

US Healthcare Premium Lead Generation Pipeline

A production-grade Google Maps scraping system delivering 67,230 verified B2B leads across four healthcare niches — dentists, medical spas, chiropractors, and veterinarians — covering 49 US states and 2,740 cities.

Scope

49 States · 2,740 Cities ·
1,528 Queries

Data Volume

67,230 Records · 27
Enriched Fields

Version

v4.4 — Live Dashboard
Edition

Date

August 2026

1. Executive Summary

This project engineered an automated B2B lead generation pipeline that scrapes, enriches, and validates healthcare business data from Google Maps at national scale. Built in Python (~650 lines), the system uses Playwright browser automation, multi-page website crawling, and DNS-level email verification to produce 27 enriched fields per record — delivering actionable sales intelligence across four healthcare verticals.

67,230 TOTAL LEADS	49 STATES COVERED	2,740 CITIES TARGETED	98.9% PHONE COVERAGE
------------------------------	-----------------------------	---------------------------------	--------------------------------

2. The Challenge

- **Fragmented data:** No single source covers all four niches nationally with consistent, comparable fields.
- **Stale contacts:** Email addresses and phone numbers decay rapidly; manual verification doesn't scale across thousands of businesses.
- **Missing qualification signals:** Raw directory listings lack treatment offerings, insurance acceptance, and booking data — the information that turns a name into a qualified lead.

The objective: a single, clean dataset with 27 enriched fields, verified contact data, and geographic density sufficient for territory-based sales campaigns.

3. Technical Architecture

1 Discovery Playwright searches 1,528 queries across Google Maps; scrolls sidebar feed to collect all place URLs.	2 Extraction Parses each Maps listing: name, category, rating, reviews, phone, address, claimed status.	3 Enrichment Crawls homepage + 3 priority internal pages; extracts emails, social links, treatments, booking/insurance/telehealth flags.	4 Verification DNS MX-record check on every email; board certification detection; composite lead score (0–100).
---	---	--	---

Resilience by design: 150-query batch processing with fresh Playwright browser contexts per batch prevents memory leaks. A JSON checkpoint file tracks completed queries for crash recovery. Cross-query deduplication by Google Maps URL and website domain. Only 1.5% of discovered websites (900 of 60,668) were dead after the full crawl.

4. Data Quality Controls

Email Filtering Pipeline

- Blocks 14 known dummy domains (email.com, example.com, etc.)
- Strips 10 public webmail providers (gmail.com, yahoo.com, etc.)
- Filters 14 asset file extensions mistakenly parsed as emails
- Validates length (6–55 characters) and syntax
- DNS MX-record verification on every candidate domain

Content Enrichment Rules

- 20+ procedure keywords across 4 treatment categories
- Regex extraction of Facebook, Instagram, LinkedIn profile URLs
- 5 booking-platform signals: Zocdoc, Calendly, "book online," "schedule online," "book appointment"
- Insurance keyword scan: Medicare, PPO, HMO, CareCredit
- Board certification verified from website license/credential text

5. Dataset Composition & Quality

Core Niche Distribution (by Source Query)

Chiropractors		18,716
Veterinarians		17,078
Dentists		16,953
Medical Spas		14,483

Lead Quality Breakdown

SCORE BAND			DESCRIPTION
90–100	42,200	62.8%	Premium: verified email, live website, board cert evidence, full enrichment
70–89	10,984	16.3%	Strong: most fields present, minor gaps only
50–69	7,732	11.5%	Moderate: core contact data available, partial enrichment
30–49	4,787	7.1%	Basic: limited website enrichment, some contact gaps
0–29	1,527	2.3%	Minimal: dead website or missing contact information

62.8% of the dataset — 42,200 records — scores 90+. These premium leads carry verified business emails, live websites, board certification evidence, and complete treatment/insurance enrichment — ready for immediate sales outreach without additional research.

Geographic Coverage — Top 10 States

RANK	STATE		RANK	STATE	
1	California	10,329	6	Ohio	2,377
2	Texas	7,237	7	Colorado	2,371
3	Florida	6,994	8	Michigan	2,333
4	North Carolina	2,665	9	Washington	2,305
5	Illinois	2,610	10	Arizona	2,265

6. Enrichment Field Performance

88.9% LIVE WEBSITES	43.9% VERIFIED EMAILS	29,519 DNS-VERIFIED	100% CLAIMED PROFILES
-------------------------------	---------------------------------	-------------------------------	---------------------------------

ENRICHED FIELD		DATA SOURCE	
Phone Number	66,472	98.9%	Google Maps listing direct extract
Google Rating	65,620	97.6%	Google Maps listing direct extract
Website URL	60,668	90.2%	Google Maps listing direct extract
Accepted Insurance	51,880	77.2%	Website content keyword scan
Facebook Profile	36,778	54.7%	Website HTML regex extraction
Verified Business Email	29,519	43.9%	Website crawl + DNS MX verification
Instagram Profile	27,110	40.3%	Website HTML regex extraction
Online Booking	15,098	22.5%	5 booking-platform signals detected
Emergency Care	8,212	12.2%	Website content keyword scan
LinkedIn Profile	6,140	9.1%	Website HTML regex extraction
Telehealth	2,329	3.5%	Website content keyword scan

7. Production Operations

1,528 TOTAL QUERIES	150 BATCH SIZE	44 AVG LEADS/QUERY	1.5% DEAD WEBSITE RATE
-------------------------------	--------------------------	------------------------------	----------------------------------

The scraper targets approximately 400 major US municipalities spanning all 50 states and DC, drawn from a curated city list embedded in the Python script. Each of the four niches is searched per city, producing 1,528 total queries (not every city+niche combination returns results). Google Maps scrolls through 15 sidebar feed iterations per query. A JSON checkpoint saves after every 150-query batch, enabling crash recovery without data loss.

8. Business Impact

Sales Enablement

- 42,200 premium leads (score 90+) with verified email and phone — outbound-ready
- Territory segmentation across 49 states / 2,740 cities for localized campaigns
- Treatment intelligence: know if a practice offers Invisalign, Botox, spinal adjustment, or pet surgery before dialing
- Insurance acceptance flagged on 77.2% of records — match payer networks immediately

Marketing Intelligence

- Social media handles: 54.7% Facebook, 40.3% Instagram — multi-channel campaigns
- 84.1% rated 4.5 stars or above — target high-reputation practices for partnerships
- 22.5% with online booking — identify tech-forward practices for SaaS/platform sales
- 100% claimed Google Business Profiles — every lead actively manages its online presence

9. Technical Stack

COMPONENT	TECHNOLOGY	ROLE
Browser Automation	Playwright (Chromium, headless)	Google Maps search, feed scroll, and listing extraction
HTTP Client	Python requests + urllib3	Website crawling with broad TLS compatibility
Email Verification	dnspython (MX record lookup)	DNS-level domain verification on every candidate email
Data Pipeline	CSV append + JSON checkpoint	Append-only output with crash-recovery checkpointing
Deduplication	In-memory set + file-based URL registry	Cross-query dedup by Maps URL and website domain
Error Isolation	Python logging + exception chains	Per-query and per-listing error containment
Output Format	CSV, 27 columns, 37.4 MB	Universal CRM, Excel, and database import

10. Deliverable

The final output — **US_Healthcare_Premium_Leads_clean.csv** (37.4 MB, 67,230 records) — contains 27 standardized, enriched fields per row. Every record includes a direct Google Maps URL for independent listing verification. The dataset requires no additional cleaning before import into any CRM, marketing automation platform, or analytics environment.

67,230 TOTAL RECORDS	27 FIELDS PER RECORD	37.4 MB DATASET SIZE	4 CORE NICHEs
--------------------------------	--------------------------------	--------------------------------	-------------------------