ORIE 3120 Final Project: NYC Airbnb Data Processing & Analysis

Matan Auerbach, Sydney Ho, Serena Huang, Owen Rector

Introduction

Over the past few weeks, our team analyzed NYC Airbnb data from the year 2019. This dataset includes 16 columns and 48,895 rows of data (excluding headers). Of the 16 columns, 5 of them are categorical (name, host_name, neighbourhood_group, neighbourhood, and room_type), 10 of them are quantitative (id, host_id, latitude, longitude, price, minimum_nights, number_of_reviews, reviews_per_month, calculated_host_listings_count, and availability_365) and 1 of them is in datetime format (last_review). For the categorical data, neighbourhood_group represents the five boroughs of NYC, the neighbourhood column represents the actual name of the neighborhood, and room_type is either classified as "Private room", "Shared room", or "Entire home/apt". The overall dataset provides comprehensive information about the various Airbnb listings in NYC, including location and features, prices, and number of reviews, which provides insight on what drives prices up and what factors determine popularity of the listing.

Our team also decided to incorporate a supplementary dataset on NYC Airbnb demand over time from 2015 to 2020. This dataset includes 7 columns and 2,193 rows of data (excluding headers). There are 6 columns with numerical data (Demand, Easter, Thanksgiving, Christmas, Temperature, and Marketing), and 1 column with datetime data (Date). For the three holidays, the possible values in these columns is 0 or 1, a binary variable to represent whether the date corresponds to one of those holidays. This dataset provides us with more information about how demand has changed over time and if there are any seasonal patterns.

In this report, our team aims to address the following questions:

1) What are the **key factors that impact the pricing** of Airbnb listings, and can we develop a **predictive model** to estimate the daily price of a listing based on these factors, such as location, room type, and review metrics? (linear regression and testing assumptions)
2) How has the demand for NYC Airbnb listings changed over **time** and can we build **forecasting models** to anticipate future demand? (SARIMA & Holt-Winters' forecasting)

All these questions relate to how to optimize listings based on different conditions in order to maximize consumer interest in the property. These insights are especially valuable to Airbnb hosts, who can use these findings to adjust their pricing strategy and make property modifications in order to maximize the number of bookings and profitability. This information may also be helpful for potential guests, who may want to use the forecasting model to determine the best time of year to book a listing. Finally, business analysts at Airbnb or competitor hospitality/travel companies can use these findings to gain a better understanding of the most influential factors on pricing and provide recommendations for revenue growth.

Question 1

In this part of our study, we aim to answer the following research question: What are the key factors that impact the pricing of Airbnb listings, and can we develop a predictive model to estimate the daily price of a listing based on these factors, such as location, room type, and review metrics?

To achieve this objective, we will use multiple linear regression to develop a predictive model for estimating the price of an Airbnb listing. By using multiple linear regression to develop a predictive model for estimating the price of an Airbnb listing, we can gain insights into the key factors that drive pricing and use this knowledge to improve our understanding of the market. This can be valuable for both hosts who are looking to set prices for their listings and travelers who are looking to find the best deals on Airbnb. The first step towards building this model was feature engineering.

Feature Engineering

The first variables we had to adjust before building a linear regression model were the categorical variables neighborhood group and room type. Since we decided that we wanted to visualize the impacts of these different categorical variables on the price variable, we chose to represent each room and neighborhood group as its own column of data, where it was a binary result, either 1 or 0. For instance, if an Airbnb listing was in Manhattan and a private room, on that row of data the Manhattan column for neighborhood group and the private room column for room type would be denoted with 1's, with all other room type and neighborhood group columns being 0.

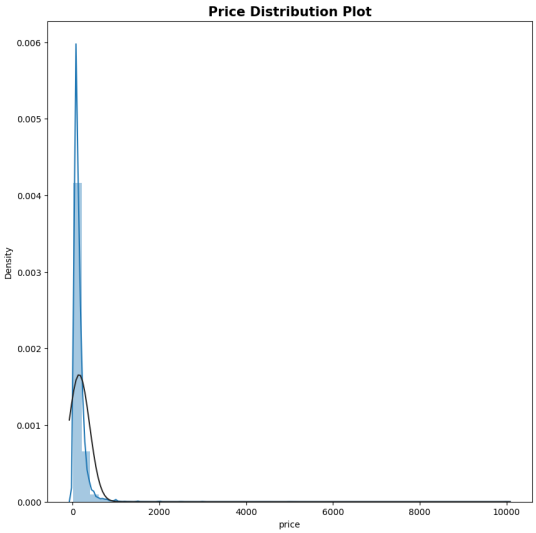
By including these dummy variables in the model, we can accurately estimate the effect of each neighborhood and room type on the price of the listing while controlling for other factors that may be associated with price. Additionally, the use of dummy variables ensures that the categorical variables are treated appropriately as discrete values, rather than being interpreted as continuous variables.

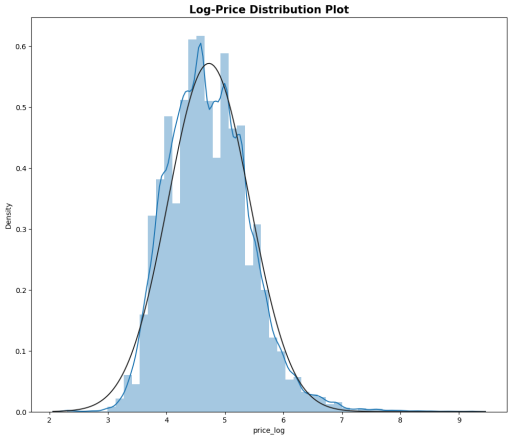
Next, knowing that location would be one of the most important features in determining price, we decided to add an additional variable, 'distance_from_manhattan'. This new variable was calculated using geographic data, including the latitude and longitude of each listing and the coordinates of Times Square to calculate the distance from Times Square of each listing, in miles. By including this new variable in our machine learning model, we can capture the impact of distance from Times Square on Airbnb prices.

Generally, the most expensive properties (capped at 1200 dollars) appear to be in Manhattan close to Times Square, and as you get further and further

away their price drops, proving to be a negative correlation. This is an interesting observation, and this engineered variable will be instrumental in the future as we attempt to answer our overarching question about being able to predict prices.

The last step before building our linear regression model is to examine our dependent variable, price. What we find is quite interesting, and is represented by the following plot:





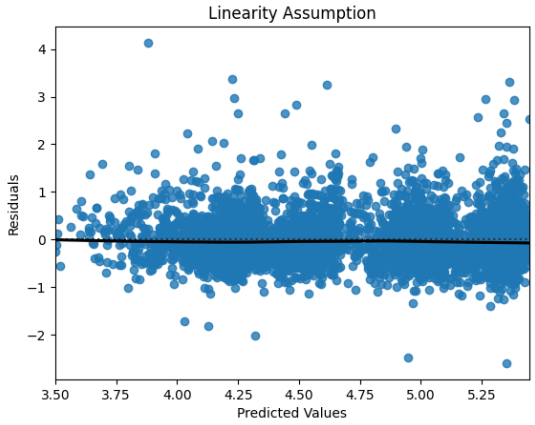
The distribution graph on the right shows that there is a right-skewed distribution on price. This means there is a positive skewness, with some large outliers with unusually large prices. To account for this, we performed a log transformation to make this feature less skewed and instead approximated a normal distribution.

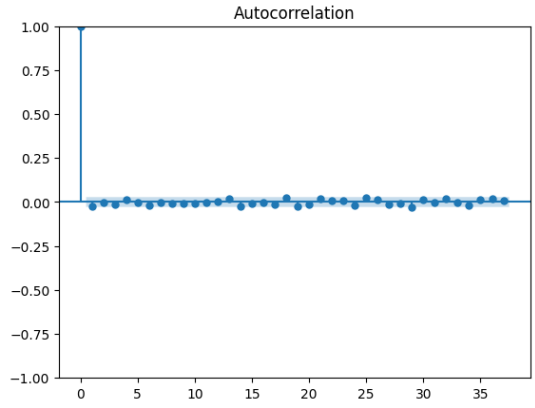
Using the log-transformed price can help to linearize the relationship between price and the predictors and make the model more accurate. In addition, it can help to reduce the influence of extreme values or outliers on the regression analysis, as the logarithm function compresses the range of values.

However, it's important to note that interpreting the results of the model based on the log-transformed variable can be a bit more challenging. The coefficients for the predictors in the model will be interpreted as the change in the response variable for a one-unit change in the predictor, holding all other predictors constant. In the case of a log-transformed response variable, the coefficient can be interpreted as the percentage change in the response variable for a one-unit change in the predictor.

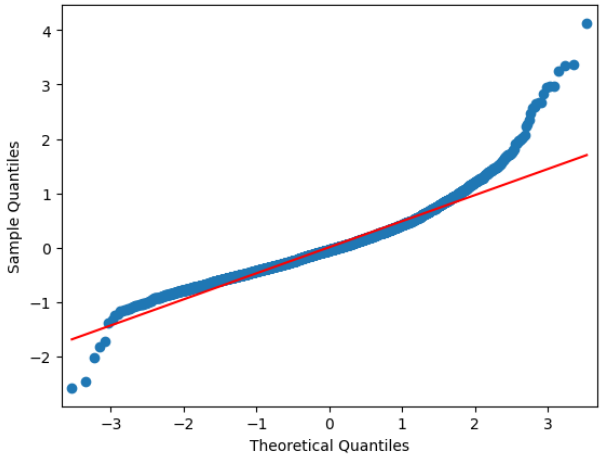
Testing Assumptions of Linear Regression

Before proceeding with our analysis, it is important to test the assumptions of linear regression to ensure that the model is valid and reliable. The following three plots were created to test these assumptions:





QQ-Plot



Assumption 1: The residuals have constant variance

Looking at the plot above titled "Linearity Assumption", we see a relatively even distribution of residual points across the range of the predictor values, and a relatively straight LOWESS line in the absolute value graph, allowing us to verify that the errors also have constant variance.

Assumption 2: The residuals have a mean of 0

Again looking at the plot above titled "Linearity Assumption" we see that the errors appear to have a constant mean of 0 in the residual plot, which means we can verify this assumption.

Assumption 3: The residuals are independent

In the autocorrelation graph shown above, we see a low correlation between the variables, with nearly all the variation at lag 0. This means that we are able to confirm that these residual errors are independent of one another, and this assumption is satisfied.

Assumption 4: The residuals are normally distributed

Lastly, looking at the QQ-plot, we can see that the plot generally follows the linear red line, suggesting a relatively good fit for this linear model. We do see a bit of a heavy-tailed distribution, but it is strong enough to proceed forward.

Linear Regression Model & Price Prediction

Now equipped with clean and prepared data, it is time to build our linear regression model. Using price_log as the dependent variable and each of the room types, neighborhood groups, distance from manhattan, and reviews per month as the independent variables, we split the data into testing and training sets and build our linear regression model. The following output was obtained:

This is the output of a multiple linear regression analysis, where the dependent variable is the natural logarithm of the price of a rental property and the independent variables are minimum_nights, reviews_per_month, dist_from_manhattan, neighborhood (categorical variable with five levels: Bronx, Brooklyn, Manhattan, Queens, Staten Island), and room type (categorical variable with three levels: Entire home/apt, Private room, Shared room).

	coef	std err	t	P>|t|	[0.025	0.975]
const	3.0611	0.008	395.644	0.000	3.046	3.076
minimum_nights	-0.0012	0.000	-9.693	0.000	-0.001	-0.001
reviews_per_month	-0.0064	0.002	-4.170	0.000	-0.009	-0.003
dist_from_manhattan	-0.0323	0.001	-43.700	0.000	-0.034	-0.031
neighborhood_Bronx	0.5040	0.015	34.706	0.000	0.476	0.532
neighborhood_Brooklyn	0.5710	0.007	86.900	0.000	0.558	0.584
neighborhood_Manhattan	0.7453	0.008	95.916	0.000	0.730	0.761
neighborhood_Queens	0.5133	0.008	62.664	0.000	0.497	0.529
neighborhood_Staten Island	0.7275	0.024	30.312	0.000	0.680	0.775
room_Entire home/apt	1.6444	0.005	305.510	0.000	1.634	1.655
room_Private room	0.8815	0.005	160.297	0.000	0.871	0.892
room_Shared room	0.5352	0.012	44.867	0.000	0.512	0.559

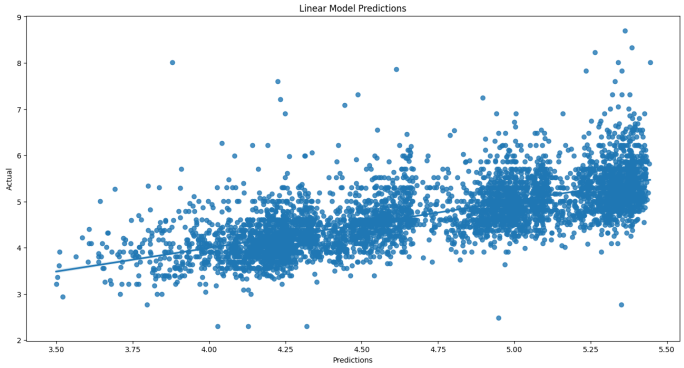
Omnibus:	14034.646	Durbin-Watson:	1.983
Prob(Omnibus):	0.000	Jarque-Bera (JB):	82780.886
Skew:	1.410	Prob(JB):	0.00
Kurtosis:	9.100	Cond. No.	3.42e+17

An R-squared output value of 0.477 indicates that the independent variables in the model explain 47.7% of the variance in the dependent variable. The adjusted R-squared value is the same, which means that the addition of each independent variable did not improve the model's fit. The F-statistic of 4451 and its associated p-value (0.00) indicate that the model as a whole is significant. We also found a Root Mean Squared Error value of 0.495. The RMSE is the square root of the MSE and is also a measure of the average difference between the predicted and actual values.

]The coefficients for each independent variable represent the estimated change in the dependent variable associated with a one-unit change in the respective independent variable, while holding all other independent variables constant. The intercept or constant (3.0611) represents the expected log price when all other independent variables are zero.

The output shows that all the independent variables in the model are statistically significant. For example, a one-unit increase in the dist_from_manhattan variable (i.e., an increase in distance from Manhattan by one mile) is associated with a decrease of 0.0323 in the natural logarithm of the price, holding all other variables constant. The coefficients for the neighborhood and room type variables indicate the differences in expected log prices compared to the reference categories (Staten Island and Shared room, respectively).

To estimate the performance of our model at predicting the prices, we can examine a plot of our y_test values against our y_pred values. If the predicted values are close to the actual values, then the points in the scatter plot will be closer to a diagonal line with a slope of 1. On the other hand, if the predicted values are far from the actual values, then the points in the scatter plot will be more dispersed and further away from the diagonal line. We obtain the following plot to the right. From this plot, we see a relatively decent fit of the predicted values against the true values, but it is far from perfect.



Conclusion

From this complete analysis, we are able to make the following conclusions regarding Question 1:

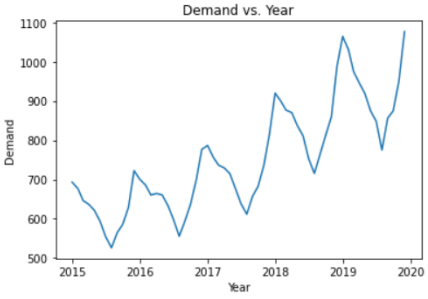
1) The distance from Manhattan has a negative relationship with the rental property price, meaning that properties located further away from Manhattan tend to be less expensive.
2) Among the room type variables, Entire home/apt has the highest coefficient, indicating that this type of rental property is more expensive than Private room and Shared room types.
3) Overall, the R-squared and RMSE metrics suggest that the linear regression model is a reasonable fit for the data, but there is still room for improvement.

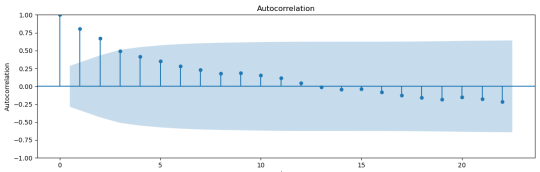
Question 2

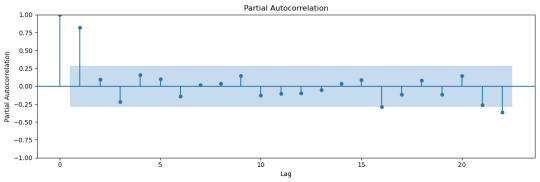
Finally, we address the last question of our study: How has the demand for NYC Airbnb listings changed over time and can we build forecasting models to anticipate future demand? For this part of the project, we chose to use ARIMA/SARIMA and Holt-Winters' forecasting models on the supplementary dataset.

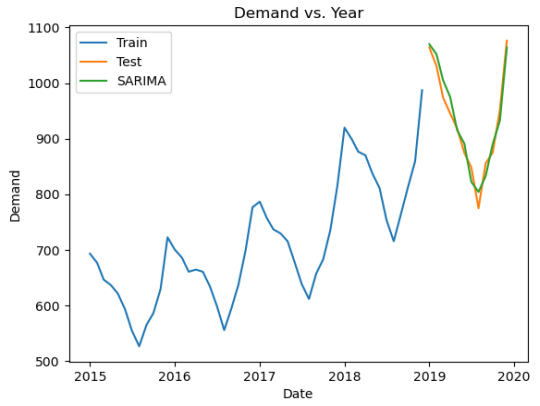
ARIMA/SARIMA Model Process

The ARIMA (Autoregressive Integrated Moving Average) model utilizes historical values to forecast future values. The SARIMA (Seasonal-ARIMA) model is similar, but it also considers the effect of seasonality. There are 3 components to an ARIMA model: p (number of autoregressive terms), d (number of nonseasonal differences needed for stationarity), and q (number of lagged forecast errors in the prediction equation). After aggregating the data to show the average demand in each month across the five years, we decided to use a SARIMA model in order to capture the clear seasonality in demand (pictured to the right). Note that we removed data from 2020 due to the COVID-19 pandemic, which is not an accurate representation of what hospitality data generally looks like.





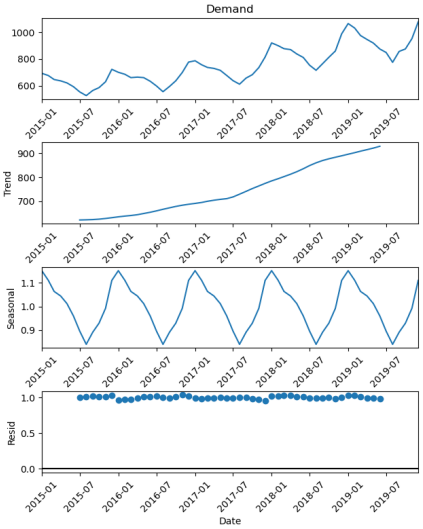




We then used the Dickey-Fuller Test to test the stationarity of the data, and we determined that the time series has a unit root due to the p-value of 0.999, indicating it is non-stationary. We then incorporated seasonal first differences by shifting the data by 12 positions and finding the difference between the original and shifted demand values. This allowed us to lower our p-value and achieve a statistically significant result. We then created autocorrelation and partial autocorrelation plots (pictured to the top left) to determine the p and q values for the SARIMA model. Looking at the partial autocorrelation plot, we set p = 2 because the "shut-off" value appears at 2, and looking at the autocorrelation plot, we see there is an exponential decrease, suggesting q = 1. We also set d = 1 because we are utilizing seasonal differencing. Finally, we chose seasonal order values of 1, 1, 1, and 12 based on the plots and to capture the 12-month lagged correlation. Using these parameters, we were able to create a SARIMA model that closely aligned with the test data (pictured to the bottom left).

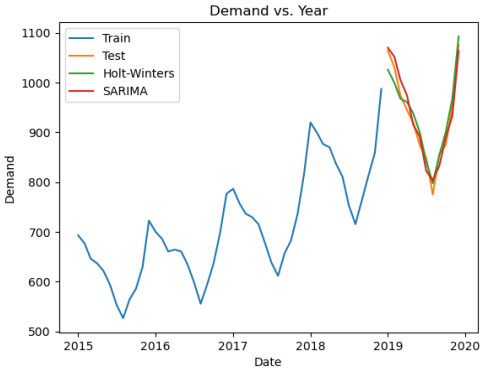
Holt-Winters' Model Process

Next, we decided to explore another forecasting model to see if we could achieve better results. We fit our data to the Holt-Winters' model because it has exponential smoothing and can highlight trends and seasonality when forecasting. The Demand vs. Year plot generated above clearly displays seasonality and the demand at each cycle seems to be increasing with each year, suggesting that the data may be multiplicative.



However, to be more certain, we decomposed the data to observe the demand, trend, seasonal, and residual time series. The four plots (pictured to the right) indicate that there is a monthly seasonality and that there is an upward trend in the demand. This verifies that the model is multiplicative and not additive as the values are not constant and there is a seasonality component as well. The residuals also show that there is very little variation in the demand across the months that are not attributed to the seasonality and trend. Now that we have decomposed the time series data, we can adjust the parameters of the Holt-Winters' model that will be fit to the data. Since there seems to be monthly seasonality, we set the seasonal periods to 12 and seasonal components to multiplicative.

Finally, we built our Holt-Winters' model by splitting the time series data into a training set and test set. We partitioned the original dataset by taking the last twelve months of the dataset as the test set and all previous data as the training set. After, we used the model to forecast the demand in the last twelve months of the dataset and plotted both the test demand and the forecasted demand of both the SARIMA and Holt-Winters' models to compare the two.



	SARIMA	Holt-Winters
RMSE	21.225034	21.363989
Mean Error	19.375007	18.760129
Max Error	30.983765	38.278571
Min Error	3.477880	1.556336

The SARIMA model predicted the demand much closer to the test data, however the Holt-Winters' model forecasted demand in a pattern much similar to the previous years. The Holt-Winters' forecast displays a decrease in demand from the beginning of 2019, a very brief period where the demand plateaus, then a longer period of decreasing demand until it reaches the minimum where demand starts to increase again for the holiday season at the end of the year. However, it is important to keep in mind that the 2019-2020 year was also the peak of the COVID-19 pandemic, resulting in an irregularity in the demand patterns when compared to the previous years. Final analysis on the effectiveness of the SARIMA and Holt-Winters' models indicated that the RMSEs were almost the same, with there being very little differences in the errors other than the Holt-Winters' model having a larger range for errors.

Therefore, there is little difference in the effectiveness of these models, but it can be concluded that the Holt-Winters' model is better for predicting the seasonal trends in the data. In the future years, when the effects of the pandemic are less prevalent, the Holt-Winters' model will be able to more accurately forecast demand. Given the overall upward trend in the demand, we anticipate that demand will continue to increase, especially as more people are willing to host/book with the relaxed COVID restrictions. There is also an overall trend where the colder months tend to have higher demand while hotter months have lower demand, with overall demand rising as the years progress. Thus, colder months in upcoming years will be the optimal time to host/book Airbnbs.

Conclusion

Through using linear regression, testing the assumptions of the linear regression model, and utilizing two different time series forecasting models, we were able to showcase the optimal conditions for Airbnb hosts, consumers, and business analysts. Using linear regression, we are able to see the optimal strategy for Airbnb hosts based on the most successful Airbnb listings using the room types, neighborhood groups, distance from manhattan, and reviews per month as variables. In addition, both hosts and consumers can leverage the forecasting data to find the best times to host/book a listing. With this information about the impact of both individual listing features and the greater seasonal trends in Airbnb demand, hosts, consumers, and analysts can make more informed decisions about their listings, bookings, and Airbnb's overall business strategy.