Process of thinking

1. First thing, I am taking time looking at the data and noting things down. Finding columns with unique identifiers, columns that contain critical information etc and columns that do not offer much information.
2. Making sure I am aware of how the correct data SHOULD look like and note any cleaning steps needed to get there. For example a column with names has numbers in it, or a column with numbers contains values that are strings, etc.
3. Cleaning the data
4. Finally, join the data on the columns that I believe are both unique identifiers for each file but are also common for all files.
5. Run analysis on the result.

I am taking each file individually and start analyzing it.

1. Website_dataset.csv

Website Dataset Overview:

- **`domain`**: The website's domain name (this will serve as a unique identifier).
- **`domain_suffix`** and **`tld`**: These are related to the top-level domain (e.g., `.com`, `.ca`), which might not be critical for analysis.
- **`language`**: The language associated with the website.
- **`company_name`**: The official name of the company (some values are missing or marked as "Unknown").
- **`city`**: The city where the company is located.
- **`country`**: The country of the business.
- **`region`**: The state or region where the business is located.
- **`phone`**: Contact phone numbers, some of which are formatted inconsistently.
- **`site_name`**: The name of the site (often overlaps with the `company_name`).
- **`category`**: The category of the business (e.g., `Real Estate Developers`).

Observations:

- **`domain`** will be important as a unique identifier for merging with other datasets.
- **`category`, `company_name`, and `phone`** are important for analysis and should be cleaned.
- **Missing Values**: Several columns, such as `company_name`, `region`, and `phone`, have missing values.
- **Phone Numbers**: These may need to be reformatted for consistency.
- **Country and Region**: These should be standardized for consistency with other datasets.

Step 2: Identify Data Issues and Cleaning Steps

Issues to address:

1. **Phone Numbers**: The phone numbers should be cleaned by removing any non-numeric characters.
2. **Country and Region**: Standardize `country` and `region` by converting to lowercase.
3. **Missing Values**: Handle missing values in `company_name`, `category`, and `region`, and fill them with "Unknown" if necessary.
4. **Language and Domain Suffix**: These columns may not be as critical for merging and analysis, but they should be checked for consistency.
5. **Category**: Make sure the `category` column is clean and contains meaningful values.

Step 3: Clean the Data

I will now proceed with cleaning the **Website dataset**, addressing the issues identified above.

Website Dataset - Cleaning Summary:

1. **Phone Numbers**: Cleaned by removing all non-numeric characters.
2. **Missing Values**: Filled missing values in `company_name`, `region`, `category`, and `phone` with `"Unknown"`.
3. **Country and Region**: Standardized by converting names to lowercase and removing extra spaces.
4. **Categories**: Cleaned for consistency.

2. Google_dataset.csv

Google Dataset Overview:

The **Google dataset** has the following columns:

- **`address`**: The business address.
- **`category`**: The category the business belongs to.
- **`city`**: The city where the business is located.
- **`country_code`** and **`country_name`**: The country of the business.
- **`name`**: The name of the business or company.
- **`phone`** and **`raw_phone`**: The contact phone numbers, some of which are formatted differently.
- **`region_code`** and **`region_name`**: The state or region where the business is located.
- **`domain`**: The business website domain (common key for merging).
- **`zip_code`**: The postal code.

Observations:

- **Phone Numbers**: The `phone` and `raw_phone` columns contain both formatted and raw phone numbers, which need to be standardized.
- **Missing Values**: Some businesses are missing values in columns like `country_name`, `region_name`, and `phone`.
- **`category`**: Categories seem to be properly populated, but the format should be standardized.
- **`country_name`** and `region_name`** will need to be standardized to lowercase for consistency.

Cleaning Steps:

1. **Handle Missing Values**: Fill missing values in important columns.

2. **Phone Number Formatting**: Standardize phone numbers across both `phone` and `raw_phone`.
3. **Standardize Country and Region Names**: Convert all country and region names to lowercase.
4. **Domain Consistency**: Ensure domain consistency for future merging.

Google Dataset - Cleaning Summary:

1. **Phone Numbers**: Standardized phone numbers by removing spaces, plus signs, and dashes.
2. **Missing Values**: Filled missing values in important columns (`name`, `category`, `address`) with `"Unknown"`.
3. **Country and Region**: Standardized country and region names to lowercase.
4. **Domain Consistency**: Ensured that the domain is consistent for future merging.

3. Facebook_dataset.csv

Facebook Dataset Overview:

The **Facebook dataset** has the following columns:

- **`domain`**: The website domain.
- **`address`**: The physical address of the business.
- **`categories`**: The categories the business belongs to.
- **`city`**: The city where the business is located.
- **`country_code`** and **`country_name`**: The country of the business.
- **`description`**: A brief description of the business.
- **`email`**: The email address of the business.
- **`link`**: A link to the business website.
- **`name`**: The name of the business or company.
- **`page_type`**: The type of page (likely indicates the kind of business).
- **`phone`** and **`phone_country_code`**: Contact phone numbers.
- **`region_code`** and **`region_name`**: The state/region where the business is located.

- **`zip_code`**: The postal code.

Observations:

- The **`categories`** column contains multiple categories, which might need splitting.
- The **`phone`** numbers have scientific notation, so they need to be reformatted.
- There are **missing values** in columns like `address`, `country_name`, `city`, and `description`.
- The **`page_type`** column indicates the type of business but might need standardization.

Cleaning Steps:

1. **Handle Missing Values**: Fill in or replace missing values where necessary (e.g., with 'Unknown' where appropriate).
2. **Phone Number Formatting**: Convert phone numbers to proper string format.
3. **Standardize Country and Region Names**: Convert all country and region names to lowercase for consistency.
4. **Split Categories**: Separate multiple categories where applicable.

Facebook Dataset - Cleaning Summary:

1. **Phone Numbers**: Reformatted phone numbers to remove scientific notation.
2. **Missing Values**: Filled missing values in columns such as `name`, `address`, and `categories` with `"Unknown"`.
3. **Categories**: Split categories into a list where multiple categories exist.
4. **Country and Region**: Standardized the names of countries and regions to lowercase.

Cleaning process

- Website Dataset

Original Issues:

- **Phone Numbers**: The phone numbers had non-numeric characters that needed to be removed.
- **Inconsistent Country and Region Names**: Similar to the other datasets, country and region names were not standardized.
- **Missing Values**: Important columns like `legal_name`, `main_city`, and `s_category` had missing values.

Cleaning Steps:

- **Standardized Country and Region Names**: Converted country and region names to lowercase and stripped any leading or trailing spaces.
- **Phone Number Formatting**: Cleaned phone numbers by removing all non-numeric characters using `str.replace(r'\D', '', regex=True)`, ensuring they contain only digits.
- **Handling Missing Values**: Filled missing values in `legal_name`, `main_city`, and `s_category` with `"Unknown"`.
- **Column Renaming for Consistency**: Renamed columns like `root_domain` to `domain`, `legal_name` to `company_name`, and `s_category` to `category` for consistency across datasets. This makes merging and analysis easier.

Final Output:

Each dataset is now standardized with cleaned columns for:

- **Company names** (`company_name`)
- **Phone numbers** (`phone`)
- **Geographical data** (`country`, `region`, `city`)
- **Categories** (`category`)

- *Google Dataset*

Original Issues:

- **Inconsistent Country and Region Names**: As in the Facebook dataset, country and region names were inconsistent.

- **Phone Numbers**: Phone numbers were inconsistently formatted, and some had extra characters like spaces or dashes.
- **Missing Values**: Some rows were missing important values like `name`, `category`, and `address`.

Cleaning Steps:

- **Standardized Country and Region Names**: Converted country and region names to lowercase and removed extra spaces, as done in the Facebook dataset.
- **Phone Number Formatting**: Cleaned both `phone` and `raw_phone` columns by removing spaces, plus signs, and dashes using `.replace()`. This ensures that all phone numbers follow a consistent format (all digits without separators).
- **Handling Missing Values**: Filled missing values in critical columns like `name`, `category`, and `address` with `"Unknown"` to ensure completeness.

- *Facebook Dataset*

Original Issues:

- **Inconsistent Country and Region Names**: Country and region names were not standardized, leading to potential inconsistencies.
- **Phone Numbers**: Some phone numbers were in scientific notation or improperly formatted.
- **Missing Values**: Important columns like `name`, `address`, and `categories` had missing values.

Cleaning Steps:

- **Standardized Country and Region Names**: Converted all country and region names to lowercase using `.str.lower()` and removed extra spaces with `.str.strip()` to ensure consistency.
- **Phone Numbers**: Reformatted phone numbers by converting them to strings and removing scientific notation (for values stored as floats). This was done with a lambda function: `str(int(x))` for valid numbers.
- **Handling Missing Values**: Filled missing values in important columns like `name`, `address`, and `categories` with the placeholder `"Unknown"`.

Step-by-Step Approach to Join the Three Datasets

1. Identify the Join Key and Common Columns

- **Join Key**: I will use `domain` as the primary key to join the datasets because it's the common identifier across all three datasets and represents the website or business.
- **Common Columns**:
    - **Category**: Present in all three datasets but might have different values.
    - **Address**: I will combine **country**, **region**, **city**, and **address** columns.
    - **Phone**: Present in all datasets but with different formatting or missing values.
    - **Company Names**: The business name (could differ slightly across datasets, so we'll prioritize one source over another if conflicts arise).

---

2. Inspect Each Dataset for Conflicts and Similarities

Before merging, I need to understand how data conflicts might arise across the three datasets:

Conflicts and Resolution Strategy:

- **Company Names**:
    - **Conflicts**: The same company might have slightly different names in each dataset (e.g., abbreviations, punctuation).
    - **Resolution**: I can prioritize data from one dataset (e.g., Facebook) but fall back to others if it's missing.
- **Category**:
    - **Conflicts**: Each dataset might categorize the business differently.
    - **Resolution**: Combine all available categories (e.g., concatenating multiple values into a list).
- **Address (Country, Region, City)**:
    - **Conflicts**: Different datasets might have incomplete or varying levels of detail.
    - **Resolution**: Prioritize one dataset but fill missing values with data from the others.
- **Phone**:

- **Conflicts**: Different formatting across datasets (e.g., country codes, spaces, or missing numbers).
- **Resolution**: Clean and prioritize one dataset (e.g., Google) but fill in missing values from others.

---

3. Cleaning and Preparing the Data for Joining

I will need to clean and standardize the data to ensure consistency during the join. I have already done a lot of cleaning earlier (standardizing country names, handling missing values, etc.), but I will review and apply final steps where needed.

4. Merging the Datasets

We'll perform a step-by-step merge:

- First, merge **Facebook** and **Google**.
- Then, merge the result with **Website**.
- As I merge, I will carefully resolve conflicts based on this strategy.

---

5. Create a Fourth Dataset

After merging, I will select the relevant columns and ensure the final dataset contains the most accurate information from the three sources.

Steps to implement the above process:

Step 1: Load the Cleaned Datasets
Step 2: Review and Prepare the Columns
Making sure that columns like `domain`, `category`, `country`, `region`, `phone`, `company_name`, etc., are correctly named in all datasets. If needed, rename them to align.
To make sure I ensure consistency across datasets:
List and Compare Columns from each dataset to see which ones need to be renamed for consistency.
Rename Columns so they match across the datasets.
Check Column Types to ensure they are the same (e.g., `str` for `domain`, `str` for `category`).

Reformat Columns where needed to make sure data types and formats align (e.g., ensure country names are lowercase across datasets).

Step 3: Merge the Datasets
Step 4: Resolve Data Conflicts and Prioritize Data
I will resolve conflicts by prioritizing one dataset over another for key columns. For example, I will prioritize facebook_dataset.csv for company name, I will prioritize google_dataset.csv for full address, I will prioritize website_dataset.csv for domain and once again facebook for phone numbers.
Step 5: Review the Merged Dataset
Step 6: Save the Final Merged Dataset

Step-by-Step Implementation for Creating `full_address`:

Step 1: Create `full_address` Columns in Each Dataset

I will concatenate `city`, `region`, and `country` into a new `full_address` field for each dataset.

This creates a `full_address` column in each dataset that combines `city`, `region`, and `country`.

Step 2: Merge the Datasets and Prioritize the Full Address

Next, I merge the datasets and prioritize the `full_address` field from Google first, then Facebook, and finally Website.

Step 3: Review and Clean the Merged `full_address` Column

After merging, I review the `full_address` column to make sure everything looks good:

Step 4: Save the Final Merged Dataset with `full_address`

Once I am satisfied with the merged `full_address` column, I will save the dataset:

Explanation:

- `full_address_fb`: Concatenates `city`, `region`, and `country` from the **Facebook dataset**.
- `full_address_google`: Concatenates `city`, `region`, and `country` from the **Google dataset**.

- `full_address_website`: Concatenates `city`, `region`, and `country` from the **Website dataset**.
- **Final** `full_address`: The merged `full_address` field takes data from **Google** first, then **Facebook**, and finally **Website**.

Final Output:

- The `full_address` column will now provide a complete address by combining `city`, `region`, and `country` from all three datasets.
- Missing or incomplete addresses from one dataset will be filled with information from the others.

We have three datasets: **Facebook**, **Google**, and **Website**. Our goal is to merge these datasets to create a **fourth dataset** that contains key business information with improved accuracy. The key columns of interest are:

- `company_name`
- `category`

- `phone`
- `full_address` (which we will create by combining city, region, and country)

Key Points to Keep in Mind:

1. **Join Column**: We'll use `domain` as the common identifier across the datasets to join them.
2. **Conflict Resolution**: We need to resolve conflicts when data is available in more than one dataset by prioritizing one dataset over another based on the perceived reliability of the data.
3. **Combining Data**: For some columns like `categories`, we will combine the values from all datasets instead of choosing just one.
4. **Creating Full Address**: We'll create a new `full_address` column that combines `city`, `region`, and `country` from all datasets.

Step-by-Step Process

Step 1: Load the Datasets

First, we load the three datasets (Facebook, Google, and Website).

Step 2: Create the `full_address` Column

In each dataset, we will combine city, region, and country to create a `full_address` column. This will give us a comprehensive address for each business.

Step 3: Join the Datasets on the domain Column

Next, we merge the datasets using domain as the key. We will perform an outer join to ensure we don't lose any data from any of the datasets.

Step 4: Resolve Data Conflicts

a. Prioritize `company_name`

- Why prioritize Website first?: The company name listed on the business's official website is likely to be the most formal and official version.
- Google second: Google business listings are generally verified, which makes them a reliable backup source.

- Facebook last: Facebook might have user-friendly or informal versions of the company name, so we use it as a fallback.

b. Combine category Values

- Why combine categories?: Businesses may be listed under different categories in different datasets. By combining all available categories, we ensure that we capture a complete set of possible categories for each business.

c. Prioritize phone

- Why prioritize Google first?: Facebook is likely to have direct business phone numbers managed by the business itself.
- Facebook second: Facebook is reliable and often verified but serves as a backup.
- Website last: Website phone numbers may be less frequently updated but are useful as a fallback.

d. Use the full_address

- Why prioritize Google first?: Google's address may be more user-managed and recent.
- Facebook second: Facebook often has reliable, verified addresses for businesses.
- Website last: Website addresses can serve as a fallback when other data is missing.

Step 5: Clean and Finalize the Merged Dataset

We now select the key columns of interest for the final dataset.

Step 6: Save the Final Merged Dataset

Finally, we save the merged dataset to a CSV file.

Explanation of Key Decisions

1. **Company Name:**
    - Facebook was prioritized because it likely contains the most formal, official version of the company name.
    - Google is a reliable secondary source due to its verified business listings.

    - Website is used as a fallback since it may contain more user-friendly or informal versions of the company name.
2. **Phone:**
    - Google was chosen first because businesses typically manage their phone numbers directly on their pages, making them accurate and up-to-date.
    - Facebook was used as a secondary source because of its verified listings.
    - Website was used as a fallback for older or less frequently updated phone numbers.
3. **Category:**
    - We chose to combine all categories because businesses may be listed under multiple categories across the datasets, and combining them gives us a more comprehensive view of the business.
4. **Full Address:**
    - Google was prioritized first for address because it's likely user-managed and recent.
    - Facebook is the second source due to its reliable listings.
    - Website is used as a fallback.

This approach ensures that I get the best data possible by carefully prioritizing the sources based on the likely reliability of the information.

Output:

- **The final dataset contains:**
    - **domain:** The unique identifier for each business.
    - **company_name:** The prioritized company name, starting with the website.
    - **category_combined:** The combined categories from all sources.
    - **phone:** The prioritized phone number, starting with Facebook.
    - **full_address:** The combined address from city, region, and country, prioritized by Facebook, then Google, and Website.

Additional Cleaning Steps:

1. Recheck for duplicates: Ensure no duplicates remain based on key columns like **`domain`**.
2. Standardize text fields: Clean up text formatting (extra spaces, case consistency).

3. Handle missing or "Unknown" values: Decide whether to fill or remove missing information.
4. Ensure correct data types: Make sure phone numbers and numerical fields are in the correct format.
5. Remove rows with missing critical information: Ensure no rows have `"No Data Available"` in both key fields **(**`domain`**,** `company_name`**).**
6. Remove any duplicates
7. Save and review the final cleaned dataset.