

Question 1

Airbnb.com could implement a system to prevent users from posting a rental if their business license number does not match the required format. This would be similar to how social media platform users cannot create an account if their password does not match the required format. If a user enters a business license number in an incorrect format, the system should not provide any feedback on how to correct the error because doing so would enable the user to falsify a business license number more easily. One potential argument against implementing this system is that some business could be lost, as some users may not want to go through the process of obtaining a legitimate business license just to post a listing on Airbnb. However, this would definitely help ensure that all business *is* legitimate. Another argument against this system is that it is a waste of resources, as a simple Google search could show users what a valid business license number looks like. However, creating this extra step still introduces some friction into the system, and users are more likely to believe they are doing something wrong if they consider falsifying this information.

Question 2

If I were a data scientist working with a housing activist organization to fight against housing insecurity, I could use the data from this project to answer the following research question: if only rentals with valid license numbers are allowed on Airbnb's platform, how would housing insecurity rates change? Would they improve? And if so, by how much? In order to find an answer to this question, I could look at the housing insecurity rate for one specific region – in the context of this project, the housing insecurity rate could be for the San Francisco Bay area in California. I could then look at the percentage of Airbnb rental listings in that area that have an invalid license number, and create a statistical model to predict the new housing insecurity rate for that region after removing the invalid listings.

Question 3

One factor that is important to consider when discussing the legality of web scraping is whether or not the data scraped is being monetized by the web scraper. If the web scraper is making money off of the information they scraped from someone else's website, this should be considered illegal because they do not have the rights to sell that information. However, if they are just giving out information for free, and if this information could have been scraped by anyone else to see for themselves, this should probably be considered legal. If companies do not want certain data to be scraped, they should allocate more resources toward making their website less susceptible to web scraping. Some practices they could employ are regularly changing the HTML structure of their website, requiring human verification (e.g. CAPTCHAs) to access certain parts of their website, and securely hosting sensitive information. Overall, companies should be held responsible for anything that is leaked to the public for free as a result of web scraping.

Question 4

When deciding whether public data should be used, one guideline to be followed is to only scrape data that is not classified as personally identifiable data. Anything that could be used to steal someone's identity, like their name, email, phone number, address, username, IP address, bank account information, credit card information, medical data, or biometric data should not be used. This would help ensure that everyone's personal data is protected. Another guideline would be to consider the scope of the scraping taking place – if scraped data is going to be distributed publicly, there should be more strict restrictions in place. For the purpose of a class project or personal research, on the other hand, there should not be nearly as many restrictions. Of course, for projects that are not supposed to be available to the public, great care should be taken to ensure that no sensitive data is accidentally released or discovered.