satisfaction and loyalty: impact of in-flight service quality. Service Business, 3(3), 293–307.