Emmaline Lehman
November 11, 2022
Project 2, SI 206
Dr. Ericson

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 ies 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 FAQ provided by the Office of Short-Term Rentals (OSTR) demonstrates the legal complexity of San Fransisco's short-term rental regulations. Though compliance with said regulations is the responsibility of Airbnb hosts and citizens of San Fransisco alike, streamlining the system by which the OSTR validates compliance efforts would prove greatly beneficial for both parties. Software developers with Airbnb and the city of San Francisco should work together to create a system that verifies policy numbers for hosts. The system would require the user input to match correctly formatted (valid) policy numbers, or opt to select pending or exemption status. Pending and exempt claimants would be issued a temporary code that would be issued after the OSTR verifies their application as legitimate, but before it is processed in full. Only after the user successfully inputs their policy number would the listing be created.
   An argument against this system is the mere presence of the invalid policy number that was found within the project. Untangling the errors that are already present, may prove to be an enormous burden on the OSTR (even if they are at fault). Another argument against this system would be the vulnerability it creates. There is a great fiscal incentive to manipulate this system and reliance on two cooperating parties to manage this volume of personal and legal data is akin to a target on their backs.

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.

Housing insecurity data is collected in doctors' offices, hospitals, schools, correctional facilities, and shelters, among many more. This data unfortunately does not always get applied in a meaningful way, due to the separation of these databases. Aggregating the data in a centralized

database would allow researchers and activists to make meaningful inferences about the unhoused 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.

Web scraping is a powerful tool used by archivists and researchers alike, but not all web scraping is performed in the name of academic pursuits. Scrapers don't adhere to the privacy policies of the websites they scrape, and they don't ask our permission to use our data. We lose power over our personal information; who it is viewed by, and what it is used for. This data can be sold or used to create biometric profiles, like Clearview AI's facial recognition database, which is built entirely from scraped photographs. The expectation of privacy on public sites is a difficult matter to deliberate, but the inherent harms that may come from web scraping are crystal clear.

d. Scraping public data does not always lead to positive results for society. While web scraping is important for accountability and open access to 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?

- Web scraping should seek to create new value from the data collected as opposed to replicating it outside of legal restrictions.
- Scraped data should also not be used for commercial purposes