**Objective**

Deliver a maintainable scraper that periodically extracts **listing cards** and **society/building pages** from **99acres, Magicbricks, NoBroker** and outputs a structured **Excel file**.  
**Scope now:** data capture + clean Excel.

**Inputs (I will provide)**

* A list of **search URLs/filters** per site and.
* Desired **run cadence** (weekly or fortnightly).
* A short **column wish-list** (you should still capture all publicly available fields).
* Any access details if a source requires login (only if applicable).

**Data to Capture (minimum required fields)**

For each **listing card** (sale/rent):

* source (99acres/Magicbricks/NoBroker)
* page\_url (listing URL)
* site\_listing\_id (as shown/embedded by the site)
* title
* description\_short (trimmed)
* price\_or\_rent (numeric)
* area\_value + area\_unit (e.g., 250, sqft)
* unit\_type (e.g., 1RK/1BHK)
* furnishing (Raw/Semi/Fully, if shown)
* society\_or\_project
* locality + city
* floor\_no / total\_floors (if shown)
* posted\_date and/or last\_updated\_on\_site (as displayed)
* seller\_type (owner/broker/builder, if visible)
* photos\_count (if shown)
* amenities\_listed (comma-separated, if shown)
* captured\_timestamp (UTC+05:30)
* run\_id (unique per execution)

For each **society/project page**:

* source
* page\_url
* society\_name
* address/locality + city
* builder\_name (if shown)
* towers / total\_units (if shown)
* typical\_unit\_sizes (text)
* amenities (comma-separated)
* avg\_price\_or\_rent\_displayed (as shown)
* coordinates (if visible)
* captured\_timestamp, run\_id

Requirement: **Capture as much public information as is available**; if a field is missing on a page, leave it blank without failing the run.

**Output (Excel only, no database)**

* **File name:** scrape\_{site\_codes}\_{YYYYMMDD\_HHMM}\_run-{run\_id}.xlsx
* **Sheets:**
  + listings\_raw — one row per listing card captured
  + societies\_raw — one row per society/project page captured
* **Column order** should be consistent across runs.
* All numeric values (prices, areas) must be **numeric types** in Excel (not text).

**Run Mode & Frequency**

* Must run **locally on Windows** with a **single click** (e.g., a simple runner).
* I will run it **weekly or fortnightly**.
* Produce a **log file** per run in a /logs folder with total pages visited, rows captured per site, warnings, and run duration.

**Politeness & Reliability (non-negotiable)**

* **Polite/slow** crawling (low load).
* Handle **JS-heavy** pages where required.
* **Graceful failure**: if any source blocks mid-run, still produce Excel for completed parts and log what failed.
* If any site fundamentally disallows automation, flag it and pause that site (don’t brute-force).

**Acceptance Criteria**

* Excel from each of the **three sites** validated against the field list above.
* Column names and data types match the delivered **column dictionary** in the README.
* Logs show page counts, row counts, and zero fatal errors.

**Deliverables**

* Source code and a minimal **config file per site** (just URLs/limits/labels—no tech details needed).
* **README** (setup, how to run, sheet/column dictionary, file naming, where logs are stored).

**Timeline & Milestones**

* Day : Confirm URLs/filters + N pages per site; finalize columns.
* Day : 99acres → sample Excel.
* Day : Magicbricks → sample Excel.
* Day : NoBroker → sample Excel.
* Day : Polish, README, handover.

**Constraints & Compliance**

* Use **only public pages**; respect site ToS and crawl politeness.
* If blocked or if major changes occur, note in logs and inform me; we’ll decide whether to pause or adjust.