Questions:

Once you've completed the coding portion of this assignment, answer the following questions using the following information.

We know we want to keep airbnb accountable by checking if an airbnb does not have a policy number (a reference to the business license San Francisco requires to operate a short-term rental). Every entry in our database has a policy number, is pending a policy number, or is exempt from having one (hotels are exempt from this law). This is because airbnb requires listers to enter this information in a text box before allowing their listing to go live. However, looking through our database, there is a policy number that doesn't look like the other policy numbers. The listing id "16204265" has an unusual policy number. Using images of the exterior of the house posted on airbnb, we can pinpoint which apartment building this rental unit is located in, and check the San Francisco Planning Office to find out if this airbnb does not have a policy number and entered random numbers, or if the lister had a typo. Through this process we found that this lister does NOT have a short-term rental business license! This is an illegal rental unit that is taking a housing unit away from the local population. We can now file a complaint with the planning office to start an investigation!

Note that the "Property Information Map" of the San Francisco Planning Office may not work on eduroam or MWireless.

Turn in your answers to these questions as well as your code.

a. Throughout this project, we acted as investigators to uphold the system of accountability created by the San Francisco lawmakers: listers must register with the city's planning office and put the business license's number on Airbnb's website, Airbnb must display some effort in validating these policy numbers, and third parties can register a complaint of illegal short-term rentals with the city planning office. We used web-scraping to do the latter using several hours of our personal time.

   Imagine you're a software developer at either the San Francisco Planning Office (SFPO) or Airbnb.com. Describe a different system that verifies that the business license is valid for short term rentals in San Francisco and list at least two arguments you might hear at your organization (either SFPO or Airbnb.com) against adopting your system.

One potential solution for the problem of validating short term rental licenses in the San Francisco area, would be for the SFPO to develop a database of approved short term rental licenses for Airbnb to utilize via API when a new listing is posted to determine the legitimacy of the license entered. Airbnb could also navigate the issue of exemptions and pending licenses by requiring these special cases to be verified by either Airbnb or SFPO representatives. One potential argument against this idea is that the database would have to be updated regularly which could delay the process of getting a license and thus negatively affect the rental market causing Airbnb to lose revenue and increasing rental prices for consumers. Another potential issue with this process is that it would require the verification of pending and exempt listings which would require attention and resources from both Airbnb and the SFPO. This also has the potential issue that either Airbnb or SFPO would be accountable for developing the verification system for new Airbnb listings and would have to develop systems to reverify licenses as they expire.

b. The database we've created through web-scraping is a great data source of information for data scientists in order to answer and explore research questions. Skim through the Housing Insecurity in the US Wikipedia page and describe at least one research question that you could answer or explore using this data if you were a data scientist working with a housing activist organization to fight against housing insecurity.

A data scientist could combine this data with data on housing insecurity in the US to research the effect of short term rental license requirements or exemptions on local housing insecurity, safety and affordability rates. They could also determine if there is a correlation between the rate of false short term rental licenses one Airbnb within a given area and the rate of housing insecurity, safety, and affordability rates.

c. As discussed in the introduction, the legality of web scraping is still uncertain in the US. Skim through the Legal Issues section of Web Scraping in the US on Wikipedia and this article about the legal issues with the Computer Fraud and Abuse Act, and describe at least one factor you believe is important to consider when discussing the legality of web scraping and why.

I think that it is important to consider why companies and commercial organizations or websites might not want people to be able to scrape their data and what they could be hiding if their sites are not allowed to be scraped. In the article about legal issues, researchers, computer scientists and journalists were looking to scrape website data in order to test the equity of outcomes when algorithms are being used. If commercial sites are able to hide this data its likely that they will not be held accountable for intentional or unintentional discrimination in algorithms. Thus, the decision regarding legality is critical for protecting beneficial research and journalism and is also important in the ongoing fight for algorithmic transparency and accountability.

d. Scraping public data does not always lead to positive results for society. While web scraping is important for accountability and open access of information, we must also consider issues of privacy as well. Many argue that using someone’s personal data without their consent (even if publicly provided) is unethical. Web scraping requires thoughtful intervention, what are two or more guidelines that must we consider when deciding to use or not to use public data?

It is important to consider the privacy of individuals when we are using public data. Although everyone has access to this information, web scraping includes taking their data to make our own inferences on it. Since they did not consent to have their data being used in that way, we must consider how we are portraying them with our findings and how our conclusions could affect them. It is also important to consider how we are going to display the information that we have collected. Passing off the data as our own would be unethical. One last guideline to consider are social implications of our discoveries. In the first project, we worked with SAT data as well as census information. If we were to publicize that information without any background context, harmful conclusions could be made against marginalized groups in society. When web scraping, we must be thoughtful in our portrayal of groups of people, and how their own information can be used against them.