# Airline Passenger Satisfaction

Liisa-Lotta Jürgenson

Raimo Köidam

Marcus Kommussaar

## Task 1. Setting up

https://github.com/kommussaar/Airline-Passenger-Satisfaction

## Task 2. Business understanding

Identifying your business goals

- Background – the reason for this project is our course Introduction to Data Science. The specific project was chosen because airlines and airports are responsible and interested in good satisfaction of its customers and should be aware of the aspects that matter and impact the experience the most.
- Business goals – understand which aspects of the travel impact the overall satisfaction the most. Predict the profile of a typical satisfied customer (e.g., gender, age, customer type) and develop strategies to improve satisfaction among groups who do not fit this profile.
  From the airlines' or airports' perspective, the ultimate business goal is to increase customer satisfaction and retain loyal customers, while attracting new ones.
- Business success criteria – the project is successful if we manage to find the aspects of the travel that clearly seem to affect the overall rating the most. Create actionable insights, such as improving satisfaction among certain demographics or customer types  In the context of an airline it could be to achieve a 15% improvement in overall customer satisfaction ratings over the next year.

Assessing your situation

- Inventory of resources – We are a team of three members, each equipped with personal computers. We are using Jupyter Notebook and Google colab as our primary tools for data analysis and the publicly available Airline Passenger Satisfaction dataset, which contains over 120,000 entries.
- Requirements, assumptions, and constraints – The project needs to be ready by Monday, Dec 9, at noon (12:00) and on Friday, Dec 13., we must present the project at a poster session, When regarding the data, we assume that the customer satisfaction ratings were given honestly based on the experience and not because of spite or other unrelated reasons. We also have limited experience and a fixed timeline to complete the project, therefore we need to manage our time efficiently.
- Risks and contingencies – As this is our first data science project, we anticipate some delays. Potential bottlenecks include data cleaning and model training. To mitigate these risks, we will conduct frequent team reviews to identify issues early and stay on schedule.

- Terminology - Feature importance: which features are most important in determining the satisfaction level of the passenger. Predictive model - a model that predicts customer satisfaction based on various factors. Customer Profile - Demographics or characteristics commonly associated with satisfied customers.
- Costs and benefits – Costs include time spent by team members and computer usage for analysis. While no direct financial benefits are expected, the project offers significant academic value and real-world experience in applying data science concepts to practical challenges. Airlines or airports implementing these findings could see improved customer retention, leading to long-term financial benefits.

Defining your data-mining goals

- Data-mining goals – 1. Identify which factors lead to satisfaction and dissatisfaction. 2. Create a predictive model to accurately predict the overall satisfaction from input factors. 3. Generate insights into the typical profile of a satisfied customer, which can help airlines target improvements toward underrepresented or dissatisfied groups.
- Data-mining success criteria – Develop a predictive model with at least 80% accuracy, Successfully identify and rank the most critical factors impacting satisfaction and provide actionable recommendations for improving satisfaction among specific demographics or customer types

## Task 3. Data understanding

Gathering data

- Outline data requirements
  The dataset must include detailed information about passenger demographics, travel details, and satisfaction ratings. Additional explanatory attributes such as service quality, delays, and in-flight services are also essential to our analysis.

- Verify data availability
  The data is publicly available on Kaggle under a public domain license, ensuring no restrictions on use. It includes 129,880 passenger records with 24 features for each passenger, offering a rich dataset for analysis. The dataset's large size and diversity ensure that it represents a broad range of passengers, making the analysis statistically robust. However, a potential limitation is that the data reflects satisfaction scores from a specific time frame and may not generalize to other airlines or situations. Nevertheless, its comprehensive nature makes it suitable for achieving our project goals.

- Define selection criteria

We will be using all the fields but dividing them into categories- responsibility of the airport, responsibility of the airline and characteristics of the passenger. This categorization allows us to focus our analysis on actionable insights for specific stakeholders, such as airport management or airline operators Only entries with complete and consistent data will be included in the analysis. Missing or

inconsistent data will be excluded, but the proportion of such exclusions will be carefully monitored to avoid removing a significant portion of the dataset.

Describing data

Our data is divided into categorical and numerical data, consisting of 24 fields and 129 881 rows and is seperated by a delimiter ",".
Fields: ID, Gender, Age, Customer Type, Type of Travel, Class, Flight Distance, Departure Delay, Arrival Delay, Departure and Arrival Time Convenience, Ease of Online Booking, Check-In Service, Online Boarding, Gate Location, On Board Service, Seat Comfort, Leg Room Service, Cleanliness, Food and Drink, In-Flight Service, In-Flight Wifi Service, In-Flight Entertainment Service, Baggage Handling, Satisfaction
Format: CSV
Source: John D, via Kaggle

Exploring data

ID is a unique number for each passenger, Age of the passenger 7 - 85 in years. Gender has two values, Male and Female, Customer Type is either returning or first-timer. Type of Travel is either business or personal. Class is the travel class Economy, Economy Plus or Business Flight distance is values from 31 - 4983 In miles, Departure Delay 0 - 1592 and Arrival Delay 0 -1584 are in minutes, Rest of the fields (except Satisfaction) have values between 0 (lowest) and 5 (highest), representing the customer satisfaction for all these fields. Satisfaction represents the overall satisfaction of the passenger – satisfied, neutral or unsatisfied.
There are 19 columns with numeric values and 5 with categorical values.

Verifying data quality

The data is of good quality with minimal missing or inconsistent values. We will encode non-numeric fields (e.g., Gender, Satisfaction) for use in the predictive model. Additional cleaning, such as handling outliers or normalizing numeric features, may be done during data preparation.

## Task 4. Planning your project

Make a detailed plan of your project with a list of tasks. There should be at least five tasks. Specify how many hours each team member will contribute to each task.

1. Data Exploration, understanding and visualizing 6h
2. Converting categorical variables into numerical variables using encoding techniques 1h
3. Cleaning up the data if necessary (remove duplicates and outliers and handle missing values) 1h
4. Splitting data into train and test sets 1h
5. Training the models and comparing the results 8h
6. Analyzing the results 5h
7. Making the poster 5h

List the methods and tools that you plan to use. Add any comments about the tasks that you think are important to clarify

- Tools
  We will use Jupyter notebook (or Google colab) to train the model(s) and document our code, also for data cleaning,
  model training, and visualization.
- Methods
  Data Preprocessing: Conversion of categorical variables, missing value handling, normalization of numerical data.
  Metrics: Accuracy, precision, recall, F1-score, and feature importance will be used to assess model performance.