Data

1. *Inside Airbnb* and Data Schema

The data for this paper originates from *Inside Airbnb*, an independent investigatory project that collects and hosts substantial AirBnB data on more than 100 cities around the world. The data collected by *Inside Airbnb* are web-scraped from the Airbnb website on a roughly monthly basis and can be found here. Although *Inside Airbnb* was originally started by its creators to investigate the effects of AirBnB on affordable housing and exacerbating gentrification, the data are made public for free and open for use.

Since the data on *Inside Airbnb* are completely collected through web scraping the Airbnb website, they contain only and all information that a visitor to Airbnb's site can see. Approximately every month, data are scraped for all listings in a particular city. For each listing, its availabilities for the next 365 days and quoted price per night over the next year are collected. In addition, information particular to the listing such as its location, host, review count, and room type are also collected.

2. Imputing Transactions

One key limitation of utilizing the web-scraped data from *Inside Airbnb* is that it does not reflect actual bookings and transactions made on the platform. For example, it is impossible to discern whether some dates for a unit may be 'blocked out' (unavailable to reserve) due to the unit being occupied, or it being unavailable on the market.

Thus, I resort to imputing the booking from the changes in availabilities across scraping dates. To impute transaction quantity and price, I rely on 2 key assumptions:

1. A unit that becomes unavailable across 2 consecutive scraping periods is due to the fact that it was booked. This assumption should generally be reasonable as hosts generally plan far ahead as to when their listings will be available; a host will not make a listing available if they know that they will be making it unavailable thereafter. This assumption allows me to impute a unit's dates that it was booked and the approximate time period, on the month level, when a booking was made.
2. The price of the booking for a set of dates is the same price as the latest scraping in which it was available on those dates. This may not be completely true: it is likely that hosts may decrease their prices very near the reservation date in order to maintain high occupancy and that this change is not detected by our monthly scraping. However, this limitation should not bias our analysis of changes in the transaction price, since we can reasonably assume that hosts adjust their prices dynamically in the same manner before and after an occupancy tax is introduced.

3. *Booking Date* vs *Date Booked*

Throughout this paper, there is an important subtle distinction between the terms *Booking Date* and *Date Booked*. Specifically, the *Dated Booked* is when the reservation is made, while the *Booking Date* is the check-in date of the reservation. For example, if a reservation for an Airbnb unit was made on April 1st for July 1st to the 7th, its booking dates are July 1st to July 7th, but its date booked is April 1st. Although in the long run the effects of taxation affect both equally, in the short run when a tax is introduced it is better to separate these and investigate them both.

This distinction becomes particularly important in investigating the treatment effect in our natural experiment. The Massachusetts state legislature enacted the legislation for all booking dates after July 1st, 2019, but this came into effect beginning on January 1st, 2019. Hence, only a reservation made on or after January 1st 2019 for a check in on or after July 1st was subject to a tax; if a reservation for after July 1st was booked before January 1st, or if a reservation was made after January 1st but for before July 1st, it was not subject to the tax. Thus, I will conduct 2 separate sets of analyses on the effects of the tax:

1. All reservations with booking dates on or after July 1st. In this scenario, the treatment was ‘administered’ on January 1st.
2. All reservations that were made after January 1st. In this scenario, the treatment was ‘administered’ effectively on July 1st.