

GET YOUR GUIDE

*Data Design and Representation
project for “Travel memories you’ll
never forget”.*

Shiyi (Ashley) Yue

Ridwan Abdusalam

Pradeepthi Mallappa

TABLE OF CONTENTS

| | |
|--|----------|
| EXECUTIVE SUMMARY | 1 |
| BACKGROUND, CONTEXT, AND DOMAIN KNOWLEDGE | 1 |
| DATA SOURCES, WEB-SCRAPING ROUTINE, AND DATABASE DESIGN CHOICES | 2 |
| DRIVING BUSINESS VALUE | 6 |
| SUMMARY AND CONCLUSIONS | 7 |
| APPENDIX | 8 |
| REFERENCES | 9 |

EXECUTIVE SUMMARY

Our project aims to extract relevant information from the "Get Your Guide" website to help UC Davis students plan their spring break vacation. We used web scraping techniques to gather data on destination names, categories, cities, guided tours, star ratings, cancellation and reservation policies, duration, estimated cost per person, price category, the validity of the ticket, guidance language, and audio guide availability. We chose to store this data in a MongoDB database for its flexibility and ability to handle large volumes of unstructured data efficiently.

This dataset will help students explore popular tourist destinations and attractions, make informed decisions based on prices, reviews, and tour guide details, and improve their overall user experience. Promoting traffic to the website and subscriptions will also benefit the company. Our project provides a data-driven approach to enhance the planning process for a relaxing and enjoyable spring break vacation.

BACKGROUND, CONTEXT, AND DOMAIN KNOWLEDGE

Background: “Get your guide” is an online tour guide created to change the way people connect to the places they visit so anyone can create their dream vacation. After the user browses and books for a journey, “Get your Guide” tells the user when to be where. The traveler can just focus on having a great time. This website gives personalized recommendations before and during the trip to have an incredible experience.

Context: Spring break is arriving and the UC Davis MSBA students are very much eager to have a fun-filled relaxed holiday and get rejuvenated for the last quarter before the job season. The

students are looking forward to making use of the information in the “Get your Guide” website to search for and plan their holidays. Hence, our project team decided to scrap the necessary information for the students to be utilized.

Domain Knowledge: This project specifically focuses on extracting usable data from “Get your Guide” to get data in a format where the students can filter for their topic of interest like Culture, Food, Nature, or Adventure related locations. They can have a view of data to explore for the period of visit, availability, ratings, reviews, approximate cost per person, amenities included, and other important information.

DATA SOURCES, WEB-SCRAPING ROUTINE, AND DATABASE DESIGN CHOICES

Introduction of Data Sources: The data is extracted from the company’s website:

<https://www.getyourguide.com>

From the main page mentioned above, we searched for 15 top popular tourist cities in the US, including San Francisco, Los Angeles, San Diego, Seattle, Las Vegas, Honolulu, Miami, Orlando, Chicago, New Orleans, Boston, New York City, Washington, DC, Nashville, and Portland. Each result page shows available activities in the city. We keep them sorted in the recommended order. One page includes 16 activities. For each city, we do web-scraping for the first three pages; therefore, 48 available activities for each of the 15 cities, which are 720 activities in total, are considered in this project.

Description of the Web-Scraping Routine(s): To get the activity pages for each city listed, we first use Selenium to automate the searching part by looping through the city list and setting each city as the input value to the home website. In order to page backward to get the second and the

third activity pages of each city, we also use Selenium by clicking on the page down button twice. We save all the first three pages for each city for future reference and a more stable web-scraping process.

Based on the three city pages for each city we saved, we extract the basic information of each available activity, including the name and the category of the activity, the link to the activity page, the number of reviews and the ratings, if any, and the pricing and the pricing category, if any. The pricing category here indicates the pricing is per person or per group. All the information is stored in the collection “top_activities” collection in the database “getyourguide” using MongoDB.

We then try to extract more detailed information about the activities based on the links to the activity pages we extracted from the city pages. Just like what we did for downloading the city pages, we use Selenium to go through all the links saved. That’s also because we found that the website has some technique against plain get requests, which means that the pages downloaded by using get requests do not include all the information we want on the website.

After downloading the 720 activity pages, we loop through each of them one by one, extracting all the listed information about the activity under the “About this activity/ticket” category. This part of the information mainly includes the cancellation policy, the estimated duration of the activity, whether a small or private group is allowed, and whether skipping the ticket line is available. To maintain the consistency and integrity of the database, we also implemented some manipulations on the detailed info extracted. These implementations will be covered in the database design choices section. The below table shows the main attributes we extracted from the city pages and the activity pages.

| Attributes | Attribute Description |
|----------------------------------|--|
| City | City of the activities |
| Rank | The rank of the activity based on the recommended order within each city |
| Name | Name of the activity on the website |
| Category of the activity | Nature of the activity including entry ticket, water activity, day trip, etc |
| Link | The link to the individual activity page |
| Reviews | Total number of reviews |
| Rating | Average star rating of the activity based on the reviews |
| Pricing | Ticket/Activity price |
| Price category | Indicates the pricing is per person or per group |
| Cancellation Policy | Norms of cancellation policies |
| Reservation Policy | Whether reserve now & pay later is an option |
| Duration | Expected duration of the activity |
| Validity for | For how many days/months the ticket is valid |
| Live tour guide | Languages the live tour guide speaks that are available |
| Optional Audio Guide | Languages the audio guide speaks that are available |
| Private or Small group available | Whether a private or small group is allowed or up to how many members |
| Skip the ticket line | Whether skipping the ticket line is available or from where to skip the line |
| COVID-19 precautions | Available precautions against COVID-19 |
| Driver | Whether local drivers are available and the languages they speak |
| Wheelchairs accessible | Whether wheelchairs are accessible at the destination |
| Pickup included | Whether pickup service is included in the activity and how does it work |
| Pickup optional | Whether pickup service is available and how does it work |
| Host or greeter | Whether a host or greeter is available and the language(s) they speak |

Database Design Choices: We chose MongoDB as the DBMS for this project because the attributes for different activities may differ a lot. If we use RDBMS, MySQL for instance, there will be a lot of null values in the table. In addition, MongoDB works much more efficiently with adding new attributes, which is inevitable for updating this project since new information about activities can always be found given the power of the internet. What's more, many kinds of information about the activities are text, which also leads to the choice of using MongoDB.

As mentioned above, we have basic information (background colored green in the table above) along with additional/detailed information (background colored blue) about each activity. The

basic information is neat and ready to use whereas the additional information is relatively messy and needs manipulation. The table above only shows the additional attributes after manipulation.

Here are the specific steps we implemented for cleaning the data.

1. We point out available services by conveying “Yes” as the values of corresponding keys.

We notice that some categories themselves are descriptive on the website and do not have corresponding values. To differentiate these attributes from the attributes the activity does not have, we manually convey “Yes” as values.

2. We replace the "Free cancellation" category with the "Cancellation policy" category and

set “Free cancellation” as the value of “Cancellation policy” since there are already the two categories, and the “Free cancellation” category only indicates the activity is free to cancel while “Cancellation policy” contains other possibilities of the cancellation policy.

3. We replace "Reserve now & pay later" with "Reservation policy" for more legibility.

Setting “Reserve now & pay later” as the value for “Reservation policy” makes more sense and makes the attribute name shorter, which increases the legibility of the doc.

4. We notice that the raw data contains many redundant attributes about durations, such as

“Durations for 2 hours”, “Durations for 3 hours” and so on. So we use regex to extract the actual durations and set them as values while leaving “Durations” as the key.

5. Just like what we do with “Durations”, we do the same with the category - “Valid”, which confronts the same redundant attribute issues.

6. We notice that there are attributes “Skip the line through...” as well as “Skip the ticket line” which basically mean the same thing but the former ones are more specific. So we combine these attributes together setting “Skip the ticket line” as the key and the other

related attributes as values while keeping the records of “Skip the ticket line” (Keep the original “Yes”s)

7. We replace the "Instructor" category with the "Host or greeter" category. We notice that these two categories deliver the same information which is the language(s) the local instructor/host/greeter speaks. Therefore, to reduce the redundancy, we leave the category “Host or greeter” only.

DRIVING BUSINESS VALUE

The data we are extracting captures the information about the tour destinations, attractions, themes, pricing, reviews and ratings, guide profiles, availability and booking details.

Advantages of Database Implementation: For database implementation, a NoSQL database such as MongoDB is selected to tackle the unstructured or semi-structured data with flexible schema format as commonly used for scraping from websites. MongoDB can handle large volumes of data and scale horizontally across multiple servers making the data suitable for high-volume scraping and for efficient execution. Mongo DB also enables efficiency in frequent and fast data updates. Collecting information about each individual tour may provide detailed insights in Mongo DB without much processing or storage space.

Business Value: Exploration of popular tourist destinations and attractions can help the users in selecting tours offered. By analyzing the prices, customer reviews, and tour guide details can optimize decision-making and improve the user experience. This is very important to the company too because this enables a batch of users at once in promoting the traffic to the website and subscribing to it.

By providing students with relevant information to plan their trips, the company can increase its revenue by attracting more customers. This data can also help the company identify popular destinations and attractions, allowing them to better focus their marketing efforts.

Our project provides a unique dataset that can be used to gain a competitive advantage. The ability to provide comprehensive information on popular tourist destinations and attractions will help the company stand out in a crowded marketplace.

The data collected can be used to improve the website's recommendation system, helping the company personalize recommendations for each user. This feature will provide a better user experience, increase subscriptions, and revenue.

SUMMARY AND CONCLUSIONS

Our project aimed to extract data from the "Get Your Guide" website to help UC Davis students plan their spring break vacation. We used web scraping techniques to gather data on destination names, categories, cities, guided tours, star ratings, cancellation and reservation policies, duration, estimated cost per person, price category, validity of the ticket, guidance language, and audio guide availability. We chose to store this data in a MongoDB database for its flexibility and ability to handle large volumes of unstructured data efficiently. Our dataset will help students explore popular tourist destinations and attractions, make informed decisions based on prices, reviews, and tour guide details, and improve their overall user experience. Promoting traffic to the website and subscriptions will also benefit the company. Our project provides a data-driven approach to enhance the planning process for a relaxing and enjoyable spring break vacation.

APPENDIX

Example of an activity:

<https://www.getyourguide.com/sassi-di-matera-l89408/matera-2-hour-guided-tour-of-sassi-t261282/>

GET YOUR GUIDE

Where are you going?

English ▼ USD (US\$) ▼ Wishlist Cart Help Log in ▼ Sign up

Italy > Basilicata > Things to do in Matera > Church of Saint Francis of Assisi, Matera,

Available today and tomorrow

Matera: Guided Tour of Sassi di Matera

GUIDED TOUR


★★★★★

4.7 / 5

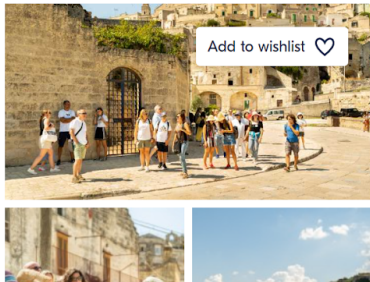
580 reviews

Activity provider: Martulli Viaggi - Matera

View all 15 images



Add to wishlist



Discover one of Matera's ancient landscapes on a guided tour of Sassi, exploring the UNESCO-listed site's alleys, churches, and historical caves on this unique cultural adventure.

About this activity

Free cancellation

Cancel up to 24 hours in advance for a full refund

Reserve now & pay later

Keep your travel plans flexible — book your spot and pay nothing today.

Covid-19 precautions

Special health and safety measures are in place. Check your activity voucher once you book for full details.

Duration 2 - 2.5 hours

Check availability to see starting times.

Live tour guide

Italian, German, French, English, Spanish

Private or small groups available

From **US\$ 48.12** per person

Book now

Reserve now & pay later to book your spot and pay nothing today

Give this as a gift

Example of an ticket:

<https://www.getyourguide.com/vatican-museums-l2738/skip-the-line-vatican-museums-sistine-chapel-ticket-t62214/>

Page 8

English
USD (US\$)
Wishlist
Cart
Help
Log in
Sign up

Italy
Lazio
Things to do in Rome
Sistine Chapel

Available tomorrow

Vatican: Museums & Sistine Chapel Entrance Ticket

ENTRY TICKET
★★★★★
4.5 / 5
56258 reviews
Activity provider: GetYourGuide Tours & Tickets GmbH

View all 16 images

See priceless works of art from the Papal collections in the Vatican Museums and Sistine Chapel. Marvel at masterpieces from antiquity to Michelangelo's legendary frescoes.

Likely to sell out

From **US\$ 28.34** per person

Book now

About this ticket

Cancellation policy

This activity is non-refundable

Covid-19 precautions

Special health and safety measures are in place. Check your activity voucher once you book for full details.

Valid 1 day

Check availability to see starting times.

Skip the ticket line

Optional audio guide

English, French, Italian, Portuguese, Spanish, Russian, Japanese, German, Chinese, Korean

Give this as a gift

REFERENCES

Main Website URL with culture:

https://www.getyourguide.co.uk/?visitor-id=3HL7WZRWCYXZOAX97QYIUCTDG2KJY08I&locale_autoredirect_optout=true

Main Website URL with Food:

<https://www.getyourguide.co.uk/?selectedTab=0a050aa1-8582-1114-8185-823361e80002>

Main Website URL with Nature:

<https://www.getyourguide.co.uk/?selectedTab=0a050aa1-8582-1114-8185-82356b310004>

Main Website URL with Adventure:

<https://www.getyourguide.co.uk/?selectedTab=0a050aa1-8582-1114-8185-8235e1900005>

About “Get Your Guide”:

<https://www.getyourguide.com/about/>