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.

The main problem I see with the current system is that it's essentially being held accountable by third parties. While having accountability via third parties is good I think it's an unfair burden for the customers to have. My system would first start with having a new division set up from within Airbnb that would be responsible for checking if a rental is illegal. Of course having the accountability division within Airbnb could be problematic, just like how HR ultimately protects the company, this new division may not care about holding listers accountable. That's why this new department would be watched over by both Airbnb *and* the San Francisco planning office. As a software developer me and my team would build a data sharing system that would make it easy for the new division at Airbnb to share all of its relevant listing information with the SFPO. This system would notify the SFPO when a new listing is put in so employees of both Airbnb and SFPO could check if a business license is valid. Furthermore the system would have a built

in reporting feature so that if a business license is invalid employees could swiftly report it to the local government. While I like this system because of how it makes the reporting process more efficient, transparent, and takes the burden away from the customers, I can see why Airbnb may be against it. Creating a new division within Airbnb could be costly and time consuming. Airbnb would have to hire and/or train new employees on how to work in this new division and how to use the newly built data sharing system. Right now it seems like Airbnb is doing just fine with the old system (customers reporting illegal listings) so I don't think they'd be enthusiastic about spending money to switch to a new system. Then of course it would also be expensive to build a data sharing system with features that make reporting/validating listings easy. The other reason against adopting my system is I dont think Airbnb would be keen on sharing its internal data with the SFPO. Remember under my new proposal these two entities would be working side by side to find and report illegal listings. Companies usually keep their data under close watch as this it's useful for determining what they should invest in next. Even though they would be partners with the SFPO I feel Airbnb would be worried about this internal data getting into a competitor's (Vrbo, Expedia, etc) hands.

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.

With the data we gathered on the Mission District I think we could explore whether a certain house type is more likely to be listed on airbnb and thus taken away from home buyers. Let me explain, our data gives us access to how many rooms are in a house, with this information we can see if a listing is a studio or multi-bedroom home. If the houses being listed in the Mission District are mainly multi-bedroom we could begin to hypothesize a home with multiple bedrooms has a higher chance of being used as a rental home. After gathering what type of housing is more likely to be turned into a rental property, me and my colleagues at the housing activist organization could advocate against the construction of these types of homes in highly urban areas. It would be more effective to build apartments/smaller living spaces that fit more people and take up less space than large family homes that just get rented out.

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 one important factor to consider when discussing the legality of web scraping is whether said web scraping crashes the website. As discussed on the wikipedia page, QVC's website was down for two whole days due to them receiving 300 search requests per minute (sometimes up to 36,000 a minute). This resulted in lost sales for QVC and damaged their reputation as a company due to not having a stable site. Don't get me wrong, I'm pro web scraping. It is a powerful tool that can be used to protect customers' interests. I just think web scrapers need to consider the capabilities of the site before web scraping. Sending 300 search requests to google may be no big deal, but for smaller businesses web scraping can wreak havoc on their online store fronts. If one wants to web scrape a smaller business they should ensure they don't send too many requests and stress the site's capabilities.

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?

One guideline I would use is whether this data is something that would cause damage to an individual in the database. This type of data is stuff that most reasonable people would know not to share/web scrape for. Things such as medical records, personal finances, and therapy records. Personally I feel this data should never be web scrapped as this is information most people don't even share with their family members. The other guideline I would use is how much benefit will a web scraper create through their actions. In other words, will this web scraping provide a utilitarian benefit to the public. If your web scraping efforts don't provide a public good I think one must reconsider if it is even worth scraping the information. To me, web scraping should be used to benefit the public and those among us who are not protected by big business interests. For example, in question B we pretended to be a part of an activist organization fighting housing insecurity. This is clearly a utilitarian endeavor and thus would pass the second guideline.