ORIE 3120 Project: NYC Airbnb Analysis

BCU Investments NYC Residential Property Prospectus

Table of Contents

Introduction

BCU Investments, a family-owned investment firm, purchases small-scale, income-producing New York City residential properties. Given that they are a small office, they have difficulty privately listing their rental properties, so they use Airbnb to maximize exposure. They have found that they can charge more on Airbnb, making up for the fees they pay for using the service.

It is 2022 in New York City, and a low-supply residential real estate market gives way to high prices. BCU Investments has not finalized a deal since late 2020 and they need a win. Rather than waiting for supply to increase and prices to drop, they are looking to close on a well-priced property, but they are not sure where to look. We are a real estate consulting group, contracted by BCU to answer a few key questions, so they can narrow in on a successful purchase, and subsequent rental.

Dataset Description

We want to utilize the most current data, to get the most accurate picture of the NYC residential real estate market. Thus, we choose to utilize two distinct datasets: one that covers Airbnb listings, and another that covers property sales (both datasets span the City).

The first dataset is from the Airbnb website under the "New York City, New York, United States" header. The dataset contains Airbnb listings along with important features such as neighborhood, borough, comments, review score, price, and many others. Using this dataset, we can understand how different features of a property may impact the bottom line. Specifically, we can understand if people will pay more for certain boroughs, or if larger listings yield a better rating.

The second dataset may be found on the New York City OpenData portal. The NYC citywide rolling calendar sales dataset shows a row for each sale, with features describing sale price, sale date, number of units, square footage, tax class, and more. Using this dataset, we can understand how sale prices, or square footage varied by borough. We can also understand periods of heightened sales across New York City.

Question 1: Which neighborhood will garner the best return on investment?

In order to garner the greatest return on investment, we want to purchase at a low sale price, but rent at a high nightly price. This will allow us to maximize our return per year, as compared to our initial investment. That said, there are over 20,000 properties for sale across over 200 neighborhoods. So we want to begin by narrowing our search down to a single neighborhood.

We want to select a neighborhood whose sale prices are low while offering a high nightly rental price. We begin by determining which neighborhoods garner the highest nightly rental prices. We can see that Jamaica, Fort Wadsworth, and Tribeca are the most expensive neighborhoods to rent (Figure 1). This does not account for the fact that the units in these neighborhoods may be larger, or more expensive for some confounding reason.

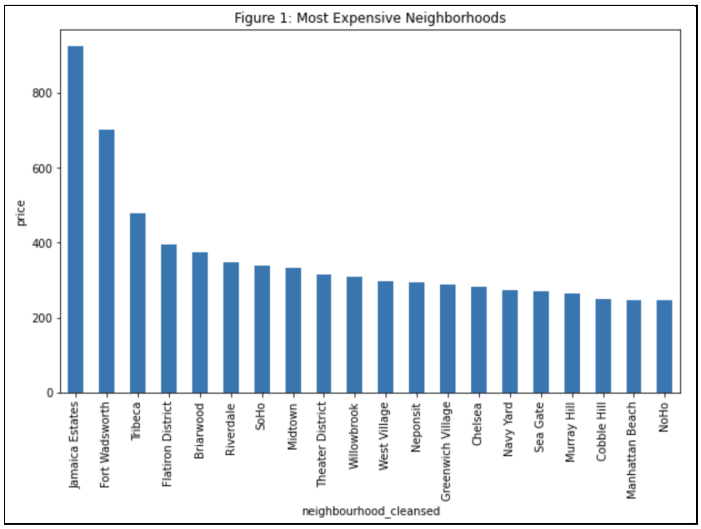

Figure 1: Neighborhoods by nightly rental price

So, we use the Airbnb listing dataset to regress nightly price on features like minimum stay, review score, room type, bed number, bath number, number of reviews, interaction terms, and, most importantly, neighborhood. By collecting the significant coefficients on the binarized neighborhood features, we can determine which neighborhoods contribute most to increased prices, holding other features constant. We find instead that Tribeca, Flatiron District, and Jamaica all contribute to higher prices, holding constant features of a listing (Figure 2). Specifically, as compared to other neighborhoods, units in Tribeca and Flatiron will be $319 and $317 more per night, respectively.

neighborhood_TRIBECA	319.201512
neighborhood_FLATIRON	317.434750
neighborhood_JAMAICA ESTATES	306.806998
neighborhood_MIDTOWN	238.171332
neighborhood_SOHO	211.993228
neighborhood_THEATER	183.252825
neighborhood_WEST VILLAGE	180.626892

Figure 2: Neighborhood coefficients on nightly price

As mentioned above, while we want heightened rental prices per night to increase our rental income, we also aim to minimize the amount BCU spends on their purchase. We want to find the

cheapest neighborhoods. Overall, we can see that the Upper East Side (from 96-110th street), Parkchester, and Belmont are the cheapest neighborhoods by unit sale price (Figure 3).

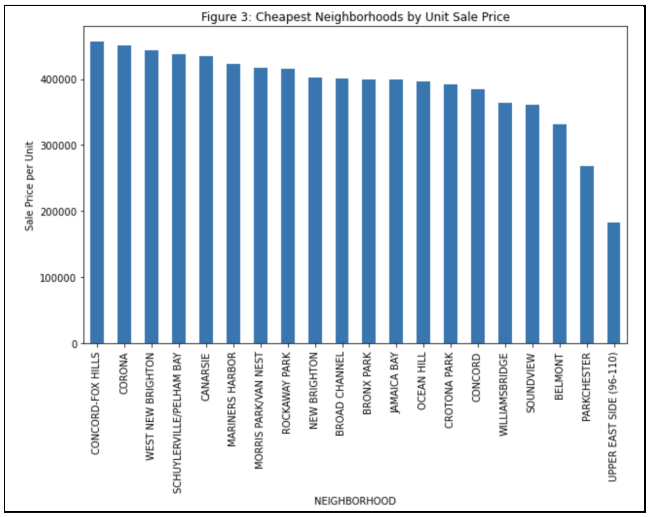

Figure 3: Neighborhoods by unit sale price

Although these three neighborhoods may have the cheapest listings, this could be a result of lower square footage, year built, or the surrounding zoning laws. We want to understand a neighborhood's relationship to the sale price, holding units, square footage, and other variables constant. So, we regress sale price on neighborhood, along with these confounding features and interaction terms, and find the neighborhoods which yield the lowest sale prices. We find with significant confidence that Downtown, Javits Center, and Flatiron all contribute to lower sale prices, holding constant features of a sale (Figure 4). Specifically, as compared to other neighborhoods, units in Downtown and Flatiron properties will sell for $3M and $1.7M less, respectively.

neighborhood_DOWNTOWN-FULTON MALL	-3.163155e+07
neighborhood_JAVITS CENTER	-1.698351e+07
neighborhood_FLATIRON	-1.651043e+07
neighborhood_EAST VILLAGE	7.524639e+06
neighborhood_HUNTS POINT	8.567422e+06
neighborhood_WILLIAMSBURG-NORTH	8.661093e+06
neighborhood_LOWER EAST SIDE	8.770358e+06

Figure 4: Neighborhood coefficients on sale price

We notice in creating an autocorrelation plot, however, that the errors are slightly correlated (Figure 5). Meanwhile, a Q-Q plot shows the data is heavily tailed (Figure 6).

We can attempt to lighten the tails to near a normal distribution while lowering the correlation in errors by squaring the dependent variable, sale price. We can see that, as a result of this manipulation, tails are lightened and error correlation is reduced (Figures 7 & 8).

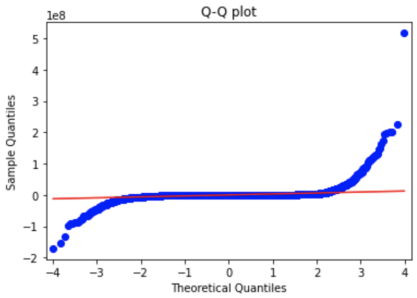

Figure 5: Q-Q Plot

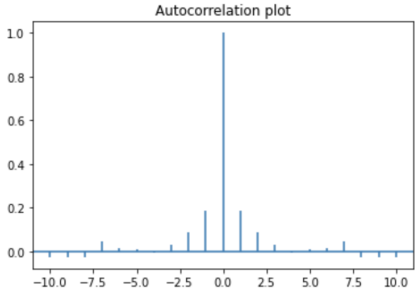

Figure 6: Autocorrelation plot

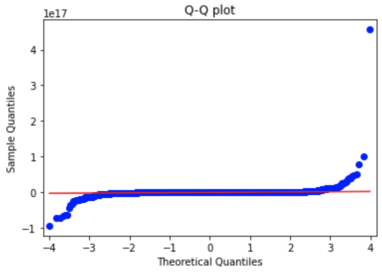

Figure 7: Updated Q-Q Plot

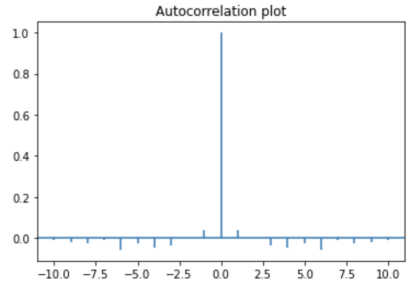

Figure 8: Updated autocorrelation plot

With this update, the Flatiron District still shows a highly negative impact on sale price, and thus, is a good selection in order to decrease BCU's upfront investment cost.

We can see that the Flatiron District, while hosting homes of a low sale price, also calls for higher nightly rental prices. This means that, given a property with certain features, the Flatiron District will be a great neighborhood to increase BCU's return on investment. Of course, there are overpriced properties in this neighborhood. That said, so long as the selection is intelligent, you can garner a lower price in Flatiron while maintaining similar features and a heightened return.

Question 2: What features of a listing are favorable from the tenant perspective?

Having narrowed in on the Flatiron district, we understand that nightly rental price will likely rely on the size of a space and the accompanying amenities. Using our regression on Airbnb nightly price from the above question, we can actually understand which non-neighborhood features have a significant impact on the nightly price (Figure 9).

bedrooms	73.139117
beds	14.823381
minimum_nights	-0.155623
room_type_Private room	-48.084026

Figure 9: Non-neighborhood coefficients on sale price

As seen above, an increased number of bedrooms or beds will increase the nightly price, whereas an increased minimum stay indicates a decreased nightly price. Likewise, a private room will be less expensive than an entire apartment, say.

What is nebulous, however, is what yields good reviews. Obviously, a pleasant landlord, quick response time, or other good behavior will likely lend itself to positive reviews. That said, we want to understand how the features of a listing impact the quality of reviews left on the listing. This is important because it is very difficult for a listing with no reviews to get booked. Moreover, if one of the first few reviews is negative, it is difficult to earn bookings and reverse the sentiment of the reviews. So, optimizing for positive reviews is especially important for an early-stage listing, and will likely allow the landlord to charge even more per night.

In order to understand reviews, we can collect review score data to understand how different features of a listing might increase the likelihood of a good score. A more implicit and important measure of review score, however, is from the review text itself. Airbnb prompts each guest to write a comment after their stay. We may analyze the sentiment of these comments in order to gather a more implicit understanding of a guest's experience.

We use the Natural Language Toolkit library to understand the sentiment of each comment. The sentiment analyzer returns a compound score between -1 and 1 for each comment (where 1 is maximally positive and -1 is maximally negative). Given that most comments are positive, the mean sentiment is around .7. We assign each listing the average sentiment of its reviews. We strive to garner comments whose sentiment is greater than the mean, so for each listing, we convert the sentiment to a binary variable, indicating that the sentiment is either above or below the mean.

We then perform logistic regression on the binary sentiment using the features of the listing such as the number of bedrooms/beds, whether the host is verified, interaction terms, and other interesting features. We collected the resulting significant coefficients (Figure 10). As displayed in the below figure, the number of beds has a negative impact on the review sentiment. So, units with fewer beds are viewed more favorably. The underlying reason for this, most likely, is that as the ratio of beds to bedrooms increases, this means more beds per bedroom–as more people share a bedroom, privacy may become an issue leading to distaste and annoyance for the tenants. Unsurprisingly, private, shared, and hotel rooms garner lower-sentiment reviews than entire apartments.

beds	-0.302222
host_is_superhost_1	0.892529
host_has_profile_pic_1	0.417532
host_identity_verified_1	0.135498
room_type_Hotel room	-1.294812
room_type_Private room	-0.724658
room_type_Shared room	-1.015607
gender_female	0.306903
gender_male	0.227516
gender_mostly_female	0.493471
gender_mostly_male	0.270672

Figure 10: Coefficients on sale price

We also wanted to explore the relationship between inferred gender and review sentiment. Using the Gender Guesser Python library, we associated a gender with each host's name. We found that commonly male and female names may perform better than androgynous or uninterpretable names.

Finally, we found that there is a positive relationship between review sentiment and the host being a superhost, having a profile picture, and verifying their identity. It is important to note that it may be that there is a limited causal relationship between these three features and the ultimate review sentiment. Rather, it is likely that a verified host or superhost will often perform better and garner higher reviews as a result. Given that we do not have access to a confounding variable that captures stay quality, we are comfortable advising that the host add a profile picture, verify, and become a superhost anyway, given that there seems to be no negative effect.

Question 3: What is our risk, given our worst and best case outcomes?

Overall it seems that we will recommend that BCU Investments purchase a property in the Flatiron District. They should proceed to seek out a property with fewer beds, for which they will list the entire apartment. Moreover, we recommend that BCU add a profile picture, verify the host, and register to become a superhost. Given they confine their listing to these criteria, it is likely that they will achieve a heightened return on their investment and garner positive reviews, which is especially important for an early-stage listing.

That said, we cannot guarantee that each review will be positive, nor do we have control over the review scores. To some extent, we can assume that lower reviews give way to lower prices. Of course, it may be that lower-priced listings are worse, and thus garner lower reviews. Although the relationship may not be perfectly causal, regressing nightly price on a listing with randomized review scores may give us a sense of price sensitivity, and thus risk.

We want to understand how the price BCU can charge might change as review quality changes. If price is highly sensitive to review quality, we may be able to identify a separate neighborhood or different specifications to lower our downside exposure.

We simulate a property in the Flatiron district with one bed and a verified superhost. We create random uniform variables of size 100,000 for review sentiment, each individual review score, and host response time. We then use these random variables to simulate 100,000 possible outcomes, depending on our randomized review sentiment, individual review scores, and host response time. In order to simulate the nightly price of each outcome, we predict price using a regression of price against our randomized variables.

We determine that the mean price is what we expect BCU to charge their tenants. Risk, however, may be defined by the standard deviation of the distribution of the nightly prices. This is pretty intuitive: if the distribution of prices is wider, we can expect a higher downside and more volatile pricing.

Different neighborhoods' nightly prices react differently to these randomized features. This is to say that although Flatiron may yield the highest expected return on investment, it may result in a riskier return. We can see the graph of the possible nightly prices in the figure below (Figure 11).

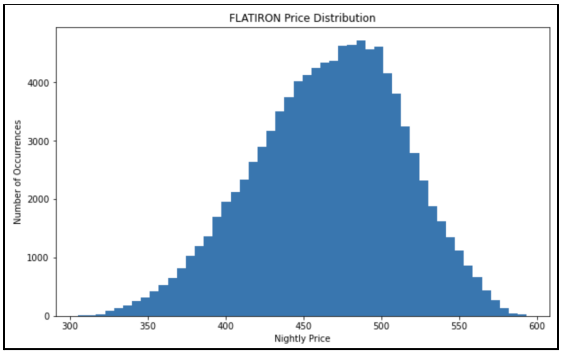

Figure 11: Flatiron scenario-dependent price distribution

Given worst-case sentiment, reviews, and response rate, the minimum possible nightly price is $305 whereas the maximum possible price is $593. So the spread of possible prices between the best case and the worst-case price is $288 per night. The standard deviation of this distribution is just under 48.

Tribeca, another neighborhood with a high, positive impact on nightly price, has a very different price sensitivity, compared to that of the Flatiron District (Figure 12). Given the worst-case sentiment, reviews, and response rate, the minimum possible price is $111 whereas the maximum possible price is $649. So the spread of possible prices between the best case and the worst-case price is $537 per night, much wider than that of the Flatiron district. It follows that the risk is significantly higher than that of the Flatiron district, with a standard deviation greater than 106.

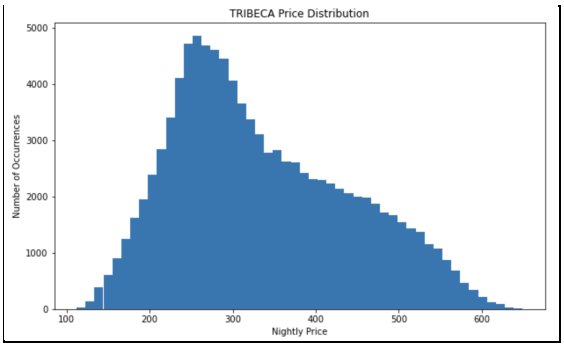

Figure 12: Tribeca scenario-dependent price distribution

Given a lower risk, we can confirm that Flatiron is a good investment from a nightly return perspective. However, in order to understand if BCU should follow through with the deal, we need to understand their risk tolerance. So long as BCU can afford to earn the worst-case prices in Flatiron (around $400), then the deal should pencil and BCU should move forward with the purchase.

Conclusion

We first found the optimal neighborhood. This was done through analyzing with linear regression which apartments are going to have high nightly rent prices and low sale prices. We also tested the assumptions of this regression. This analysis, which oriented us toward the neighborhood with the greatest return on investment, directed us toward Flatiron.

In the second phase, we realized positive reviews and sentiment were important for early-stage listings. We utilized logistic regression on binarized review sentiment to identify features that positively impacted sentiment. For example, we found that if the “room type” was "entire apartment" (rather than "shared room", for example), review sentiment tended to be higher. In addition, keeping fewer beds per bedroom was also beneficial. Identifying these features that positively impacted review sentiment greatly informed us for phase three of our analysis.

In the third and final phase, we considered that most investments are not risk-free. As a result, we found how the guest experience influenced the rent pricing of the property with the specification determined in the second phase (Flatiron, fewer beds, verified host, etc.). Thus, we simulated an abundance of review-related variables in order to understand the standard deviation of the predicted nightly price of a property in Flatiron. Compared to Tribeca, another neighborhood with high nightly price, Flatiron showed a significantly lower price sensitivity.

From identifying the best neighborhood, optimal listing features within the correct neighborhood, and risk factors given our final specification, our end-to-end analysis will greatly help BCU Investments identify profitable properties across NYC.

Appendix

Bibliography

Airbnb. “Get the Data.” *Inside Airbnb*, http://insideairbnb.com/get-the-data/.

City of New York, NYC Open Data. “NYC Open Data.” *NYC Open Data WP Engine*, https://opendata.cityofnewyork.us/.

Follow-up questions

In order to determine the most profitable real estate investment, we may have wanted to analyze year-over-year trends in price. That is to say, if we regress sale price on many features of a property, across multiple years, we may analyze how a neighborhood's effect on property price changes over time. This analysis would allow us to understand neighborhoods with appreciative value. In order to arrive at this information, we would need data across multiple years.

Should we aim to optimize for review number in addition to review quality? We realize that guests may prefer multiple four-star reviews over very few five-star reviews, for example. Should we begin with a lower price or minimum stay in order to invite reviews and incite trust in an early-stage listing? This could be a valuable addition to the analysis.

Is there a way to understand price history? It is hard to disambiguate if low prices are indicative of a worse stay, and thus lower reviews, or if a lower review calls for a lower listing price. If we understood how prices change as reviews are submitted, we would be able to trace better a causal relationship. We could append historical data to help with this.