**Summary, Scope and Findings**

**Project Summary**

The purpose of this project was to conduct analysis on flight data collected through a survey and identify the primary drivers of passenger likelihood to recommend a given airline partner. In short, this analysis attempts to determine how best to improve customer satisfaction. The survey recorded 10282 observations across 32 variables pertaining to the customer demographics, behavior, flight information and most importantly, their likelihood to recommend the airline. We converted the likelihood to recommend values into detractor, passive and promoter categories in order to derive a Net Promoter Score (NPS). A Net Promoter Score is an index ranging from -100 to 100 that measures the willingness of customers to recommend a company’s services to others. Southwest Airline’s current score is a 7, which means there is room for improvement, but the score is on the positive side of the scale. Through descriptive statistics and modelling, we identified key differentiating features between customers, airline partners, and routes and how they relate to NPS. This analysis provides insights that may prove effective in improving customer satisfaction, and in turn, improve collective NPS.

**Business Questions Addressed**

Although many questions were asked throughout this analysis in an attempt to uncover as many insights as possible, the following questions guided the direction and framing of our analysis:

1. What is unique about highly satisfied customers (Age, gender, loyalty status, price sensitivity)? Can we re-create this for more customers?
2. What is unique about our unhappy customers? What is the driving factor of low satisfaction? What’s required to move them to become a promoter?
3. What is the current percentage breakdown of our partners regarding Net Promoter Score? Who is the best, who is the worst (as a percentage of their flights)?
4. Which routes/destination/departure locations are leading to low satisfaction?

**Key Findings**

An extensive exploration of the analysis can be found in later sections of this report, but below are the executive level summary insights we believe are most salient to the analysis.

### **Customers**

* The promoters group follows a relatively normal distribution with a high count of passengers aged from 36 to 42 years old and the number gradually decreases toward both ends.
* The count of passives appears to be randomly scattered, we observed high ends and high counts of middle-aged passengers.
* The detractors are overwhelmed with teenagers and the elderly
* Find out young and elderly passengers’ wants and needs, fulfill them to improve their NPS

### **Partners**

The top performing partners regarding the highest average likelihood to recommend are: West Airways at 1st, Fly Here at 2nd, and Fly to Sun at 3rd.  The worst performers are (from worst to better) are Fly Fast, Going North and Cheapseats. As seen in the boxplot below, many partners have the same median score of 8, but there is some difference in the variance of their rating volatility.

A screenshot of a cell phone

Description automatically generated

## **Recommendations to Improve Satisfaction**

**Customer**

* Continue paying closed attention to middle-aged passengers to maintain and improve their NPS

**Partner**

* Northwest Airways and FlyFast should look into the reasons behind their passengers’ unlikelihood to recommend
* FlyWest should work on improving the needs of its large portion of passive passengers to improve their NPS

**Routes**

* Airline company should pay most attention on the Low Satisfaction Route.
* For Cheapseat Airline Inc, it should try to solve these three routes first: San Jose to Los Angeles, Houston to Chicago, and Orlando to San Juan. Why customer is not satisfied with these airlines.

**Methodology**

The methodology for approaching this problem was to break the task into a series of steps that allowed us to methodically explore the data and identify the appropriate analysis with the time available for the study. Each part of the analysis was assigned a primary analyst and a secondary analyst. This measure was taken to eliminate the chance of error, and to explore the data in multiple ways. Furthermore, in order to focus our analysis on likelihood to recommend, we organized our analysis along three lines of effort that relate directly to the three main players of air travel: the passengers, the airline, and the route. To that end, we used this three-way breakdown as a lens from which to view the dataset and provide context to the analysis. When appropriate, we grouped all observations by their net promoter stance (detractor, passive, promoter) and studied each subset individually to gain a better appreciation for each subset. This allowed us to compare the things that are going well, satisfied customers, satisfied partners, and satisfied routes, with those that were not.

**Data Cleaning**

**Dataset Description**

The data for our data set was collected via survey where 10282 observations across 32 variables pertaining to the customer demographics, behavior, flight information and most importantly, their likelihood to recommend the airline we recorded during a time frame covering roughly a two month period from 01 January 2014 until 09 March 2014.

**Treatment for Missing Data**

The following tables contains the variables in the dataset that had no recorded values. The table also shows what measures we used to replace these missing values. In all cases except free text, the mean was selected to replace the missing values. The median would have been a more conservative estimate of central tendency to use, specifically because the distribution of each of these variables was right skewed, however, we selected the mean because the values were still relatively low, and were under the benchmark we have established as a flight qualifying as delayed, which is 15 minutes or less is not considered late, and a flight being +/- 15 minutes is both not unreasonable, and not uncommon. Furthermore, any risk in using the mean over the median is mitigated by the fact that this replacement method impacted so little data (1.9% to 2.2%). No approximation measure was taken for the free text data, the black entries were simply ignored. This approach is reasonable as this is a character field containing stings of text from customer reviews, and no arithmetic was performed on these values.

|  |  |  |  |
| --- | --- | --- | --- |
| **Variable** | **Number of Occurrences / % of Observations** | **Replaced with what measure** | **Value of Replacement** |
| Departure.Delay.in.Minutes | 200 / 1.9% | Mean | 15.64233 |
| Arrival.Delay.in.Minutes | 226 / 2.2% | Mean | 15.29881 |
| Flight.time.in.minutes | 226 / 2.2% | Mean | 15.8218 |
| FreeText | 10000 / 97.3% | None | None |

**Data Transformation**

During our analysis we constructed several subsets of data aligned with specific analysis tasks. For example, we made subsets to organize data by numeric types of values, category (factor) type values, and a free text dataset to capture the free text reviews (only 2.75% of the data had these reviews). Additionally, we made subsets of the data that included all the variables—both original and engineered—but partitioned the data for our three main groups of analysis: Detractors, Passives, and Promoters. We did this so we could better examine the uniqueness of each of these groups to better determine what makes them different. When appropriate (primarily during modeling) we used the complete set of data in order to provide a more robust models that are more prepared to make predictions.

**Feature Engineering**

The original dataset had 32 feature variables but to improve our analysis and make it easier to understand, we built an additional 4 variables. The most important transformation was transforming each customer’s likelihood to recommend into a Promotion Stance variable by assigning all likelihood to recommend values of 6 and below as “Detractors”, 7 and 8 as “Passives”, and 9 and 10 as “Promoters.” The next two variables, Departure Delay Severity and Arrival Delay Severity, represent the transformation of time values slit into delay time increments: “0 to 15 minutes”, “16 to 30 minutes”, “31 to 45 minutes”, “46 to 60 minutes”, “Greater than one hour”. Although this comes at a cost to precisely measuring time delays, it groups the occurrences into logical categories and makes it more usable for caparisons. We also made a variable titled “Routes” to aggregated origin locations with destination locations. We built this feature so we could assess if particular routes are more likely to yield low or high satisfaction. Finally, we created a feature titled Year of Users that captured the time between the survey (2014) and the year of the customer’s first flight. We did this to help determine how long each customer has been a customer.