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.

As an Airbnb software developer, I would verify that the business license is valid for short-term rentals in San Francisco by automatically verifying the business license against the San Francisco Planning Office database, ensuring that hosts can only post their rentals if they have a valid short-term rental license for their property. However, an argument against this could be that the San Francisco Planning Office database might not be accessible or readily updated, so this could result in issues for hosts not being able to post their listings even if they have obtained a license. In addition, it could be argued that this will discourage people with exceptions for hosting listings as if the hosts rent for less than 90 days in a year, they are exempt from a license, but this may cause issues for the hosts actually posting the listing.

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.

One research question that I could explore would be finding if a certain demographic of people (i.e. based on race) are disproportionately affected by housing insecurity compared to others. The way I would do this is to first scrape for data regarding the racial concentration of each city in every state. Then I’d use other data sources to find the racial composition of those who identify as facing housing insecurity for each city/area, and then calculate the percentages of those affected by housing insecurity for each race. That way, I could compare the percentages to see if some races experience a significantly higher rate of housing insecurity compared to other races.

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 violating a website’s terms of service not being a crime makes sense, but there are web scraping cases that I feel should be punishable. One factor important when discussing the legality of web scraping is what web scraping is used for. If the web scraping is used for malicious purposes or actions that negatively affect others, such as auction sniping (as seen in the Wikipedia article), this should be deemed illegal as it puts those who are not using web scraping at a disadvantage. In addition, issues of plagiarism should be considered. In the US, Meltwater U.S. Holdings was held liable for screen scraping and republishing Associated Press news information, which I think is a valid case to have legal repercussions for as it allows entities to pass off content as their own.

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 to using public data, especially when modelling social phenomena, is to consider the bias of the dataset. When processing and analyzing this data, it is important to consider who is represented in the dataset, who made the dataset, and how the biases of the creators of the dataset affect the results. Secondly, don’t reveal identifiable personal information when presenting data. Even if this data was from a public dataset, identifiable information should be redacted as it can compromise an individual’s safety, lead to identity fraud, and is an overall invasion of privacy. In addition, verify the validity of this information as incorrect information can lead to incorrect conclusions that could be potentially harmful.