Airline Satisfaction Analysis

By: Daniel Gallo

Agenda

- Problem Statement
- Data Processing
- Exploratory Data Analysis
- Modeling & Evaluation
- Key Findings

Problem Statement

The goal of this project is to determine which factors most strongly influence airline passenger satisfaction and to predict whether a passenger will be satisfied based on their flight experience.

Data Processing

Data Overview

This dataset provides both subjective experience ratings and objective delay information, making it well-suited for predictive modeling.

Rows	103,904
Columns	25
Target	Satisfaction
Features	Demographics, Service Ratings, Delays

Missing Values

This dataset was clean for the most part, but there were 310 values missing in the arrival delay (minutes) column. This column was very right skewed, so I utilized the median to fill in the missing values to avoid inflatings typical delay values.

Prior to imputing the column, I performed the technique “Missingness indicator feature”, as this will help the machine learning models to know that the value was originally missing and can help alleviate bias caused by the imputing process.

Outliers

Although several columns contained outliers, these values appeared to represent valid observations. As a result, they were retained in the dataset, since they may be correlated with passenger satisfaction or dissatisfaction.

Because extreme delays and ratings reflect real passenger experiences, removing them could bias satisfaction predictions.

Feature Engineering

Age Groups

I created an age group column, that grouped flyers into categories such as children, teenagers, young adults, adults, and seniors.

Children	Age 12 and below
Teenager	Age between 13 and 19
Young adult	Age between 20 and 25
Adult	Age between 26 and 65
Senior	Age 65 and above

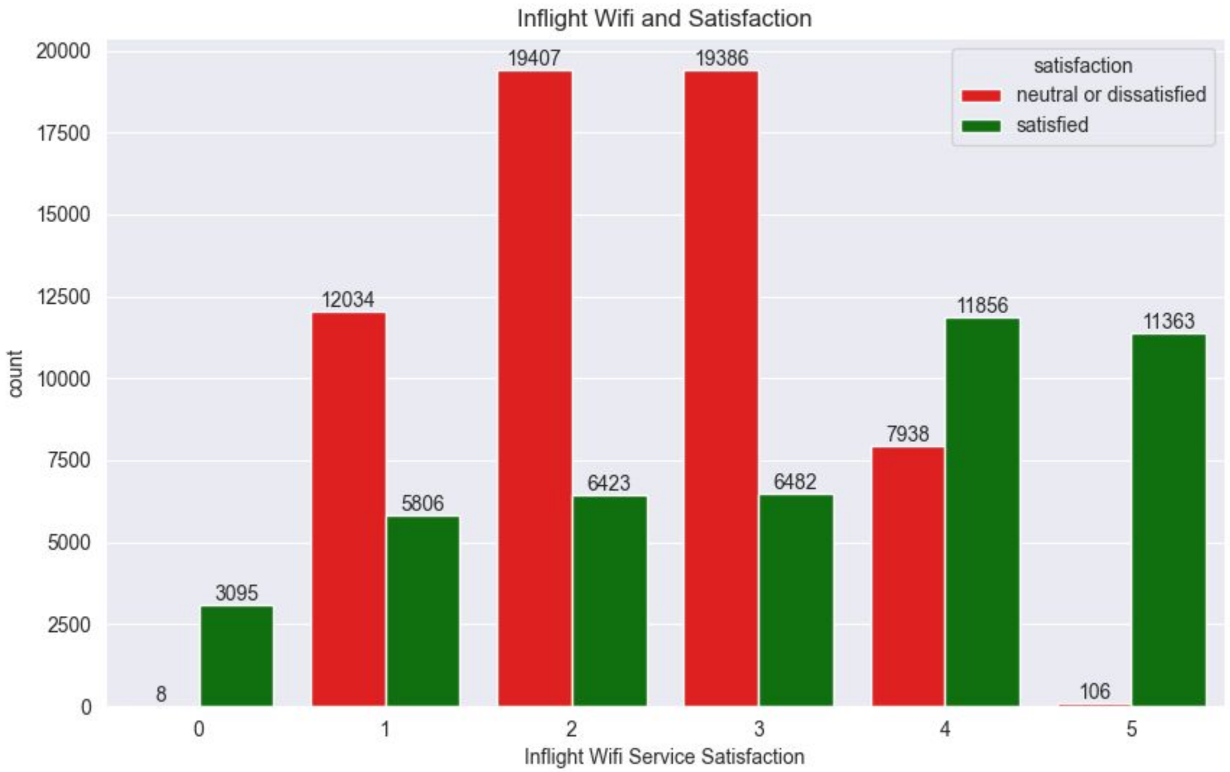
Delay Groups

I also created columns for delay groups, the departure and arrival. These contain the following categories:

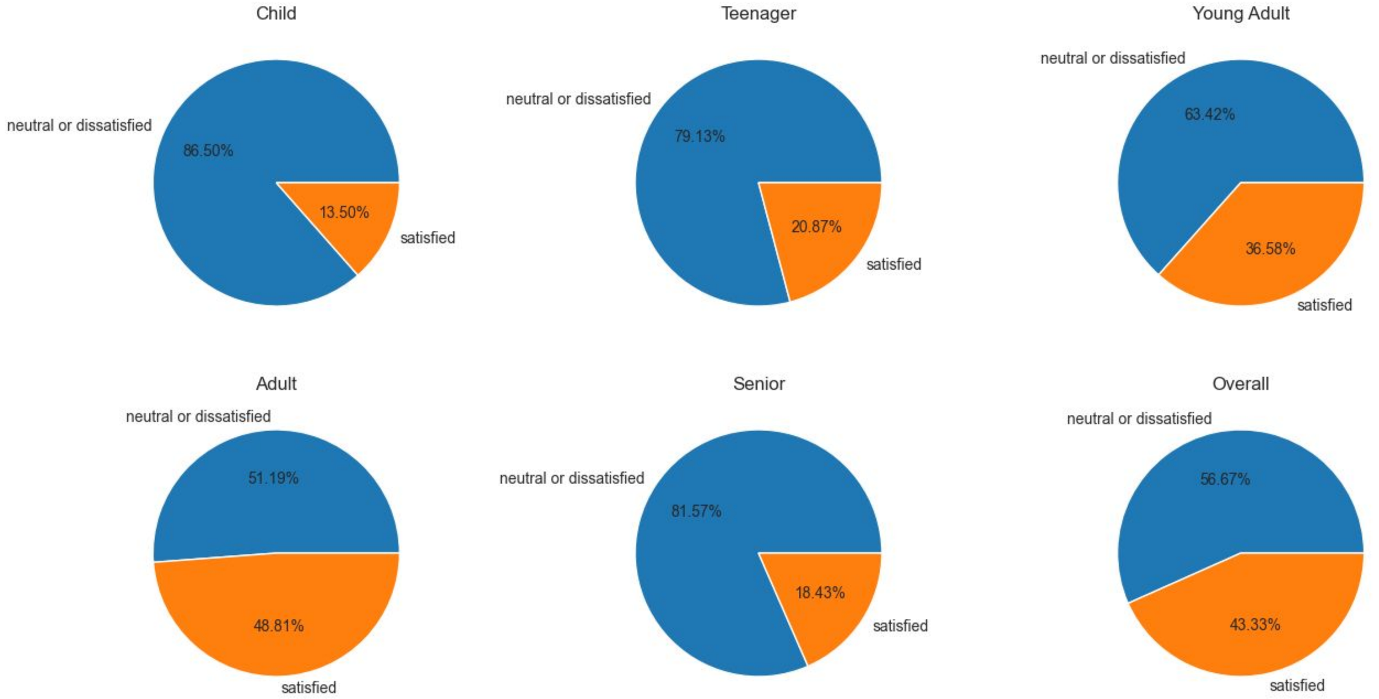
On Time	No Delay
Small Delay	15 minutes or less
Medium Delay	16 - 45 minutes
Large Delay	46 minutes to 2 hours
Very Large Delay	More than 2 hours

Exploratory Data Analysis

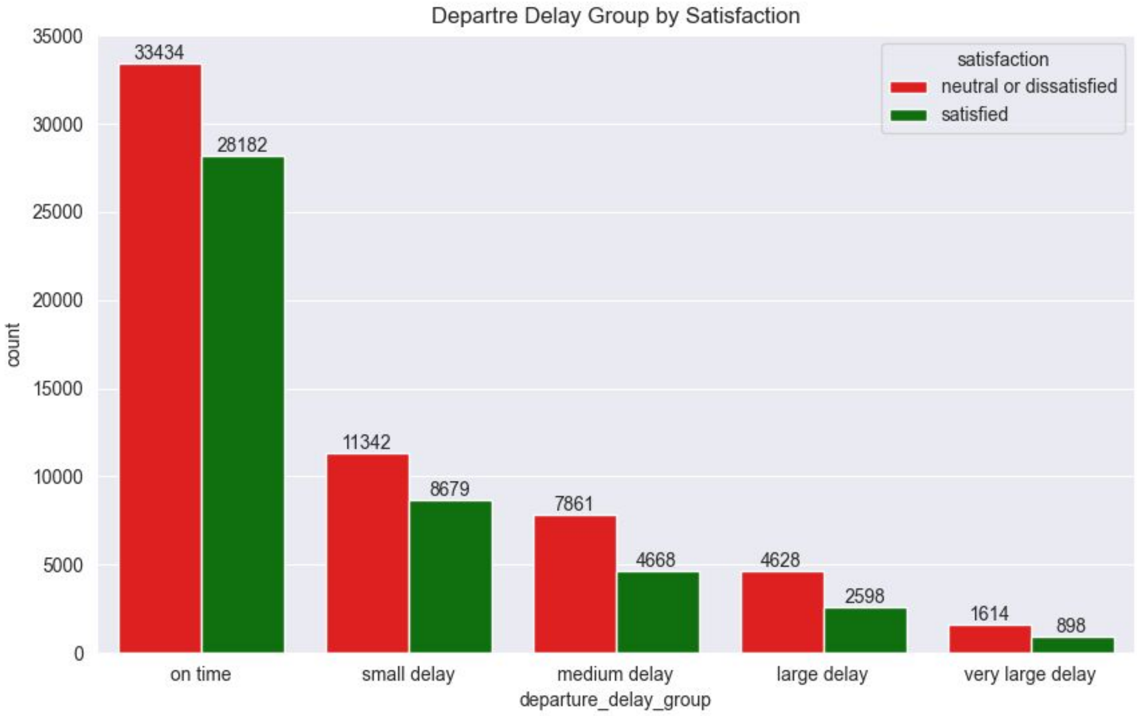
How does inflight wifi relate to the satisfaction?



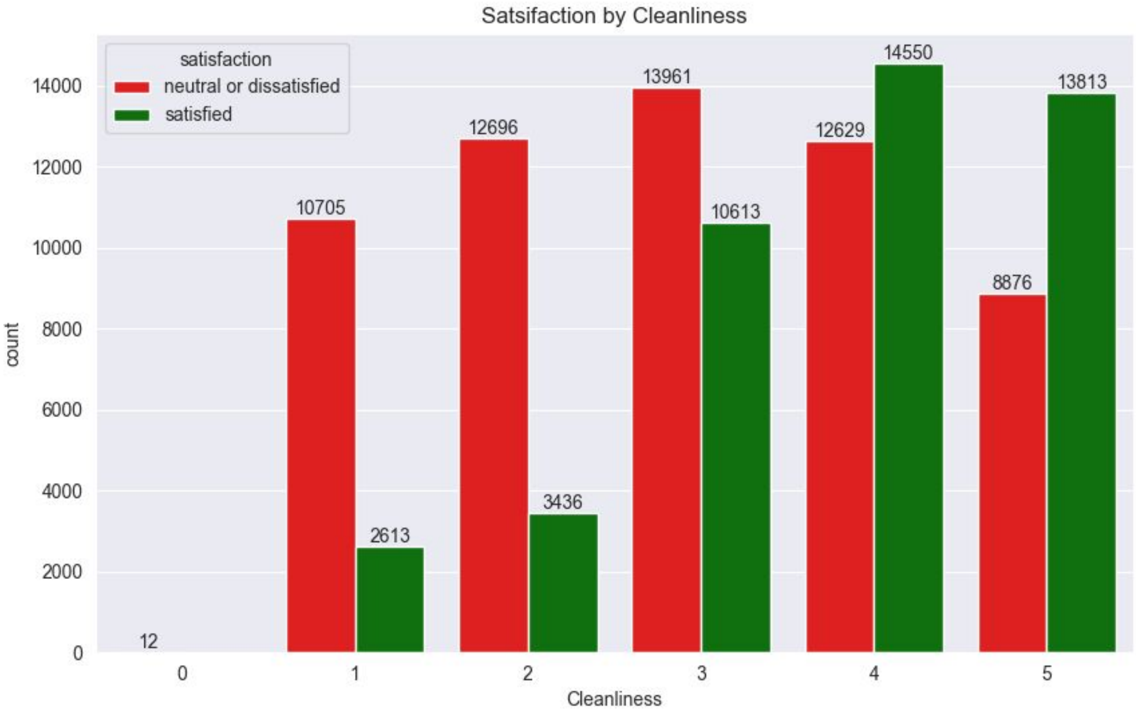
How does the satisfaction proportion differ by Age?



Does Departure Delay have an effect on Satisfaction?



Does cleanliness affect satisfaction?



Machine Learning

Model Preprocessing

Before I could start training different models, I needed to make sure the dataset was made up of numerical values.

Data was split in 70/30 manner, where 70% of data was for training, and 30% was for testing.

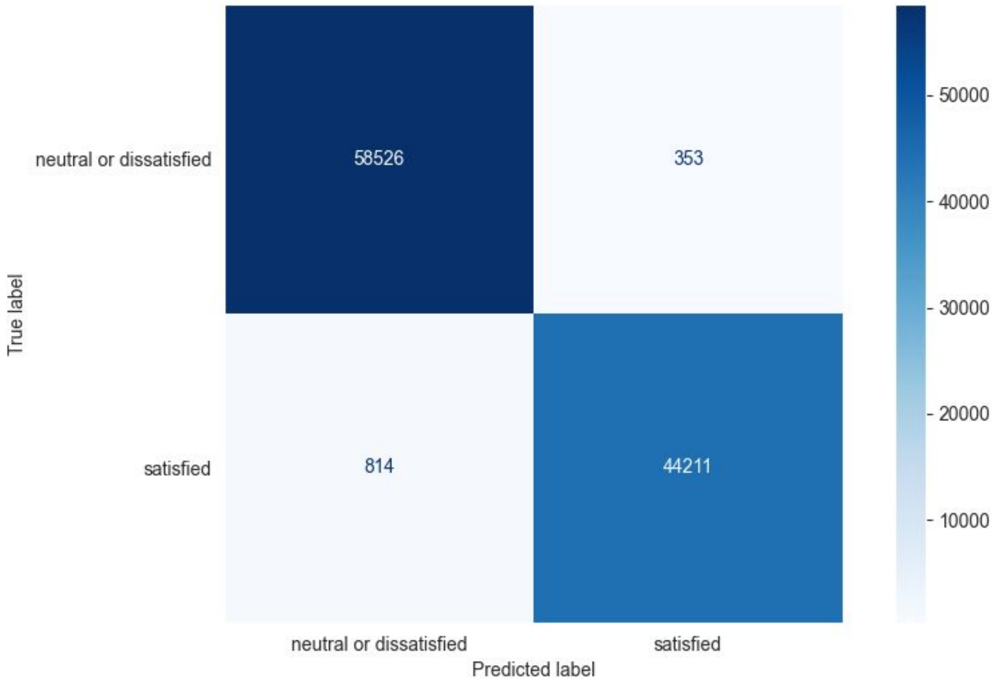
Model Results

I trained 3 models: Logistic Regression, Random Forest, and K Nearest Neighbors (KNN).

The results of the accuracy are below.

Model	**Accuracy**
Logistic Regression	87.54%
Random Forest	96.06%
KNN	92.79%

Random Forest Confusion Matrix



Conclusion

- Satisfaction is influenced by *multiple* service areas
- A single poor experience can outweigh multiple positive ones
- Improvements should prioritize reliability and consistency

Questions?