Geocoding Truck Stops Documentation

William Co

2025-08-12

This report documents the geocoding of U.S. truck stop data, addressing challenges from inconsistent address formats. Using phone number matching and structured data from Truck Stops and Services, Yelp, Yellow Pages, and iExit, we achieved a 99.19% match rate. A custom interface was developed to support manual verification and ensure data accuracy.

Table of contents

1 Setup

We are working with a truck stop directory dataset that includes truck stop details. However, the dataset lacks geographic coordinates (latitude and longitude). The objective of this project is to extract and assign accurate coordinates to each entry.

The main challenge arises from the inconsistent formatting of addresses. Some entries contain full addresses, while others include only road names, highway exits, or mile markers. The lack of standardization complicates automated geocoding.

The dataset also required extensive data cleaning to be useful.

2 Challenges

2.1 Inconsistent Addresses

Addresses in the dataset fall into the following categories:

- **Standard addresses:** These include a street number and road name, allowing for straightforward geocoding.
- **Exit-based addresses:** These reference a highway and exit number, but may lack a full street address.
- **Non-standard addresses:** These do not conform to either of the above formats, such as entries that only specify the intersection of two streets or other ambiguous location descriptions.

This variability in address formats presents a significant challenge for automated geocoding.

2.2 Age

Some locations in the dataset are quite old, making it difficult to find current information using modern mapping services. In these cases, it is often necessary to consult dated websites to locate and verify these places. Human judgment is required to determine whether such locations still exist or have been repurposed for other uses.

3 Data Scraping

We use the following websites to scrape geographic coordinates. Each time we scrape these websites, we extract the latitude and longitude of the truck stops.

3.1 Truck Stops and Services Website

The first step leverages a website aimed at truckers and RV travelers called Truck Stops and Services and the RV and Travelers Directory. Data from these websites is scraped due to their consistent formatting and the extensive truck stop information they provide.

This would be an example of the webpage:

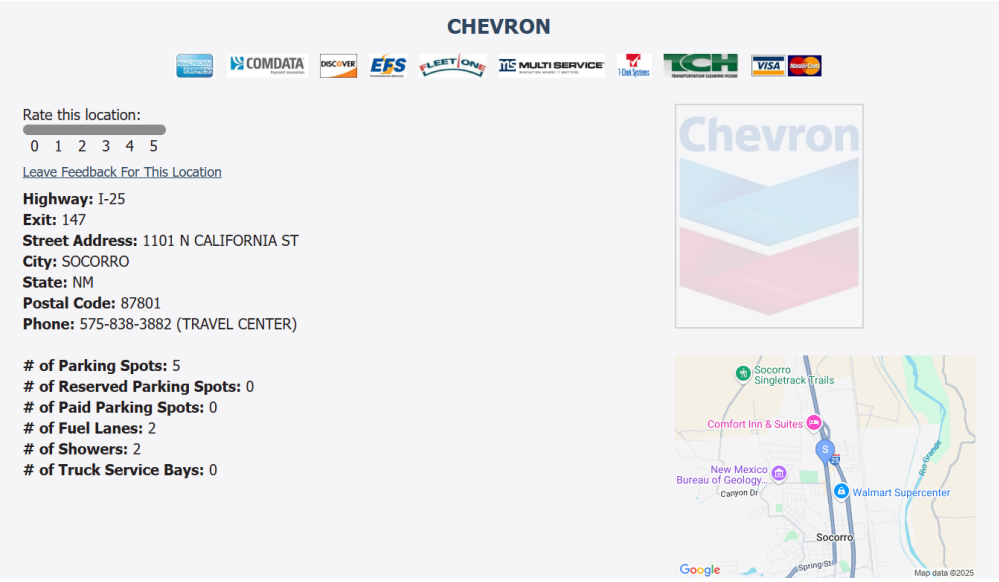

Figure 1: Example of the Truck Stops and Services website interface

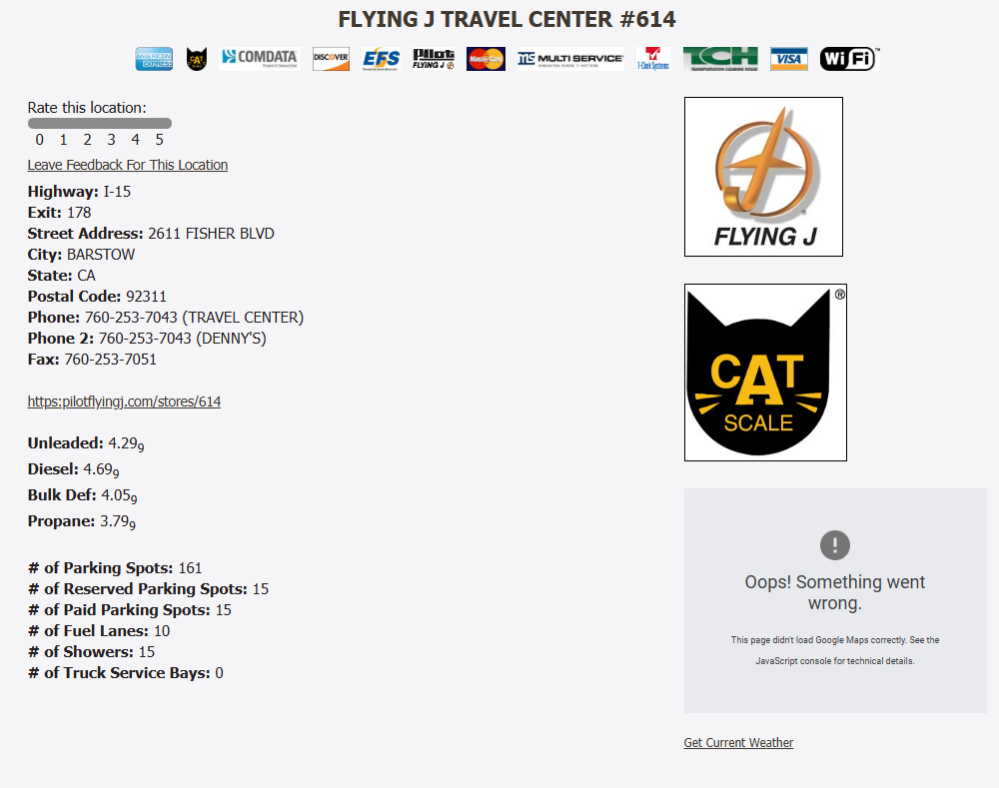

Figure 2: Example of the RV and Travelers website interface

Another example of the Truck Stops and Services website interface.

The complete data available from this scraping is available in Appendix: Truck Stops and Services/ RV and Travelers Data Dictionary.

3.2 Yelp

Next, we leverage the Yelp API, which is unique in allowing phone number-based search queries. This feature makes it particularly well-suited to our dataset, which contains standardized phone numbers. An example of the Yelp website interface is shown below:

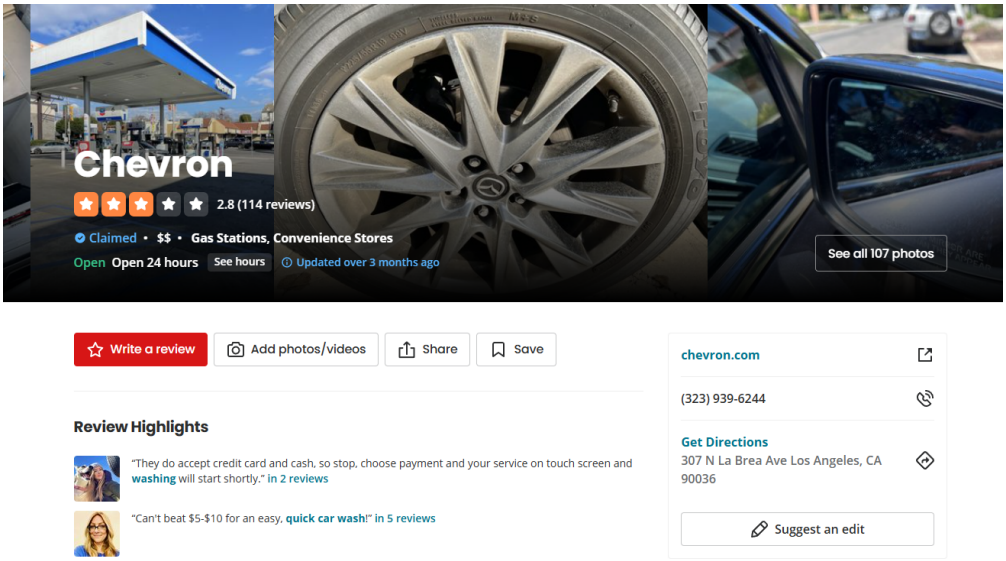

Figure 3: Example of the Yelp website interface

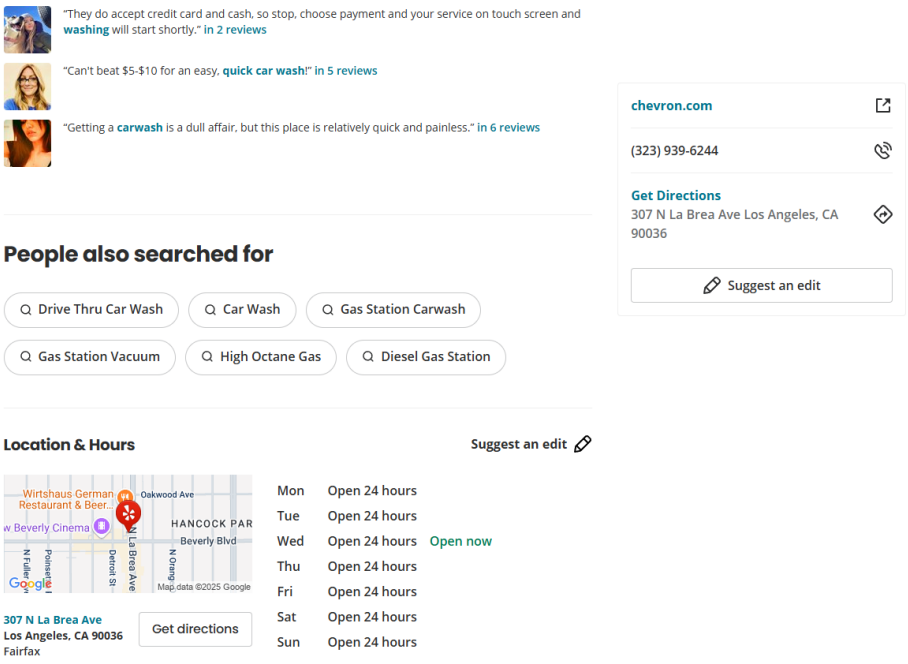

Figure 4: Another example of the Yelp website interface

The full list of data fields extracted from Yelp is available in the Appendix: Yelp Data Dictionary.

3.3 Yellow Pages

We then scrape YellowPages, which also allows direct phone number queries through web scraping. Examples of the website interface are shown below:

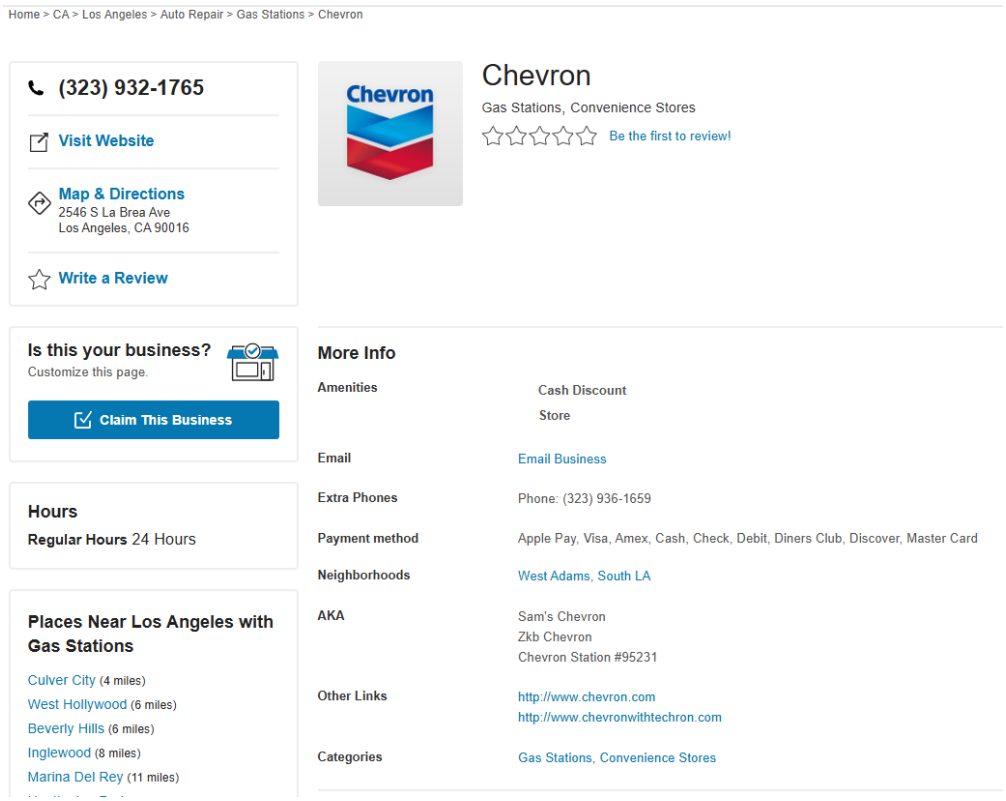

Figure 5: Example of the YellowPages website interface

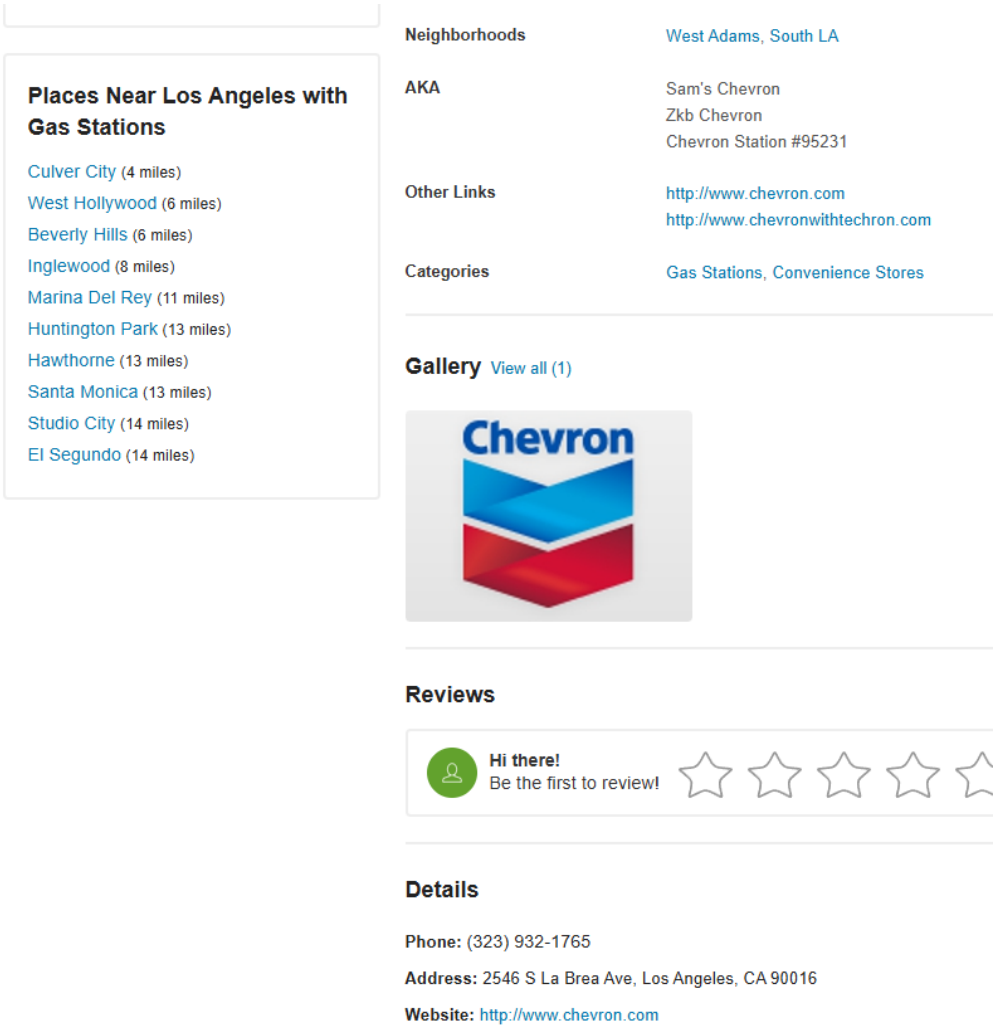

Figure 6: Another example of the Yellow Pages website interface

The full list of data fields extracted from Yellow Pages is available in the Appendix: Yellow Pages Data Dictionary.

3.4 Entry Matching

After scraping from these three websites, we now have three reference datasets to match against our original truck stops directory.

We developed two main methodologies for matching.

3.4.1 Phone Number Matching (Truck Stops and Services):

Directly matching entries based on phone numbers. We match the phone numbers from Yelp, Yellow Pages, and Truck Stops and Services to the original truck stop directory.

3.4.2 Place Name to ZIP Code Matching

This strategy involves a hierarchical matching process. First, the state or ZIP code must match. Next, the city or highway exit must match. After that, the road name must match. Finally, the business or place name must match.

Using these matching processes, each entry in the original truck stop directory can be matched in four ways: three by phone number matching (from Yelp, Yellow Pages, and Truck Stops and Services) and one by place name to ZIP code matching.

3.5 Post Matching

3.5.1 False Positives

We encountered numerous false positive matches after the initial matching process. Yelp and Yellow Pages are known to have inconsistent phone numbers, and all the scraped data sources have documented examples of incorrect latitude and longitude coordinates. This leads to multiple possible latitude and longitude values for a single truck stop directory entry.

To address this, we developed the following coordinate matching methodology.

3.5.2 Coordinate Matching

Suppose we have a Yelp source, where the phone number matches to one unique latitude and longitude. Next, the Truck Stops and Services (Phone) source matches to two different latitude and longitude pairs, while Truck Stops and Services (Place Match) matches to one distinct coordinate. There is no match for Yellow Pages. Finally, we have the original entry_n from the truck stop directory. In this scenario, we observe several possible coordinates for a single truck stop, but there is no reliable method to determine which coordinate is correct.

Suppose we also entry_{n+1} where we have multiple coordinates from the Truck Stops and Services (Place Match) source but no matches from other sources.

Let C_i represent coordinate pairs (latitude, longitude) where $C_i = (\text{lat}_i, \text{lon}_i)$.

Data Source	Yelp	Truck Stops and Services (Phone)	Truck Stops and Services (Place Match)	Yellow Pages
entry_n	C_1	C_2, C_3	C_4	$\emptyset$
entry_{n+1}	$\emptyset$	$\emptyset$	C_5, C_6, C_7	$\emptyset$

For entry_n, we have the set of possible coordinates $\mathcal{C}_n = \{C_1, C_2, C_3, C_4\}$, where each coordinate represents a potential location for the same truck stop entry.

In order to discern the correct coordinate, we make the assumption that a true coordinate C^* exists where if two different sources agree on the same coordinate, it must be correct.

Distance-Based Validation Approach:

Using this approach, we calculate the Euclidean distance between all coordinate pairs. For any two coordinates C_i and C_j, the distance is defined as:

$$D_{i,j} = \sqrt{(\text{lat}_i - \text{lat}_j)^2 + (\text{lon}_i - \text{lon}_j)^2}$$

For entry_n, we compute all pairwise distances:

- $D_{1,2}$: distance between Yelp and first Truck Stops coordinate
- $D_{1,3}$: distance between Yelp and second Truck Stops coordinate

- $D_{1,4}$: distance between Yelp and Place Match coordinate
- $D_{2,4}$: distance between first Truck Stops and Place Match coordinates
- $D_{3,4}$: distance between second Truck Stops and Place Match coordinates

The general distance matrix for n coordinates is:

$$\mathbf{D} = \{D_{i,j} : i, j \in \{1, 2, \ldots, n\}, i \neq j\}$$

We then identify the minimum distance $\min\{D_{i,j}\}$ in $\mathbf{D}$ and record this value in a dedicated column called `min_distance`. This enables systematic manual correction and error analysis.

Afterward, we flag entries where `min_distance` exceeds 200 meters for manual review and correction. If two coordinates C_i and C_j from different sources are within a small distance (i.e., $\min\{D_{i,j}\}$ is less than 200 meters), we treat their midpoint location as the final coordinate for that entry. If manual correction is required, the updated coordinates C_i^* and C_j^* are recorded as the corrected location.

4 Appendix

4.1 Truck Stops and Services/ RV and Travelers Data Dictionary

The following table summarizes the data fields used in the truck stop dataset:

4.1.1 General Information

Column Name	Description
`state_id`	State identifier
`state`	Name of the U.S. state
`name`	Truck stop name
`href`	Relative URL path
`full_url`	Full website URL
`stop_type`	Type of stop (e.g., fuel, full)
`Chain`	Company or chain name

4.1.2 Location Details

Column Name	Description
`Latitude`	Latitude coordinate
`Longitude`	Longitude coordinate
`Highway`	Associated highway
`Exit`	Exit number
`Mile Marker`	Highway mile marker
`Street Address`	Street address
`City`	City name
`State`	State abbreviation
`Postal Code`	ZIP/postal code

4.1.3 Contact Information

Column Name	Description
`Phone`	Main contact number
`Phone 2-5`	Additional phone numbers
`Fax`	Fax number
`Mailing Address`	Mailing address

4.1.4 Amenities & Services

Column Name	Description
`Hours of Operation`	Operating hours
`# of Parking Spots`	Total parking spaces
`# of Reserved Parking Spots`	Number of reserved spaces
`# of Paid Parking Spots`	Paid-only spots
`# of Fuel Lanes`	Fuel pump lanes for trucks
`# of Showers`	Total shower stalls
`# of Men's Showers`	Men's shower stalls
`# of Truck Service Bays`	Truck repair/service bays

4.1.5 Fuel Types & Links

Column Name	Description
`Unleaded`	Unleaded gasoline available (Y/N)
`Diesel`	Diesel fuel available (Y/N)
`Bulk Def`	Diesel exhaust fluid (DEF) availability
`Propane`	Propane fuel available (Y/N)
`https`	HTTPS version of site URL

Column Name	Description
`http / htp`	Alternate/incomplete protocols

4.2 Yelp Data Dictionary

4.2.1 General Business Information

Column Name	Description
`Original_Phone`	The phone number used as input for the Yelp phone search
`Name`	The official name of the business
`Rating`	Yelp rating (e.g., 4.5 stars)
`Review_Count`	Total number of Yelp reviews
`Is_Closed`	Boolean indicating if the business is permanently closed
`URL`	Full Yelp business listing URL

4.2.2 Location Details

Column Name	Description
`Address`	Street address of the business
`City`	City where the business is located
`State`	State (abbreviation)
`Zip_Code`	Postal or ZIP code
`Latitude`	Latitude coordinate
`Longitude`	Longitude coordinate

4.2.3 Contact & Business Attributes

Column Name	Description
`Phone`	Official business phone number returned by Yelp
`Categories`	List of categories (e.g., “Coffee & Tea”, “Gas Station”)
`Price`	Price level indicator (`$`, `$$`, etc., if available)

4.3 Yellow Pages Data Dictionary

4.3.1 General Business Information

Column Name	Description
ADDRESS	Full address of the business as listed on Yellow Pages
AKA	Alternate names or aliases for the business
BUSINESS_NAME	The primary name of the business
BUSINESS_URL	URL to the Yellow Pages business listing
CATEGORIES	Business categories (e.g., "Restaurants", "Auto Repair")
STATUS	Business status (e.g., "Open", "Closed")
WEBSITE	Official website of the business, if available

4.3.2 Location Details

Column Name	Description
JSONLD_CITY_1	City extracted from the embedded structured JSON-LD data
JSONLD_STATE_1	State extracted from the embedded structured JSON-LD data
JSONLD_STREET_1	Street address from JSON-LD
JSONLD_ZIP_1	ZIP code from JSON-LD
JSONLD_LAT_1	Latitude coordinate from JSON-LD
JSONLD_LNG_1	Longitude coordinate from JSON-LD

4.3.3 Contact & Business Attributes

Column Name	Description
ORIGINAL_PHONE	Phone number used to initiate the Yellow Pages lookup
FORMATTED_PHONE	Formatted business phone number as displayed
JSONLD_PHONE_1	Phone number from the structured JSON-LD data
EXTRA_PHONES	Any additional phone numbers found
PHONE	Phone number listed in the primary Yellow Pages HTML content
JSONLD_NAME_1	Business name from structured JSON-LD data

4.3.4 Metadata

Column Name	Description
SCRAPED_AT	Timestamp of when the data was scraped
SEARCH_URL	URL used to perform the Yellow Pages phone-based search

4.4 iExit Data Dictionary

Column Name	Description
state	U.S. state abbreviation (e.g., TX, CA) where the highway exit is located
highway	Name or number of the highway (e.g., I-10, US-101)
exit_id	Unique identifier for the highway exit as used in iExit
title	Display title or name of the exit
exit_name	Name of the exit (may include road or location name)
exit_description	Additional descriptive text about the exit or nearby services
exit_location	Textual representation of the exit's location
iexit_detail_link	URL link to the iExit detailed page for the exit
latitude	Latitude coordinate of the exit
longitude	Longitude coordinate of the exit
google_maps_link	Direct link to the exit location on Google Maps
direction	Direction of travel (e.g., Northbound, Eastbound)