We were interested in understanding whether as an Airbnb host or a potential investor might be able to ascertain certain, meaningful insights from the Boston market data. Specifically, we were interested in understanding what the data communicate with respect to the Overall Ratings of given listings and Projected Listing Revenues.

Why these two metrics? To a certain extent, these metrics represent a tension between supply and demand. On the demand side, with reviews, users are communicating their relative satisfaction with given listings and stays at those listings in order to collectively guide future travelers (e.g., wisdom of the crowds). Future Airbnb hosts would be wise to learn about the types of features that satisfy or dissatisfy travelers. On the supply side, owners of existing properties and future investors wanting to join the Airbnb marketplace would want to know what types of features are most highly correlated and associated with higher revenues. Ultimately, for Airbnb’s model to sustain there must be an intersection between users’ willingness to pay and hosts’ willingness to provide those listings.

If Airbnb were a perfectly balanced market, containing the entirety of users demands at prices that they would be willing to pay (e.g., balanced supply and demand), one would expect to see a clear and strong alignment between the overall ratings in the Boston market and the projected revenues. Meaning, as

**Approach**

-Number of potential variables at play:

-**Geography:** To treat Boston as a monolithic market would overly simplify the understanding of the market. In any given market there are more desirable areas and less desirable areas. In an effort to better control for that we created dummy variables to understand the given effect that a neighborhood would have on the expected outcomes.

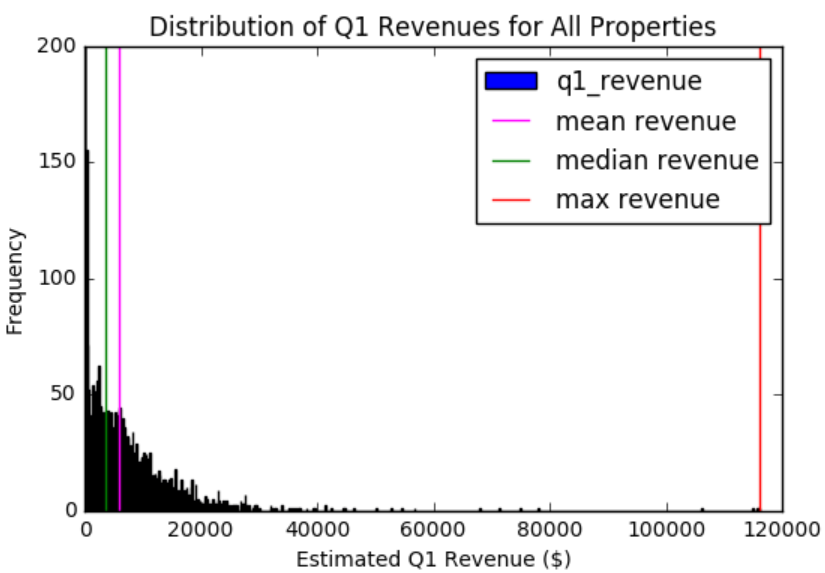
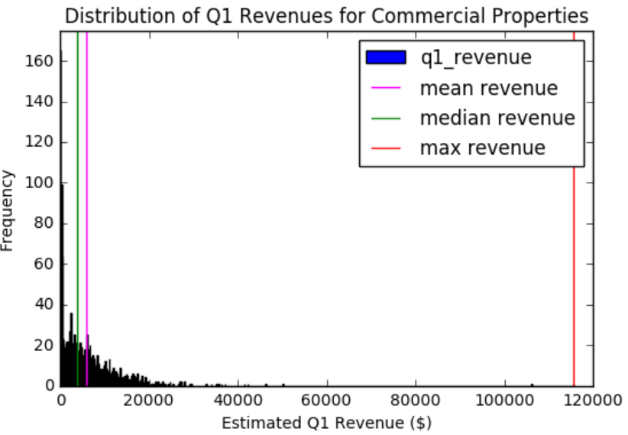
-**Amenities:**  There are 44 unique amenities provided by hosts. These are the “included” features that would be included with the listing and would likely factor into a traveler’s decision to stay at a given property.

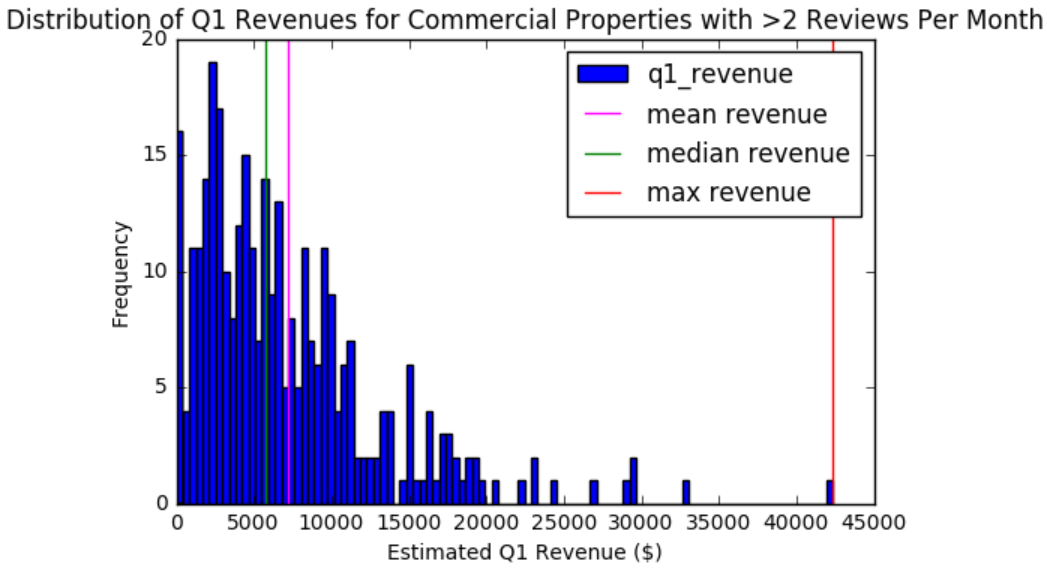
**-Calendar data:** As aforementioned, our data is forward looking and so to project revenues we had to make a key assumption to better ascertain whether the unavailability was driven by someone who had already booked the property or whether the unavailability was simply coincidental.

**Key Assumption:** We decided to segment our hosts and their associated properties into two groups- “commercial” and “hobbyist”. We defined commercial hosts as those who have greater than or equal to 3 listings. The logic behind this decision was to reflect the reality that a couple or individual could conceivably have two properties and split their time between the properties. An additional filter explored in our analysis that is considered for some of our analysis is identifying whether a property has more than 2 reviews per month.

The pivotal assumption in our analysis is that revenues can be calculated by taking the mean of the listing price of each property by day of the week and imputing those values onto the occupied days for the commercial hosts in order to calculate the projected revenue. One could imagine applying a statistical distribution to this data set to control for that based on assumed occupancy rates or other advanced techniques, however, we went for the more straightforward approach in this analysis. We limited the scope of our calendar projections to just the first quarter as 3-months advance booking does not seem to be an unreasonable projection of the overall total revenue for that quarter.

Below are charts indicating the various cuts in the data





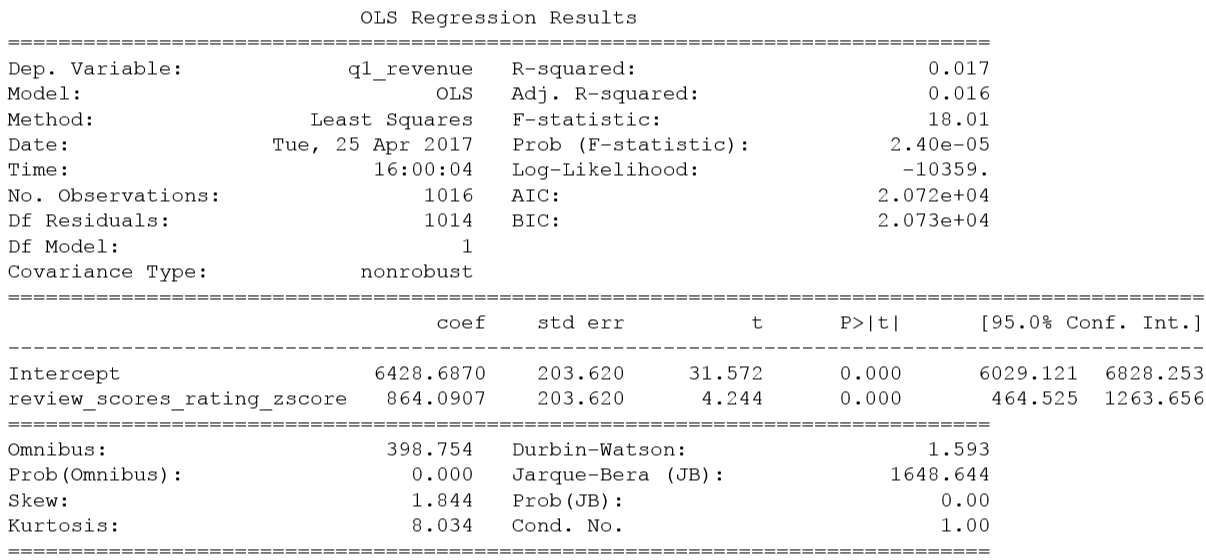
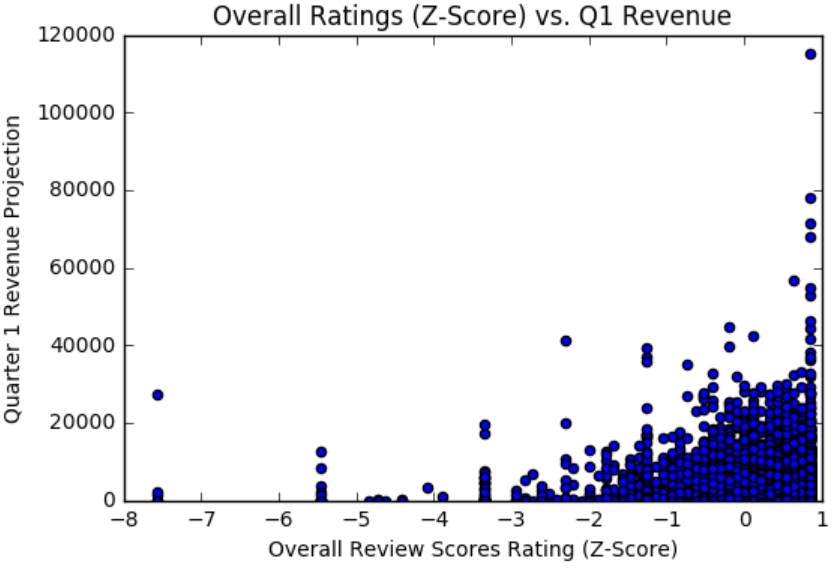
**Data Challenges:**

-**Seasonality:** our data is only a year’s worth of Airbnb data and is forward looking. Meaning, we do not have the data available to go back and retrospectively understand how the output would change through time. Intuitively certain features and areas would be more or less desirable at given times (e.g., proximity to Fenway during baseball season).

-**Presence of highly correlated amenities:**

**Initial Considerations**

Before we began this analysis we wanted to understand the relationship between our two dependent variables that we were exploring. Intuitively we thought there would be a relationship between overall ratings in the Boston market and revenue projections. After all, when travelers make a decision to book a property one would think the relative rating of a listing would shape the future decisions of travelers. Our analysis confirmed that there is a statistically significant relationship, however, it is not very explanatory (only 1.6% of projected quarter 1 revenue can be explained by the normalized scores). It is also interesting to note that there is a non-normal distribution of ratings- meaning that there is an inflationary impact at work.

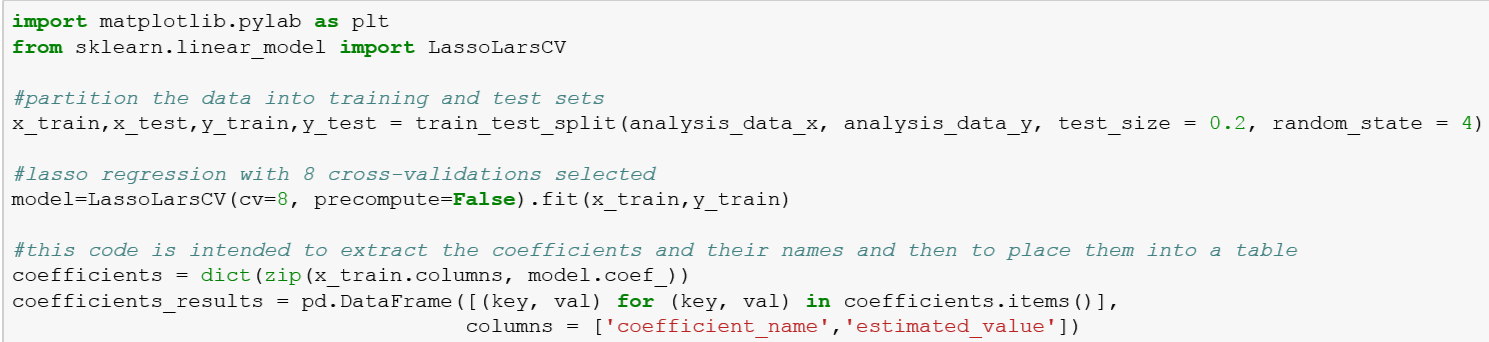


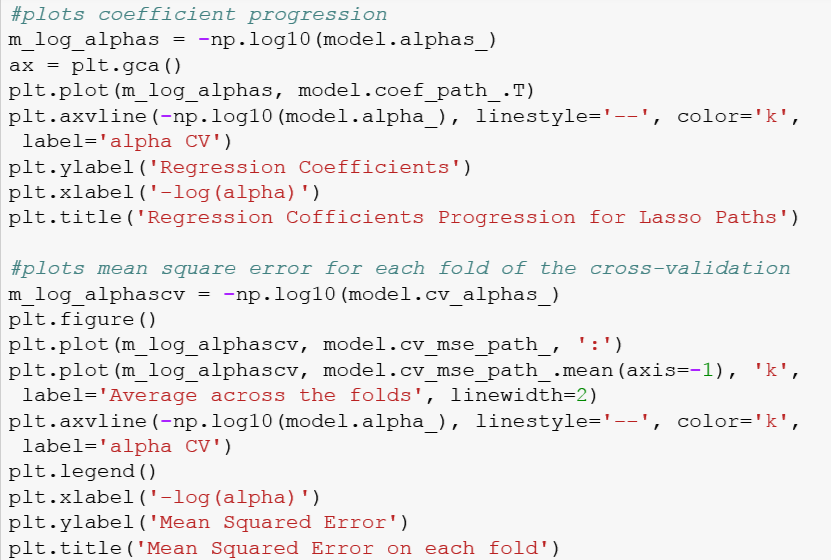
Thus, in light of this initial analysis we decided to take a two-fold approach- first, to understand the factors most correlated with relative ratings in the Boston market and the revenue projections for the first quarter, and, in light of that analysis, arrive at some potential recommendations for Airbnb hosts or potential investors.

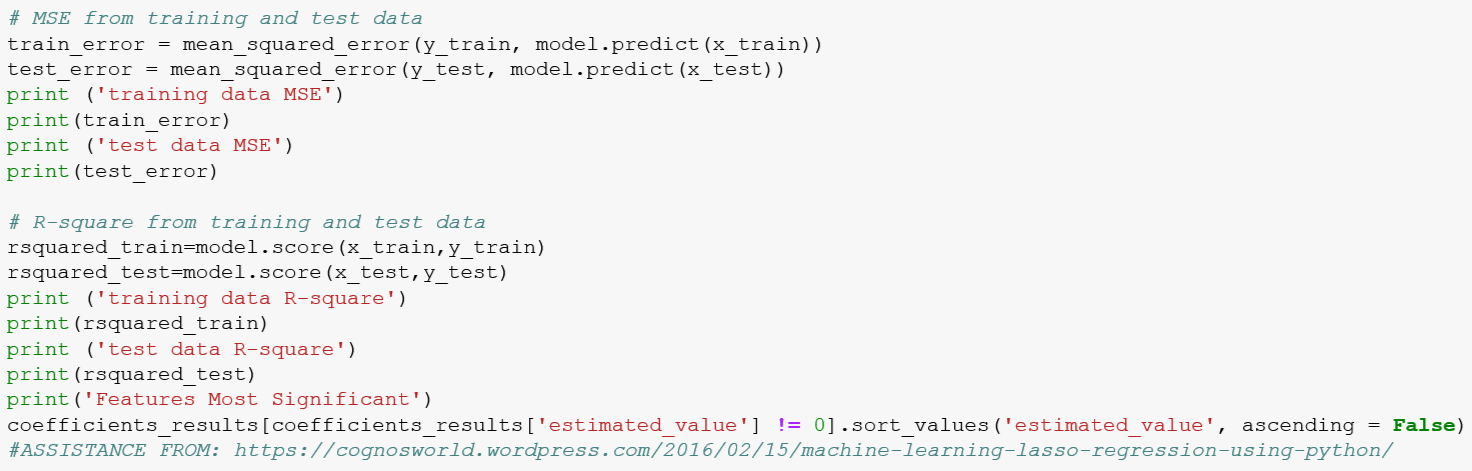
**Approach**

In total there were 99 variables that we were exploring. Given the combinatorial challenges associated with finding the most relevant predictors, we opted for a feature selection regression called the LASSO (Least Absolute Shrinkage and Selection Operator). By constraining the absolute size of the coefficients we are able to arrive at some of the most significant predictors of ratings as well as revenue projections. This LASSO will be applied to each of the questions below.

Detailed code below explains the approach we took and note the citation at the bottom of the code. **NOTE:** the utilization of dummy variables requires n-1 variables and a base case. Our base is an (property type) 'Apartment' with a (amenity) 'Carbon Monoxide Detector' and an (room type) 'Entire home/apt'.

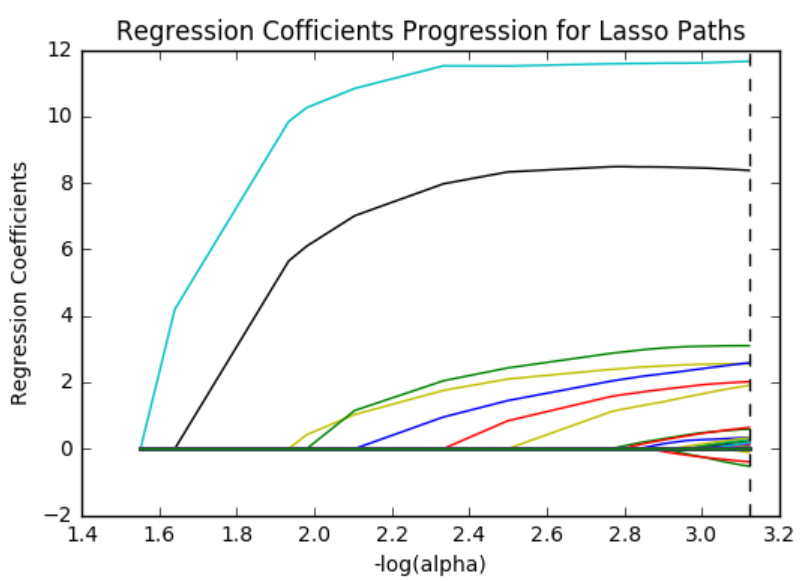
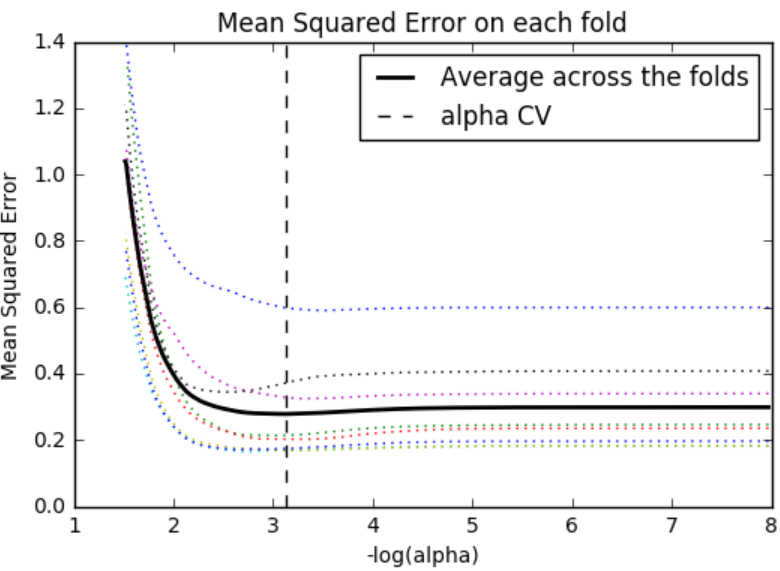






**FACTORS IMPACTING RELATIVE PERCEPTIONS IN RATINGS**

After running the cross-validated LASSO models on both segments combined we generate this output and identify these variables as significant:

training data MSE

0.245037755093

test data MSE

0.147256055422

training data R-square

0.764120601836

test data R-square

0.825754345001

|  |  |  |
| --- | --- | --- |
| **Type of Variable** | **coefficient\_name** | **estimated\_value** |
| Rating | review\_scores\_value | 0.364668 |
| Rating | review\_scores\_cleanliness | 0.240849 |
| Neighborhood | West End | 0.172262 |
| Amenity | Other pets | 0.150278 |
| Amenity | Dryer | 0.143614 |
| Rating | review\_scores\_communication | 0.124529 |
| Neighborhood | Mattapan | 0.118821 |
| Rating | review\_scores\_checkin | 0.100702 |
| Rating | review\_scores\_accuracy | 0.082706 |
| Amenity | Doorman | 0.073673 |
| Rating | review\_scores\_location | 0.068892 |
| Neighborhood | West Roxbury | 0.066577 |
| Neighborhood | Brighton | 0.063317 |
| Amenity | Pets live on this property | 0.044513 |
| Amenity | Suitable for Events | 0.018491 |
| Amenity | Pets Allowed | 0.016994 |
| Property Type | Bed & Breakfast | 0.011869 |
| Amenity | Cable TV | 0.010303 |
| Amenity | Shampoo | 0.008127 |
| Neighborhood | Chinatown | 0.00592 |
| Amenity | Laptop Friendly Workspace | 0.004803 |
| Amenity | Smoke Detector | 0.004662 |
| Neighborhood | Jamaica Plain | -0.017787 |
| Amenity | Kitchen | -0.066066 |
| Neighborhood | Mission Hill | -0.083571 |

The fact that ratings are extremely significant predictors of the normalized ratings is unsurprising since these parts factor into the overall ratings score. What is significant to note is their relative importance and weight. Across the entirety of listings, the value and cleanliness are the two most significant predictors at roughly 4X-6X relative weight as compared to the review score of the location. Also, unsurprisingly, location, as defined by certain neighborhoods, is an important factor towards people’s overall satisfaction as roughly 25% of the total predictors in this model relate to that variable.

In terms of amenities, what is quite interesting is to note that pet friendliness (as identified by the three categories of ‘Other Pets’, ‘Pets Allowed’ and ‘Pets Live on this Property’) would amount to the third strongest coefficient in identifying overall relative ratings. Also, quite curiously the amenity of ‘kitchen’ appears negatively impactful towards overall satisfaction. In order to fully understand the impact of these categorical variables it is conceivable that sentiment analysis could better uncover the underlying associations that are making these positive or negatively impactful.

**Projected Revenue**

Below are the results from the commercial properties (>= hosts with 3 listings).

|  |
| --- |
| **training data MSE** |
| **25846266.39** |
| **test data MSE** |
| **27641996.71** |
| **training data R-square** |
| **0.402413082** |
| **test data R-square** |
| **0.321640062** |

|  |  |  |
| --- | --- | --- |
| **Category** | **coefficient\_name** | **estimated\_value** |
| Neighborhood | South End | 3247.992058 |
| Neighborhood | Beacon Hill | 3101.895147 |
| Neighborhood | Downtown | 2689.539947 |
| Neighborhood | Condominium | 2355.920559 |
| Neighborhood | North End | 2337.882215 |
| Neighborhood | Back Bay | 2077.058264 |
| Amenity | Wireless Internet | 1753.831483 |
| Amenity | Wheelchair Accessible | 1373.121942 |
| Property Type | Loft | 1303.084007 |
| Amenity | Pool | 894.023352 |
| Description | accommodates | 863.432209 |
| Amenity | Dryer | 853.199288 |
| Neighborhood | Roxbury | 749.714951 |
| Amenity | Iron | 704.107888 |
| Amenity | Air Conditioning | 567.776616 |
| Neighborhood | Charlestown | 560.400031 |
| Amenity | Smoke Detector | 515.445234 |
| Amenity | Breakfast | 512.000381 |
| Fees | cleaning\_fee | 204.853541 |
| Reviews | review\_scores\_cleanliness | 165.266846 |
| Reviews | review\_scores\_communication | 150.852153 |
| Reviews | reviews\_per\_month | 150.448082 |
| Reviews | review\_scores\_accuracy | 121.026667 |
| Amenity | Doorman | 10.636626 |
| Reviews | number\_of\_reviews | 0.15728 |
| Description | calculated\_host\_listings\_count | -39.449594 |
| Neighborhood | South Boston | -45.077115 |
| Neighborhood | Jamaica Plain | -78.975221 |
| Description | Other | -83.050179 |
| Neighborhood | Fenway | -123.931928 |
| Reviews | review\_scores\_value | -125.643053 |
| Neighborhood | Dorchester | -174.306466 |
| Amenity | Gym | -271.959582 |
| Amenity | Lock on Bedroom Door | -320.284592 |
| Amenity | Pets live on this property | -335.340299 |
| Amenity | Buzzer/Wireless Intercom | -336.389677 |
| Neighborhood | Brighton | -442.198838 |
| Amenity | Other pets | -910.560084 |
| Amenity | Shampoo | -1022.944102 |
| Neighborhood | East Boston | -1079.627709 |
| Amenity | Dogs | -1323.67834 |
| Amenity | Free Parking on Street | -1512.956534 |
| Property Type | Boat | -1671.994148 |
| Neighborhood | South Boston Waterfront | -1929.708139 |
| Amenity | Free Parking on Premises | -1933.584277 |
| Room Type | Private room | -1986.674635 |
| Neighborhood | Roslindale | -2117.963458 |
| Room Type | Shared room | -2203.705522 |

Overall there are an incredible amount of variables that are flagged as predictors to generate this model. Some key takeaways we have from this first iteration was the impact that neighborhood has on the expected revenue. This suggests that there are some very desirable locations for travelers to Boston. Now, as a potential host this does not necessarily take into account the cost for entry- meaning that some of these areas are likely more expensive to host and thus generate higher revenues, but not necessarily greater profitability.

In terms of reviews, it is interesting to note that cleanliness as a metric that generates additional revenue. What is somewhat fascinating is that the presence of a cleaning fee appears to be

**Analysis**

Businesses exist to deliver value to customers who are willing to pay. Therefore when exploring these results we used the following mental model to approach the analysis.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Impact to Revenue | |
|  |  | + | - |
| Impact to Reviews | + | Sweet Spot | Are travelers making a killing? |
|  | - | Are buyers making a killing? | Less Please |

**Sweet Spot**

The first category is comprised of the items that customers rate highly and also exhibit a willingness to pay. Items in this category are things that hosts ought to consider focusing their efforts on as travelers appear willing to pay for these and also are willing to provide higher ratings for these types of features, thereby creating a virtuous cycle of higher ratings and higher revenues.

The biggest feature we would like to highlight is the fact that cleanliness consistently rates among traveler’s most desired features and, also, they appear willing to pay for it. Surprisingly in this data, it suggests that the presence of a cleaning fee actually has a positive impact on revenue (NOTE: cleaning fees were not used in calculating the revenues). This data suggests that hosts should most pay attention to this feature of their listing.

|  |  |  |
| --- | --- | --- |
| **Category** | **Predictor** | **Effect** |
| It's the extras that count | Breakfast | 512 |
| Bed & Breakfast | 0.01187 |
| Security | Doorman | 10.6366 |
| Doorman | 0.07367 |
| It's the extras that count | Dryer | 853.199 |
| Dryer | 0.14361 |
| Connected | Wireless Internet | 1753.83 |
| Laptop Friendly Workspace | 0.0048 |
| Clean is King | cleaning\_fee (presence of) | 204.854 |
| review\_scores\_cleanliness | 165.267 |
| review\_scores\_cleanliness | 0.24085 |

**Less Please**

Interestingly enough, this analysis did not uncover any strong predictors of items that people have expressed a desire for less of and are not willing to pay for. This intuitively would make sense as these items would become filtered out of the data through time as travelers don’t consume the feature, they don’t pay for it, and eventually hosts stop offering it.

**Are Travelers Making a Killing?**

|  |  |  |
| --- | --- | --- |
| **Category** | **Predictor** | **Effect** |
| Location, location, location | Brighton | -442.198838 |
| Brighton | 0.063317 |
| Who doesn’t love dogs? | Dogs | -1323.67834 |
| Other pets | -910.560084 |
| Other pets | 0.150278 |
| Pets Allowed | 0.016994 |
| Pets live on this property | -335.340299 |
| Pets live on this property | 0.044513 |
| Overall good value? | review\_scores\_value | -125.643053 |
| review\_scores\_value | 0.364668 |

Three interesting divergences arrive in the data: two localized and one perhaps with broader implications. The first one has to do with the neighborhood of Brighton, MA. This neighborhood is adjacent to Boston University. Travelers who stay in this locale tend to have more favorable ratings, however, based on projected revenues of our commercial listers, staying in Brighton appears to be correlated with lower revenues. There are a number of potential explanations beyond the limits of this analysis. Could Brighton be a pretty well received location and the hosts of Airbnb haven’t caught onto their ability to effectively price this neighborhood? Could the bias in our segmentation approach negatively impact the revenue projections for hosts in this area? Perhaps Brighton is filled with positive and optimistic college students who just don’t have the means to pay? Or even, is it possible that it is the next hipster paradise just waiting to be discovered?

The second divergence and perhaps the most suprising, in terms of magnitude has to do with the impact of pets on the revenues. In terms of reviews, the presence and association with pets appears to have an overwhelmingly positive correlation with relative review ratings. However, when it comes to the revenue projections the story could not be quite the opposite. In fact, it appears that pets have a roughly -$1245 combined effect on a properties projected revenues in our data set. Is this because hosts just aren’t aware of how much people value their pets? Are pets a polarizing issue when it comes to making an Airbnb stay decision? Once again, does the cut in our data negatively impact our projections of expected revenue?

In both of these situations it is quite intriguing to consider the possibility that travelers are free-riding on hosts’ listings. In both situations, the data might possibly suggest that hosts have an opportunity to charge more than what they are currently charging. Or, it may also be possible that an overage of supply exists driving down the total revenues. Either way, it appears that travelers are definitely getting a good deal with respect to these two categories.

**PLEASE CONFIRM THIS LOGIC/THOUGHT PROCESS**

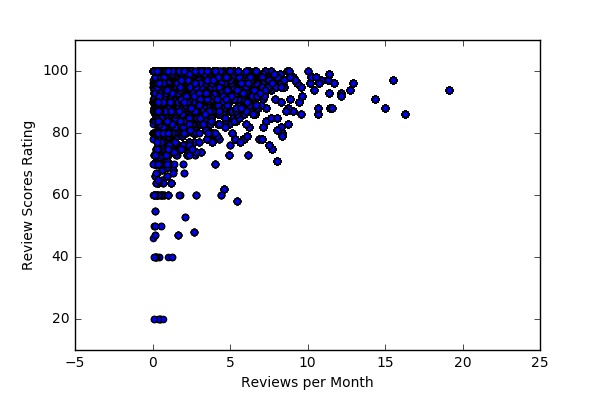
The macro-issue perhaps uncovered in the data possibly suggests that overall the Boston Airbnb market is favorable to travelers to the detriment of hosts. Theoretically, one would expect for these two metrics to be perfectly balanced and have zero impact on revenue and reviews. Now, part of this is fantasy as indicated that overall review ratings are partially calculated using these values, however, the magnitude of their delta might suggest the overall market’s favorability of one group vs another. Such implications would be very interesting and could be extremely impactful for future hosts. If this divergence could theoretically point to underlying conditions of the marketplace, then this divergence could serve as a metric for indicating which markets hosts could make money more easily.

**LIMITATIONS**

Overall this analysis was intended to be a starting point on an extremely robust and interesting data set. There are many limitations in this analysis, but nevertheless, it is a fascinating starting point.. Foremost is the opacity in the calendar data which was discussed at length earlier. This invariably impacted the outcomes of our analysis in ways that were beyond our reach. Another key limitation is the strong underlying correlations in our data. Within our analysis we treated every amenity as an independent variable, but there are likely clusters of offerings that form within the data. Meaning, there are commercial properties in downtown Boston that have doormen, gyms, pools, and other amenities that are treated as independent observations that likely skew the outcomes of the feature selection. Future analysis may consider a form of cluster analysis to build distinct groupings to understand the ideal “type” of property as opposed to raw features.

Another limitation is the fact that Boston is somewhat treated as a monolithic entity. As the data indicates, revenues and ratings necessarily vary quite dramatically between different neighborhoods, different urban/suburban contexts, different types of properties, etc. We attempted to normalize our data by neighborhoods, however, the data set was not large enough to effectively conduct that analysis. We did not have exact addresses either, which hindered our ability to effectively combine square footage and other data to come up with efficiency metrics which would be very meaningful to scale the data for future hosts.

Also, perhaps the most key limitation in this analysis is that it does not deal with the imagery nor the unstructured textual data associated with the listing and its reviews. These items are central to the decision making process of future travelers (customers) and would necessarily be a part of future analysis in trying to better characterize the market and the impacts of different features on revenues and reviews.



Interesting chart showing that higher reviews per month appear to lead to repeated high reviews

