

Master of Science (MS) in Data Science

Module: ITC6004A1 – Data Visualization

Instructor: Konstantinos Bloutsos

Students’ names:

Alkiviadis Kariotis 241735

Konstantinos Megaritis 271868

Konstantinos Prozymas 27365

Term: Winter Term 2023

Type: Group Assignment

Investment Intelligence Division of XYZ REITs company

Submission Date: Friday, April 7

Words: 4487

Table of Contents

[Abstract 3](#_Toc131746576)

[Introduction 3](#_Toc131746577)

[Objectives 4](#_Toc131746578)

[Data Structure 4](#_Toc131746579)

[Type of data 4](#_Toc131746580)

[Description of dataset 5](#_Toc131746581)

[Data Quality Assessment 6](#_Toc131746582)

[Dataset Preview 6](#_Toc131746583)

[Data Cleaning 7](#_Toc131746584)

[Data Analysis 9](#_Toc131746585)

[Graphs – Insights – Recommendations 9](#_Toc131746586)

[Regression - Forecasting 16](#_Toc131746587)

[Conclusion 18](#_Toc131746588)

[References 19](#_Toc131746589)

[Appendix 20](#_Toc131746590)

# Abstract

This comprehensive study aims to provide valuable insights that will support XYZ REITs company in making informed decisions in the competitive NYC rental market.This report presents a comprehensive exploratory data analysis (EDA) of the New York City Airbnb Open Data available on Kaggle. Utilizing Power BI as the primary analytical tool, the report examines detailed information on Airbnb listings and host profiles in New York City. The analysis identifies trends, patterns, and factors that influence user behavior and preferences, providing valuable insights and recommendations for Airbnb hosts and potential investors. The findings can inform pricing strategies, areas of opportunity, and overall understanding of the platform's dynamics in one of the world's most popular travel destinations.

# Introduction

As part of the Department of Investment Intelligence Division, we were assigned to research the specific dataset and come up with recommendations regarding the characteristics of the amenities that our REITs (Real estate investment trusts) company that would be the most profitable for our organization to invest onto.

So, Airbnb, an online marketplace for short-term rentals, connects hosts with travelers seeking various types of accommodations. The platform offers diverse rental properties, including apartments, homes, and boats, located in neighborhoods around the world. With a booking fee structure that charges hosts 3% of the booking value and guests between 6% to 12%, depending on the booking type, Airbnb has become a popular choice for both hosts and guests.

New York City, a major global tourist destination, attracts a significant number of Airbnb users. As such, the New York City Airbnb Open Data provides a rich dataset for understanding trends and patterns within the platform. This report aims to conduct an in-depth EDA of the dataset, focusing on features such as listing price, location, availability, and reviews. By leveraging Power BI's visualization and exploration capabilities, the analysis offers valuable insights and recommendations that can guide decision-making for Airbnb hosts, potential investors, and other stakeholders in the short-term rental market.

# Objectives

Our department will perform a comprehensive research to identify key insights and trends in the New York City (NYC) rental market, aiming to benefit XYZ REITs company's strategic decision-making process. We will begin by determining the number of neighborhood areas in NYC and examining the distribution of rental properties across these neighborhoods. This analysis will help us understand how rental prices vary with respect to neighborhoods, rental property types, and rental amenities. We will also investigate the volume of reviews given to individual properties and identify the top 10 hosts in NYC, along with the types of properties they offer. This information will allow us to gain insights into the popularity and reputation of various properties and hosts. Additionally, we will explore the average prices of different property types and the number of properties available in each area. To provide a better understanding of the rental market, we will assess the availability of each host and examine whether a listing's price varies according to its location. Furthermore, we will calculate the average price for a listing based on neighborhood and neighborhood area, as well as how it varies depending on room type and the minimum number of nights required for a booking. Our research will also delve into the relationship between the price of a listing and its availability, as well as the median price of properties in each neighborhood. We will identify the neighborhood area with the most reviews and determine the last date of review for a particular price range. Lastly, by examining how the average price varies based on neighborhood and the number of reviews, we will be able to uncover trends and patterns in the market.

# Data Structure

## Type of data

Cross-sectional data is a type of observational data collected from a sample of individuals, households, companies, or other units at a specific point in time. In cross-sectional data, each unit is observed only once, and the data collected pertains to the characteristics of the units at that particular point in time. Cross-sectional data can be contrasted with longitudinal data, which is collected from the same units over multiple time periods, allowing for the analysis of changes or trends over time. Cross-sectional data is often used in descriptive statistics and in analyzing the relationship between different variables at a specific point in time (Beck & Katz, 1995).

Regarding our case, the dataset about New York Airbnb has various information regarding the year of 2019 and can be found in this link: <https://www.kaggle.com/datasets/dgomonov/new-york-city-airbnb-open-data>.  
The dataset contains each apartment available for rent, having each one of them individually and only once shown in the dataset, having for each one seventeen different characteristics.

## Description of dataset

The New York City Airbnb Open Data is a comprehensive dataset containing information about Airbnb listings in New York City. The data was collected from the Inside Airbnb website and is available on Kaggle. It contains information on over 48,000 listings, including the following:

1. id: This column contains a unique integer identifier for each Airbnb listing. It can be used to uniquely identify and reference each listing in the dataset.
2. name: This column contains the name of each Airbnb listing. The name can provide some insight into the type of listing, its location, or other details that may be useful in understanding the data.
3. host\_id: This column contains a unique integer identifier for the host of each Airbnb listing. This can be useful for grouping listings by host or analyzing host-specific trends in the data.
4. host\_name: This column contains the name of the host of each Airbnb listing. This can be useful for identifying popular hosts or for analyzing the relationship between the host's name and the popularity or success of their listings.
5. neighbourhood\_group: This column contains the borough of each Airbnb listing, providing information on the geographic location of the listing.
6. neighbourhood: This column contains the neighborhood of each Airbnb listing, providing further detail on the location of the listing within the borough.
7. latitude: This column contains the latitude coordinates of each Airbnb listing, which can be useful for visualizing the locations of the listings on a map.
8. longitude: This column contains the longitude coordinates of each Airbnb listing, which can be useful for visualizing the locations of the listings on a map.
9. room\_type: This column contains the type of room being listed, which can be either a private room, an entire home/apartment, or a shared room. This can be useful for analyzing the popularity or success of different types of listings.
10. price: This column contains the price per night for each Airbnb listing, providing insight into the pricing trends and variability in the data.
11. minimum\_nights: This column contains the minimum number of nights required for each Airbnb listing, which can be useful for analyzing booking trends and preferences.
12. number\_of\_reviews: This column contains the number of reviews for each Airbnb listing, which can be useful for analyzing the popularity or success of the listings.
13. last\_review: This column contains the date of the last review for each Airbnb listing, which can be useful for analyzing trends in review activity over time.
14. reviews\_per\_month: This column contains the average number of reviews per month for each Airbnb listing, which can be useful for analyzing the frequency and consistency of review activity.
15. calculated\_host\_listings\_count: This column contains the number of listings for the host of each Airbnb listing, which can be useful for analyzing host-specific trends in the data.
16. availability\_365: This column contains the number of days each Airbnb listing is available for booking within the next 365 days, which can be useful for analyzing booking trends and availability.
17. rating: Rating from the renters for each amenity listed in the dataset.

# Data Quality Assessment

Data quality theory is the concept that data should meet certain standards in terms of accuracy, completeness, consistency, relevance, and timeliness to be considered high quality. Good data quality ensures that information is reliable and can be effectively used for decision making and analysis (Pipino et al., 2002).

So, in order to able to proceed and analyze our data we moved on towards cleaning the dataset with the following actions:

## Dataset Preview

To have a quick look and preview some basic statistics from the original data obtained we started with the following steps:

1. Add the data source to our report.
2. In the "Visualizations" pane, we selected the "Table" visualization.
3. Drag and drop all the variables that were meaningful for summarization into the "Values" field.
4. In the "Visualizations" pane, we selected the "Summary Statistics" card.
5. In the "Fields" section of the "Summary Statistics" card, we selected the respected variables.
6. The summary statistics for each of the variable that we have selected was displayed in the "Values" section of the "Summary Statistics" card, providing the numbers depicted in the table below:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **count** | **mean** | **std** | **min** | **25%** | **50%** | **75%** | **max** |
| id | 48895 | 19017143.236 | 10983108.386 | 2539.000 | 9471945.000 | 19677284.000 | 29152178.500 | 36487245.000 |
| host\_id | 48895 | 67620010.647 | 78610967.033 | 2438.000 | 7822033.000 | 30793816.000 | 107434423.000 | 274321313.000 |
| latitude | 48895 | 40.729 | 55 | 40.500 | 40.690 | 40.723 | 40.763 | 40.913 |
| longitude | 48895 | -73.952 | 46 | -74.244 | -73.983 | -73.956 | -73.936 | -73.713 |
| price | 48895 | 152.721 | 240.154 | 0 | 69.000 | 106.000 | 175.000 | 10000.000 |
| minimum\_nights | 48895 | 7.030 | 20.511 | 1.000 | 1.000 | 3.000 | 5.000 | 1250.000 |
| number\_of\_reviews | 48895 | 23.274 | 44.551 | 0 | 1.000 | 5.000 | 24.000 | 629.000 |
| reviews\_per\_month | 48895 | 1.373 | 1.680 | 10 | 190 | 720 | 2.020 | 58.500 |
| calculated\_host\_listings\_count | 48895 | 7.144 | 32.953 | 1.000 | 1.000 | 1.000 | 2.000 | 327.000 |
| availability\_365 | 48895 | 112.781 | 131.622 | 0 | 0 | 45.000 | 227.000 | 365.000 |
| rating | 48895 | 3.003865 | 1.415054 | 1.000 | 2.000 | 3.000 | 4.000 | 5.000 |

To get a quick visual view of what happening in the dataset we proceeded with the following steps:

1. From the marketplace of Power BI, we downloaded the "scatter chart matrix".
2. Calendar

   Description automatically generated with low confidenceWe drag and drop all the variables in the aforementioned onto the "Values" field of the "Scatter Chart Matrix" visual and fine-tuned in the "Format" pane in terms of adjusting markers, axis titles, and font sizes.

## Data Cleaning

* 1. Duplicates

To drop and check for duplicates in our dataset in Power BI, we followed these steps:

1. Open the Power Query Editor by selecting "Edit Queries" from the "Home" tab in the Power BI Desktop.
2. Selected the table that contains the duplicates you want to remove.
3. From the "Transform" tab, selected "Remove Rows" and then choose "Remove Duplicates".
4. In the "Remove Duplicates" dialog box, we selected all the columns of our dataset in order to identify duplicates.
5. Once we finalized the procedure and the duplicates have been removed, we selected "Close & Apply" from the "Home" tab to apply the changes and close the Power Query Editor.

Following that procedure, we reached the conclusion that the dataset had no duplicates.

* 1. Missing values

To identify missing values, we used the "Data Profiling" feature:

1. From the "View" tab in the Power Query Editor, we selected "Data Profiling" and then choose "Column Quality".
2. In the "Column Quality" pane, we saw a summary of the quality of each column in the dataset, including the number of missing values.
3. To preview the missing values for every column separately, we clicked on the desired column name in the "Column Quality" pane. This opened a detailed view of the column quality, including the number and percentage of missing values.

After that process we reach in the following results:

|  |  |
| --- | --- |
| **Columns** | **Missing values** |
| name | 16 |
| host\_name | 21 |
| last\_review | 10052 |
| reviews\_per\_month | 10052 |
| rating | 10052 |

* 1. Replace 0 values with NaN (with explanation)

Because Power BI does not show NaN (Not a Number) values in a graph. When you create a visual in Power BI, such as a chart or a table, any NaN values in the data are automatically excluded from the visual. So, we decided to proceed and replace all the missing values from all the columns mentioned before with NaN values, by following this procedure:

1. Opened the Power Query Editor by selecting "Edit Queries" from the "Home" tab in the Power BI Desktop.
2. Selected the four column that contains the missing values.
3. From the "Transform" tab, we selected "Replace Values" and then choose "Replace Errors".
4. In the "Replace Errors" dialog box, we selected "null" from the "Error Values" dropdown menu. This selected all the missing values in the column chosen, and then in the "Replace with" field we typed NaN, and finally Close & Apply the Power Query Editor.

After completing this procedure, we had done again the process of identifying missing values just to be sure everything worked out correctly.

* 1. Column Formats

From the "Modeling" tab, we selected "Data Type" and then choose the appropriate data type for each column of the dataset. Even though Power BI tried to automatically detect the data type based on the values in the column, we needed fix some of them.

# Data Analysis

## Graphs – Insights – Recommendations

For the analysis, we used Power BI, which is a business analytics service that provides interactive visualizations and business intelligence capabilities. Through basic observations we anticipated the following:

1. In New York City, there are five neighborhood groups: Bronx, Brooklyn, Manhattan, Queens, and Staten Island. Airbnb property listings for these areas are as follows:

|  |  |
| --- | --- |
| **Neighborhood Group** | **Properties** |
| Bronx | 1,091 |
| Brooklyn | 20,104 |
| Manhattan | 21,661 |
| Queens | 5,666 |
| Staten Island | 373 |
| Total Properties | 48,895 |

1. Approximately 37,500 hosts operate in NYC, with some properties belonging to the same hosts. The number of hosts in each area are:

|  |  |
| --- | --- |
| **Neighborhood Group** | **Hosts in Each Area** |
| Bronx | 789 |
| Brooklyn | 15,966 |
| Manhattan | 16,578 |
| Queens | 3,983 |
| Staten Island | 256 |
| Total Hosts in NYC | 37,572 |

1. The most popular areas in NYC with the highest number of Airbnb properties include:

|  |  |
| --- | --- |
| **Popular Areas in NYC** | **Properties** |
| Williamsburg, Brooklyn | 3,920 |
| Bedford-Stuyvesant, Brooklyn | 3,714 |
| Harlem, Manhattan | 2,658 |
| Bushwick, Brooklyn | 2,465 |
| Upper East Side, Manhattan | 1,971 |

1. Brooklyn and Manhattan have the highest concentration of Airbnb properties, with the most popular areas being Williamsburg, Bedford-Stuyvesant, and Harlem.
2. There are three types of properties in NYC: Entire home/apt, Shared room, and Private room. The number of each property type is as follows:

* Entire Home/Apartment: 25,409
* Shared room: 1,160
* Private room: 22,326

The average prices for these room types are:

|  |  |  |
| --- | --- | --- |
| **Room Type** | **Number of Properties** | **Average Price** |
| Entire Home/Apartment | 25,509 | $212 |
| Shared Room | 1,160 | $70 |
| Private Room | 22,326 | $90 |

1. The top 10 hosts for NYC Airbnb with the most reviews are: Michael, David, John, Alex, Daniel, Sarah, Maria, Jessica, and Sonder(NYC).
2. The following hosts have received more than 2000 reviews, indicating they are among the most reviewed hosts: Danielle, David, Eric, Jason, John, Kevin, Maya, Michael, and Sarah. Our observation suggests a strong correlation between the number of properties and the number of reviews received by hosts, as some hosts own multiple properties and receive a large number of reviews.Approximately 37 hosts own properties that are available for at least 150 days a year, with these properties primarily located in Brooklyn and Manhattan. These properties are available for rent for half of the year.
3. The table shows the average price of listings grouped by neighborhood and neighborhood group. For example, the average price in Tribeca is $490.6, while in Fort Wadsworth, Breezy Point, Sea Gate, Riverdale, and Staten Island it ranges from $442.1 to $800.

|  |  |  |
| --- | --- | --- |
| **Neighborhood** | **Neighborhood Group** | **Average Price** |
| Tribeca | Manhattan | $490.6 |
| Fort Wadsworth | Staten Island | $800 |
| Breezy Point | Queens | $213.3 |
| Sea Gate | Brooklyn | $487.9 |
| Riverdale | Bronx | $442.1 |

1. The median price of listings is also shown based on their availability, with values ranging from 72 to 361.

|  |  |
| --- | --- |
| **Availability** | **Price** |
| 109 | $132 |
| 361 | $80 |
| 72 | $94.5 |
| 134 | $94.5 |

The results indicate that there is no significant relationship between the price and availability of a listing.

1. Furthermore, the table shows the number of reviews received for each neighborhood group, with Manhattan having the highest number of reviews at 21,661, followed by Brooklyn at 20,104, Queens at 5,666, the Bronx at 1,091, and Staten Island at 373.

|  |  |
| --- | --- |
| **Neighborhood Area** | **No. of Reviews** |
| Bronx | 1,091 |
| Queens | 5,666 |
| Staten Island | 373 |
| Manhattan | 21,661 |
| Brooklyn | 20,104 |

1. Finally, the table shows the best listings regarding the rating but also with the highest number of reviews, because we had more than five amenities in our dataset that achieved a rating of 5 stars:

|  |  |  |
| --- | --- | --- |
| **id** | **rating** | **No. of Reviews** |
| 8168619 | 5 | 543 |
| 5061121 | 5 | 451 |
| 31994 | 5 | 426 |
| 699472 | 5 | 404 |
| 195233 | 5 | 401 |

Then, we moved forward making more complex graphs trying to compare different values and columns of the dataset, in order to obtain some insights and finally be able to make recommendations for our company.

We followed the following steps:

1. Insight: Observe the overall distribution of listing prices and identify the most common price ranges.   
     
   We proceeded with making a histogram with X-axis the price and Y-axis the numbers of listings (count of id). After observing the distribution of all the listing prices we identified that 92% of our dataset lies below the price of 300$. We used the following Measure to build the percentage that we needed to show with a Card Visual:  
     
   Below\_300\_Percentage = DIVIDE(COUNTROWS(FILTER(AB\_NYC\_2019, AB\_NYC\_2019[price] < 300)), COUNTROWS(AB\_NYC\_2019))\*100  
     
   Recommendation: Focus on listings within the most common price ranges to appeal to a larger audience. By pricing the listings competitively and within the market's preferred range, this will increase the likelihood of attracting more guests and achieving higher occupancy rates.
2. Insight: Identify neighborhoods with the highest and lowest average prices.  
     
   We created a Bar chart with X-axis the neighborhood and Y-axis the average price. By visualizing the Average price by neighborhood and by neighbor\_group we found out that even though Manhattan has the biggest average price the top two most expensive, with significant difference neighborhoods, come from Staten Island.

Recommendation: Consider setting prices competitively based on the neighborhood's average price. Research the local market to understand the price expectations of potential guests and adjust the pricing strategy accordingly. This will help you remain competitive and attractive to potential guests.

1. Insight: Determine which neighborhoods have the highest and lowest number of listings.   
     
   Through visualizations we understood that 85,42% of the number of listings come from Brooklyn and Manhattan, sharing approximately the same number of listings each one of them, with Manhattan having 1500 more listings. Even though that's the case Brooklyn has the two neighborhoods with the highest number of listings. To reach that conclusions we also used the following measures:

Manhattan\_Brooklyn\_listing\_Percentage = DIVIDE(

CALCULATE(COUNT('AB\_NYC\_2019'[id]),'AB\_NYC\_2019'[neighbourhood\_group] = "Manhattan" || 'AB\_NYC\_2019'[neighbourhood\_group] = "Brooklyn"),

COUNT('AB\_NYC\_2019'[id]))\*100

Manhattan\_listing\_Count = CALCULATE(COUNT('AB\_NYC\_2019'[id]), 'AB\_NYC\_2019'[neighbourhood\_group] = "Manhattan")

Brooklyn\_listing\_Count = CALCULATE(COUNT('AB\_NYC\_2019'[id]), 'AB\_NYC\_2019'[neighbourhood\_group] = "Brooklyn")

Recommendation: Target neighborhoods with fewer listings to face less competition. Entering markets with less competition increases the chances of your listing standing out and receiving more bookings. Consider analyzing the unique selling points of these neighborhoods to tailor your listing to potential guests' preferences.

1. Insight: Understand the proportion of different room types (e.g., Entire home/apt, Private room, Shared room) in the dataset.  
     
   Through depicting with a pie chart the room type across all the listings in our dataset, we found out that Shared room type covers only 2.3% of the whole listings, while Entire home/apt is 51.9% and Private room are the 45.7% of our listings.  
     
   Recommendation: Offer the most popular room types to attract more bookings. By catering to the majority of guests' preferences, you increase your listing's appeal and potential for bookings. Analyze the room type distribution in your target area and adjust your offerings to match popular demand.
2. Insight: Identify neighborhoods with the highest and lowest average availability throughout the year.  
     
   Through the Treemap we can see which Neighborhood groups have greater average availability dates through the year 2019, having Staten Island which has the lowest number of listings in combination with the 2 most expensive neighborhoods.  
     
   Recommendation: Choose neighborhoods with higher average availability to maximize booking opportunities. High availability indicates that there may be a steady demand for accommodations in that area. By listing your property in such neighborhoods, you increase your chances of receiving more consistent bookings.
3. Insight: Explore the relationship between price and minimum nights to see if there’s any correlation.  
     
   Through visualizing the scatter plot and restricting it to price<300 and minimum nights below 30, which represent approximately 91% of the dataset and calculated by:

Min\_Nights\_Below\_30\_Count = CALCULATE(COUNT('AB\_NYC\_2019'[id]), 'AB\_NYC\_2019'[minimum\_nights] < 30)/48895\*100  
  
And by also adding a Trend Line to inspect the tendency we can see that as the number of minimum nights increases the price falls.  
  
Recommendation: Set minimum nights based on the observed trend to match market expectations. Set up appropriate minimum stay requirements through the analyzation of the relationship between price and minimum nights. This can help you cater to guests' preferences and improve booking rates.

1. Insight: Determine which neighborhoods have the most reviews, indicating higher popularity.  
     
   The sum of reviews provided with almost the same image as the number of listings, with the only difference having Brooklyn 30k ahead of Manhattan.   
     
   Recommendation: Focus on popular neighborhoods with more reviews to benefit from increased demand or focus to a more niche market by targeting the non-popular one’s. A higher number of reviews may indicate a higher level of interest in a particular neighborhood. By listing the property in these areas, you can capitalize on this demand and potentially receive more bookings.
2. Insight: Explore the relationship between price and review scores to see if higher prices correlate with better reviews.  
     
   We can see that the number of reviews is getting lower as the price is increasing, which is logical because customers tend to prefer lower prices. Adding a Trend Line to inspect the tendency we can see clearer the aforementioned statement.  
     
   Recommendation: Offer competitive prices while maintaining high-quality accommodations to achieve high review scores. Striking a balance between affordability and quality can make the listing more attractive to potential guests. This can lead to higher review scores, which can further enhance the listing's appeal.
3. Insight: Identify the hosts with the most listings, indicating successful Airbnb businesses.  
     
   By creating a table with row the host\_name and column the count of the listings (id), and at the same time filtering the visual to get us back only the top 10 based on the listing’s count, we get back the 10 names of the most successful hosts.   
     
   Recommendation: Study the strategies of successful hosts to learn from their experiences and improve your own listings. Examine their pricing, amenities, and customer service practices to identify areas where you can enhance your own offerings and better compete in the market.

1. Insight: Compare the proportion of room types across different neighborhoods.  
     
   Through the stacked bar chart we can see that In Brooklyn there is the highest number of private rooms and in Manhattan the higher number of Entire home/apart.   
     
   Recommendation: Offer room types that are popular in specific neighborhoods to cater to local demand. By tailoring the offerings to the preferences of guests in a particular area, increases your listing's appeal and potential for bookings. Research the most popular room types in the target neighborhood and adjust the property accordingly.
2. Insight: Analyze the relationship between the rating of amenities and listing prices.

We created a scatter plot with the X-axis representing the rating and the Y-axis representing the price, and added a trend line to examine the correlation between the two variables. The plot revealed that as the rating of the amenities increases, the listing price tends to increase as well, indicating that guests are willing to pay more for better-rated amenities. We also observed that listings with higher-rated amenities typically received more bookings, further emphasizing the importance of quality amenities in driving demand.

Recommendation: Invest in high-quality amenities to attract more bookings and justify higher listing prices. By providing top-rated amenities, you can not only enhance the overall guest experience but also position your listing as a premium offering within the market. Regularly review and update the amenities to maintain high ratings and remain competitive in the market.

## Regression - Forecasting

To start our  linear regression analysis , we obtained sample data from the original dataset in order to make a forecast. The data used in this analysis consists of 5000 rows which were produced randomly by using Python code.The python code used is the following:

Text

Description automatically generated

In order to perform the linear regression analysis, we started with creating two new columns (the (X^2) and the(xy))

Created a column xy = specified\_rows[Y]\* specified\_rows[X] to calculate the multiplication on row level between the two variables.

Created a column X2 = specified\_rows[X]^2 to calculate the exponent of each element of the variable X

Then we calculated the slope and intercept(a and b) for the linear regression model as shown below:

b = (count(specified\_rows[minimum\_nights])\*sum(specified\_rows[xy])-sum(specified\_rows[minimum\_nights])\*sum(specified\_rows[price]))/(count(specified\_rows[minimum\_nights])\*sum(specified\_rows[X2])-sum(specified\_rows[X2]))

a = (sum(specified\_rows[price])\*sum(specified\_rows[X2])-sum(specified\_rows[minimum\_nights])\*sum(specified\_rows[xy]))/(COUNT(specified\_rows[minimum\_nights])\*sum(specified\_rows[X2])-sum(specified\_rows[minimum\_nights])^2)

The results of the linear regression analysis are as follows:

* Slope: 0.64
* Intercept: 143.76

Based on these results, we can forecast using the following model:y=143.76+0.64x, where y represent the price and x represent the minimum\_nights. It is important to note that this forecast is based on the assumption that the underlying relationship between the variables is linear. Now we proceeded on attaching the aforementioned model on the PowerBI containing all the original data. In order to input the model, we created a column with the following command:

y\_fit=143.76+0.64\*'AB\_NYC\_2019'[minimum\_nights]

To view forecasted values in comparison with the real ones, we created a Line Chart putting in the X-axis the minimum nights and, in the Y-axis, the original price. Then on the Secondary Y-axis we assign the y-fit model that we have created and preview the results. Following the exact same procedure on another sample that we produced from python, we performed another forecast for price and rating this time. Then we calculated the slope and intercept (a and b) for the linear regression model as shown below:

b = (count(specified\_rows(1)[Rating])\*sum(specified\_rows(1)[X\*Y])-sum(specified\_rows(1)[Rating])\*sum(specified\_rows(1)[price]))/(count(specified\_rows(1)[Rating])\*sum(specified\_rows(1)[X^2])-sum(specified\_rows(1)[X^2]))

a = (sum(specified\_rows(1)[price])\*sum(specified\_rows(1)[X^2])-sum(specified\_rows(1)[Rating])\*sum(specified\_rows(1)[X\*Y]))/(COUNT(specified\_rows(1)[Rating])\*sum(specified\_rows(1)[X^2])-sum(specified\_rows(1)[Rating])^2)

The results of the linear regression analysis are as follows:

* Slope: -0.25
* Intercept: 152.75

Based on these results, we can forecast using the following model:y=152.75-0.25x, where y represent the price and x represent the rating. It is important to note that this forecast is based on the assumption that the underlying relationship between the variables is linear. Now we proceeded on attaching the aforementioned model on the PowerBI containing all the original data. In order to input the model, we created a column with the following command:

y\_fit=152.75 – 0.25\*'AB\_NYC\_2019'[rating]

By integrating our two forecasts, the company can determine the optimal price to set for a newly acquired amenity, as well as the minimum number of nights required for successful renting. Coupled with our other research, the company can identify key specifications and characteristics that contribute to amenity success, and adjust the minimum nights and price accordingly to achieve the desired rating and vice versa.

# Conclusion

In conclusion, our analysis has provided valuable insights and recommendations for maximizing the success of Airbnb listings in New York City. By focusing on the most common price ranges, targeting neighborhoods with fewer listings or higher average availability, and offering popular room types, hosts can appeal to a larger audience and increase their occupancy rates. Additionally, understanding the relationship between price, minimum nights, and review scores can help our company set competitive prices while maintaining high-quality accommodations, leading to higher guest satisfaction, in their amenities or future investments.

Our linear regression analysis and forecasting models offer a data-driven approach to determining optimal pricing and minimum stay requirements for new listings, based on factors such as rating and neighborhood characteristics. By integrating these forecasts, our company can make informed decisions on how to adjust their listings to achieve the desired ratings and occupancy rates.Studying the strategies of successful hosts and analyzing the relationship between the rating of amenities and listing prices can further inform our company on how to enhance their offerings and remain competitive in the market.

# References

Beck, N., & Katz, J. (1995). What To Do (and Not to Do) with Time-Series Cross-Section Data. American Political Science Review, 89(3), 634–647. <https://doi.org/10.2307/2082979>

Pipino, L. L., Lee, Y. W., & Wang, R. R. (2002). Data quality assessment. Communications of the ACM, 45(4), 211–218. <https://doi.org/10.1145/505248.506010>

Su, X., Yan, X., & Tsai, C. (2012). Linear regression. Wiley Interdisciplinary Reviews: Computational Statistics, 4(3), 275–294. <https://doi.org/10.1002/wics.1198>

# Appendix

*This form should be filled out by all team members after the completion of the group assignment. The team leader should be chosen upon agreement and is responsible to upload the group assignment after its completion and deal with any technical and other issues that might arise during the submission process.*

Team Leader name: Alkiviadis Kariotis

Team Leader ID: 241735

|  |
| --- |
| Team member name: Alkiviadis Kariotis  Team member ID: 241735  ***I herewith express my agreement with the submission of this final version of the group project by the team leader.***  A pair of glasses  Description automatically generated with medium confidence  Date: 11/03/2023  Team member Signature: |
| Team member name: Konstantinos Prozymas  Team member ID: 27365  ***I herewith express my agreement with the submission of this final version of the group project by the team leader.***    Date: 11/03/2023  Team member Signature: Konstantinos Prozymas |
| Team member name: Konstantinos Megaritis  Team member ID: 271868  ***I herewith express my agreement with the submission of this final version of the group project by the team leader.***    Date: 11/03/2023­  Team member Signature: Konstantinos Megaritis |