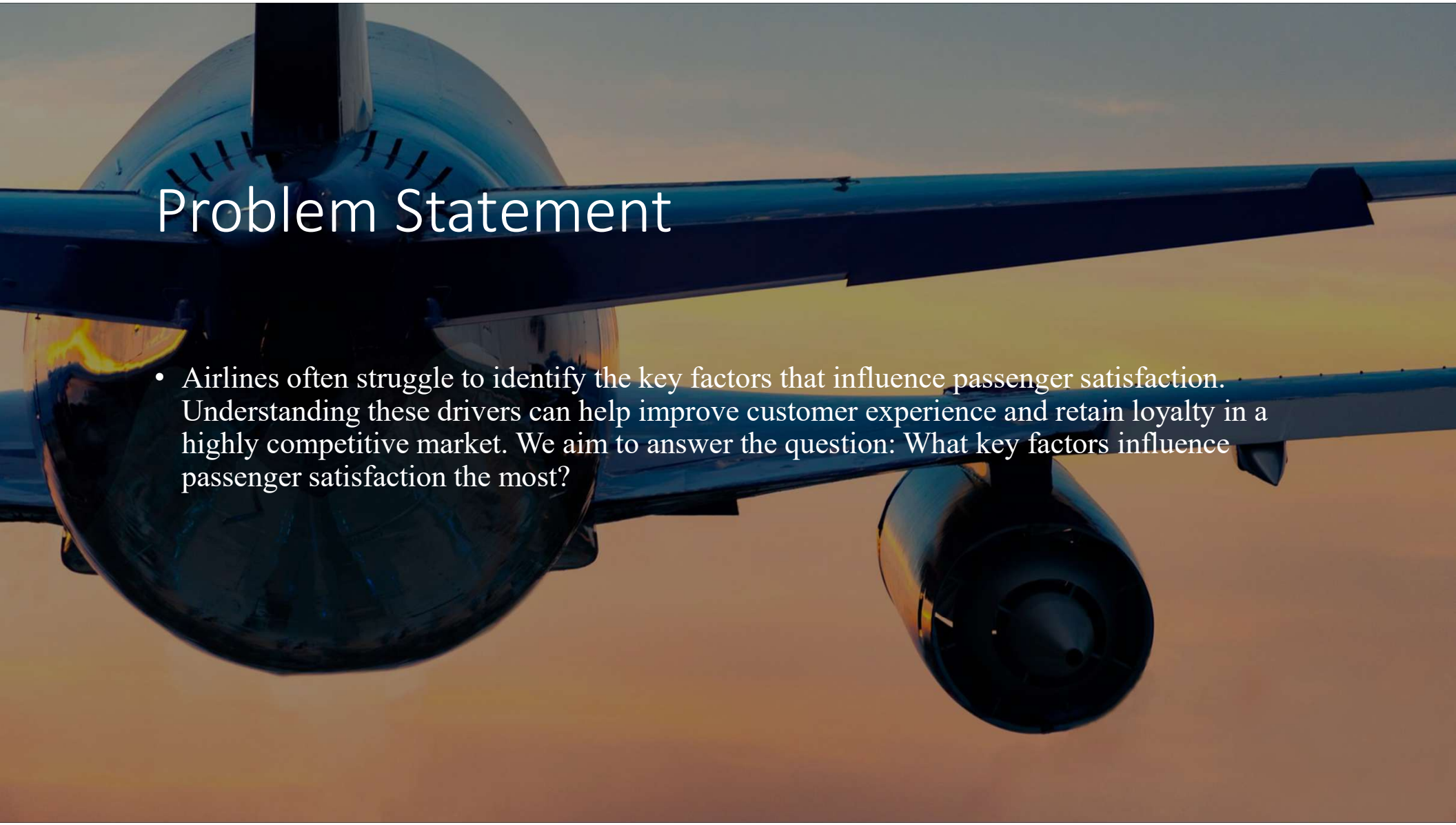
What Drives Airline Passenger Satisfaction?

Micro-Project #1

https://github.com/KcRyan7487/ANA500

Kasey Ryan

2025_May



Problem Statement

- Airlines often struggle to identify the key factors that influence passenger satisfaction. Understanding these drivers can help improve customer experience and retain loyalty in a highly competitive market. We aim to answer the question: What key factors influence passenger satisfaction the most?

Hypothesis Formulation

- Our hypothesis is that certain variables in the dataset will serve as strong and statistically significant predictors of passenger satisfaction, while others may not show a meaningful relationship. Specifically, we anticipate that the following variables will be among the strongest predictors:
  - Class
  - On-board service
  - Inflight service
  - Arrival Delay in Minutes

Acquire

- Dataset: airline.csv
- Description: The purpose of this file appears to be predicting target variable of “satisfaction” based on the other variables present in the dataset. A longer more description name might be something like “Airline Passenger Satisfaction”. A slightly flashier name might be something like “What Drives Airline Passenger Satisfaction” which we have chosen as our official name for this project/analysis/paper/presentation.
- High Level Dataset metrics:
- Source: Retrieved from the provided list of datasets/.csv’s to use as part of course materials for ANA-500
- Raw data: 129,880 rows, 25 columns

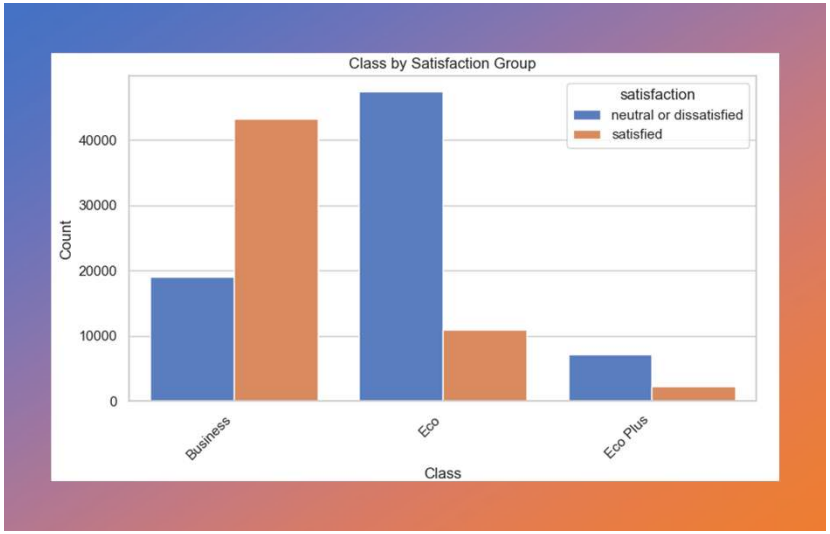
Column	Populated Values	Missing Values	Distinct Values	Data Type
Gender	129880	0	2	object
Customer Type	129880	0	2	object
Age	129880	0	75	int64
Type of Travel	129880	0	2	object
Class	129880	0	3	object
Flight Distance	129880	0	3821	int64
Inflight wifi service	129880	0	6	int64
Departure/Arrival time convenient	129880	0	6	int64
Ease of Online booking	129880	0	6	int64
Gate location	129880	0	6	int64
Food and drink	129880	0	6	int64
Online boarding	129880	0	6	int64
Seat comfort	129880	0	6	int64
Inflight entertainment	129880	0	6	int64
On-board service	129880	0	6	int64
Leg room service	129880	0	6	int64
Baggage handling	129880	0	5	int64
Checkin service	129880	0	6	int64
Inflight service	129880	0	6	int64
Cleanliness	129880	0	6	int64
Departure Delay in Minutes	129880	0	466	int64
Arrival Delay in Minutes	129487	393	472	float64
satisfaction	129880	0	2	object

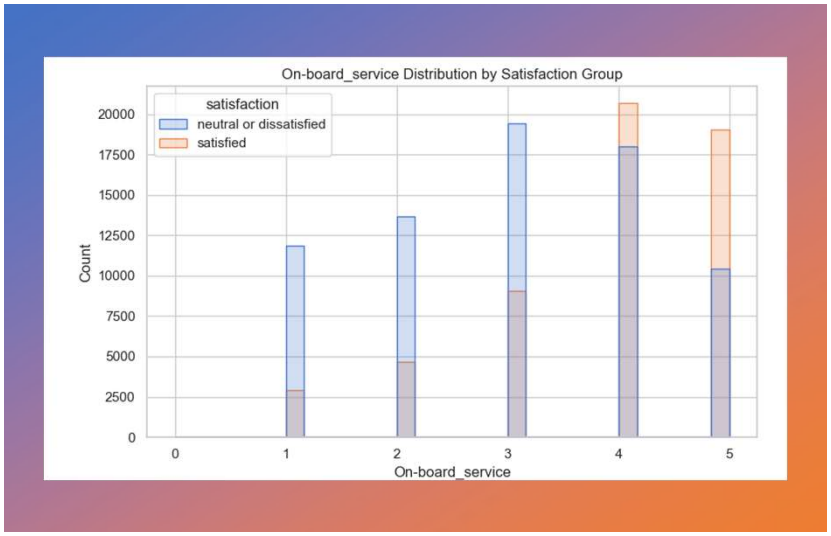
Prepare

- Initial transformations: Drop the blank column at the beginning of the file (hard coded with name "Unnamed: 0“. Also drop the “id” column.
- Initial data integrity checks:
  - 0 duplicate rows
  - 393 missing values for “Arrival Delay in Minutes”

Prepare (additional considerations)

- Some of the variables have quite a few distinct values. It may behoove us to categorize these to a higher aggregation/granularity: Age, Flight Distance, Departure Delay in Minutes, and Arrival Delay in Minutes
- We'll have to decide what to do with the blank values for Arrival Delay in Minutes. We suspect this may be a strong predictor of satisfaction and so, if initial modeling supports that hypothesis and we suspect the variable should remain in the final model, it'll be very important we align on the best approach. Perhaps multiple approaches could be compared such as:
  - Dropping missing values
  - Imputing them based on the global average,
  - Imputing them based on various nearest neighbor methods
- Our target variable satisfaction only contains two possible values, a combination of neutral or dissatisfied or satisfied. Since our desired outcome is satisfied we could recode these to satisfied as a 1 else 0 for ease of processing in future steps if/when needed.





Analyze data

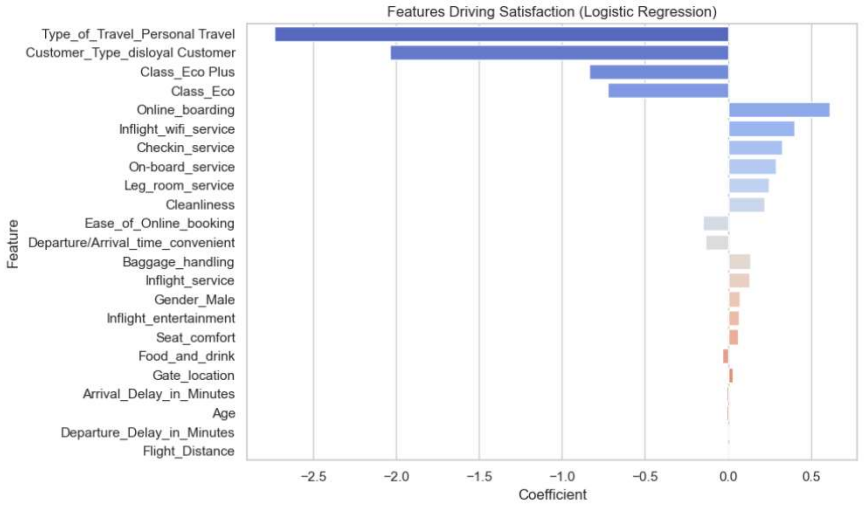
- Exploratory analysis examining the distributions/frequencies for the various variables as well as subset comparisons of said distributions/frequencies by the two satisfaction groups
  - Some hypothesized variables' examples are shown here; more information available in the supporting knitted pdf of the full analysis

Analyze data (cont)

- We opted to drop the observations with missing values for our model building for now, though will be gathering feedback from stakeholders regarding similar analyses conducted using other imputation methods
  - We suggest proceeding with dropping because of the low volumes involved (393 out of 129,880 i.e. approx. 0.3%) and with it the advantage of our model being based upon truly observed data
- 129,487 observations remain, used to train a logistic regression model
  - 80/20 training/testing split
  - 103,589 training / 25,898 testing

Report

- Some interesting findings:
  - Some of our hypothesized variables do indeed appear to have a statistically significant relationship with satisfaction
  - Others such as type of travel, customer type, and the "Online_boarding" variable appear to have a much stronger relationship as well
  - Therefore, additional exploration and refinement will have to be done to select our final model's variables
  - Some features found to be most influential are shown here, more information available in the supporting knitted pdf of the full analysis
- Solid performance on the testing subset of data, BUT complexity of the current model may still be an issue
  - 87% accuracy, 86% precision, 84% recall
  - In future iterations we plan to explore further refinements to balance explanatory power with reduced complexity/parsimony/elegance



Act

- <apply results, connect results with the chosen problem statement or business question>