

Data Cleaning Documentation – Apartments.com

Sample Data

For this phase, our sample data is generated directly by running `apartment_finder.py` on Apartments.com search results page. The script outputs a CSV file with one row per unit or single-property listing.

Example rows:

```
Street,City,State,Zipcode,Unit,Rent,SqFt,Bedrooms,Bathrooms,Available
"123 Main St","Portland","ME","04101","1A",2100,650,1,1,1 (multi-unit)
"456 Oak St","Portland","ME","04102","","2500,1200,3,1,1 (single-unit)
```

- **Multi-Unit Properties** (e.g., apartment complexes) produce multiple rows with different unit values
- **Single-Unit Properties** (e.g., townhouses or houses) produce a single row with an empty unit field

Data Cleaning Documentation

We applied basic cleaning steps to the scraper to make sure the data is consistent and useful

- **Address Normalization:** The raw address is cleaned and split into Street, Zipcode, City and State. This makes sure each component can be indexed and filtered into MainePad Finder.
- **Multi-Unit vs Single-Unit:** We validated Multi-Unit and Single-Unit separately:
 - If the page contains a multi-unit pricing view, it is treated as a multi-unit property. Each row is parsed into a separate CSV row.
 - If no multi-unit pricing view is found, the listing is treated as a single-unit property and written as one row with an empty unit value
- **Availability:** Availability is mapped to a binary field. Any values such as 'Now' or 'Available' are recorded as a '1', whereas anything else is '0'.
- **Duplicates:** Before inserting, each row is checked if there are duplicates.
- **Conversions:** Rent, square footage, bedrooms and bathrooms are parsed from text; dollar signs, commas and labels are removed so values can be stored as integers.

Validation

Each team member validated their own parsing logic: one focused on multi-unit tables, the other on individual listings, and both cross-checked outputs against the live site.

- Verified that addresses, unit labels, and rent values match Apartment.com
- Confirmed that single-unit listings are recorded once with a blank 'unit' field and multi-

units are expanded into multiple rows

