Project 2 Questions

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.

If I was working as a software developer for Airbnb.com, the system I'd create to verify one's business license is by creating a joint-webpage connection in some way to SFPO's validation system. Through this, rather than simply being able to enter a policy number or something of the sort in the textbox, you have to enter *a policy number that has been validated by SFPO*. Maybe this could also be done by having accounts that link to one another, similar to how many websites allow you to sign-in with alternative accounts like Facebook. Here you would use your SFPO account to sign in to Airbnb, and it would check your SFPO account to make sure that your policy number is valid (maybe it doesn't have to override the Airbnb account, as in it only needs to apply for this one step). This would require both entities (Airbnb and SFPO) to coordinate so that there is actual validation between what the user enters and what is true (to show that they either have a real policy number, or that there is actually one pending or exempt).

Some arguments against this would involve the effort that would be required to make this system initially work. The SFPO and Airbnb would have to work together in some way to make sure that the system functions, and there would have to be an entirely new process implemented. This type of coordination takes time and effort, and creating overlapping accounts in some way or form might be difficult.

Another argument that Airbnb might mention is how it puts a lot more effort on the user to go through the process of connecting the accounts and validating them, whilst many of the users have a validate policy number. It would delay the process and might cause users to be frustrated that "they are being punished due to the few people who type in false or invalid policy numbers." This in turn could drive people away from Airbnb and cause them to lose profit, which is something they'd want to avoid. Another related argument is that the new system would be confusing because of the added complexity between linking accounts, where one might not know how to link the accounts and thus more time and resources would need to be devoted towards helping with such a problem.

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.

Looking at the Wikipedia page (as well as other sources), we can see that California has one of the highest levels of housing insecurity of rented households, as well as one of the highest overall levels of housing insecurity in the United States. Using this data, we can explore how house renting, particularly those who commercialize such (those with multiple houses who are then able to rent theirs) like those on Airbnb, limit many from being able to secure a house to rent. We could delve deeper into the problem at hand - that of affordable housing, and see if the strategy trying to manage short-term rentals led to a decrease in housing insecurity, making an emphasis on those who do not have an actual license!

Another question worth exploring would be the pricing of short-term rentals, and by using such data we have from our database, comparing it to the trend of house affordability and price changes. We can also see, particularly that of the Mission District, how the population of houses that are used as short-term rentals for Airbnb and the like affect the availability of housing in the areas (which in turn would harm housing affordability), another aspect of Housing Insecurity.

Through the exploration of these research questions we could most likely create a strong case against short-term housing and the idea of Airbnb as a whole, in which the commercialization of renting creates large amounts of housing insecurity whilst only benefiting those up the upper class (those that can afford to rent/own multiple properties). Changing the rules about short-term rentals and the concept of such as a whole, working with a housing activist organization, would be a great way to fight against this source of housing insecurity.

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 factor that I believe is important to consider when discussing the legality of web scraping is discussing the who of the entities doing the web scraping and who is being protected. In a lot of instances, those who are attempting to conduct research on unethical actions taken by corporations and the like are the ones being targeted and blocked as to violating policy of web scraping. A lot of the instances depend on who is doing the web scraping and for what purpose, though obviously one cannot look at every single one as a case-by-case basis.

That being said, it seems most implementations are in order to protect the large corporations rather than the individual. It's important to note that many of these sites already collect our data and sell it without a care, and the government does the same as well. Not to say that simply because one side already does it, does it make it right, but I think it's important to recognize that *this is already being done whether we want it to or not* and we need to go back and solve this

issue to discuss the idea of web scraping - in which usually the intention behind such policies by the company is not to protect the individuals, but to protect themselves. That being said, there is still also a worry of bad-faith groups then abusing web scraping to harm, that's why the factor of who and intention is important when discussing such. It is a very complicated issue to say the least.

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?

1) One guideline we should consider is anonymity. Can the use or intended use/result of the research we are ultimately trying to obtain from this data be done in such a way that protects the user whose data we are publicly taking from. If we are unable to protect the anonymity of the data in a meaningful capacity so that it is not linked back to said individual, then it should not be used.

2) Another guideline is intent. What is the intent behind our research? Why do we need their data? We should be able to justify the use of their data through the overall good and intent we wish to provide by using their data for research purposes. This can be somewhat viewed in a utilitarian way, in which there is benefit to this data…though this is a complex issue.

3) Scope. To what end are we using this data and how much of it? There are varying degrees of information that is public that people are more accepting of as common knowledge, versus unknowing public information that could be interpreted in some way that the user did not intend to be as such. How much data and from whom are we taking it from.

4) Another guideline that could be considered is consent. Is there a way, if at all, to get the actual consent of the user? We must exhaust this guideline when possible. Simply because someone has their data publicly online does not mean intention was behind such, and informed consent is important in the use of their data.