Table of contents

1 Data Report - TravelHunters Project

This document describes all data used in the TravelHunters project and ensures traceability and reproducibility of results.

The TravelHunters project collects and processes travel data from various sources to create a comprehensive dataset for hotels, activities, and destinations. Data is collected through web scraping, cleaned, and stored in a structured database.

1.1 Raw Data

1.1.1 Overview of Raw Datasets

Name	Source	Storage Location
Booking.com Hotels	Booking.com via Scrapy Spider	`data_acquisition/json_backup/booking_worldwide.json`
Activities Data	GetYourGuide via Scrapy Spider	`data_acquisition/json_backup/activities_worldwide.json`
Destinations Data	Tourism Websites via Scrapy Spider	`data_acquisition/json_backup/destinations_worldwide.json`
Accommodations Data	Various Sources (Hostelworld, Airbnb)	`data_acquisition/json_backup/hotels.json`

1.1.2 Details Booking.com Hotels Dataset

- **Description**: Hotel data from Booking.com including names, ratings, prices, locations and image URLs
- **Data Source**: Booking.com website via Scrapy Spider (`scraping_data_files/spiders/booking.py`)
- **Data Acquisition**: Automated web scraping with pagination for worldwide coverage
- **Legal Aspects**: Publicly available data, respecting robots.txt
- **Data Classification**: Public business data
- **Variables**: Hotel name, rating, price, location, image URLs, description

1.1.2.1 Data Catalogue - Hotels

Column Index	Column Name	Datatype	Values	Description
1	name	TEXT	String	Hotel name
2	rating	REAL	0.0-10.0	Hotel rating
3	price	TEXT	String/NULL	Price information
4	location	TEXT	String	Hotel location
5	image_urls	TEXT	JSON Array	Hotel image URLs
6	description	TEXT	String/NULL	Hotel description
7	source	TEXT	String	Data source (booking.com)

1.1.3 Details Activities Dataset

- **Description**: Activities and tours from GetYourGuide with prices, ratings and locations
- **Data Source**: GetYourGuide website via Scrapy Spider (`scraping_data_files/spiders/activities.py`)
- **Data Acquisition**: Web scraping with pagination for various destinations
- **Data Classification**: Public business data

1.1.3.1 Data Catalogue - Activities

Column Index	Column Name	Datatype	Values	Description
1	title	TEXT	String	Activity title
2	price	TEXT	String/NULL	Price information
3	rating	REAL	0.0-5.0	Activity rating
4	location	TEXT	String	Activity location
5	image_urls	TEXT	JSON Array	Activity image URLs
6	source	TEXT	String	Data source (getyourguide.com)

1.1.4 Entity Relationship Diagram

The ER diagram is located at `docs/er_diagramm/er_diagramm-ER.svg` and shows the relationships between Hotels, Activities and Destinations.

1.1.5 Data Quality

- **Completeness**: >95% of hotels have names and locations, ~80% have ratings
- **Consistency**: Uniform data formats through validation and cleaning

- **Currency**: Data regularly updated through re-scraping
- **Accuracy**: Automatic validation of URLs and data formats

1.2 Processed Data

1.2.1 Overview of Processed Datasets

Name	Source	Storage Location
Merged Travel Data	Consolidation of all raw data	`data_acquisition/merged_travel_data.json`
TravelHunters Database	Cleaned and structured data	`database/travelhunters.db`

1.2.2 Details Merged Travel Data

- **Description**: Consolidated and cleaned data from all sources in uniform format
- **Processing Steps**:
  - Data loading from JSON files
  - Validation and cleaning (URL validation, duplicate removal)
  - Data format normalization
  - Consolidation into SQLite database
- **Access**: Via `data_cleaning_and_db_integration.py` script
- **Statistics**:
  - 2,000 hotels from Booking.com
  - ~90 activities from GetYourGuide
  - ~671 destinations
  - Total: >2,700 records

1.2.2.1 Data Catalogue

1.2.2.2 If applicable: Entity Relationship Diagram

1.2.3 Details Processed Dataset 2

...