Geocoding Truck Stops Report

William Co

2025-07-16

This is the abstract of the academic report. Summarize the main points and findings here.

Table of contents

1 Introduction

This document is prepared for Prof. Ron Yang, who will be returning to meet with us. It outlines the current progress of the geocoding task.

2 Setup

We are working with a 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.

3 Approach

3.1 Google (Initial) Approach

The first approach involved directly submitting address strings to the Google Maps APIs (Geocoding and Places). However, this method encountered several limitations:

- The API struggled with inconsistent or non-standard address formats.
- We could not reliably verify the accuracy of returned locations.
- Many addresses returned irrelevant or incorrect coordinates.

3.1.1 Methods Tried With Google (Initial) Approach

- Constraining search to specific regions (both hard and soft constraints).
- Using the Nearby Search function (by place type and radius).
- Cleaning and standardizing address strings before submission.
- Exploring and comparing the Google Geocoding and Places APIs.

Although the Google API offers convenience, it lacks reliability in the context of our dataset.

3.2 Truck Stops and Services Website (Second) Approach

The second approach leverages a standardized and reliable feature of our dataset: phone numbers. These are consistently formatted and can be used to match entries with external data sources. We utilized two key websites for this task:

- Truck Stops and Services
- RV and Travelers Directory

This would be an example of the webpage:

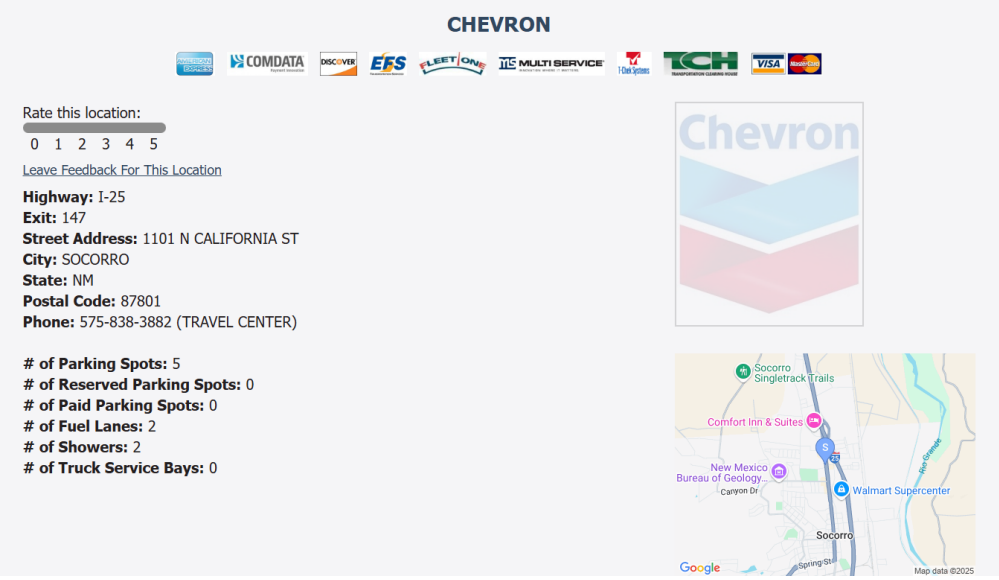

Figure 1: Example of the Truck Stops and Services website interface

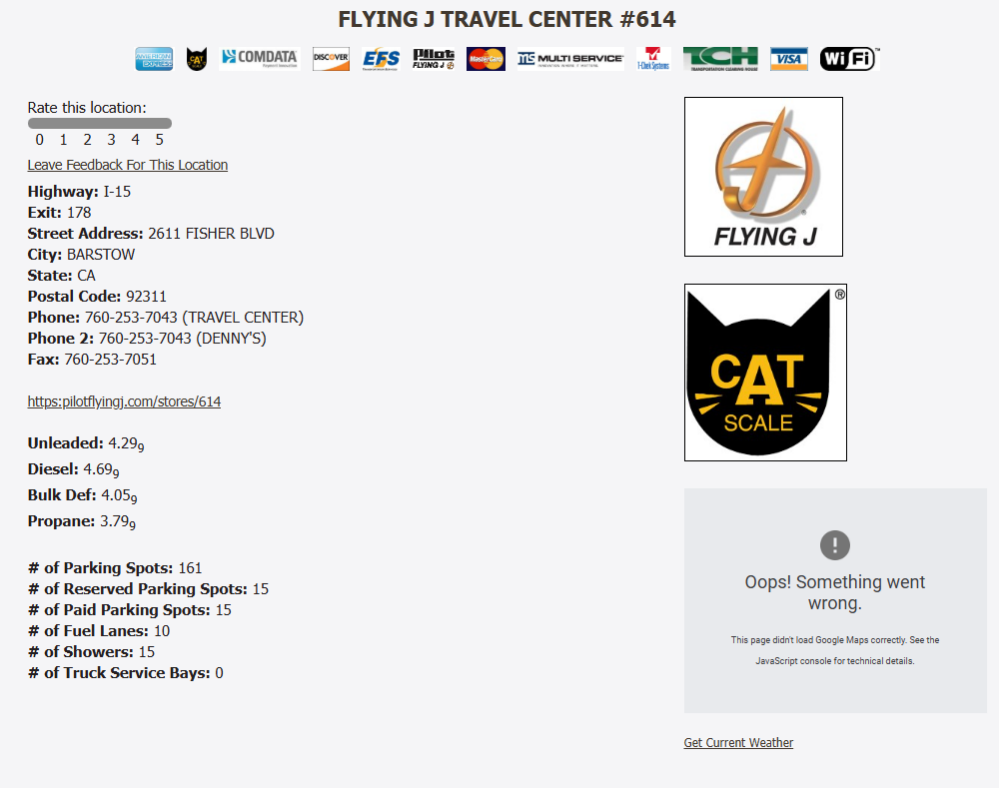

Figure 2: Example of the Truck Stops and Services website interface

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. Most notably, we observe directly the latitude and longitude, directly along with phone and the name of the stop.

Both websites maintain structured records that include phone numbers, making them effective for cross-referencing. Using these sources, we employed two main matching strategies:

1. **Phone Number Matching**: Directly matching entries based on phone numbers.
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 align. After that, the road name must match. Finally, the business or place name must match.

3.2.1 Results

- Phone number matching alone successfully resolved 1,795 out of 2,333 entries (76.94%), 538 missing entries.
- Matching based on ZIP code and place name resolved 1,900 out of 2,333 entries (81.44%), 433 missing entries.
- By combining both methods, we were able to resolve 2,090 out of 2,333 entries (89.58%), 243 missing entries.

This approach improved match rates compared to the first method and provided a more verifiable matching process.

3.3 Yelp (Third) Approach

The third approach leverages 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:

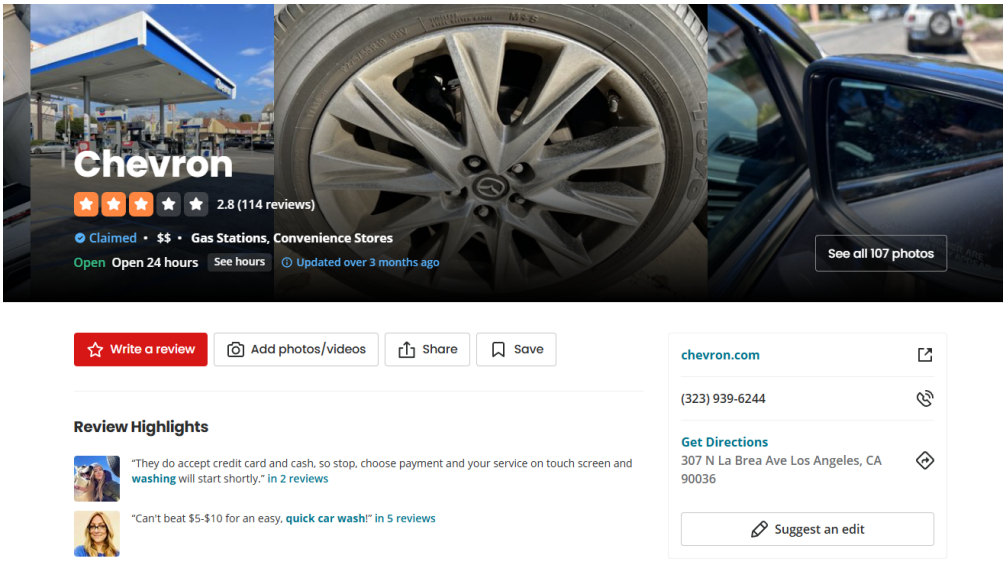

Figure 3: Example of the Yelp website interface

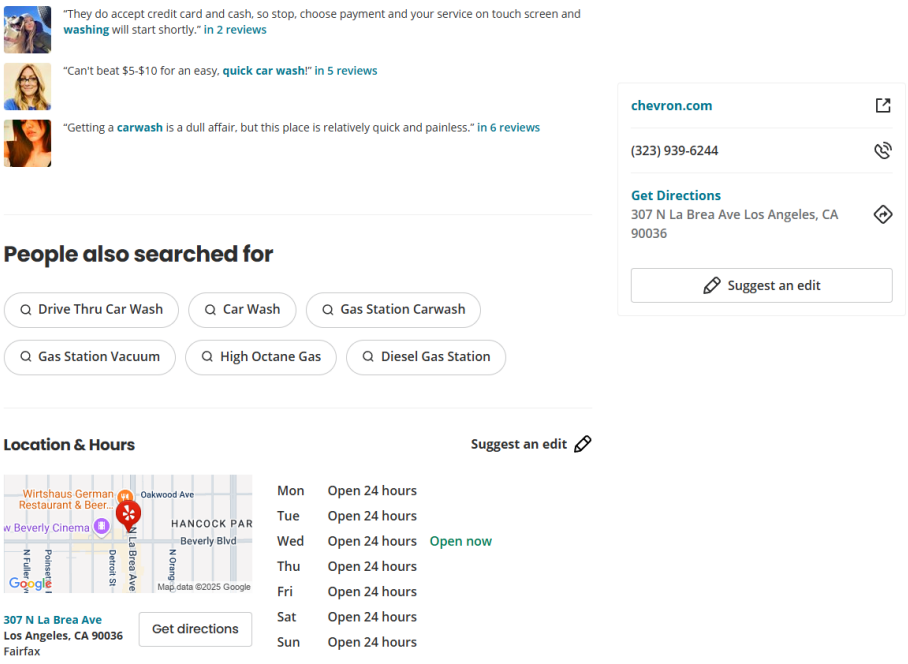

Figure 4: Another example of the Yelp website interface

The full list of data fields extracted from Yelp is available in the Appendix: Yelp Data Dictionary. Most notably, we observe, place name, latitude, longitude and weather a place is still active.

3.3.1 Results

- Matches before Yelp (Second Approach): 2,090 out of 2,333 entries (89.58%), 243 entries missing
- Matches after Yelp integration (Second and Third Approach): 2,189 out of 2,333 entries (93.83%), 144 entries missing

3.4 Yellow Pages (Fourth) Approach

We further improved our match rate using YellowPages, which also allows direct phone number queries through web scraping. Examples of the website interface are shown below:

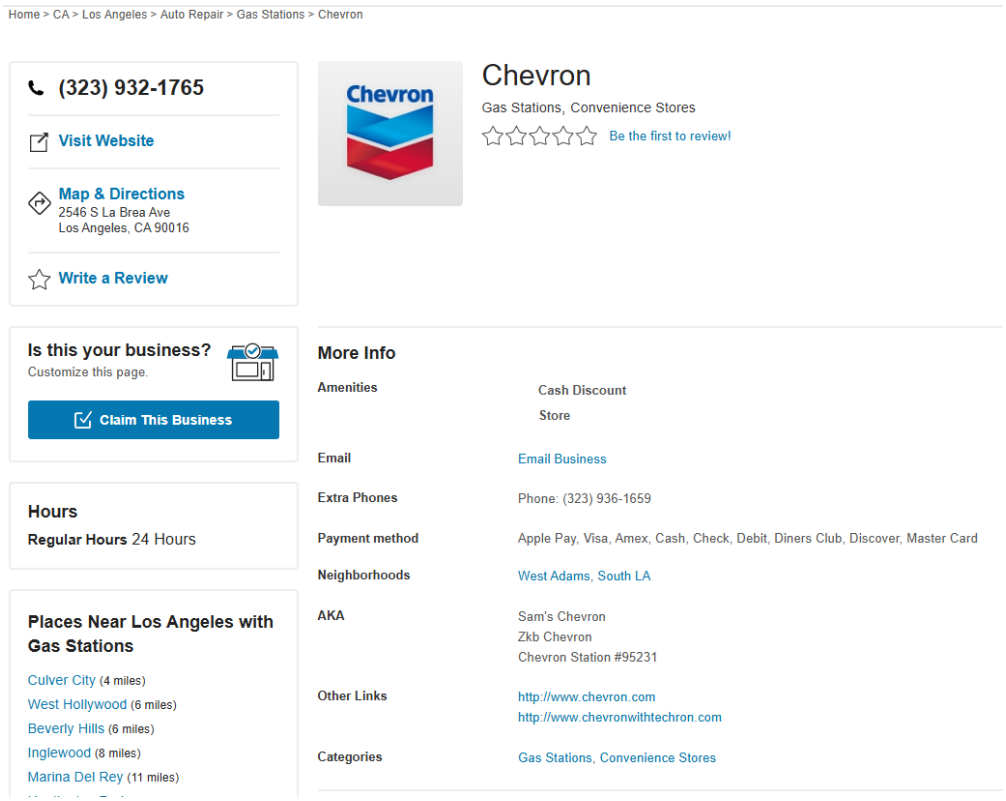

Figure 5: Example of the YellowPages website interface

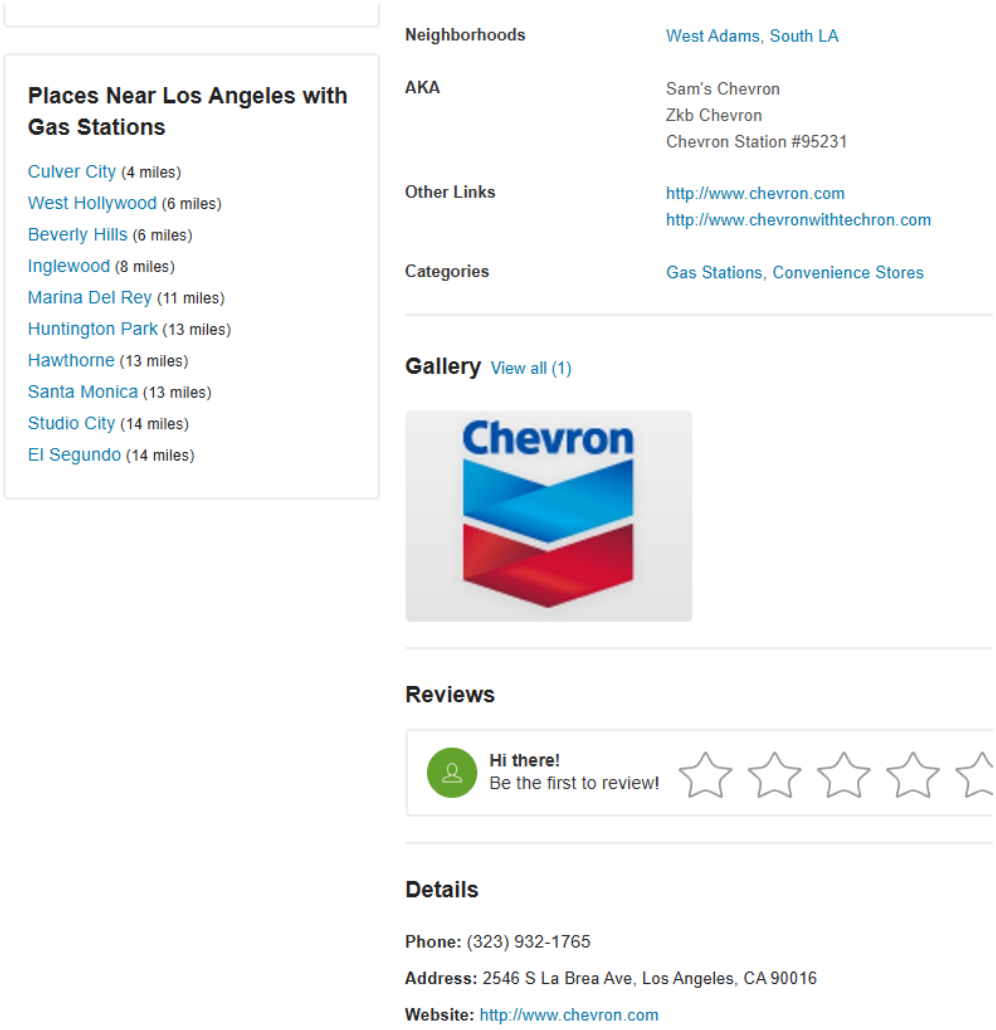

Figure 6: Another example of the Yellow Pages website interface

The full list of data fields extracted from Yellow Pages is available in the Appendix: Yellow Pages Data Dictionary. Most notably, we observe, place name, latitude, longitude and phone.

3.4.1 Results

- Matches after Yelp integration (Second and Third Approach): 2,189 out of 2,333 entries (93.83%), 144 entries missing
- Matches after YellowPages integration (Second,Third and Fourth Approach):2,314 out of 2,333 (99.19%), 19 entries missing

Given the small number of remaining unmatched entries, manual verification is now a feasible next step.

3.5 Summary of Approaches

After implementing considering all(2,3,4) approaches, the consolidated output can be viewed here: Final Output.

The entries that remain unmatched are available here: Missing Outputs.

3.5.1 Matching Summary Table

Matching Method	Matches	Total Entries	Match Rate	Missing Entries
Truck Stops and Services Phone number matching	1,795	2,333	76.94%	538
Truck Stops and Services ZIP code and place name matching	1,900	2,333	81.44%	433
Truck Stops and Services All matching	2,090	2,333	89.58%	243
Truck Stops and Services + Yelp Approach	2,189	2,333	93.83%	144
Truck Stops and Services + Yelp + Yellow Pages	2,314	2,333	99.19%	19

3.6 Current Project: iExit Integration

Since the project's goal is to match latitude and longitude data to exit locations, a comprehensive dataset of U.S. highway exits is essential. To my knowledge, iExit is the most complete single source currently available online. Example of the Exit image website is as follows.

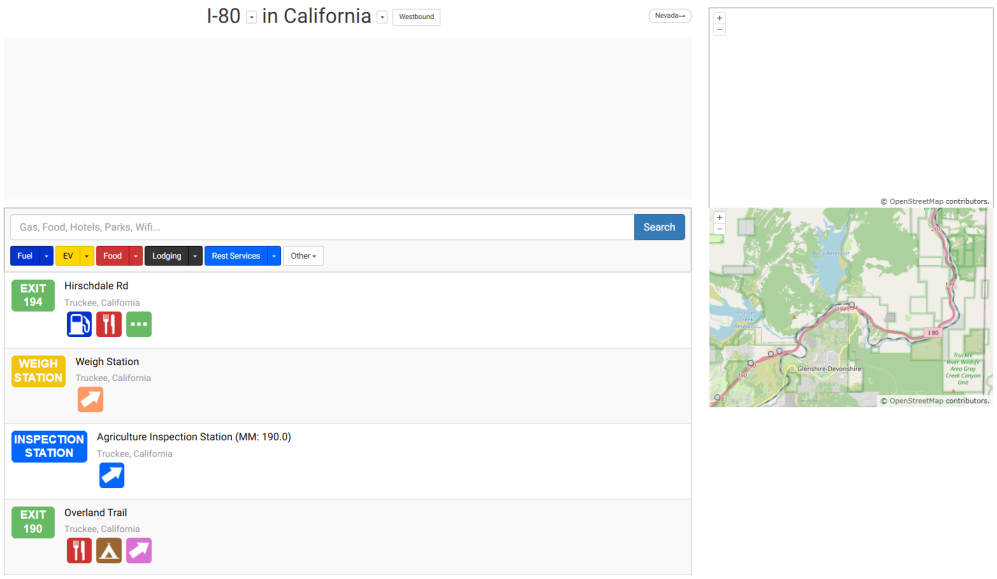

Figure 7: Example of the iExit website interface

The full list of data fields is available in the Appendix: iExit Data Dictionary. Most notably, we observe highway, exit, latitude, longitude and **direction of travel**. Direction of travel is to be emphasized as this is a unique piece of information not available elsewhere.

3.7 Utilities Developed

To support manual validation, we developed a custom GUI tool to assist future research assistants. This interface displays two columns—original dataset vs. matched results—side-by-side, allowing quick visual inspection and correction without requiring coding experience. *(Attach GUI screenshot here)*

3.8 Next Steps

- Finalize code cleaning for future reproducibility.
- Conduct manual verification of the remaining 19 unmatched entries.

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

Column Name	Description
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
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)

Column Name	Description