

Data Appendix

Dataset: airline.csv

- A dataset consisting of all the reviews and corresponding data for various airlines from Skytrax, a website for reviewing airlines and airports.
- The unit of observation is an airline review.

Column	Description	Example
airline_name	Name of the airline being reviewed.	Adria Airways
link	URL path to the airline's Skytrax review page.	adria-airways/airline-reviews/adria-airways
title	Title or headline of the review.	Adria Airways customer review
author	Name of the reviewer.	D Ito
author_country	Country of the reviewer.	Germany
date	Date the review was posted (YYYY-MM-DD format).	2015-04-10
content	Full written text of the review.	Outbound flight FRA/PRN A319. 2 hours 10 min flight. I thought drinks/snacks for sale but sandwich soft drinks were served complimentary.
aircraft	Aircraft type mentioned in the review (if provided).	A319
type_traveller	Category indicating reason for travel.	Couple Leisure
cabin_flown	Cabin class flown by the reviewer.	Economy
route	Route flown as described in the review.	ZRH-LJU and LJU-VIE
overall_rating	Numerical rating of the airline (1–10).	7
seat_comfort_rating	Rating for seat comfort (1–5).	3

cabin_staff_rating	Rating for cabin staff service (1–5).	4
food_beverages_rating	Rating for food and beverages (1–5).	2
inflight_entertainment_rating	Rating for inflight entertainment (1–5).	4
ground_service_rating	Rating for ground service (1–5).	3
wifi_connectivity_rating	Rating for WiFi/connectivity (1–5).	1
value_money_rating	Rating for value for money (1–5).	4
recommended	Whether the reviewer recommends the airline (1 = Yes, 0 = No).	1

Dataset: airline_cleaned.csv

- A cleaned dataset consisting of only the relevant columns for our project.
- The unit of observation is an airline review.

Column	Description	Example
content	Full written text of the review.	Outbound flight FRA/PRN A319. 2 hours 10 min flight. I thought drinks/snacks for sale but sandwich soft drinks were served complimentary.
type_traveller	Category indicating reason for travel.	Couple Leisure
cabin_flown	Cabin class flown by the reviewer.	Economy
overall_rating	Numerical rating of the airline (1–10).	7

Dataset: airline_cleaned_with_sentiment.csv

- A dataset consisting of only the relevant columns for our project with sentiment scores for the reviews.
- The unit of observation is an airline review.

Column	Description	Example
content	Full written text of the review.	Outbound flight FRA/PRN A319. 2 hours 10 min flight. I thought drinks/snacks for sale but sandwich soft drinks were served complimentary.
type_traveller	Category indicating reason for travel.	Couple Leisure
cabin_flown	Cabin class flown by the reviewer.	Economy
overall_rating	Numerical rating of the airline (1–10).	7
sentiment_score	A decimal value representing the sentiment of the content (Between -1 and 1 inclusive)	0.9460179209709167

Dataset: airline_cleaned_with_mismatch.csv

- A dataset consisting of only the relevant columns for our project with sentiment scores for the reviews and sentiment_score-overall_rating mismatch indicators.
- The unit of observation is an airline review.

Column	Description	Example
content	Full written text of the review.	Outbound flight FRA/PRN A319. 2 hours 10 min flight. I thought drinks/snacks for sale but sandwich soft drinks were served complimentary.
type_traveller	Category indicating reason for travel.	Couple Leisure
cabin_flown	Cabin class flown by the reviewer.	Economy
overall_rating	Numerical rating of the airline (1–10).	7
sentiment_score	A decimal value representing the sentiment of the content (Between -1 and 1 inclusive)	0.9460179209709167
mismatch	Whether there is a mismatch between the sentiment of the content and the overall rating (0 = no mismatch, 1 = mismatch)	0, 1