SAN FRANCISCO AIRPORT PASSESNGER SATISFACTION: DATA INTERPRETATION AND MODELLING

Kulanika Gnanaratna

1. Introduction

This report presents an analysis of passenger satisfaction at San Francisco Airport (SFO) based on survey data. The objective is to understand factors influencing whether passengers rate the airport as "**good**" and to provide a predictive model for future use.

The factors analysed include,

- reported dirty locations (**dirty**) - a count of the number of locations which the passenger felt were dirty.
- wait times (**wait**) - the number of hours the passenger spent at the airport between arrival and flying.
- travel frequency (**lastyear**) - a count of the number of times the passenger flew out of SFO in the previous 12 months.
- destination type (**usa**) - whether the passenger was flying to a destination in the USA or another country.

Findings and a predictive logistic regression model are discussed below.

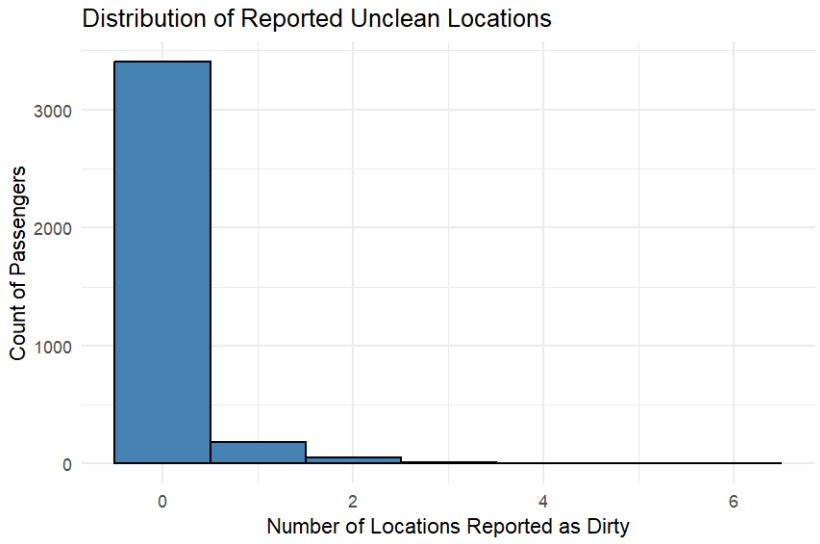
2. Data Exploration and Key Insights (Descriptive Statistics)

Passenger Satisfaction Levels (good):

- **57.71%** of passengers rated their experience as "Good" (1).
- **42.29%** rated their experience as "Not Good" (0).

Over half of the passengers found their experience satisfactory, which is a positive outcome. However, still a significant portion (42.29%) was dissatisfied, signaling areas that need attention.

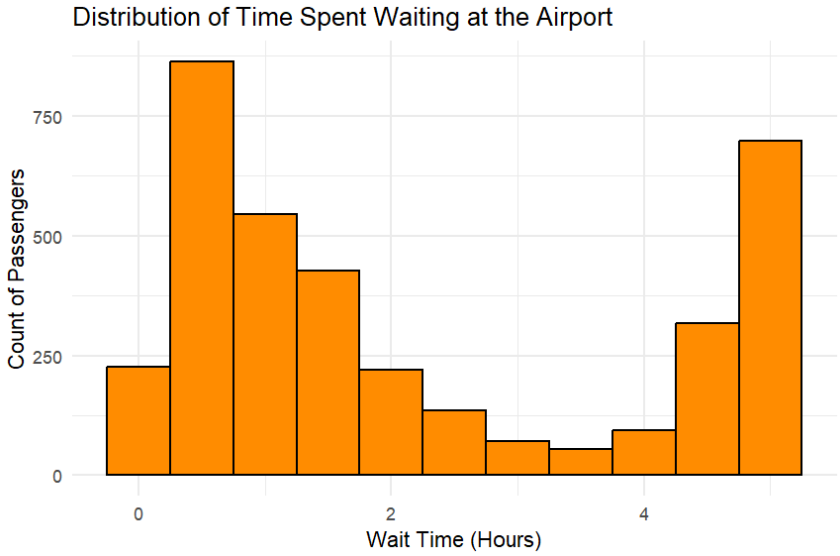
Reported Unclean Locations (dirty):



- The vast majority of passengers did not report any concerns about dirtiness in the airport (25% and 75% of passengers reported **0 dirty locations**).
- Some passengers reported up to **6 dirty locations**, indicating isolated but severe cases.

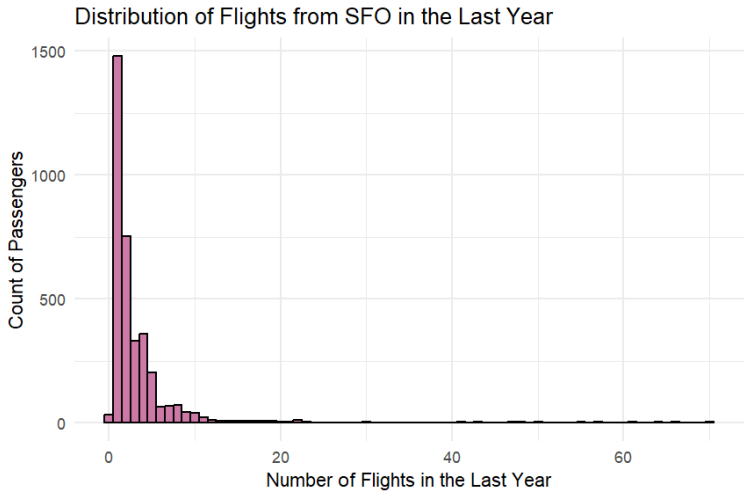
Most passengers reported no dirty locations, but a small group reported several, showing that cleanliness complaints are isolated but significant.

Wait Time (wait):



- Typically, most passengers spent between **0.65 and 4.45 hours** at the airport, with half of the passengers spending less than **1.41 hours**.

Travel Frequency (lastyear):



- Half of the passengers flew out of SFO fewer than **2 times** in the past year, while the most frequent passengers (75% of passengers) flew out less than **4 times**.
- Most passengers are occasional flyers, though some flew out of SFO as many as **70 times**.

The data is highly variable, with some passengers flying frequently and others rarely.

Destination Type (usa):

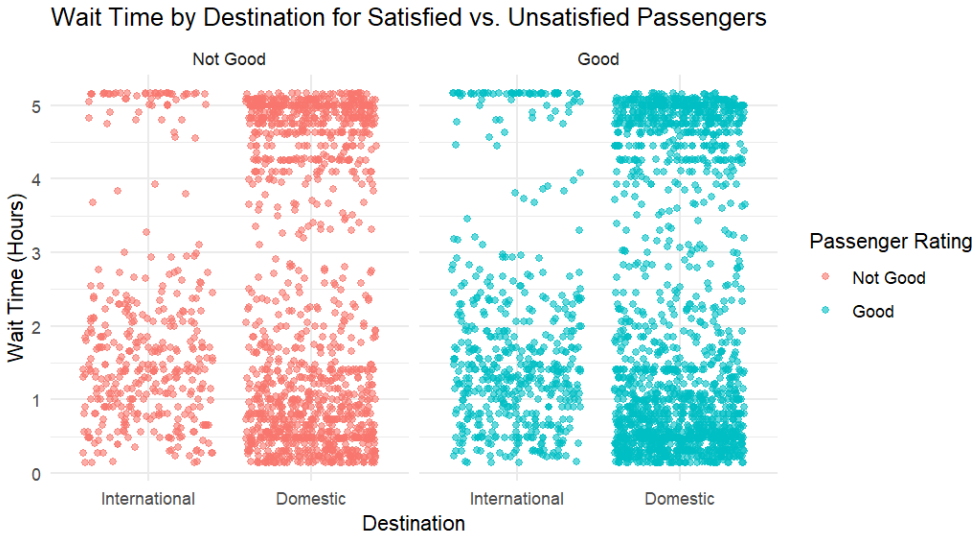
- **75.57%** of passengers were traveling within the USA.
- **24.43%** were traveling to an international destination.

Percentages clearly show the majority of passengers were traveling within the USA.

3. Visualization of wait and usa

A scatter plot was created to examine the relationship between wait time, flight destination, and passenger rating/satisfaction level. The points are slightly spread out in the scatterplot to avoid them overlapping which makes it easier to see each individual point and understand the patterns in the data.

Another plot was visualized for clear depiction.



Key observations:

- Passengers with wait times under **2 hours** showed **higher approval rates**.
- Satisfaction decreased as wait times increased, particularly for wait times exceeding **4 hours**, where a higher proportion of disapprovals was evident.

- **Domestic flights** accounted for **approximately 75.6%** of passengers and exhibited slightly **higher approval rates** compared to international flights.
- Despite this, both groups experienced lower satisfaction at longer wait times (≥ 4 hours).

Interpretations:

- Passengers who spent less time waiting (**closer to 1 hour**) were more likely to rate SFO as "good," particularly for **domestic flights**.
- A **sharp decline** in satisfaction was observed for **domestic travelers** as wait times exceeded 4 hours.
- **International travelers** showed **greater tolerance** for longer wait times but still exhibited dissatisfaction as wait times increased beyond this threshold.

4. Model Exploration - Logistic Regression Analysis

To predict passenger satisfaction, a series of models were built to find the best combination of factors. Logistic regression is a type of model that helps estimate the probability of an outcome (here, whether passengers approved of the airport). Different combinations of variables were tested to find the best fit for the data. The model's effectiveness was measured using a statistic called AIC (Akaike Information Criterion). A lower AIC value indicates a better model.

Initial Model (All Predictors / Factors):

- An initial predictive model was built using all four variables, reported unclean locations (**dirty**), wait time (**wait**), flight frequency (**lastyear**), and flight destination (**usa**). This model examined how each variable contributed to predicting passenger approval of the airport.

- **Model Statistics: -**

Variable	Coefficient	Std. Error	z-Value	p-Value	Significance
Intercept	0.5169	0.078	6.610	3.85e-11 (<<0.001)	***
dirty	-0.8262	0.108	-7.635	2.27e-14 (<<0.001)	***
wait	-0.0746	0.019	-4.006	6.19e-05 (<0.001)	***
lastyear	-0.0030	0.004	-0.686	0.492	Not significant
usa	0.0607	0.079	0.764	0.445	Not significant

- **<u>Interpretation</u>: -**

Variable	Effect on Satisfaction	Explanation
dirty	Strong negative impact	More dirty locations significantly reduced satisfaction.
wait	Negative impact	Longer wait times lowered satisfaction.
lastyear	No significant impact	Flying more often didn't change satisfaction levels.
usa	No significant impact	Domestic vs. international flights didn't matter much.

Procedures and Model Refinement

1. The Initial Model with all the 4 predictors, was taken into consideration.
    a. **Model 1**: All Predictors

        Predictors: dirty, wait, lastyear, usa

2. Variables were assessed for statistical significance, and insignificant variables (**lastyear** and **usa**) were removed step-by-step to simplify the model. AIC values were compared at each step to evaluate model performance.
    a. **Model 2**: Excluding lastyear

        Predictors: dirty, wait, usa

    b. **Model 3**: Excluding lastyear and usa

        Predictors: dirty, wait

        Since **lastyear** and **usa** were not significant predictors, we removed them and built a simplified model using only **dirty** and **wait**. This reduced model had a lower AIC, meaning it explained the data more efficiently and made it the best choice at this point.

3. Further exploration was done into whether the effect of **dirty** and **wait** was interrelated (e.g., whether passengers' satisfaction dropped even more when they experienced both long wait times and unclean areas.)
    a. **Model 4**: Interaction Between dirty and wait

        Predictors: dirty, wait, dirty:wait

However, adding this interaction term (**dirty** X **wait**) did not improve the model, as its AIC was slightly higher than the reduced model (**Model 3**).

Model Comparison

Models were evaluated using AIC (a measure of model quality; lower is better).

Model	AIC
Model 1	4893.52
Model 2	4891.99
Model 3	4890.52
Model 4	4892.41

Chosen Model

The chosen "best" model is the **Reduced Model** (**Model 3**), which includes only the variables **Dirty** and **Wait**. This model was selected because:

- It had the lowest AIC value compared to other models, indicating better performance while maintaining simplicity.
- Removing the insignificant factors (**lastyear** and **usa**) improved the model's efficiency without losing predictive accuracy.
- Adding an interaction term (**dirty × wait**) did not lower the AIC, meaning the interaction between these variables does not provide additional explanatory power.

5. Conclusion

Passenger satisfaction is primarily influenced by cleanliness and wait times. Enhancing cleaning procedures and minimizing wait times, especially for domestic passengers, can significantly improve approval ratings. The predictive model helps identify areas for improvement and measure the impact of changes. By addressing these factors, SFO can enhance passenger experiences and overall satisfaction.