King Saud University
College of Computer and Information Sciences
Department of Information Technology

كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

Phase 1 - Logbook: Detailed Overview

Date of Collection: February 1-6, 2025

Data Collection Methods Overview:

1. **TripAdvisor and Booking.com (Feb 1 and Feb 4, 2025):** Utilized Apify for web scraping, focusing on hotel reviews across Saudi Arabia. Targeted data included customer reviews, ratings, and accommodation specifics.
2. **Ministry of Tourism Open Data Portal (Feb 2, 2025):** Direct download of tourist statistics including overnight stays and spending patterns.
3. **Survey Distribution (Feb 4-6, 2025):** Implemented through Google Forms, targeting both domestic and international tourists' preferences and satisfaction levels.

Tripadvisor

Date of Collection: Feb 1, 2025

Data Collection Methods:

We utilized **Apify**, a web scraping and automation platform, to collect data from TripAdvisor via its API. Our goal was to extract relevant information on hotels across Saudi Arabia to analyze customer reviews, ratings, and other relevant attributes.

Future Tasks:

Data preprocessing:

- Data Processing and Cleaning: Plan to handle missing values and standardize formats in the analysis phase.
- Data Analysis: Utilize the collected data to determine visitor satisfaction trends and rating distributions.

Challenges & Solutions:

- Rate Limiting: Encountered API rate limits which slowed down data collection.
  - **Solution**: Implemented a staggered data retrieval approach to comply with rate limits while ensuring complete data collection.
- Data may not be sufficient for the analysis & research.
  - **Solution**: Web scraped booking.com using the same tool to compensate for any possible data insufficiencies

Decisions Made and Rationale:

The TripAdvisor dataset, while rich in hotel reviews, lacked coverage of diverse accommodation types and broader geographic areas. To address this and enhance our analysis:

Rationale:

- **Diverse Accommodations**: Including Booking.com provides insights into a wider range of lodging options.
- **Wider Geographic Reach**: Booking.com offers data from more varied locations across Saudi Arabia.
- **Comprehensive Analysis**: This expansion ensures a richer, more detailed study of the tourism sector.

Web Scraping Tools:

Apify

وزارة السياحة
Ministry of Tourism

Date of Collection: Feb 2, 2025

Data Collection Methods:

1. Got access to the Ministry of Tourism dataset without the need of a key or registration.
2. Query Design: No need for a query.
3. Data Retrieval: download data straightforward from the website, including fields like Overnight Stays, Tourists Number (Overnight Visitors) , Tourists Spending, and more.

Future Tasks:

Data preprocessing:

- Data Processing and Cleaning: Plan to handle missing values and standardize formats in the analysis phase.
- Data Analysis: Utilize the collected data to determine spending patterns, identify trends, and analyze tourism behavior.

Challenges & Solutions:

- No challenges were encountered.

Date of Collection: Jan 31, 2025

Data Collection Methods:

1. Needed to register to get access to the X developer platform and then got the bearer token.
2. Query Design: used a python code (the bearer token was used in the code) to collect the data and save it in a csv file.
3. Data Retrieval: after running the code the data was retrieved, it includes the date of the post and the content of the post.

Future Tasks:

Data preprocessing:

- Data Processing and Cleaning: Plan to handle missing values and standardize formats in the analysis phase.
- Data Analysis: Utilize the collected data to determine tourists reviews and preference, identify trends, analyze hashtags and more.

Decisions Made and Rationale:

Abandoning the Dataset: After reviewing the data limitations, specifically the cap of 100 posts per query and the extensive preprocessing required to make the data usable, the decision was made not to proceed with this data source.

Rationale:

- **Limited Data Scope:** The cap of 100 posts was not sufficient to provide a comprehensive view of public sentiment and trends.

- **High Preprocessing Effort:** The data required significant cleaning and preprocessing efforts, which were not feasible within the project's timeline and resource constraints.

Challenges & Solutions:

- Key generation: There were a lot of keys generated like API Key, API Key Secret, and last Bearer token therefore it was confusing which key to use.
    - **Solution**: Searched for the proper key to use, using Google, FAQ X developer platform.
- **API Rate Limiting and Data Cap:** Faced challenges with the API's limit on the number of posts retrievable, which restricted the depth of data analysis.
    - **Solution:** Explored alternative data sources that could offer more extensive data without such restrictions.

API QUERY:

https://colab.research.google.com/drive/1Ebcqm33pGVXIy9MuCCY89keJx7o7iDmy?usp=drive_link

- Surveys

Date of Collection: Feb 4 - 6, 2025

Data Collection Methods:

1. Developed a comprehensive questionnaire to capture tourists' preferences, satisfaction levels, and spending habits. The survey was designed using **Google Forms** to facilitate easy distribution and collection of responses.
2. The survey was shared via email and social media platforms to reach a wide range of domestic and international tourists who visited Saudi Arabia in the past year.
3. Responses were automatically collected in Google Forms, which were then exported to a googles sheet.Which was then exported as a CSV. The data includes timestamps of responses, demographic information, and detailed answers to each survey question.

Future Tasks:

Data preprocessing:

- We Plan to clean the data by removing any incomplete responses and standardizing text inputs into categorical data.
- Develop strategies to handle missing data, either by inputting mean/mode values or excluding incomplete entries based on the analysis requirements.
- Any other data preprocessing methodologies

Data Analysis:

- Utilize the collected data to analyze tourist satisfaction levels, spending patterns, and preferences.
- Identify Trends: Use statistical methods to identify trends and patterns in the survey data.
- Segmentation Analysis: Perform segmentation analysis to understand different tourist behaviors based on demographic and trip characteristics.

Challenges & Solutions:

- Initially faced a lower response rate than expected response rate which could affect the data's representativeness.
    - **Solution**: Searched for the proper key to use, using Google, FAQ X developer platform.

- Ensuring all survey questions comply with privacy laws and ethical standards.
  - **Solution:** Designed the survey to be anonymous, including a consent form at the beginning explaining the purpose of the study and how the data will be used.

Booking.com

Date of Collection: Feb 4, 2025

Data Collection Methods:

We used **Apify**, a web scraping and automation platform, to collect data from Booking via its API. This enabled us to analyze customer reviews, length of stay, and the origin of guests for various hotels in different cities across Saudi Arabia.

Future Tasks:

- **Data preprocessing:** Further cleaning and structuring of the scraped data to ensure consistency, remove duplicates, and handle missing values.
- **Data Analysis:** Performing in-depth analysis to extract meaningful insights from customer reviews, stay durations, and guest origins.

Challenges & Solutions:

- Some hotel listings may have missing customer reviews, incomplete stay duration details, or unstructured guest origin data.
  - **Solution**: Implement data preprocessing techniques such as imputation for missing values, data validation checks, and filtering out unreliable entries.
- The initial scraping attempt retrieved only a small subset of the available data, limiting the scope of analysis..
  - **Solution**: Expanded the list of URLs being scraped to cover more hotels across different cities, ensuring a more comprehensive dataset.

Web Scraping Tools:

Apify