Project 2 Proposal: Web scraping and Regression
AirBnB Arts & Crafts Class Experiences

Background
Airbnb Experiences offers a hosting platform for locals to advertise and schedule activities that introduce travelers to their city, craft, cause, or culture. Visitors can use this system to find, schedule, and rate their experiences.

Problem
The floral industry is facing disruption—through Costco, grocery stories, online deliveries, DIY culture, etc. A local flower shop owner wishes to diversify her business to reinvigorate customers' appreciation for the art of flower arranging and share her gift in a new way with those who love to learn crafts. In consideration of the market demand for flower arranging classes and features to include to make her class more popular, I want to analyze what are the most important characteristics of an Airbnb Experience listing, and how do they influence average rating?

Data
Web scraped Airbnb Experience listings that are returned in a search of "Crafts":

Independent Variable	**Type**	**Description**	**Use for Model?**
Location	String/Object	→ turn into zip? Region? Country?	If use, may need to limit to avoid having too many variables
Language(s)	String/Object	→ turn into # of lang?	Y
Duration of experience	Float		Y
Cost	Int		Y
Number of people per experience	Int		Y
~~Number of reviews~~	Int		Can't use if pre-listing (not useful)
Includes -drinks	N/Y (0/1)		Y
Includes -food	N/Y (0/1)		Y
~~Avg past ratings?~~	Float		Business can't influence Existing listings vs. pre-listing Can't use if pre-listing (not useful)
Host - # guests hosted	Int		Y
Number of photos on listing	Int	Sometimes photos, other times galleries	Y
~~Number of guest photos~~	Int		Can't use if pre-listing (not useful)
Number of tags	Int	Not always present	Y
~~How long been listed~~			Business can't influence Can't use if pre-listing (not useful)
Dependent Variable	**Type**	**Description**	**Use for Model?**
Avg (current) rating	Float	Not always present	Y

Kristen Tokunaga

Project 2 Proposal: Web scraping and Regression
AirBnB Arts & Crafts Class Experiences

Known Unknowns/Barriers

- Timing of listings
    - Note: duration of listing may not be meaningful, but time of day/week/month/year that the host offers the experience may influence guest interest
- Frequency of classes
    - Is this experience available all the time? Or limited based on host availability?
- Host's descriptions of the experience may have a significant effect on whether guests sign up for the experience
- Average review ratings are the only available indicator of success/popularity of an experience—number of bookings, revenue, and profit are not publicly available
- Limitations of variables that can be used in this analysis, based on features that can be influenced pre- vs. post-listing (ex. guest photos, number of reviews, average of past ratings)
    - Also important to note that the host may not have direct influence over these variables and so even if these variables had high impact, they would not lead to actionable recommendations for the hosts to increase their reviews
- The Airbnb Experiences website may not display all information consistently:
    - Ex. may omit sections that are N/A, and I will need a way to gracefully deal with those observations/variables
    - Ex. placement and type of "images" or "image galleries" may vary—e.g., as "profile picture" in Section 16 or banner at top of the page
- Handling the infinite scrolls feature on the starter page/link
- Some "links" don't actually have a hyperlink
- Starting list of features identified for analysis = 9… may not have enough to meet the 10 minimum & no wiggle room to substitute if the features are not feasible to scrape, but will search for more if time permits

Potential Resources

- https://www.airbnb.com/help/article/1581/what-are-airbnb-experiences
- https://fortune.com/2017/10/23/airbnb-ceo-experiences-new-york/
- https://towardsdatascience.com/web-scraping-a-simple-way-to-start-scrapy-and-selenium-part-i-10367164c6c0
- https://towardsdatascience.com/web-scraping-a-less-brief-overview-of-scrapy-and-selenium-part-ii-3ad290ce7ba1
- Selenium, Scrapy
- NumPy, Pandas
- Statsmodels, SK Learn
- Possibly other data sources or scraping resources, if time permits/available

MVP

Linear regression model using variables that are "clean" (i.e., consistently represented across all observations in my sample, such as duration and price).
Hopefully have a way to overcome missing avg. reviews or skip over listings that don't have avg. reviews?

Questions for Instructors

1. What is "Project Luther"? Why is it referenced in the Project 2 Intro?
2. Do we have to use Selenium/Scrapy? Or can we use ready-made scrapers that don't require coding (e.g., Google Chrome Extension – Web Scraper, ParseHub)?

Kristen Tokunaga