DSC 530

Joel McMillin

Term Project

November 20, 2021

My EDA sought to explore the question of what impact Covid-19 had on Airbnb rentals in New York City. My hypothesis was that Airbnb rentals would drop in early 2020, due to the initial stages of the pandemic. I also believed that prices would fluctuate to reflect decreased demand and availability due to travel restrictions and increased safety standards throughout the hospitality industry. In order to explore this topic, I used the variables: room type, price, total number of reviews, the date of the last review, and reviews per month. There could be a change in preference of room type (fewer instances of shared rooms, for example), but this variable accounted for less than 1.3% of total rentals. Hotel rooms were an even smaller proportion, at 0.7% of total rentals. Prices would drop to reflect demand, even as the supply itself would drop due to decreased demand. The total number of reviews would stagnate as fewer rentals were used. Last reviews would also drop in early 2020. Finally, reviews per month, as with the other variables, would drop as fewer rentals were used.

After the vaccine became widely available in the first half of 2021, my hypothesis was that we would then see a slow increase as society returns to ‘normal’. I expected to see all of these variables increase, with the exception of room type, which would potentially reflect more shared rooms, but as previously noted, the variable is in and of itself an extreme outlier. I did not expect what the data showed: Number of reviews, date of last review and reviews per month all increased dramatically starting in late 2020. Not only did these variables reach pre-pandemic levels by early 2021, by September 2021, they all greatly exceeded their pre-pandemic rates. There is a great deal of fluctuation in price, and there always has been, with more properties at lower prices, and those having more (and more recent) reviews than more expensive properties. I chose to keep outliers in my visuals as they paint the picture of extreme variance in these variables. This is to be expected in a location like New York City, which has some of the world’s most expensive real estate. The other side of that coin is that New York City also has a great deal of more affordable options.

While it is easy to look at this data and assume all of these variables are closely related in some causal way, there are other things that should be considered. How much of the increase in occupancy could be attributed to something else not shown in the data? Is the extreme increase due to consumers getting stir-crazy and needing some kind of escape? With all of the global supply chain issues and flight cancelations, could it be that there was some back-log of reviews that customers didn’t write until they had free time during the pandemic? Does the dataset actually represent what my assumptions lean on: that ‘last reviews’ are an accurate reflection of popularity of particular properties?

Working through this project, my greatest challenge was cleaning the data. Converting room types to numbers, replacing null values, deleting extreme outliers ($26,000/night hotel room, $0/night homes, for example), those were all time consuming, but I was pleased with the end result. I didn’t expect to come to any deep conclusions, and I greatly enjoyed this opportunity to put the EDA skills learned in this class to use in a real world example.