Homework-4-1

March 31, 2017

1 Scraping Houses in Redfin

To be completed INDIVIDUALLY and due on April 7 at 3pm.

In this homework, we will practice web scraping. Let's get some basic information for each house/apartment in Massachusetts.

You can see houses around Boston University here

On each house page, scrape the information of this house, including * full address * price * number of beds * number of baths (full or not) * Sq. Ft (square feet) * price per Sq. Ft * property type * County * community * built * Property details (e.g. dishwasher, elevator, parking, heating, air conditioning, maintenance, etc.)

Try to extract as many features as possible. You will use this information in future homework.

Save the data in "house.csv" in the following format:

full address, price, beds, baths, ...

To receive credit, you must commit house.csv to Github. (20 pts)

In [ ]:

Now that you can scrape and extract information of a house, scrape the information of houses in Massachusetts (as of today there are around 10000 houses in MA listed in Redfin). To see all the listings, you can query all the zip codes in Massachusetts. You can see all the zip codes in Massachusetts here. Save the data in "MA_houses.csv", one line for each house: **To receive credit, you must commit MA_houses.csv to Github (20 pts)**

In [ ]:

Note that many websites will block you if you are sending too many queries. To prevent getting block, you should wait between queries (typically 2-5 seconds).

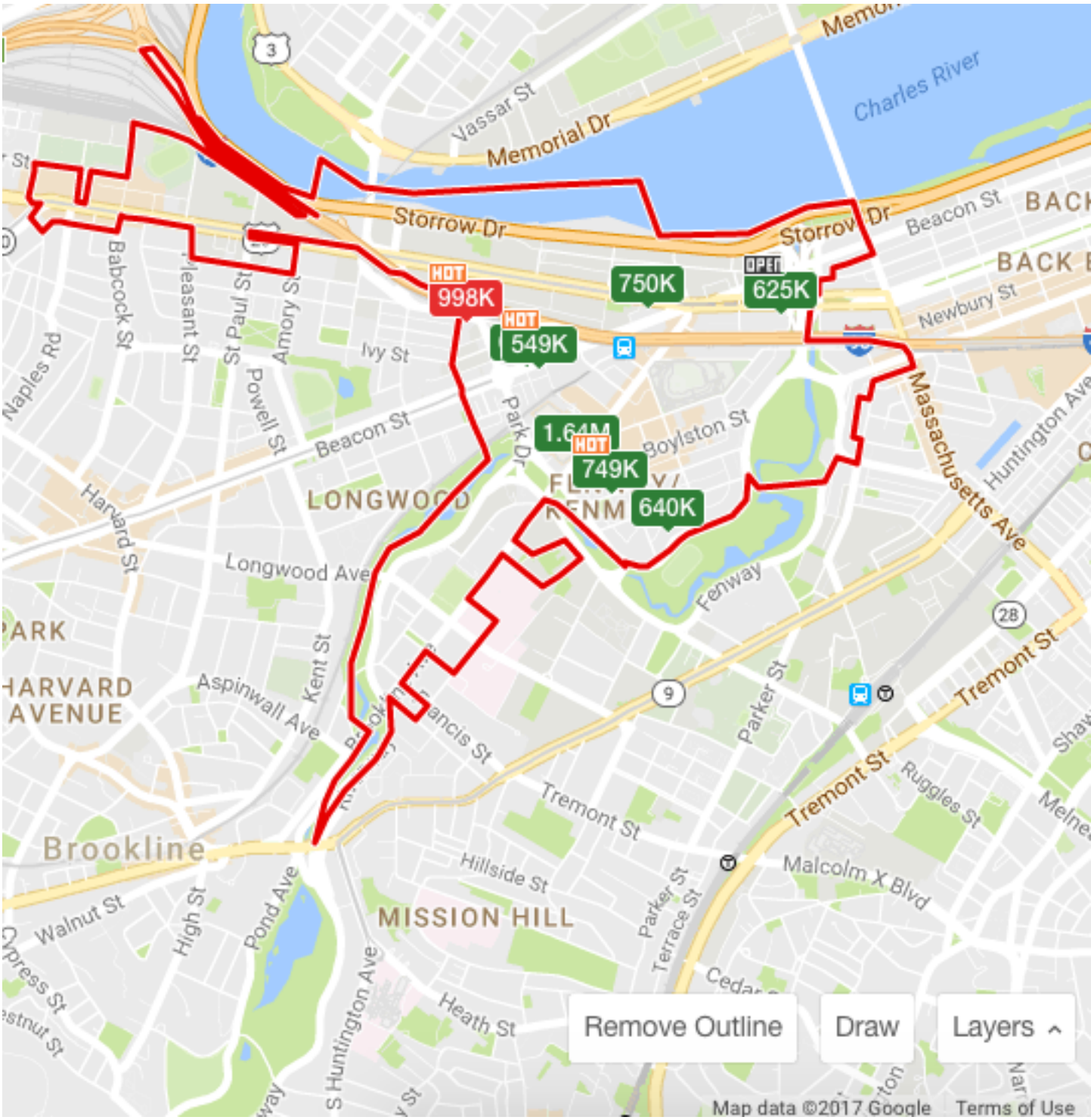

Figure 1: Information to be scraped