**Document Outlining ETL Approach, Data Cleaning Strategy, and Joining Logic**

**ETL Approach Overview**

The ETL process consists of the following key steps:

* **Extract:**

- Data is extracted from multiple Excel files, each containing information about agents, transactions, and other related data.

- The primary data files are:

- `agents.xlsx`: Contains agent details such as name, contact information, and office details.

- `agents\_2.xlsx`: Contains additional agent-related details.

- `all\_agents.xlsx`: Contains a comprehensive list of agents with further details.

- `transactions.xlsx`: Contains transaction records, including sale price, property type, and agent information.

* **Transform:**

*Data Cleaning:* We clean the data by standardizing column names, handling missing values, and normalizing textual data (e.g., email, names, addresses).

Standardization: Columns such as emails and names are standardized to ensure consistency across data sources.

Missing Data: Missing values are handled by filling with relevant values from other data sources or applying default values.

Address Parsing: Addresses are split into standardized components (e.g., street, city, state, zip) to ensure consistency in address formatting using the library USAdress.

Name Splitting: Full names are split into first, middle, and last names using a custom function (`split\_name`).

Data Merging: Data from different sources (agents, agents\_2, all\_agents, transactions) are merged using common keys (e.g., `email`, `agent\_id`, etc.).

* **Load:**

- The final, cleaned, and merged data is loaded into multiple output Excel files:

Team Performance: Summarizes the performance metrics at the team level.

Agent Performance: Summarizes the performance metrics at the agent level.

Normalized Brokerage: Contains normalized brokerage data to facilitate comparison across brokerages.

**Data Cleaning Strategy**

Name Standardization: Agent names are standardized to title case (e.g., “ashutosh singh” becomes “Ashutosh Singh”).

Email Normalization: Emails are converted to lowercase to ensure consistent matching.

Office Name Standardization: Office names are converted to uppercase for consistency.

Address Cleaning: Inconsistent or incomplete address data is cleaned by concatenating address components and filtering out invalid data (e.g., "nan" values).

**Joining Logic**

- Data is merged using a left join or inner join, based on the availability of the email address or agent identifier.

- Where necessary, data from different sources (e.g., `agents\_2.xlsx`, `all\_agents.xlsx`) is merged on common columns like `email`, and missing values are filled from alternative sources.

- If the agent's email address is missing, a fallback matching mechanism (using agent names and office names) is implemented.

- For merging with transactions, a *temporary key* is generated to facilitate joining when direct matches are unavailable.

---

**Python Scripts to Demonstrate Key Parts of the ETL and Data Normalization Process**

Below is the Python code that demonstrates the core parts of the ETL pipeline, including data extraction, transformation, and normalization.

* **Data Extraction**:

- The data is read from Excel files using `pd.read\_excel()`.

* **Data Transformation**:

- The `clean\_dataframe` function standardizes column values (e.g., titles the names, converts emails to lowercase).

- The `get\_address` function parses addresses into separate components (e.g., street, city, state, zip). We are using usaddress library here.

Merging: The `pd.merge()` function merges different datasets on the `email` column.

* **Data Normalization**:

- The `MinMaxScaler` from `sklearn.preprocessing` is used to normalize numeric data, such as sales, to a 0-1 scale.

* **Performance Calculation**:

- The performance metrics are calculated by grouping the data by `team\_name` or `fullname` and aggregating various statistics (e.g., total sales, average sale price).

* **Loading**:

- The final cleaned and aggregated data is saved to Excel using `pd.ExcelWriter()`.

* **Final Output**:

The final output consists of multiple Excel files:

Team Performance: Aggregated team metrics such as total sales and average sale price.

Agent Performance: Aggregated agent-specific metrics.

Normalized Brokerage Data: Normalized performance metrics for brokerages.