

# Table of contents

<b>1</b>	<b>Project Charta - TravelHunters</b>	<b>2</b>
1.1	Problem Definition . . . . .	2
1.2	Situation Assessment . . . . .	2
1.3	Project Goals and Success Criteria . . . . .	3
1.4	Data Mining Goals . . . . .	4
1.5	Project Plan . . . . .	4
1.6	Roles and Contact Details . . . . .	6

# 1 Project Charta - TravelHunters

## 1.1 Problem Definition

**Domain:** Travel and Tourism Data Aggregation

**Problem Statement:** Travelers today face information overload when planning trips, with relevant data scattered across multiple platforms (booking sites, travel guides, activity platforms). There is no centralized system that provides comprehensive, up-to-date travel information including accommodations, activities, and destinations with pricing, ratings, and visual content.

**Expected Benefits:** - Centralized travel data repository for better decision-making - Automated data collection reducing manual research time - Comprehensive dataset enabling travel analytics and recommendations - Foundation for future travel recommendation systems

**Target Users:** - Travel enthusiasts seeking comprehensive trip planning information - Travel agencies requiring aggregated data for client recommendations - Data analysts researching travel trends and pricing patterns - Developers building travel-related applications

**Stakeholders:** - **Primary Users:** Travel planners and tourism professionals - **Data Providers:** Booking.com, TripAdvisor, Lonely Planet, activity platforms - **Development Team:** Data engineers, web scrapers, database administrators - **End Beneficiaries:** Travelers seeking better trip planning tools

## 1.2 Situation Assessment

**Available Resources:** - **Personnel:** Development team with expertise in web scraping, data processing, and database management - **Tools:** Python ecosystem (Scrapy, SQLite, Pandas), Quarto for documentation - **Infrastructure:** Local development environment with capability for web scraping - **Data Sources:** Public travel websites (Booking.com, TripAdvisor, Lonely Planet)

**Time Constraints:** - Project duration: July 2025 (Summer School timeframe) - Limited to proof-of-concept and initial data collection

**Restrictions and Constraints:** - **Legal:** Must comply with website terms of service and robots.txt - **Technical:** Rate limiting to avoid overwhelming target websites - **Data Quality:** Dependent on source website structure and content quality - **Scalability:** Limited by computing resources for large-scale scraping

**Identified Risks:** - Website structure changes affecting scraper reliability - Rate limiting or IP blocking by target websites - Data quality inconsistencies across different sources - Storage limitations for large datasets with images

## 1.3 Project Goals and Success Criteria

**Primary Objectives:** 1. **Data Collection Success:** Collect comprehensive travel data from multiple sources - Target: >2,000 hotel listings with pricing and ratings - Target: >500 activity listings with descriptions and images - Target: >200 destination entries with coordinates and descriptions

### 2. Data Quality Standards:

- 90% of entries have complete name and location information
- 80% of hotels have pricing information
- 70% of activities have rating information
- 60% of entries include image URLs

### 3. System Reliability:

- Scrapers handle pagination and multiple pages successfully
- Robust error handling for website changes
- Data cleaning pipeline removes duplicates effectively

### 4. Database Integration:

- All data successfully stored in structured SQLite database
- Proper indexing for efficient queries
- Data integrity constraints maintained

**Success Metrics:** - **Quantity:** Total items collected across all categories - **Coverage:** Number of unique destinations covered - **Completeness:** Percentage of fields populated per category - **Accuracy:** Manual validation of sample data entries

**Out of Scope:** - Real-time data updates or live synchronization - Advanced recommendation algorithms - User interface development - Commercial deployment or monetization

## 1.4 Data Mining Goals

### Primary Data Mining Task: Data Integration and Information Extraction

**Technical Objectives:** 1. **Web Scraping and Data Extraction:** - Extract structured data from unstructured web content - Handle dynamic content and pagination - Parse and normalize pricing, rating, and location data

#### 2. Data Cleaning and Preprocessing:

- Remove duplicate entries based on name/location similarity
- Standardize pricing formats and currency conversions
- Validate and clean URLs, ratings, and text content

#### 3. Data Integration:

- Merge data from multiple sources (hotels, activities, destinations)
- Create unified schema across different data types
- Establish relationships between accommodations, activities, and destinations

#### 4. Feature Engineering:

- Generate unique identifiers for entities
- Extract location coordinates where available
- Categorize and tag content appropriately

**Quantitative Success Criteria:** - **Data Completeness Score:** >75% for critical fields (name, location, price/rating) - **Duplicate Reduction:** <5% duplicate entries after cleaning - **Data Integration Success:** Successfully merge >95% of collected data into unified schema - **Processing Accuracy:** <2% data corruption during cleaning and transformation

**Technical Requirements:** - Scalable data processing pipeline handling 10,000+ records - Robust error handling with <1% data loss due to processing errors - Database performance supporting sub-second queries on full dataset

## 1.5 Project Plan

### TravelHunters Data Collection and Integration Project

The project follows the CRISP-DM methodology adapted for web scraping and data integration:



Figure 1.1: TravelHunters Project Timeline

gantt

```

title TravelHunters Project Timeline - July 2025
dateFormat YYYY-07-DD
tickInterval 1day
section Project Setup
    Define problem and goals      :done, a1, 2025-07-01, 1d
    Setup development environment :done, a2, 2025-07-01, 1d
    Initial spider development    :done, a3, 2025-07-02, 1d
    Project foundation: milestone, done, m1, 2025-07-03, 0d
section Data Acquisition
    Booking.com spider development :done, a4, 2025-07-02, 2d
    Activities spider implementation :done, a5, 2025-07-03, 1d
    Destinations data collection :done, a6, 2025-07-03, 1d
    Worldwide data scraping :done, a7, 2025-07-04, 2d
    Data collection complete: milestone, done, m2, 2025-07-06, 0d
section Data Processing
    Data cleaning pipeline :active, a8, 2025-07-03, 2d
    Database integration :a9, 2025-07-03, 1d
    Duplicate removal and validation :a10, 2025-07-04, 1d
    Quality assurance :a11, 2025-07-04, 1d
section Analysis & Documentation
    Exploratory data analysis :a12, 2025-07-05, 2d
    Documentation completion :a13, 2025-07-06, 2d
    Final evaluation :a14, 2025-07-07, 1d
    Project completion : milestone, m3, 2025-07-08, 0d

```

**Phase Descriptions:** - **Project Setup (Days 1-3):** Environment setup, initial spider development, proof of concept - **Data Acquisition (Days 2-6):** Large-scale data collection from multiple sources with pagination - **Data Processing (Days 3-5):** Cleaning, integration, and database storage - **Analysis & Documentation (Days 5-8):** Analysis, documentation,

and final evaluation

## **1.6 Roles and Contact Details**

List the people involved in the development work here with their role titles, tasks and contact details