By: Cassandra McDaniel, Rachel Abellera, Cristina Costin

Questions:

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.

Airbnb could cooperate with the city planning office and create a database that contains all valid policy numbers. When a host inputs their policy number, it would automatically be cross-referenced with the database of valid policy numbers. Invalid policy numbers would automatically register a complaint with the city planning office. People might argue against the plausibility of this plan, because Airbnb would have to get the city planning office to cooperate by giving access to their database, which may require extra resources, such as time and money, that Airbnb is unwilling to give up. Another argument against this system would be the possibility of false complaints being registered if the policy number has not yet been uploaded by the city planning office.

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.

According to the housing insecurity Wikipedia page, in almost every state, over 10% of total households are insecure. The one state with under 10%, is Wyoming, which has a 9% of insecure housing over all the households in Wyoming. A data scientist could ask the question "Is there an increase in total insecure households in states with more long term airbnb rentals?".

The data scientist could investigate how many households are currently being used for long term airbnb rentals and the total number of households with housing insecurity across all states. In addition, data scientists could further research the percentage of long term airbnb rental households that could house local residents and families for reasonable prices and see whether there is a correlation to the percentage of housing insecurity in that state.

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.

It is important to consider personal and intellectual property when discussing the ethics of web scraping. Although there may be information that is publicly available and legal to webscrape, it is important to consider what kind of information is being scraped and the ethical issues that come with it. When web scraping is done in large volumes, there is a possibility that this information is coming from small businesses, non-profit organizations, or researchers. If this information was sold to third parties or used for commercial purposes, it could result in the harm of these businesses or organizations. In these scenarios, it is important to consider whether it is actually fair to take this data.

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?

The first guideline that we must consider is how personal this data is considered to the user. For example, some sites may release individuals' names, or phone numbers, which is information that is fairly easy to find for most individuals with a presence online. However, there are databases that include detailed personal information, such as home addresses, schedules, or a list of family members. If these public databases were scraped, that could be a huge violation of privacy to the users. If we do need to scrape these kinds of databases, it is important to consider if we should seek consent from the individuals we are scraping data from, since there may be individuals who feel that this data is too personal to be scraped regardless of whether it is public or not.

Another guideline that we must consider when deciding to use or not to use public data is whether it is absolutely necessary to gather this data based on what our intention is with this data. We must consider the owner of the data and how it will affect them if their data is gathered. For example, if the data is being collected for research purposes and the owner has consented for the data to be collected for this specific purpose, it is most likely safe to continue scraping this public data. However, if it is discovered that scraping and using this public data may negatively impact the owner, it would most likely be a good idea to stop scraping and find another way to go about finding the necessary data.

Another guideline that could be considered is verifying that the data being scraped isn't copyrighted, in order to ensure that we have a fair use to scrape it.