**Data gathering**

My exploratory data analysis spanned a single month of data for all available US cities from the Inside Airbnb website. The dataset was initially gathered through a iteration step that pinged the website for data for a single month (July 2019). This was limited to a single month to avoid being blocked by the server for too many calls for data.

**Statistical/Hypothetical Question and EDA**

* Do hosts with less response time tend to get worst reviews than high response time hosts?
  + There is minimal correlation between the review score and the host response rate.
  + The model on the right has a very low R^2 and a very high p-value, indicating that host response rate is not a good predictor of the review score rating.
* Do Real Beds cost more or get better reviews than non-real beds?
  + There is correlation between the number of real beds and the price, but this makes intuitive sense as more beds would mean more space, and more real estate would most likely cost more to rent. Also, “other” beds include pullout sofas and couches. So yes, to sleep in a real bed you will most likely pay more money.
* What impacts a location’s cleaning cost?
  + There are several variables that have a high correlation to the cleaning fees associated with an Airbnb. Most make intuitive sense (living quarters, the number of people it accommodates, etc.)
  + The number of beds accounts for roughly 30% of the variability in the cleaning costs.
  + The number of people the location accommodates accounts for roughly 30% of the variability in the cleaning costs.
  + Its pretty clear there is some collinearity between the variables, but it at least shows the number of bathrooms, bedrooms, beds, and whether the location is open year round all play a role in the cleaning fees
* Do locations with higher cleaning fees have better cleaning reviews?
  + This was a challenging question to answer as the cleaning reviews are on a 1-10 scale and the majority of the reviews were at a 10. Even removing the 10 rated scores, cleaning fees appear to take no significant role in how well people rate the room’s cleanliness.
* Do long time hosts tend to get better reviews?
  + There was some correlation (.19) between the number of days a host has been hosting an Airbnb and the ratings received, but there doesn’t appear to be anything significant that would indicate this is a good predictor for the data.
* Does the number of days the venue is open impact the price points?
  + When the availability\_365 variable was reviewed in the histogram, it appeared to be highly volatile. It was challenging to see skewness in the shape of the data as the days open year round were highly spread out.
  + This variable is not a good predictor of the overall price of the location.
  + However, the cleaning fee does have some correlation, even if it only accounts for 4% of the cleaning fee variability.
  + o What do you feel was missed during the analysis?

**What do you feel was missed during the analysis?**

* Amenities were missing. I know if I would’ve taken the time to parse out the data within the amenities field, I would’ve been able to tell a much better story about the overall price. Each one of these categories was in a dictionary, so it would’ve been doable and probably could’ve helped to answer some of my hypothesis questions.
* I also think a more robust multilinear regression model would’ve helped with some of the multicollinearity within the cleaning data. The cleaning fees seemed to be the easiest to predict.
* Finally, I would like a larger sample with data that spans further than just July 2019. However, I didn’t want to lock my access to the website by pinging their server too many times.

**Were there any variables you felt could have helped in the analysis?**

* + I left off property\_type after realizing there were so many categories (Yurt, Apartment, Apartmenthotel, etc.) to analyze. This would’ve been a great category to include, but I think would’ve prompted other hypothesis questions and would’ve required a significant amount more time to analyze.

**Were there any assumptions made you felt were incorrect?**

* + Yes several. I categorized two types of hosts: younger “newer to Airbnb” and older “experienced Airbnb hosts”. Some of my questions seem to imply that older hosts and younger hosts would have different outcomes because the older hosts would have the Airbnb process “figured out”. However, this doesn’t appear to be the case. In fact, older hosts simply have more opportunities for worse reviews, whereas younger hosts have less reviews to sample from.

**What challenges did you face, what did you not fully understand?**

* + There is a bit of a learning curve with Python, which means the hardest part was trying to recreate what we learned from the text with industry standard libraries. I couldn’t figure out how to get a logistic regression to work in the time allotted, because the text’s example was specific to the data within the text.