

- 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 system that would verify a valid business license could be to require rentals to submit their license (or exempt status) before they are able to list their property on Airbnb.com, leading to only validated rentals available on their site. This way, every property a viewer sees would be legal for rent and problems relating to illegal short-term rentals would disappear. One argument against this from the rental company's perspective would be that their income could be severely limited if their business license is pending, which is often out of their control, due to the SFPO process which I could imagine taking quite a while. Another argument against this stems from the Airbnb saturation that occurs in this area, since requiring everyone to have a permit is impossible now that the city is cracking down on rental properties in order to regulate affordable options on the market. By requiring these valid licenses in order to list on Airbnb.com, there are only a finite amount of listings possible which excludes rental companies and essentially puts them out of business.

- 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.

Using this data, one research question I could explore would be to try and identify if there is a relationship between the price of a house and the structural conditions listed. In this project, we categorized listings via keywords such as number of bedrooms or private/shared spaces, and seeing the different criteria on the Wikipedia page about different structural inadequacies (such as leaks, cracks, holes, etc.) reminded me of the ability to categorize and filter the results in this way. I think it would be interesting to see if there are specific conditions that affect the price more than others, or if the location plays the largest role in determining housing insecurity for a certain sub-population.

- 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.

One important factor to consider when discussing the legality of web scraping is the type of data that is being scraped. In many instances from both sources, the data was either related to a specific person (name, characteristics, etc.) or copyrighted material from other websites (American Airlines, eBay, etc.) which was problematic for other users to be collecting. Just because this type of information is on a publicly viewable website does not mean that it is free to use and scrape for someone else's benefit. Copyright exists for a reason, so companies have proprietary rights over their technology or data that they worked hard to gain, and personal information used in harmful ways can lead to lengthy court cases or even identity theft. Not all web scraping has malicious intent, but certain types of data scraped are definitely more high-risk than others.

- 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 consider when using public data is to put yourself in the shoes of the people whose data you are collecting. It is important to consider whether or not you would want someone else to use your data, such as search history, personal metrics, or online preferences, so keeping in mind the humanity behind the information is critical to remaining ethical and using the data for a good cause. Another guideline could be to ensure that the data in question is an accurate representation of the public; if we are using public data to generate models and suggestions for the future, it is imperative that the sample from which that data is collected accurately represents the target population, otherwise the results will only be generalizable to a portion of the public. If we want to restructure our society to be more inclusive and accepting then we need to ensure that our modes of information analysis and modeling reflect those ideals as well. We cannot learn from our mistakes and do better if we continue to make them based on historically exclusive methods in the realm of data collection and interpretation.