SI 206 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.

Assuming that the information can be accessed via such, I would recommend implementing a system that relies on the usage of an API, which in turn avoids needing to web scrape all together. APIs allow for the extraction of a specific dataset, which is perfect for obtaining only the information that will determine whether the business license is valid or not for short term rentals. Since APIs can function on an automatic schedule, they can simply be used to get the necessary information efficiently to be then plugged into the algorithm that will check the license number with a stored database of valid numbers to determine whether or not a business license is valid. However, there are some cons to using APIs that may present as conflicts. One concern that may arise are security risks when implementing APIs as it makes the site potentially more vulnerable to these threats. Since the potential of a company experiencing a security breach is only increasing, it may be risky to utilize APIs. Another potential argument to make against the implementation of APIs into this potential system is that, if a data leak does occur and the supplier of the API alters, or removes all together, the API, there will no longer be API access and updates will no longer be coming through for the algorithm to analyze. Thus, APIs are controlled by the provider and the developers have little handle over the mutation of the API, thus impacting the data that has been collected.

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.

By looking at the Housing Insecurity in the US Wikipedia page, it is interesting to note that California (along with New York) has the highest total percentage of households that are considered as "housing insecure" at a rate of 20%. Thus, this dataset is extremely helpful in analyzing research questions as the location of focus is San Francisco, California. Using this data, as well as other data that may need to be collected in order to answer this question, I would pose the following: What are the aspects of housing that make certain properties more expensive than others, and are these prices variable? Assuming all properties have the same listing title, number of bedrooms, location, etc. examining what exactly are the components of properties that raise the price when compared to others is an important question to consider.

These factors may contribute to housing insecurity as some aspects of properties may not be enough to raise the price significantly when compared to others. This can also examine what goes into the process of deciding on a price, and why it is so variable across the board. If there was something more concrete, this could reduce the amount of insecurity that exists if everything was more uniform.

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.

It is truly fascinating the amount of uncertainties that surround the concept of web scraping and its legality, as seen by the history and controversies of its use. I found the article titled “Federal Judge Rules It Is Not a Crime to Violate a Website’s Terms of Service” from the Electronic Frontier Foundation to be intriguing as it brought up a valid point: web scraping is crucial for journalism and research efforts. The article references the case of Sandvig v. Barr, which was the lawsuit brought forth by those wanting to research whether or not discrimination is present in online algorithms. This information is to be obtained through web scraping, however, which is considered to be a violation of certain websites’ terms of service. However, after the ruling, it became more clear that violating a website’s terms of service was not exactly a criminal offense as private entities cannot enact criminal law. Thus, researchers and journalists were able to continue in their efforts to analyze whether or not algorithms have integrated implicit or explicit biases, which alone has been an ongoing topic in the world of technology. By allowing scraping, crucial conclusions (that would not ordinarily be disclosed to the public or noticed without the use of such) can be discovered. Furthermore, as stated in the wikipedia article, the protection of a website’s content should also be put into the context of the amount of harm the access will bring to the site’s system. Yes, in the context of disclosing information about whether or not a site has implemented discriminatory algorithms into its makeup is harmful to the website, but having the ability to discover such information via web scraping should be allowed, as every site should be transparent enough to disclose whether they are using biased systems.

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?

Web scraping, in addition to posing a threat to privacy, also brings up concerns surrounding copyright. If the data is known to be public, then there is not a concern. However, it is important to ensure that all aspects of the data are available to the public. In order to check for copyright, one must manually do so, which is crucial to avoiding legal allegations in the scraping process. Copyright alone in regards to web scraping is also complicated, as there are ways to get around it; by definition, outright duplication of content by the original author or creator is considered to be copyright infringement. Ensuring that the data is public is crucial to any web scraping

procedure as checking for copyright is a serious guideline that must be followed during the process. Furthermore, another guideline that must be considered when engaging in web scraping is the rate at which the scraping is done. If the web scraping is done at an exceedingly high rate that is not controlled (in measuring the requests per second), the website owner may think they are being hacked and will be concerned by the web scraping. Thus, this is an important guideline to follow when scraping public data as the site may become overloaded with requests, potentially causing the site to crash and directing traffic away from the site.