DATA Analyst Assignment Pawzz

Q1: The Data Pipeline: From Chaos to Clean

1. The Problem, As It Actually Hits Us
The initial data input is messy and unstandardized, arriving from various sources:

- An Excel sheet with 30 vets scraped from Google Maps, featuring wildly **inconsistent phone number formats** ("9823045678", "+91 98230 45678", "9823-045-678").
- A WhatsApp message with **raw, unstructured data** (a blurry photo and an unverified location).
- A Google Form submission with **vague, free-text entries** like "near the temple" and "good people" for a phone number.

This chaotic input necessitates a robust, automated, and human-checked system to clean and standardize the data before it can be used.2. Step 1: Three Doors, One House

Handling Data Ingress with Source-Specific Tools

Different data sources require tailored input mechanisms, but all must lead to a single central management system (Airtable).

- **Door 1: The Intern Door (Airtable)**
    - **Purpose:** Enforces *pre-cleaning* and *data validation* on structured input.
    - **Mechanism:** Interns use an Airtable form with built-in validation (e.g., cell turns red if phone number is invalid, dropdowns for consistent category selection).
    - **Required Fields:** Business name, Category dropdown, Phone field with format validation, Address field, Source, Attachments.
- **Door 2: The WhatsApp Door (Twilio + Make + Python)**
    - **Purpose:** Captures raw, unstructured data from the "wild west" of user communication.
    - **Mechanism:** Twilio receives messages. Voice notes are transcribed, and photos are run through **OCR (Python)**. A **regex bomb** extracts potential phone numbers, emails, and names.
    - **Storage:** Data is dumped into a "Raw_WhatsApp" table for daily human review.
- **Door 3: The Community Form Door (Google Forms → Airtable)**
    - **Purpose:** Keeps submission simple to maximize user adoption.
    - **Mechanism:** Simple Google Form feeds directly to Airtable as "Raw_Community." Minimal required fields (Business name, Category, Best contact).

Step 2: The Address Problem

Automated Geocoding for Standardization and Duplicate Detection

Human-inputted addresses are unreliable, leading to duplicate entries for the same location. The fix is to automate address standardization:

- **Process:** Every listing runs through the **Google Geocoding API** within an hour of entry.
- **Data Stored:**
  - `raw_address` (user input)
  - `formatted_address` (Google's standardized result)
  - `latitude`, `longitude` (for mapping and matching)
  - `geocoding_status`
- **Duplicate Check:** Listings with coordinates within 50 metres and similar names are flagged as potential duplicates.
- **Failure Handling:** If geocoding fails, the listing is flagged "Address_Needs_Review" for manual intern fixing on Google Maps.

Step 3: Required Fields & Proof Standards (Per Category)

Implementing Guardrails for Quality Control

Proof standards vary by category (e.g., a Vet Clinic needs a License Number; a Feeder does not).

- **Mechanism:** A separate "Category_Standards" table defines rules for each category.
- **Guardrail:** If a listing does not meet the category's **Required Fields**, it is blocked from entering the pipeline and remains in a "Draft" status.

Category	Required Fields	Proof Needed for Verification
Vet Clinic	Name, Phone, Address, License Number	Photo of signboard OR website with address OR phone call confirming license
Shelter/NGO	Name, Phone, Address, Registration Number (if registered)	Registration document OR two independent online mentions OR referral from known partner

Feeder	Name, Feeding Location, Phone	Photo of feeding activity OR referral from a rescuer
...and so on		

Step 4: The Pipeline Stages

The Listing Lifecycle and Status Management

Every listing moves through a defined set of statuses, with automated and human-based checks at each stage.

1. **Submitted:** Initial entry.
    - *Automated Actions:* Normalise phone, geocode address, check for obvious duplicates.
    - *Exit:* Moves to "Needs Fix" (if issues) or "Ready for Cleaning" (if passes).
2. **Needs Fix:** The intern's queue for manual correction of errors (e.g., "Phone invalid," "Address geocoding failed"). If fixed, returns to "Submitted." If unfixable, moves to "Rejected."
3. **Ready for Verification:** Automated checks passed; waiting for human verifier.
4. **Verification In Progress:** A verifier applies a checklist (call, check online, verify license).
    - *Exit:* Moves to "Verified," "Needs Fix" (if issues found), or "Pending" (if unable to reach).
5. **Verified:** Listing is confirmed real. Ready for daily batch publication.
6. **Published:** Live on the directory. Can move back to "Needs Fix" if a user reports an issue.

Step 5: The Dashboard

Metrics for Team Lead Management

A Power BI dashboard provides critical oversight metrics to reduce founder dependency and manage operations.

- **The Funnel:** Tracks Incoming (by source) and the **Conversion Rate** to "Verified." A narrow "Verification" stage indicates a bottleneck.
- **Quality Metrics:** Tracks Verification Pass Rate, Rejection Rate by Category, and Duplicate Rate (spikes indicate detection is too loose).
- **Backlog:** Monitors counts and age of items in "Needs Fix" and "Verification" to ensure nothing is forgotten.
- **Source Performance:** Identifies which source (Intern, WhatsApp, Form) produces the highest verification rate to optimize input methods.

Step 6: Handling the Really Messy Stuff

Technical Solutions for Unstructured Data

Specific scripts and tools address the most common, complex data issues:

- **WhatsApp Voice Notes:** Use Make.com to trigger transcription API, extract text, and create an Airtable record with the transcription and attached audio file.
- **Screenshots: OCR** is used, but the original image is stored for human checking in case of errors.
- **Inconsistent Phone Formats:** A Python script runs hourly to:
    - Strip all non-digits.
    - Add +91 for 10-digit numbers.
    - Flag numbers over 12 digits for review.
- **Inconsistent Addresses:** The system maintains an **"Address Alias" table**. When an intern manually fixes a vague address (e.g., "near the temple"), the alias is stored to suggest the fix for future submissions of the same vague term.

Q2: Red Flags, Verification & Duplicate Prevention

1. The Problem: People Lie (Or Just Get It Wrong)
The Need for Rigorous Verification

Unorganised businesses and potential scams necessitate a system to flag suspicious listings before publication.2. Top Red Flags (What Makes Us Suspicious)

Automated and Manual Warning Signs

1. **Phone Number Smells:** Less than 10 digits, more than 12 digits, same number in 3+ listings (aggregator/scam), or use of a known disposable number.
2. **Address Smells:** Geocoding fails completely, geocoding returns a residential building for a commercial category, or address contains only landmarks.
3. **Online Presence Smells:** No online footprint for a commercial business, a new social media account with few posts, template website, or a Google Maps listing with only one suspicious review.
4. **Name Smells:** Generic name ("Animal Clinic"), misspellings, or name matches a chain but phone does not.
5. **Proof Smells:** Attached photo is a **stock image** (checked by reverse image search), photo is of a different location, or missing a typical photo for that category.
6. **Behavioural Smells:** Submitting 10+ listings in 5 minutes (bot/scraper), using a temporary email, or multiple listings from the same IP with slight variations.

3. Verification Checklist: Vet Clinic Example

The Human Verifier's Process

A verifier follows a detailed five-step process to confirm a listing's authenticity:

1. **Step 1: Call**
    - Dial the number, confirm address and hours.
    - If no answer, try 3 times at different times before moving to **"Pending."**
2. **Step 2: Check Online**
    - Verify the pin/reviews on Google Maps.
    - Check for online presence (website/social media); cross-check the address.
3. **Step 3: Verify License**
    - Check a state's public registry if available.
    - If no registry, ask for the license number during the call (for internal verification only).
    - **Note:** If unconfirmed, the listing can still be published with a "License not confirmed" note.
4. **Step 4: Cross-Check with Proof**
    - Ensure any attached photos (e.g., signboard) match the name and address. Use Google Street View if needed.
5. **Step 5: Decision**
    - **Verified:** If all checks pass.
    - **Verified with note:** If minor issues (e.g., phone works, address confirmed, but no website).
    - **Reject:** If major issues (e.g., phone disconnected, address doesn't exist).
    - **Pending:** If unclear (e.g., phone rings but no answer after 3 attempts).

4. Handling Conflicts: Two Different Numbers for Same Business

Duplicate Prevention and Resolution

The nightly duplicate script flags listings with high **Name Similarity (> 90%)** and close **Coordinates (< 100m)** but different phone numbers.

1. **Step 1: Detection:** Nightly script flags "Potential duplicate with phone conflict."
2. **Step 2: Investigation:** Verifier calls both numbers to determine which is primary, old/disconnected, or if the business has two lines.
3. **Step 3: Resolution:**
    - **Merge:** If they are the same clinic, records are merged. The working number is kept as primary, the other as secondary.
    - **Reject:** If one number is confirmed as "wrong number."
    - **Audit Trail:** The merged record is logged with the merge date, verifier, original

entries, and conflict resolution details.

5. Audit Trail: Who Did What, When, and Why

Ensuring Accountability and Compliance

Every change to a critical field (verification status, phone, address) is logged in a separate "Audit" table.

- **Tracking:** Logs capture Field Changed, Old Value, New Value, Reason, and Proof (e.g., "Automated script," "Screenshot_of_call_log.png").
- **Purpose:** Allows tracing back actions if a user complains about a fake listing, ensuring compliance and distinguishing between mistake and malice.

6. Escalation Rules: When to Say Yes, No, or Maybe

Publishing with Honesty and Transparency

- **Reject Immediately:** Phone disconnected after 3 attempts, address geocodes to a vacant lot, business confirmed closed, or malicious intent is confirmed.
- **Re-Check Later:** No answer after 3 calls (**Pending**), or License not found (**License Pending**).
- **Publish with Warnings (Badges):** Provides honesty to the user about the listing's verification level.
    - **"Limited Info"** (Works but no online presence).
    - **"Community-Sourced"** (Submitted but not yet fully verified).
    - **"Incomplete Profile"** (Missing optional fields).
    - **"Last Verified [date]"** (For old listings).

The End-to-End Flow (Putting It All Together)

This example traces a single listing, "Dr. Mehta Veterinary Clinic," from a raw WhatsApp message to becoming a "Phone-Verified" listing on the directory.

Timeframe	Action	Status	Notes
Day 1, 10:32 AM	User sends WhatsApp with name and phone, but no address (only a photo).		
Day 1, 10:33 AM	Twilio/OCR extracts name and phone. Creates Airtable record.	Submitted	Name: Dr. Mehta Veterinary Clinic; Address: null.
Day 1, 10:35 AM	Automated script runs. Phone normalised. Geocoding fails (no address).	Needs Fix	Tagged: "Missing address."
Day 2, 9:15 AM	Intern manually finds address on Google Maps using the photo's context.	Ready for Verification	Address updated to "Shop 5, Shanti Nagar, Malad West."
Day 2, 11:30 AM	Verifier calls, confirms details, checks Google Maps (matches). No website found.	Verified	Note added: "Phone verified, address confirmed. No online presence."
Day 3, 2:00 AM	Nightly publish script runs.	Published	Goes live with **Badge: "Phone-Verified."**
Two weeks later	Another user submits a different phone for "Dr. Mehta."		Duplicate detector flags it. Verifier investigates, finds the old number disconnected, merges records, and updates the audit log.