Project 2

Members: Isabella Smith, Naren Edara

1. 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.
   a. One method that Airbnb can implement is an automated validation system that works in cooperation with the SFPO that checks if the license number inputted on the Airbnb website matches a legit license number in the SFPO. One possible argument that could be made against this system is the fact that SFPO and Airbnb may not want to cooperate and share their data with each other due to security and privacy concerns. If one company's data gets compromised it could lead to risk for the other company as well which will be seen as a big negative. Another argument that could be heard at the organization is that implementing a system like this would require too much time and resources. Building an entirely new validation system is not easy or cheap, especially when cooperation between two different companies is needed. This could lead to extra costs and resources devoted to keeping data safe and secure.

2. 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. Are illegal renting units one of the driving causes of inadequate housing in big cities in the US?
3. 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.
   a. One factor that is important when discussing the legality of web scraping is using web scraping to gain an unfair advantage over other users of the website. The Wikipedia article discussed a case with eBay in which people were using web scraping to create scripts that would automatically bid for them also known as auction sniping. Acts like this go directly against certain websites' terms of

service and should be seen as illegal because they are directly hurting the website along with other consumers. Another factor that can be important to consider in web scraping is scraping which causes a website to crash due to too many requests. This happened in 2014 with QVC when Resultly crashed the QVC site and led to a loss in sales.

4. 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?
    a. The first guideline that should be followed when deciding to use or not to use public data is to clearly specify the reason the data is being collected and why. The second guideline is to try and only collect nonidentifiable data and try to minimize the amount of personal data that is being collected. By following these two guidelines, we try to consider the privacy of individuals and keep the personal information that is collected to a minimum.