# Real Estate Crawling Specs

By Yi-Lin Tsai at the University of Delaware

1. I have a list of 86,910 real estates. You will need to search for the corresponding webpage in the two websites: (1) Redfin.com (2) Zillow.com.
2. Datasets and the information to crawl. You can store each dataset as individual CSV or Excel sheets. I will use the keys to merge them later on. You can find an example in the “Crawling\_Example.xlsx.”
   1. Crosswalk: House ID (I will provide you with the yellow columns), Redfin URL, Zillow URL. For now let’s treat the Redfin URL and Zillow URL are the keys to linked to the other datasets.
   2. Redfin datasets
      1. Redfin\_Main: each row represents a real estate (Redfin\_URL), and the corresponding fields that I want to crawl (in green). For the columns in orange, the two numbers are the ones I want to crawl, but I do understand it might not be doable.
      2. Redfin\_PriceHistory: Each row represents a combination of the Redfin\_URL and Event. The information includes Date, Record, Source, and Price.
      3. Redfin\_TaxHistory: Each row represents a combination of the Redfin\_URL and Year.
      4. Redfin\_SimilarHomesNearby: Each row represents a combination of the Redfin\_URL and a similar home. We only need the URL for the “Similar Home Nearby” now and can crawl the details for those similar homes later. Usually for each real estate, Redfin lists up to 6 similar homes nearby.
      5. Redfin\_Schools: Each row represents a combination of the Redfin\_URL and a school.
   3. Zillow datasets
      1. Zillow\_Main: each row represents a real estate (Zillow\_URL), and the corresponding fields that I want to crawl (in green). For the columns in orange, the two numbers are the ones I want to crawl, but I do understand it might not be doable.
      2. Zillow\_PriceHistory: Each row represents a combination of the Zillow\_URL and Event. The information includes Date, Event, Price, and Source.
      3. Zillow\_TaxHistory: Each row represents a combination of the Zillow\_URL and Year.
      4. Zillow\_SimilarHomesNearby: Each row represents a combination of the Zillow\_URL and a similar home (Zillow\_Similar\_URL). We only need the URL for the “Similar Homes Nearby” now and can crawl the details for those similar homes later.
   4. Example (see the attachment)
3. Deliverables include the program(s) you use to crawl the websites and export the dataset. I might need to crawl additional real estates in the future.
4. Timeline: (expected) 1 to 2 months after we agree on the starting time.
5. Weekly update: weekly update by emails or conference calls.
6. Quality check: I will do random check on the accuracy of the data.