Project 2 - Reflection Questions **Emma Peterson**

Turn in your answers to these questions as well as your code.

1. As a software developer at Airbnb, a system that could be used to verify the validity of the business license for short term rentals could be parsing through policy numbers of the listings and flagging all policy numbers that are not in the format “STR-0000XXX”. This would cause a problem for any listing that has a pending or exempt policy number, which might raise some resistance to adopting this system at Airbnb. Airbnb benefits when they have as many rental properties as possible, and therefore would not want restrictive laws or policies that might prevent ineligible rental properties from slipping through the cracks. Among my organization, I might hear the argument that excluding listings with a “Pending” policy number is excluding possible eligible listings, therefore eliminating a sector of possible business. Another possible reason for objection to this system is that it flags properties that are legally not required to have a valid policy number, and will create an extra headache in trying to differentiate between these such properties and properties that are not legally allowed to be Airbnb rentals.

2. Based on the data in the “Housing Insecurity in the US” Wikipedia page, a data scientist could answer the following research question: How do the prices of both a bedroom’s rent and the average to purchase a home correlate with the prevalence of housing insecurity statewide? The maps showing the average price for rent for a one bedroom apartment and average price of a house are nearly identical, showing a trend of housing being expensive no matter what kind of housing you are looking for in a given area. I think it would also be important to note how in these expensive housing states, the data is likely heavily swayed by highly and densely populated areas that are desirable locations to live. This influx of people who want to live in a given area driving up the prices could have a correlation with destabilizing the housing market and causing more insecurity, a phenomenon that, with the data to back it up, would be very important to a housing activist organization. It would give the organization an idea of where to focus their energies, as well as who is suffering the most at the hands of the housing insecurity crisis.

3. One factor I believe is important to consider when discussing the legality of web scraping is information that companies are trying to prevent people from

scraping. In many of the cases outlined in the “Web Scraping in the US” Wikipedia page, it seems like companies are trying to prevent people from quickly and efficiently synthesizing publicly available data that is so muddled in their respective websites that people would never be able to effectively access it without the help of web scraping. For example, in the dispute between American Airlines and FareChase, FareChase was accessing public data, but in a way that saved the consumer money and lost American Airlines profit. I think this should be completely legal, and it is important to be specific when writing legislation regulating the use of web scraping, as there is a big difference between accessing the private servers of a company and simply taking their publicly available data and putting it in a more digestible format for the consumer.

4. Although web scraping utilizes public data, it is also important to consider when the level of planning and thought that goes into web scraping is an invasion of privacy and therefore unethical. There are multiple guidelines you can follow to gauge this, and decide if your web scraping endeavor is ethical or not. I think the most important guideline for this quandary is considering who would be affected by the data that you are scraping, specifically by a mass of people knowing the data. Would anybody be uncomfortable, or feel violated? In the case of scraping people’s public social media accounts, like Facebook and Instagram, even though they consciously posted the content, it might make the person behind the account feel like they are being watched, which, in my opinion, is an unethical and “shady” use of web scraping. However, if the data being scraped belongs to a corporation or large entity, where it is not necessarily one person being affected, I think this is much more ethical. In some cases, the companies feel their privacy was “violated” because their data was used to give better rates for their services to consumers, which I am less inclined to feel sorry for at large corporations with millions of dollars. An example of this is airline companies American and Southwest Airlines, who were involved in lawsuits with companies that aimed to scrape web data to compare airline prices. Another guideline one could use as a baseline when deciding whether or not to use public data is if the amount that you are web scraping is affecting the performance of somebody’s site. It is possible to scrape the web in such high volumes that it will affect the performance of a website, which is not ethical as you are disturbing the function of somebody’s work for your own personal gain. Communicating with the owner of the website if something like this were to

occur, as well as respecting the boundaries and security concerns that they bring up, can help keep your web scraping practices ethical.